ЕРЕВАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Агаджанян Рубен Борисович

МОДЕЛИРОВАНИЕ СЛОЖНЫХ СИСТЕМ МЕТОДОМ СТОХАСТИЧЕСКИХ ВЫЧИСЛЕНИЙ

Диссертация

на соискание учёной степени кандидата технических наук по специальности 05.13.04

“Математическое и программное обеспечение вычислительных машин, комплексов, систем и сетей»

Научный руководитель:

доктор физико-математических наук,

Нигиян С. А.

Ереван – 2018

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Актуальность темы. На практике, управление сложными производственно-экономическими системами в ряде отраслей осуществляется на основе нормативных требований и показателей, устанавливаемых соответствующими органами надзора, контроля качества и другими регулирующими организациями. Оценка стабильности подобных систем и прогнозирование устойчивости их фунционирования с точки зрения соответствия нормативным отраслевым требованиям является первостепенной задачей для многих предприятий и

организаций. Под **стабильностью** будем понимать такое состояние системы, при котором ее ключевые показатели находятся в пределах заданных допустимых значений. Под **устойчивостью** будем понимать способность системы возвращаться в равновесное (стабильное) состояние при выходе из него внутренними или внешними воздействиями. Особенностями этих систем являются многообразность их структуры, стохастичность поведения их компонент, наличие объектов с большим числом характеристических признаков и сложных многопараметрических связей между звеньями системы, динамичность изменения и многокритериальность оценки состояния системы, наличие обратных связей в разных звеньях управления и т.д. Задача анализа и выявления нарушений функционирования в звеньях сложных систем представляет собой длительный и трудоемкий процесс, прежде всего из-за необходимости исследования и идентификации большого числа взаимосвязанных случайных ключевых параметров и оценки их воздействия на работоспособность системы в целом. Для проведения соответствующих исследований в настоящее время используются разные методы и подходы, зависящие в первую очередь от предметной области и специфики задач. Большинство решений, реализованных в существующих программных продуктах управления и контроля работы сложных производственно - экономических систем представляют собой технологические инструменты для протоколирования фактов обнаружения нарушений, проведенных исследований, результатов выполненных корректирующих и превентивных действий. Эти учетные данные вводятся пользователями и компетентными специалистами в систему с помощью предусмотренных соответствующих электронных форм. На основе этих данных системами создаются различные аналитические отчеты и протоколы по выполненным процедурам и проведенным изменениям. Однако, уровень автоматизации обнаружения, распознавания типов нарушений и принятия решений по проведению адекватных корректирующих и превентивных действий все еще не достаточен для удовлетворения требований пользователей этих систем. В связи с этим, в настоящее время задача автоматизации процессов идентификации случайных несоответствий и определения адекватных корректирующих и превентивных действий является актуальной.

Для получения более обоснованной оценки стабильности сложных систем и прогнозирования их устойчивости необходимо принять во внимание целый ряд факторов, среди которых можно выделить следующие: сложность структуры системы и ее компонент, множество взаимосвязанных ключевых параметров и

функциональных признаков объектов системы, закономерности отображения входных параметров в выходные показатели и другие. Основными проблемами в рассматриваемом классе систем являются разработка методов и алгоритмов обнаружения несоответствий ключевых показателей их нормативным значениям, определение причинно-следственных связей между объектами, классификация несоответствий и принятие решений на основе установленных правил. Указанные задачи являются особенно актуальными для предприятий таких отраслей, как фармацевтическое производство, здравоохранение, пищевая промышленность и многих других, в которых предъявляются жесткие требования к соблюдению нормативных стандартов производства продукции. Для оценки ключевых параметров производственно-экономических систем применяются различные методы, такие как аналитические методы, моделирование, методы эмпирических оценок, методы проб и ошибок, статистические методы, стохастические вычисления и другие. Выбор конкретного метода зависит от типов параметров и метрик его оценки. В фармацевтической промышленности одним из основных международных отраслевых стандартов является “Надлежащая производственная практика GMP (Good Manufacture Practice)”[1-6]. Подобные стандарты охватывают все аспекты производства: от контроля исходных материалов, помещений и оборудования, включая технологические этапы производства, до обучения и личной гигиены персонала. Нарушение стабильности при несоответствиях ключевых показателей отраслевым нормативным стандартам может повлечь за собой останов производства, финансовые потери и полное прекращение деятельности до восстановления стабильности. Для организации мониторинга производственных и управленческих процессов компаний на соответствие отраслевым стандартам, в частности стандарту GMP, широко используется система качества CAPA (Corrective and Preventive Actions, корректирующие и превентивные действия) [7-10]. Согласно методологии CAPA при поступлении информации об имеющихся отклонениях ключевых показателей, система должна идентифицировать возникшую проблему, провести анализ несоответствий и определить корректирующие и превентивные действия, направленные на устранение этих несоответствий и возвращение системы в стабильное состояние. В настоящее время, задача анализа и обнаружения несоответствий в рассматриваемом классе сложных систем представляет собой длительный и трудоемкий процесс, прежде всего из-за неструктурированности сообщений о несоответствиях и необходимости исследования и идентификации

большого числа взаимосвязанных случайных ключевых параметров. Между тем невозможность своевременного устранения несоответствий отрицательно сказывается на многие финансовые и экономические показатели, а также создает сложности в организации производства и многих других процессов, среди которых, в частности, своевременная поставка конечной продукции потребителю. Без технологических инструментов не представляются возможными оперативный анализ большого числа динамических показателей, оценка стабильности системы и принятие своевременного решения о проведении процедур CAPA. Для проведения соответствующих исследований в настоящее время используются разные методы и подходы, зависящие в первую очередь от предметной области и специфики задач. Данная задача обусловлена не только потребностями в автоматизации учета проведения процедур CAPA и протоколировании производственных этапов. Основными проблемами являются разработка методов и алгоритмов распознавания входных сообщений о несоответствиях, оценка ключевых показателей, определение причинно-следственных связей и классификация несоответствий на основе некоторых установленных правил. При этом необходимо обеспечить автоматизированный анализ информации о возникших несоответствиях для оперативного принятия решений о проведении процедур CAPA.

Проблемам исследования сложных производственно – экономических систем и разработке программного обеспечения контроля функционирования этих систем посвящено много работ ведущих специалистов в области стратегического и оперативного менеджмента и информационных технологий. Ведутся исследования по моделированию производственных и управленческих процессов, автоматизации системы мониторинга и повышению качества продукций компаний. Разрабатываются программные комплексы, предназначенные для мониторинга всего цикла производственных и управленческих процессов. В научном плане проводятся семинары, конференции, выпускаются журналы, посвященные данной тематике. Все это доказывает, что имеется большой научный и практический интерес к исследованиям и разработкам по данной тематике.

Для рассматриваемого нами класса сложных систем отметим следующие проблемы, которые не позволяют на практике эффективно проводить процедуры CAPA для соблюдения правил GMP и других отраслевых стандартов:

- сложность вычислений и существенный рост в потребности вычислительных ресурсов при увеличении объемов хранения и обработки данных;
- необходимость учета корреляций и взаимозависимостей между большим количеством случайных параметров;
- трудоемкость процесса оценки степени соответствия ключевых показателей заданным эталонным значениям, требующего большого числа наблюдений;
- необходимость принятия ряда допущений и упрощений для возможности организации вычислений;
- отсутствие развитых методов и алгоритмов классификации неструктурированных сообщений.

Преодоление указанных проблем возможно путем разработки единой консолидированной системы автоматизации процессов обработки неструктурированных входных данных об отклонениях, их классификации, генерации экспертных оценок стабильности системы и заключений для проведения точных и своевременных процедур CAPA. В основе данной системы должна быть некоторая модель [11-13], описывающая закономерности отображения входных сообщений о несоответствиях в определенные выходные классы, идентифицирующие данные несоответствия. В настоящее время ведутся активно исследования по указанным направлениям, в частности в таких областях как машинное обучение, семантический и синтаксический анализ неструктурированных текстов, разрабатываются новые подходы в вопросах моделирования текстовых коллекций и интеллектуального анализа данных [14-16]. Особенно бурно развиваются методы системного и вероятностно-семантического анализа [17-21] и генеративных вероятностных моделей с различными распределениями для дискретных данных [22,23]. Методологической основой для данной диссертационной работы послужил ряд исследований в области системного анализа, моделирования, машинного обучения и теории проектирования информационных систем [24,25]. Актуальность представленных выше задач обусловлена, с одной стороны, недостаточным применением на практике теоретических исследований и работ, консолидирующих данные, процессы и ресурсы управления процедурами CAPA, с другой стороны, отсутствием развитых инструментов программного обнаружения и идентификации несоответствий.

Целью работы являются проектирование стохастической модели сложных систем и разработка программного обеспечения управления стабильностью их функционирования.

Достижение данной цели предполагает решение следующих задач:

1. Исследование методов управления стабильностью сложных стохастических систем.
2. Построение модели обнаружения случайных отклонений ключевых параметров (показателей) от нормативных значений и определения состава и последовательности проведения корректирующих и превентивных действий (CAPA).
3. Разработка метода классификации случайных несоответствий и алгоритма динамического построения базы данных эмпирической выборки.
4. Разработка программного обеспечения управления стабильностью сложных стохастических систем.

Объектом исследования являются:

1. Сложные стохастические системы и методы управления их стабильностью;
2. Алгоритмы классификации объектов на основе метрических параметров и перекрестного контроля значений признаков объектов эмпирической и контрольной выборок;
3. Метод оценки погрешности в алгоритмах классификации случайных несоответствий.

Методы исследования. В диссертационной работе использованы методы математического моделирования стохастических систем, теория вероятностей и случайных процессов, теория множеств, методы системного анализа, теория и методы построения информационных систем.

Научная новизна диссертационной работы заключается в следующем:

1. Разработана модель управления стабильностью стохастических систем, основанная на применении процедур CAPA и эмпирической выборке, представляющей собой отображение множества входных показателей ключевых параметров в классы корректирующих и превентивных действий.
2. Разработан метод динамического определения эмпирической выборки и ее модификации путем оценки степени погрешности классификации вектора значений признаков объектов.

3. Предложен алгоритм эффективного управления системой путем контроля параметров объектов, обнаружения и классификации случайных несоответствий на основе заданных метрик и данных из эмпирической выборки.
4. По результатам исследований, полученных в диссертационной работе, разработано программное обеспечение автоматизации процесса управления стабильностью сложных стохастических систем.

Практическая значимость работы. Разработанная математическая модель может быть использована в реализации и построении программной системы управления процедурами CAPA для различных задач и областей применения. Разработанные методы и алгоритмы ориентированы на решение комплекса практических задач контроля ключевых показателей и управления стабильностью сложных стохастических систем. На основе метрических алгоритмов и эмпирической выборки разработаны методы программного обнаружения несоответствий и их классификации. Практические примеры проектирования программного обеспечения управления процедурами CAPA для ряда областей, в частности здравоохранения, фармацевтики и инженерного обслуживания показывают практическую ценность результатов исследования и принятых решений.

Внедрение разработанного программного обеспечения позволило значительно сократить:

- Текущие расходы за счет автоматизации процесса контроля стабильности и применения адекватных процедур по предотвращению ее нарушения.
- Количество фактов нарушения стабильности функционирования системы в следствие проведения предупреждающих действий.

Степень достоверности результатов. Достоверность полученных результатов данной диссертационной работы подтверждаются многократными испытаниями и внедрениями разработанного программного обеспечения на конкретных предприятиях.

На защиту выносятся следующие основные положения:

1. Модель управления процедурами CAPA в сложных стохастических системах, включающая в себя процессы идентификации случайных несоответствий и источников их возникновения и метод определения корректирующих и

превентивных действий на основе классификации объектов и их ключевых показателей;

2. Методология построения и дальнейшей адаптации эмпирической выборки объектов и соответствующих классов для последующей идентификации несоответствий;
3. Метод контроля показателей объектов, включающий в себя функции формирования исходных данных, регистрации текущих значений параметров контролируемых объектов и идентификации несоответствий;
4. Метод и семейство алгоритмов классификации обнаруженных несоответствий на основе стохастических вычислений их близости к элементам и кластерам эмпирической выборки;
5. Разработанное программное обеспечение управления стабильностью сложных стохастических систем как инструмент оперативного контроля и принятия решений по проведению процедур CAPA.

Внедрение результатов работы. Результаты исследований, разработанных методов и алгоритмов прошли промышленную апробацию и на их основе разработано программное обеспечение управления процедурами CAPA, которая внедрена и в настоящее время находится в эксплуатации в следующих организациях:

1. Фармацевтическая компания “ЛИКВОР”.
2. Фармацевтическая компания “Фарматек”.
3. Медицинский диагностический центр “Норк-Мараш”.

В настоящее время на основе результатов данной диссертационной работы совместно с датской компанией “ZEVIT” ведется разработка нового программного обеспечения для инженерно-сервисной компании.

Апробация результатов работы. Основные результаты диссертации были представлены на семинаре факультета прикладной математики и информатики ЕГУ, на семинаре фармацевтов и врачей (Ереван, 2017), организованном международной организацией USAID, на 14-ой международной научно-практической конференции “Advances in Science and Technology” (Москва, 2018).

Использование результатов диссертационной работы подтверждено соответствующими документами.

Публикации. Основные результаты диссертационной работы представлены в 7 научных работах (один доклад в международной научной конференции и 6 печатных работ в научных журналах).

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы (112 источников) и приложения. Содержание изложено на 121 страницах основного текста, включая 30 рисунков.

ГЛАВА 1: ИССЛЕДОВАНИЕ И МОДЕЛИРОВАНИЕ СЛОЖНЫХ СТОХАСТИЧЕСКИХ СИСТЕМ

Исследования в области моделирования стохастических систем (стохастическое моделирование) на сегодняшний день представляет собой бурно развивающее направление, прежде всего из-за большого числа предметных и прикладных задач, связанных с изучением и управлением случайными процессами. Стохастические модели строятся в основном на выдвигаемых гипотезах в отношении

закономерностей поведения изучаемых систем и предсказании значений, зависящих от вероятностных показателей. Как правило, описание стохастических систем включает в себя использование методов системного и статистического анализов большого числа объектов, для каждого из которых задан многомерный вектор характеристических признаков. Задачи, решаемые при моделировании сложных систем хотя и могут отличаться друг от друга, в частности, особенностями предметной области, в большинстве случаев их объединяет некоторый общий подход к оптимизации процессов, оценки их эффективности и выработке определенного воздействия для достижения установленных целей (значений ключевых показателей). Сложность стохастических систем обусловлена не только слабой предсказуемостью ее поведения, но и большим числом взаимодействий и причинно-следственных связей между отдельными ее элементами. Распространенными методами исследования и математического описания подобных систем являются методы системного и статистического анализа.

1.1. Обзор методов исследования сложных стохастических систем

Мы будем рассматривать сложные стохастические системы с целью последующего построения их классификационных моделей, которые позволят контролировать ключевые показатели стабильности при различных случайных процессах. Представим рассматриваемый класс сложных стохастических систем в виде следующего формального описания.

Пусть задано множество X , состоящее из m подсистем объектов:

$$X = \{\mathrm{X}_1, \mathrm{X}_2, \dots \mathrm{X}_\mathrm{m}\},\ \ \mathrm{X}_\mathrm{i} = \{\mathrm{x}_\mathrm{ij}\},\ \mathrm{j} = (1,2, \dots, \mathrm{n}(\mathrm{X}_\mathrm{i})),$$

для подсистемы X_i имеется некоторое множество характеристических свойств (признаков) $\mathrm{F}^{\mathrm{X}_\mathrm{i}} = \{\mathrm{f}_1^{\mathrm{X}_\mathrm{i}}, \mathrm{f}_2^{\mathrm{X}_\mathrm{i}}, \dots, \mathrm{f}_{\mathrm{l}(\mathrm{X}_\mathrm{i})}^{\mathrm{X}_\mathrm{i}}\}$, которые принимают определенное значение для каждого элемента x_ij множества X_i . Таким образом, элементу x_ij будет соответствовать вектор признаков $\mathrm{F}^{\mathrm{X}_\mathrm{i}}(\mathrm{x}_\mathrm{ij}) = (\mathrm{f}_1^{\mathrm{X}_\mathrm{i}}(\mathrm{x}_\mathrm{ij}), \mathrm{f}_2^{\mathrm{X}_\mathrm{i}}(\mathrm{x}_\mathrm{ij}), \dots, \mathrm{f}_{\mathrm{l}(\mathrm{X}_\mathrm{i})}^{\mathrm{X}_\mathrm{i}}(\mathrm{x}_\mathrm{ij}))$.

Для всех объектов x_ij семейства X_i получим матрицу функциональных признаков:

$$\begin{cases} f_1^{X_i}(x_{i1}), f_2^{X_i}(x_{i1}), \dots, f_{l(X_i)}^{X_i}(x_{i1}) \\ f_1^{X_i}(x_{i2}), f_2^{X_i}(x_{i2}), \dots, f_{l(X_i)}^{X_i}(x_{i2}) \\ \text{----------} \\ f_1^{X_i}(x_{im}), f_2^{X_i}(x_{im}), \dots, f_{l(X_i)}^{X_i}(x_{im}) \end{cases}$$

$F^{X_i} = ||f_j(x_{ij})||_{m(X_i)\times l(X_i)} =$

Для всей системы получим вектор подмножеств функциональных признаков :

$$F = (F^{X_1}, F^{X_2}, \dots, F^{X_m})$$

Для упрощения обозначений, мы иногда будем рассматривать объекты только одного подмножества X_i, которое в дальнейшем обозначим X и соответственно, исключим один из индексов в обозначении элементов этого подмножества: $X = \{x_1, x_2, \dots, x_m\}$, а также введем упрощенное обозначение признаков объекта x_i:

$$f(x_i) = \{f_1(x_i), \dots f_n(x_i)\} .$$

Пусть в системе регистрируются значения признаков $f(x_i)$ объекта x_i в некоторый момент времени t, которые обозначим как $f^t(x_i) = \{f^t_1(x_i), \dots f^t_n(x_i)\}$.

Через x_i^t обозначим объект, признаки которого измерены в момент времени t.

Между элементами системы имеются связи (отношения) $R = \{r_{ij}\}$, которые можно записать как отношение между двумя произвольными объектами: $r_{ij} =< x_i, x_j >$.

Зададим множество классов $Y = \{y_1, y_2, \dots, y_k\}$. В работах [26-30] сложная система рассматривается как совокупность пар $< x_i, y_i >$ входных объектов x_i и значений выходного результата (класса) $y_i \in Y$.

Имеется неизвестная зависимость $A: X \to Y$, которую можно записать также в виде $A: F \to Y$, представляющую собой классификацию путем отображения объектов множества X или признаков объектов F в определенные классы Y. Введем понятие “исследователя” – субъекта (обозначим через I), в сознании которого отражена система в виде совокупности объектов, их признаков и отношений. На необходимость учета взаимодействия изучаемой системы и “исследователя” впервые указал Эшби [31]. Для рассматриваемого класса сложных стохастических систем характерно также взаимодействие с внешней средой [32,33], представленной множеством W в виде правил, норм и эталонных значений. Эти правила в совокупности с исследователями направлены на достижение системой определенных целевых конечных результатов, системообразующих критериев P [34]. Формально, модель стохастических систем как объекта управления можно записать в виде кортежа: $S = \{X, F, R, Y, A, I, W, P, t\}$.

На рисунке 1.1 представлен пример системы из фармацевтической отрасли:

Рисунок 1.1. Пример классификации признаков объектов.

Построение системы классификации необходимо для исследования и последовательного решения следующих задач:

- идентификация несоответствий по регистрируемым и эталонным значениям признаков объектов;
- определение информативных параметров;
- анализ корневых причин случайных процессов;
- классификация объектов и определение корректирующих и превентивных действий.

Наряду с понятиями стабильности и устойчивости, введем еще несколько определений, которые приняты в теории систем и будут использованы в настоящей диссертационной работе. Система является **самоорганизующей** или **саморегулирующей**, если имеется обратная связь, обозначающая обратное воздействие выходных результатов системы на его процессы. При этом различают **отрицательную связь**, при которой имеет место коррекция процесса, направленная на возвращение системы в стабильное состояние и **положительную связь**, при которой выходные данные способствуют процессу дальнейшего отклонения от нормы. На этих понятиях базируется, в частности, теория имитационного

моделирования [35]. Приведем еще одно важное свойство сложных систем – способность к **самообучению**. Современные программные системы обладают свойством самообучения (машинного обучения) [36], которое отличается от саморегулирующих свойств системы. Самообучение представляет собой метод интеллектуальной обработки данных, представленнных в виде пар <объект, класс> из некоторой заданной выборки, при котором “обучающий” алгоритм извлекает знания из этой выборки.

При этом, обучающий алгоритм путем определения закономерности отображения “объект->класс” оптимизирует параметры решающих функций для последующей правильной классификации поступающих новых сообщений, распознавания образов и генерации прогнозных заключений.

Приведем несколько известных подходов, которые применимы для исследования вышеописанного класса сложных стохастических систем. Одним из самых распространенных методов являются энтропийные методы исследования [37-39], в которых обобщено понятие информационной энтропии и описано семейство функционалов количественного представления множества случайных процессов системы. Другим методом исследования стохастических систем является метод так называемых когнитивных карт [40-47]. Согласно этому методу, представлен некоторый интерпретирующий инструмент и соответствующий алгоритм принятия решений. Посредством когнитивных карт строят некоторую интерпретацию причинно-следственных связей и определяют правила состояний когнитивной системы в рассматриваемой предметной области [48-50].

В основе многих исследований находятся задачи выявления общих закономерностей или гипотезы о закономерностях, взаимосвязей элементов системы и построения соответствующих решающих функций на основе интеллектуального анализа данных. Решение данных задач позволит построить модель управления стабильностью сложных стохастических систем и исследовать закономерности отображения входных объектов в выходные классы.

1.2. Моделирование процессов управления стабильностью стохастических систем

В настоящей работе рассматриваются модели стохастических сложных систем, целевая функция которых заключается в обеспечении стабильного состояния путем оценки устойчивости контролируемых ключевых параметров, идентификации

случайных несоответствий и управления корректирующими и превентивными действиями.

В основе моделей стохастических систем находятся вероятностная теория оценки состояния системы и прогнозирование ее поведения в будущем. Данная оценка производится на основе анализа ключевых параметров системы и их взаимосвязи, классификации объектов по степени близости их признаков элементам эмпирической выборки и прогнозировании наступления несоответствий в изучаемых процессах [36],[51],[52]. На рисунке 1.2 представлена структура модели управления стабильностью сложной стохастической системы, которую можно представить в виде трех подсистем – учетной, управляющей и классификационной:

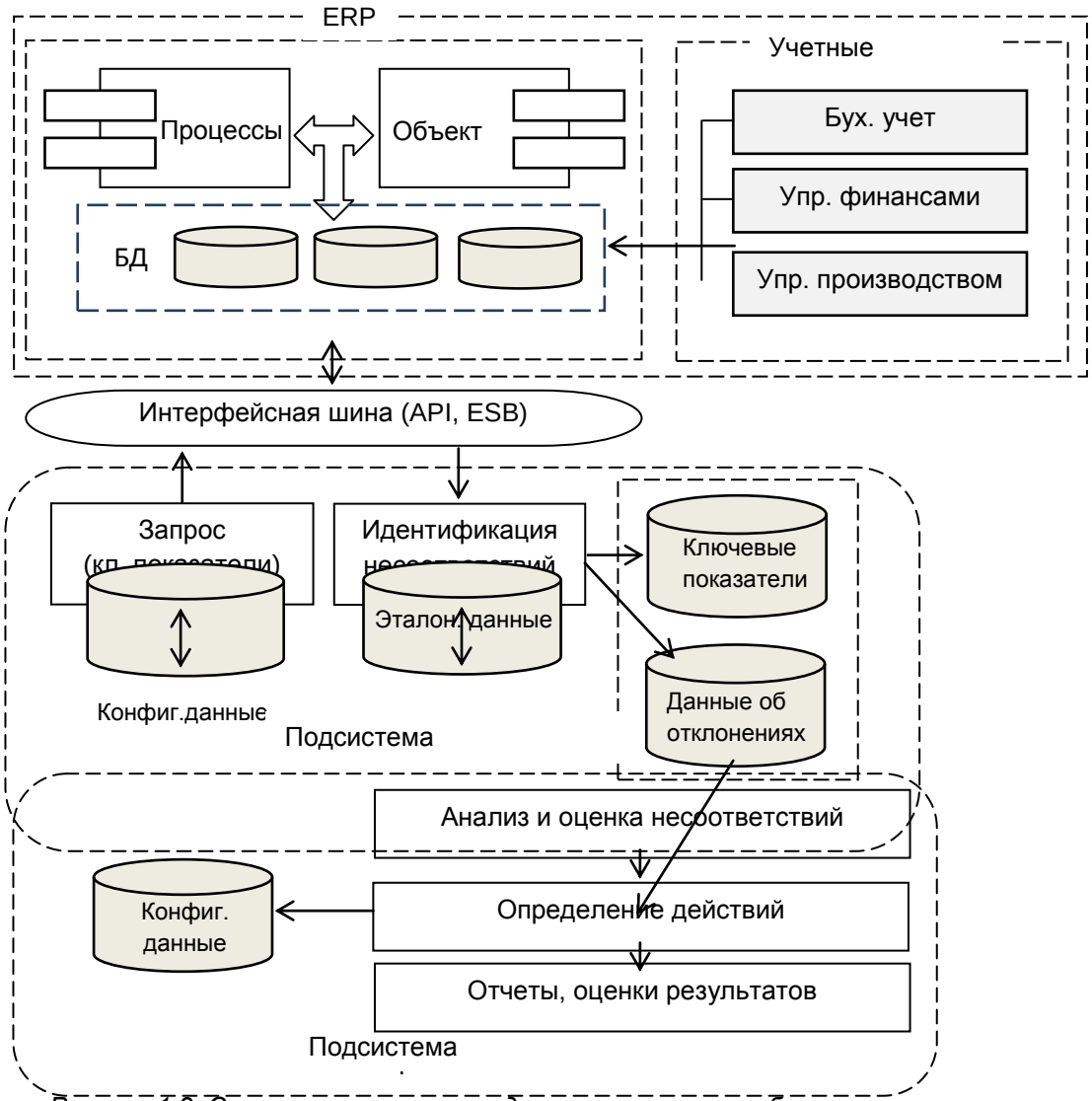


Рисунок 1.2. Структурная схема модели управления стабильностью системы.

Моделирование процесса контроля стабильности включает в себя выполнение следующих основных шагов:

Шаг 1. Вся корпоративная информация формируется в учетной подсистеме в результате функционирования и взаимодействия различных компьютерных программ и систем.

Шаг 2. В контрольной подсистеме на основе конфигурационных данных генерируется запрос на выборку тех значений из всей базы данных, которые относятся к ключевым показателям. Для обработки запроса используется интеграционная шина между учетной и контрольной подсистем.

Шаг 3. Данные выборки значений ключевых показателей в контрольной части анализируются и протоколируются в виде базы данных, содержащей значения всех ключевых показателей в некоторый момент времени t и базы данных, содержащей только те ключевые показатели, которые находятся вне заданных эталонных значений.

Шаг 4. Определение класса из заданного конечного множества классов, содержащего в себе информацию о корректирующих и превентивных действиях, необходимых для устранения обнаруженных несоответствий.

Шаг 5. Генерация следующих выходных данных:

- экспертные заключения и соответствующая отчетность, включающие в себя полную информацию об обнаруженных несоответствиях, в частности, оценку рисков влияния на различные подсистемы и объекты;
- корректирующие и превентивные действия;
- требуемые ресурсы для проведения этих действий.

Приведем несколько общеизвестных подходов к исследованию и моделированию сложных стохастических систем. Классическими методами изучения стохастических систем являются методы оценки корреляции, массового обслуживания и имитационного моделирования [53]. В то же время, как показано в работе [54], для объектов со сложной структурой классический подход к описанию процессов, в частности, методами массового обслуживания, не всегда дает ожидаемых результатов из-за вынужденной чрезмерной абстракции описания взаимозависимости переменных. Преодоление данного недостатка требует новых подходов с применением различных методов аппроксимации и аналитических инструментов точного описания случайных процессов. Заслуживают внимания новые методы, позволяющие преодолеть излишнюю абстракцию параметров стохастической модели и учесть динамику процессов в конкретной предметной области. В статье [55], в качестве метода моделирования случайных процессов

рассматривается некоторая предельная переходная матрица, позволяющая описать закономерности случайных процессов марковского типа. Заслуживают внимания вопросы моделирования сложных систем на основе многомерной классификации объектов и методов оценки их эффективности [56]. Данная оценка производится на основе метода анализа данных и построения граничных значений эффективности (DEA, Data Envelopment Analysis). Однако, как отмечено в работе [56] данный подход требует дополнительных исследований, в частности в вопросах учета особенностей предметной области и формализации метода определения входных и выходных переменных в алгоритмах поэтапного формирования многомерной границы эффективности.

В последующих главах мы рассмотрим задачу идентификации несоответствий ключевых показателей сложных систем путем кластеризации эмпирической выборки и классификации новых объектов на основе оценки близости к заданным кластерам.

1.3. Программные методы управления процедурами CAPA

В настоящее время имеется широкий спектр программных решений, включающих в себя технологические инструменты управления процедурами CAPA. Основными функциями указанных систем являются протоколирование всех этапов процедур CAPA, включающих в себя данные о выявленных отклонениях, проведенных процедурах для их устранения и предотвращения появления в будущем. Одной из основных задач при проведении процедур CAPA является составление адекватной документации о предпринятых действиях, включая важные исторические данные для непрерывного улучшения качества конттролируемых параметров. Такие известные системы, как MasterControl® и Infotehna® позволяют протоколировать следующие основные процедуры CAPA:

1. идентификация проблемы, несоответствия или инцидента;
2. оценка степени риска и потенциального воздействия на производственные и другие процессы;
3. проведение процедур исследования обнаруженных проблем;
4. создание плана действий, в котором перечислены все ресурсы и задачи, которые необходимо выполнить, чтобы исправить и / или предотвратить проблему;
5. проведение работ согласно плану действий
6. оценка стабильности ситемы по результатам выполненных работ.

Рассматриваемые программные средства позволяют вести учет конкретного источника происхождения информации, который инициировал это действие. Документирование источника важный этап в процедуре CAPA, необходимый для последующего исследования проблемы и реализации соответствующего плана действий. Эта информация может поступать как из внешних так и внутренних источников. Примерами источников информации о несоответствиях могут быть: запрос на обслуживание, внутренний аудит качества, жалоба клиента, контроль качества, наблюдение персонала, данные о тенденциях, оценка рисков, мониторинг производительности процесса, обзор управления, анализ режима сбоя. Возможны и другие источники, зависящие от конкретной предметной области.

Одним из основных достоинств системы MasterControl® является способность оценки уровня риска, в зависимости от источника несоответствия и характера самой проблемы. Так например, система может назначить высокий приоритет для отклонений, связанных с задачами безопасности производимого продукта, требуемыми

немедленного проведения мер по исправлению отклонений. Аналогично, система может назначить низкий приоритет для обнаруженного ежемесячного простоя некоторого оборудования.

Следующим достоинством рассмотренных программных систем является наличие развитого пользовательского интерфейса, включающего в себя множество электронных форм, необходимых для ввода данных и просмотра отчетов. Стандартный пользовательский интерфейс системы Infotehna® позволяет вести учет плана действий, включающего в себя общую цель и инструкции по проведению исследования отклонений, несоответствий и инцидентов, а также назначать ответственных лиц, и ожидаемые даты исполнения.

Основными недостатками существующих систем являются то, что все данные о планируемых корректирующих и превентивных действиях генерируются на основе детерминированной статической информации, хранимой в базе данных, взаимодействующей с рассматриваемыми программными приложениями. Однако, основной проблемой рассматриваемых сложных систем является стохастичность процессов возникновения отклонений показателей ключевых параметров от заданных нормативных значений. Другим недостатком является невозможность автоматической идентификации случайных несоответствий с недетерминированными значениями параметров. Также отсутствует в указанных

системах функции автоматического определения признаков оценки стабильности систем после проведения корректирующих и превентивных действий. Таким образом, в рассмотренных системах не в полной мере решены задачи автоматического распознавания несоответствий и определения корректирующих и превентивных действий, необходимых для постоянного контроля ключевых параметров и управления стабильностью системы. В последующих главах будет рассмотрен подход и предложены методы автоматической идентификации случайных несоответствий и определения корректирующих и превентивных действий, основанных на анализе заданной эмпирической выборки данных.

1.4. Классификация объектов и восстановление зависимости по исходной выборке

Одной из основных задач в представленной модели контроля стабильности систем (см. рис. 1.2) является задача классификации множества объектов сложной стохастической системы.

Введем следующие обозначения:

пусть задано $\{x_1, \dots x_l\} \subset \mathrm{X}$ множество объектов и $f_1(x_i), \dots, f_n(x_i)$ множество признаков произвольного объекта x_i, которые в зависимости от принимаемых значений могут быть бинарными $\{0,1\}$, действительными $\{\mathrm{R}\}$ и качественными $\{\mathrm{w}\}$.

Зададим множество $\mathrm{Y} = \{\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_\mathrm{m}\}$ непересекающихся классов $\mathrm{y_i} \neq \mathrm{y_j} | \mathrm{y_i}, \mathrm{y_j} \in \mathrm{Y}$.

Пусть задана некоторая эмпирическая выборка с неизвестной зависимостью $\mathrm{X} \to \mathrm{Y}$. Нам необходимо построить алгоритм (решающую функцию) a, который аппроксимировал бы зависимость $y(x)$ на всем заданном пространстве. Необходимо, чтобы этот алгоритм мог бы классифицировать любой новый объект (признаки объектов) $x \in X' \not\subset X$ путем определения соответствующего класса. Иными словами, алгоритм, “обученный” на эмпирической выборке должен любому новому объекту x поставить в соответствие определенный $\mathrm{y_i} \in \mathrm{Y}$. Это задача восстановления зависимостей по эмпирическим данным.

Таким образом, для построения модели классификации необходимо определить закономерность отображения $\mathrm{X} \to \mathrm{Y}$ и составить соответствующую решающую функцию от x. Если эта закономерность строго подчиняется некоторому известному

распределению и все характеристические признаки измеряются в единой шкале измерений, то здесь применимы известные методы математической статистики построения решающих правил [57-59]. Часто в параметрической функции аргументами являются параметры различных единиц измерения. Это накладывает определенные сложности при выборе соответствующей модели. Данную проблему преодолевают путем первоначального назначения переменных и произвольного выбора конкретной шкалы измерения исходя из соображений удобства. Это в свою очередь приводит к построению некоторой, возможно даже не совсем адекватной модели, но в которой можно манипулировать результатами анализа и изменять размерность переменных. Естественным является постановка задачи с требованием независимости функции и соответствующей модели от выбора конкретной шкалы. Это требование в теории информации и измерений известно как требование адекватности (“meaningfulness”), которое неформально означает, что справедливость суждения (гипотезы, вывода) относительно некоторых аргументов функции не зависит от выбора допустимых шкал и/или их преобразований, в которых измеряются эти аргументы. В работе [60] на основе результатов теории функциональных уравнений исследуется задача характеризации параметрических семейств функций, замкнутых относительно допустимых преобразований шкалы измерения. Актуальность задачи обусловлена тем, что в определенных теоретических моделях необходима конкретизация размерности используемых переменных и типов шкал. Среди шкал по типам допустимых преобразований (классификация С.Стивенса) [61] наибольший практический интерес представляют номинальные, порядковые и интервальные шкалы, а также шкалы отношений.

На рисунке 1.3 представлена иллюстрация задачи, состоящей из двух выборок: эмпирической X и тестовой (контрольной) X', на которых, соответственно, происходят процессы самообучения (саморазвития) и применения алгоритма классификации:

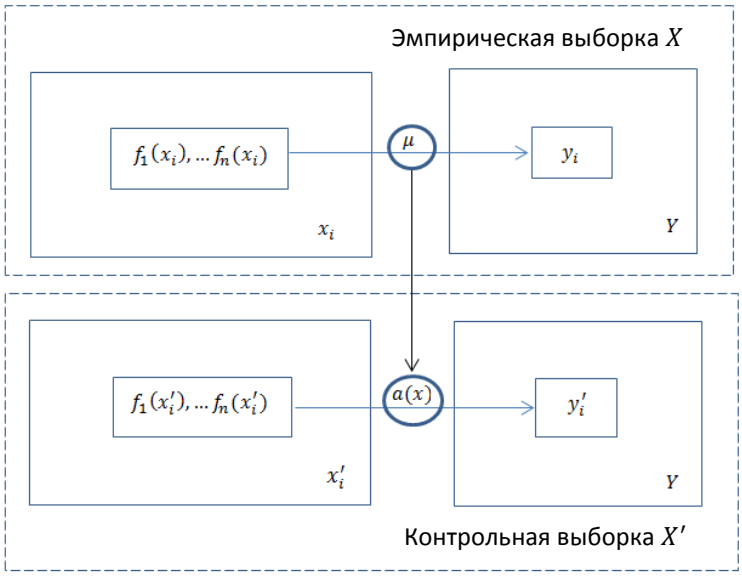


Рисунок 1.3. Классификация контрольной выборки на основе эмпирических данных.

Самонастраиваемый алгоритм $a(x)$, получая на входе новый объект x', отличный от эмпирической выборки, определяет соответствующий этому входному объекту класс из множества Y. Согласно гипотезе компактности, близкие по признакам объекты в основном принадлежат одному и тому же классу. Поэтому, задача состоит в нахождении для нового объекта $x \in X'$ из контрольной выборки X' близких объектов $x \in X$ из обучающей выборки X и согласно закономерности отображения μ выбору соответствующего класса y_i. Этому классу будет соответствовать и новый объект $x \in X'$, согласно гипотезе компактности ($y_i' := y_i$).

1.4.1. Метрики расстояний между объектами эмпирической и контрольной выборок

В предыдущем разделе мы представили задачу классификации объектов на основе оценки их близости к объектам эмпирической выборки. В данном разделе мы рассмотрим алгоритм вычисления этой близости. Данный алгоритм основан на определении расстояния между объектами посредством метода нахождения k ближайших соседей. Объект контрольной выборки относится к тому классу, на котором определены его k ближайших соседей из эмпирической выборки. Метод ближайших соседей заключается в выполнении следующих шагов:

1. вычисление расстояния от объекта контрольной выборки $x' \in X'$ до каждого объекта из обучающей выборки $x_i \in X$;

2. определение подмножества $X^k \subset X$ k объектов эмпирической выборки с минимальными расстояниями R_{min} ;
3. выбор наиболее часто встречающегося класса $y_i \in Y$ в подмножестве X^k ближайших соседей ;
4. установление соответствия объекта x' классу y_i ;

Ниже на рисунке 1.4 представлена иллюстрация метода классификации на основе определения k “ближайших соседей”.

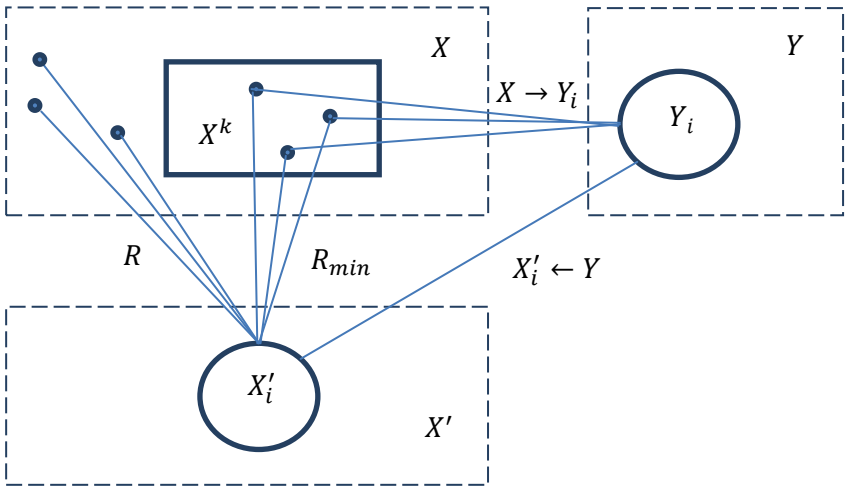


Рисунок 1.4. Классификация методом k “ближайших соседей”.

Введем следующие обозначения: $x_i \in X$, $x'_j \in X'$- объекты эмпирической (обучающей) и контрольной (тестовой) выборок, соответственно, с признаками $f(x_i) = \{f_1(x_i), \dots f_n(x_i)\}$ и $f(x'_j) = \{f_1(x'_j), \dots f_n(x'_j)\}$. Тогда, расстояние между объектами x_i и x'_j согласно евклидову расстоянию можно вычислить по формуле

$$\rho(x_i, x'_j) = \sqrt{\sum_{k=1}^{n} (f_k(x_i) - f_k(x'_j))^2}$$

Приведем еще несколько общеизвестных метрик измерения расстояний между объектами, на которых базируется концепция нахождения ближайших соседей [59] :

метрика Минковского (манхэттенское расстояние): $\rho(x_i, x'_j) = \sum_{k=1}^{n} |f_k(x_i) - f_k(x'_j)|$;

расстояние Чебышева: $\rho(x_i, x'_j) = \max(|\sum_{k=1}^{n} |f_k(x_i) - f_k(x'_j)|)$;

степенное расстояние $\rho(x_i, x_j) = \sqrt[r]{\sum_{k=1}^{n} (f_k(x_i) - f_k(x'_j))^p}$

Введем дополнительно следующие обозначения:

$x_i(x'_j) \in X$ – i-ый сосед объекта $x'_j \in X'$, где X'- контрольная выборка.

Отсортируем расстояния объектов следующим образом:

$\rho(x_j', x_1(x_j')) \leq \rho(x_j', x_2(x_j')) \leq \cdots \rho(x_j', x_l(x_j'))$, тогда оценка близости объекта x_j' к произвольному классу $y \in Y$ можно представить выражением:

$\sum_{i=1}^{l}[y_i\,(x_j') = y]|w(i, x_j')$, где $w(i, x_j')$- числовое значение, характеризующее степень важности i-го соседа.

Согласно гипотезе компактности, объект будет ближе к тому классу, к которому в основном относятся его ближайшие соседи, т.е. метод классификации по значениям метрик можно записать в следующем виде:

$a(x_j', X^l) = \arg\max_{y \in Y} \sum_{i=1}^{l}[y_i\,(x_j') = y]|w(i, x_j')$.

В частном случае, если в качестве весов $w(i, x_j')$ принять значение 1 только для самого ближайшего соседа, а для всех остальных 0, то приведенное выше выражение формально будет представлять расстояние только до одного ближайшего соседа.

Далее, рассмотрим задачу оценки точности алгоритма классификации. Естественно, возникает задача оценки точности алгоритма классификации, определяющего пару <“Объект”,“Класс”> в тестируемой (контрольной) выборке. При этом, погрешности алгоритма могут быть вызваны как неточностью выбора значений параметров функции отображения, так и неполнотой обучающей выборки. Для проведения оценки алгоритма введем понятие функции потерь $\varepsilon(a, x)$, с помощью которой будем определять величину погрешности $\varepsilon(a, x) = [a(x) \neq y(x)]$.

Примем следующее, если значение $\varepsilon(a, x) = 1$, то имеет место ошибка классификации.

Если просуммировать по всем объектам $\varepsilon(a, x)$, то получим общее число ошибок алгоритма $a: X \to Y$ на обучающей выборке, которую так же называют эмпирическим риском или функционалом качества алгоритма на X^l:

$$Q(a, x) = \frac{1}{l}\sum_{i=1}^{l} \varepsilon(a_i, x_i)$$

Задача сводится к минимизации эмпирического риска $\mu(X^l) = \arg\min_{a \in A} Q(a, X^l)$. Нам необходимо выбрать тот алгоритм $a: X \to Y$, при котором значение Q минимально.

Во второй главе будет представлен метод определения оптимального алгоритма с точки зрения минимизации эмпирического риска.

1.4.2. Кластеризация объектов

В подразделе 1.4.1 мы ввели в рассмотрение эмпирическую выборку для оценки близости новых объектов и вычисления расстояния между объектами. Однако, на практике не всегда представляется целесообразным рассматривать сходство нового объекта со всеми объектами эмпирической выборки. Мы рассмотрим вопросы выбора критерия разбиения эмпирической выборки на отдельные кластеры. Пусть заданы объекты эмпирической выборки и расстояние между ними. Необходимо их разнести на компактные группы путем определения множества кластеров Y и составления алгоритма отображения $a: X \to Y$, при котором в каждом кластере будут находиться объекты максимально близкие друг к другу, а объекты разных кластеров максимально удаленными друг от друга. Далее, каждый из кластеров можно рассматривать как единый объект и применять к нему уже известные методы классификации и прогнозирования. Упрощение модели с применением метода кластеризации эмпирической выборки возможно путем нахождения типичного представителя для каждого кластера и применения известных расчетов к произвольному кластеру как объекту.

В основе представленных выше методов лежит концепция компактности множеств объектов. Одной из основных целей кластеризации является метод построения иерархии объектов. Иерархическая кластеризация – это один из основных способов проведения систематизации множества объектов. На рисунке 1.5 представлены визуально несколько вариантов расположения объектов по кластерам.

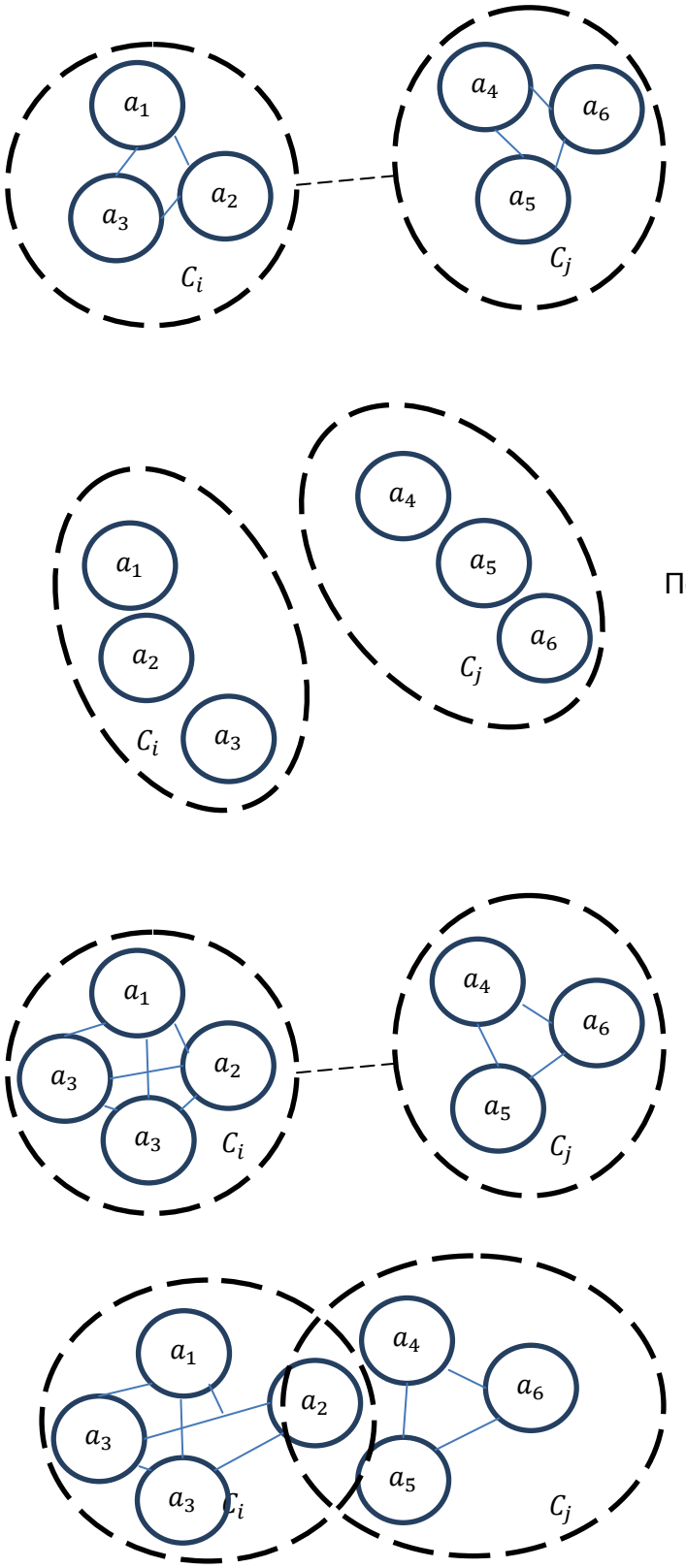


Рисунок 1.5. Графическое представление типов кластерных структур

Как видно из рисунка, первый вариант – это выраженные внутрикластерные и межкластерные расстояния, второй – ленточные (последовательные), третий -

кластеры с выраженными объектами в центре. Результат кластеризации зависит как от выбора функции расстояния (критериев вычисления близости) так и от их нормировки. Результаты кластеризации могут быть достаточно чувствительны к выбранным методам: используя взвешенную эвклидову метрику, но назначая различные значения весовым коэффициентам возможно получение различных результатов кластеризации.

Приведем методы построения иерархической кластеризации [36],[62]:
для начала примем, что все объекты образуют элементные кластеры, т. е. сколько объектов, столько и кластеров. Возьмем два самых ближайших друг другу объектов и объединим их в один кластер и далее мы будем их считать нераздельными и рассматривать как один объект или один кластер. Для этого нового кластера мы должны будем подсчитать расстояние до других кластеров, чтобы опять были бы нам известны все попарные расстояния. Если в начале итерации у нас были попарные расстояния между объектами, то далее нам придется поддерживать попарные отношения между существующими и вновь создаваемыми кластерами, т. е. перенести функцию расстояний с объектов на кластеры. Далее, процедура повторяется $(l-1)$ раз, где l- размер выборки (количество объектов). Следующим шагом находим во всем множестве кластеров опять два ближайших. Это должна быть функция расстояния между двумя кластерами. Объединяем эти кластеры в один кластер, в результате чего у нас на один кластер становится меньше. Аналогично для нового кластера мы должны вычислить расстояние между этим кластером и всеми остальными. Расстояние можно подсчитать, в частности, методом ближайшего соседа. Разумно также использовать расстояние до дальнего соседа. Если точки представляют собой элементы линейного векторного пространства, которые можно складывать как векторы или умножать на число, то мы можем найти центр каждого кластера.
Введем следующие обозначения:
$R(C_i, C_j)$- расстояние между кластерами, полученными на n –ом шаге итерации,
$C_i = C_{i1} + C_{i2}$ – новый кластер, объединяющий кластеры C_{i1} и C_{i2} (на $n-1$ шаге),
$R(C_{i1}, C_j), R(C_{i2}, C_{ij}), R(C_{i1}, C_{i2})$- попарные расстояния между точками C_{i1}, C_{i2}, C_j .
Тогда, формулы классификации можно записать следующим образом:

$$R(C_{i1} \cup C_{i2}, C_j) =$$

$$\lambda_1 R(C_{i1}, C_j) + \lambda_2 R(C_{i2}, C_{ij}) + \beta R(C_{i1}, C_{i2}) + \gamma |R(C_{i1}, C_j) - R(C_{i2}, C_j)|, \quad (1)$$

где $\lambda_1,\ \lambda_2,\ \beta,\ \gamma$ - числовые параметры.

На рисунке 1.6. приведена графическая интерпретация формулы

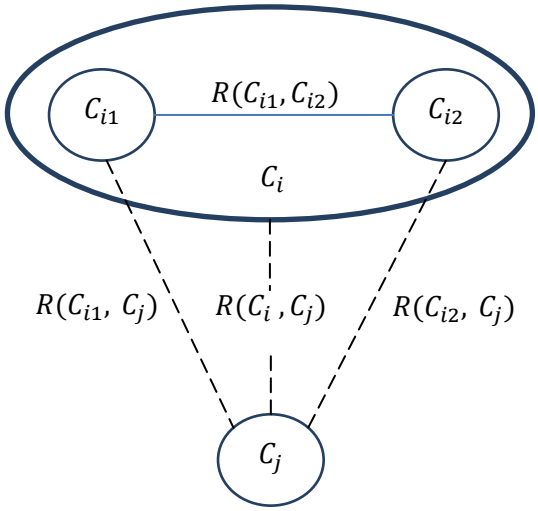


Рисунок 1.6. Расстояния между кластерами $(n-1)$ *и* (n) *итерациями.*

Приведем частные случаи формулы (1):

- расстояние “ближнего соседа” ($\lambda_1 = \ \lambda_2 = \ \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$):

$$R_1(C_i, C_j) = \min_{c(C_i)\in C_i, c(C_j)\in C_j} \rho(c_i, c_j)\ , \qquad (2)$$

где $c(C_i), c(C_j)$- точки (объекты) кластеров C_i, C_j, $\rho(c_i, c_j)$- расстояние между этими точками

- **расстояние “удаленного соседа”** ($\lambda_1 = \ \lambda_2 = \ \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}$) :

$$R_2(C_i, C_j) = \max_{c(C_i)\in C_i, c(C_j)\in C_j} \rho(c_i, c_j) \qquad (3)$$

На рисунке 1.7 представлены графически расстояния между ближними и удаленными точками разных множеств:

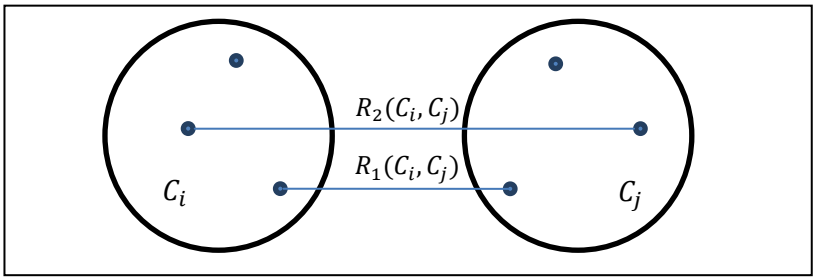


Рисунок 1.7. Расстояния между ближними и удаленными точками.

- **среднее расстояние между всеми точками** двух кластеров

$$(\lambda_1 = \frac{|C_{i1}|}{|C_j|}\ \lambda_2 = \frac{|C_{i2}|}{|C_j|}, \beta = \gamma = 0):$$

$$R_3(C_i, C_j) = \frac{1}{|C_i||C_j|}\sum_{c(C_i)\in C_i}\sum_{c(C_j)\in C_j)} \rho(c_i, c_j), \qquad (4)$$

На рисунке 1.8 представлено графически попарное вычисление всех расстояний между всеми точками двух разных множеств.

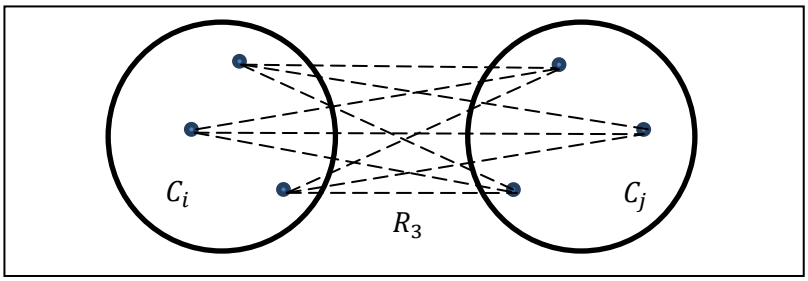


Рисунок 1.8. Среднее расстояние между точками кластеров.

Представленные случаи можно также описать следующей формулой Колмогорова, определяющей степень сходства объектов [63]:

$$K_\eta([C_i, C_j], c) = \left(\frac{n_i c(C_i,c)^\eta + n_j c(C_j,c)^\eta}{n_i + n_j}\right)^{\frac{1}{2}} \qquad (5)$$

где, c - объекты, у которых определяется сходство (близость) с кластерами C_i и C_j, n_i, n_j- количество объектов в классах C_i и C_j соответственно.

- расстояние между центрами множеств ($\lambda_1 = \frac{|C_{i1}|}{|C_j|}, \lambda_2 = \frac{|C_{i2}|}{|C_j|}, \beta = -\lambda_1\lambda_2,\ \gamma = 0$)

$$R_4(C_i, C_j) = \rho^2\left(\sum_{c(C_i)\in C_i}\frac{c(C_i)}{C_j}\sum_{c(C_j)\in C_j)}\frac{c(C_j)}{C_j}\right) \qquad (6)$$

Так как точки множеств представляют собой элементы линейного вектора, то для таких множеств складывая эти элементы можем найти центр (весовое значение)

каждого множества, рассчитав среднеарифметическое значение между точками. На основе этих среднеарифметических значений можно получить расстояние между двумя множествами. На рисунке 1.9 представлено расстояние между средне-арифметическими значениями множеств [36],[64].

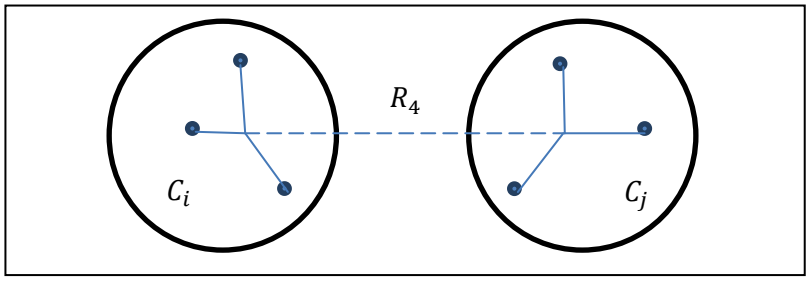


Рисунок 1.9. Расстояние между центральными позициями множеств.

- расстояние Уорда ($\lambda_1 = \frac{|C_j|+|C_{i1}|}{|C_j|+|C_i|}$, $\lambda_2 = \frac{|C_j|+|C_{i2}|}{|C_j|+|C_i|}$, $\beta = \frac{-|C_j|}{|C_j|+|C_i|}$, $\gamma = 0$)

$$R_5(C_i, C_j) = \frac{|C_i||C_j|}{|C_i|+|C_j|}\rho^2\left(\sum_{c(C_i)\in C_i}\frac{c(C_i)}{C_j}\sum_{c(C_j)\in C_j}\frac{c(C_j)}{C_j}\right) \quad (7)$$

здесь и дальше под обозначением $|C_i|$, $|C_j|$ будем понимать мощности соответствующих множеств.

Все перечисленные расчеты расстояний представляют собой частные случаи формулы Ланса-Уильямса (1) при заданных соответствующим образом коэффициентах $\lambda_1, \lambda_2, \beta, \gamma$. Проиллюстрируем на следующем графике объединение точек (объектов) в группы (кластеры) согласно вычислениям расстояний “ближнего соседа”.

Согласно иерархической классификации последовательно объединяются в вложенные группы точки исходного множества.

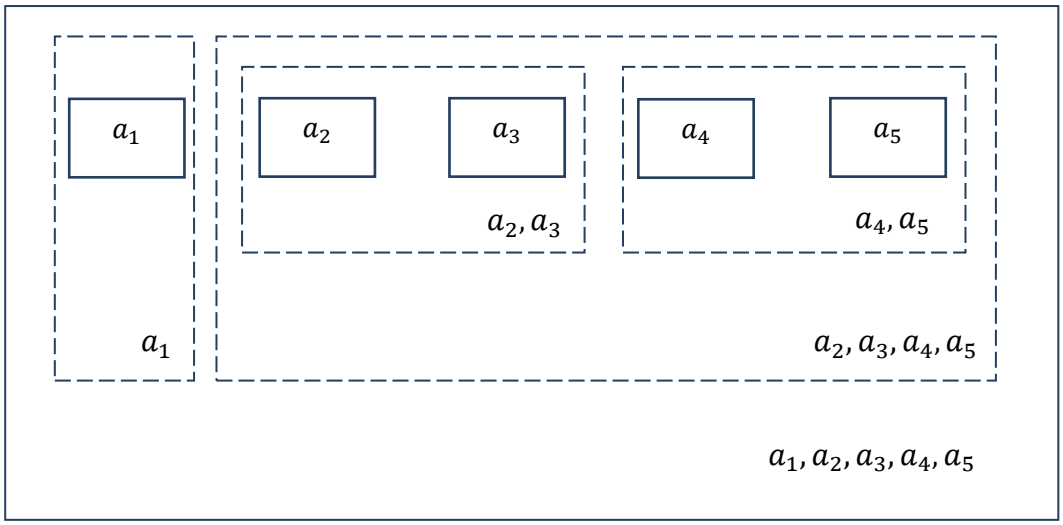


Рисунок 1.10. Объединение объектов и кластеров согласно расстояниям ближайших соседей.

Как видно из рисунка, на первом шаге итерации согласно расстояниям ближних соседей объединяются в пары самые близкие точки a_2 и a_3, a_4 и a_5. На следующем шаге эти два подмножества объединяются в подмножество (a_2, a_3, a_4, a_5). На завершающем шаге происходит объединение созданного кластера с существующим одноэлементным кластером a_1 в новый кластер a_1, a_2, a_3, a_4, a_5.

Представленные пошаговые объединения объектов и кластеров можно изобразить в виде следующего графика-дендрограммы [65]

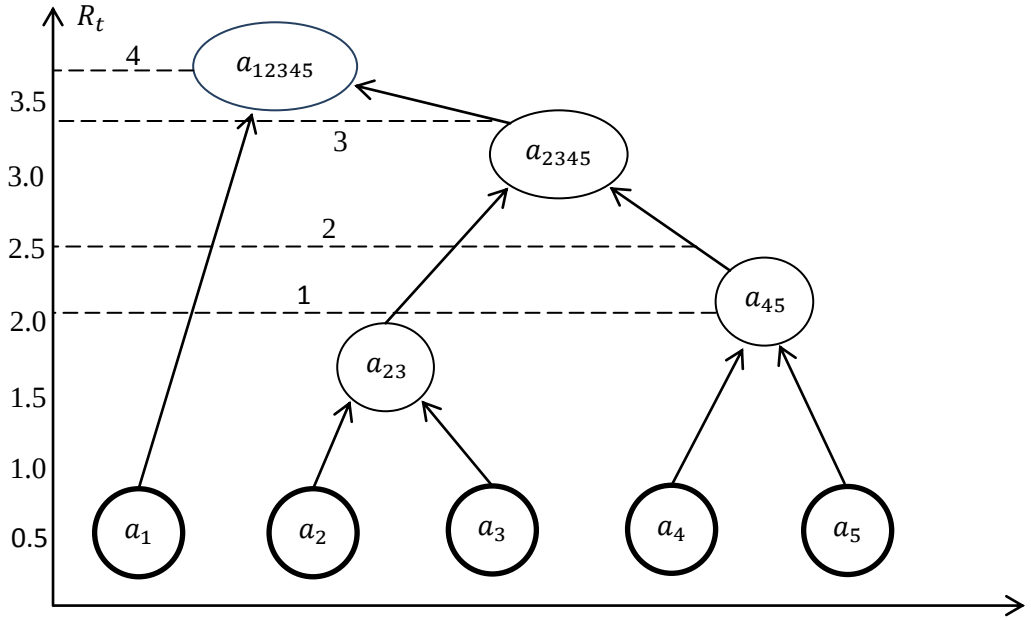


Рисунок 1.11. Дендрограмма процесса последовательного объединения точек в группы.

Линии 1,2,3,4 показывают количество кластеров (пересекающие линию дуги) на данном этапе итерации. При небольших значениях R_t мы имеем больше кластеров, при увеличении значения R_t число кластеров уменьшается. Оптимальным является то разбиение при котором $\Delta = (R_t - R_{t-1}) \to max$, как видно из графика в рассмотренном подмножестве из пяти элементов наибольшая разница достигается в итерации при переходе с 2-го шага к 3-му, когда мы получаем два кластера, элементы которых максимально разнесены по соответствующим подмножествам. Заметим, что при построении дендрограммы важно, чтобы на любом шаге кластеризации выполнялось бы свойство монотонности $(R_t - R_{t-1}) \geq 0$, в противном случае будет иметь место хаотичное разбиение на подмножества. Согласно теореме Миллигана [65] кластеризация монотонна, если для параметров $\lambda_1, \lambda_2, \beta, \gamma$ выполняются условия:

$$\lambda_1 \geq 0, \lambda_2 \geq 0, \ \lambda_1 + \lambda_2 + \beta \geq 1, \min(\lambda_1, \lambda_2) + \gamma \geq 0$$

При выполнении условия монотонности в дендрограмме не будут самопересечения. Исключением могут быть дендрограммы, построенные на расстояниях между центрами кластеров.

Необходимо учесть, что графики, построенные по разным функциям вычисления расстояния до ближних точек могут отличаться друг от друга, если не совсем точно определены критерии кластеризации. При выборе критериев, которые хорошо разбивают точки на подмножества, все рассмотренные выше функции вычисления расстояний дадут приблизительно одинаковые дендрограммы.

Одной из проблем реализации алгоритмов вычисления межкластерных расстояний является задача нахождения ближайших кластеров $R(C_i, C_j) \to min$. Попарное сравнение всех вычисленных расстояний довольно длительный процесс при большом количестве кластеров. Поэтому, иногда прибегают к реализации функции перебора только близких пар, расстояние между которыми ограничивается некоторым заданным значением δ [36]:

$$R(C_i, C_j) \to \min_{R(C_i, C_j) < \delta}$$

Имеется много методов выбора значений δ. В частности, при наличии очень большой выборки задается некоторое количество или процент элементов с выборки, который можно выбрать случайно, например, программой-генератором случайных чисел. И последовательно вычислить для всех пар расстояния между точками. Наименьшее из них выбирается в качестве δ . Далее, мы можем из всей выборки рассматривать только те объекты, расстояние между которыми не превышает δ [36].

Приведем пошаговый алгоритм кластеризации:

1. Инициирование начальных значений кластеров C, равных выборке объектов (точек) множества A:

   $\mathrm{N_C} = \mathrm{N_A}$, где $\mathrm{N_C}, \mathrm{N_A}$ – соответственно количества кластеров и объектов.

   Также, совпадают расстояния между кластерами с расстояниями между точками этих кластеров:

   $\mathrm{R}(\mathrm{C_i}, \mathrm{C_j}) = \rho(\mathrm{c}(\mathrm{C_i}), \mathrm{c}(\mathrm{C_j}))$.

2. Вычисление одним из приведенных выше методов расстояния между кластерами и нахождение пары $\mathrm{C_i}, \mathrm{C_j}$ с наименьшим расстоянием.

   Если значение $R(C_i, C_j) \leq \delta$, то оно присваивается значению итерации R_t :

   $R_t := \min\ \mathrm{R}(\mathrm{C_{it}}, \mathrm{C_{jt}}) \leq \delta$, где t- номер итерации.

3. Объединение кластеров $C_i,\ C_j$ в новый кластер C_{ij} .

4. Повторение пунктов 2 и 3 для нового кластера методом так называемого редуктивного алгоритма Ланса-Уильямса [59].

1.5. Обоснование цели работы и формулирование задач исследования

Обобщая результаты исследований сложных стохастических систем, приведенные в главе 1, приходим к следующему выводу: автоматизация процессов обнаружения и классификации несоответствий возможна путем рассмотрения эмпирической выборки и вычисления близости контрольного объекта к элементам этой выборки. Под объектом мы будем понимать элемент подсистемы (см. раздел 1.1) с вектором значений признаков, регистрируемых в некоторый момент времени. Добавим также, что без надлежащих методов и программных средств автоматизации не представляется возможным проведение оперативного анализа большого числа динамических показателей, оценки стабильности системы и принятие своевременного решения о выполнении процедур CAPA.

Основная цель диссертационной работы заключается в проектировании стохастической модели управления стабильностью сложных систем и разработке на ее основе информационной системы идентификации несоответствий и управления корректирующими и превентивными действиями.

Для достижения данной цели необходимо решение следующих задач:

1. Определение структуры и особенностей функционирования сложных стохастических систем рассматриваемого класса. Для решения данной задачи необходимо представить формальное описание компонент системы, взаимосвязей между объектами (признаками объектов) и источниками входных сообщений о возможных несоответствиях ключевых показателей заданным значениям.
2. Построение модели управления процедурами CAPA, реализующей функции регистрации сообщений и обнаружения несоответствий путем контроля ключевых показателей. Для построения данной модели необходимо определение решающего правила отображения этих несоответствий в множество установленных классов, определяющих основные характеристики несоответствий и отклонений по отношению к заданным нормативным

значениям. На основе метода классификации сообщений необходимо разработать алгоритм определения последовательности проведения корректирующих и превентивных действий для возвращения системы в стабильное состояние.

3. Проектирование методов автоматизации процедур CAPA как основы для построения информационной системы контроля стабильности. Решение данной задачи для рассматриваемого класса стохастических систем возможно путем построения эмпирической (обучающей) выборки и разработки метода оценки близости контролируемого объекта к этой выборке. Данный метод позволит решить задачу классификации несоответствий путем обнаружения схожих объектов между контрольной и эмпирической выборками.
4. Разработка информационной системы контроля стабильности сложных стохастических систем и управления процедурами CAPA. В основе данной разработки должны быть методы и алгоритмы, позволяющие реализовать следующие функции:

- обнаружение и идентификация несоответствий ключевых параметров заданным нормативным стандартам;
- классификация несоответствий и определение корректирующих и превентивных действий, направленных на устранение обнаруженных отклонений и несоответствий;
- оценка степени влияния обнаруженных несоответствий на валидированные процедуры, оборудование, процессы и системы до и после выполнения процедур CAPA;
- протоколирование и документирование всех этапов CAPA, от обнаружения отклонений и несоответствий до планирования, исполнения действий CAPA и оценки их эффективности.

Выводы к главе 1

1. Дано определение сложных стохастических систем, целевая функция которых заключается в управлении ключевыми параметрами, характеризующими стабильность состояния систем.
2. Представлены методы моделирования процессов управления стабильностью сложных стохастических систем.
3. Рассмотрен алгоритм классификации объекта на основе схожести с объектами эмпирической выборки. В основе алгоритма – метод вычисления расстояния до k ближайших объектов эмпирической выборки.
4. Рассмотрены методы оптимизации эмпирической выборки посредством алгоритмов объединения объектов в кластеры и построения их иерархических групп с помощью функционала определения близости расстояния между объектами в заданном многомерном метрическом пространстве.
5. Обоснованы цель и задачи диссертационной работы, включающих в себя построение стохастической модели и программных средств управления стабильностью сложных систем.

ГЛАВА 2: МОДЕЛЬ КЛАССИФИКАЦИИ НЕСООТВЕТСТВИЙ И УПРАВЛЕНИЯ КОРРЕКТИРУЮЩИМИ ДЕЙСТВИЯМИ

Метрический подход в задачах классификации основан на гипотезе компактности, согласно которой схожие объекты в основном находятся в одном компактно локализованном подмножестве (классе) пространства объектов [66],[67]. Если каждый объект описывается k признаками, то он может быть представлен как точка в k-мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние между их признаками (сходство по степени близости признаков).

Под терминами “несоответствие”, “отклонение” или “инцидент” мы будем понимать объекты x_i, которые определены в некотором многомерном метрическом пространстве и имеют множество признаков, характеризующих степень сходства этих объектов между собой, но при этом, значения этих признаков выходят за пределы заданных допустимых значений. Под метрическим пространством, согласно классическому определению, будем понимать такое множество, в котором каждой паре объектов $(x_i, x_j) \in X$ ставится в соответствие некоторая числовая функция $\rho(x_i, x_j)$, называемая расстоянием между x_i и x_j , удовлетворяющая следующим аксиомам метрических пространств [66],[67] :

$\rho(x_i, x_j) \geq 0$, для любых x_i и x_j;

условие тождественности: $\rho(x_i, x_j) = 0$, если $x_i = x_j$;

условие симметричности: $\rho(x_i, x_j) = \rho(x_j, x_i)$, для любых $x_i, x_j \epsilon X$;

условие неравенства расстояний (аксиома треугольника): $\rho(x_i, x_k) + \rho(x_k, x_j) \geq \rho(x_i, x_j)$.

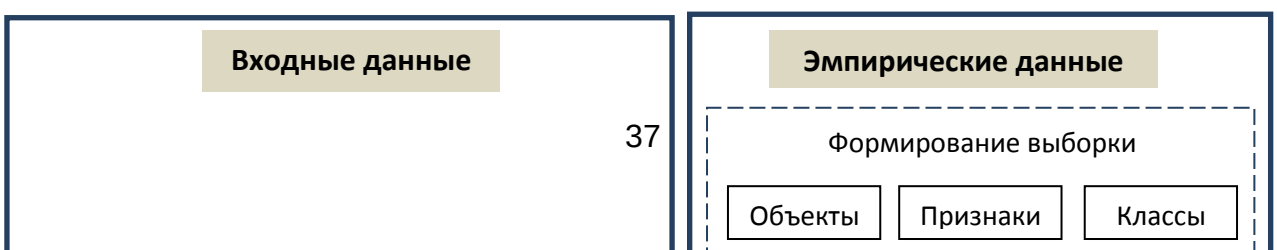
В данной главе мы рассмотрим алгоритмы и методы автоматизации процедур CAPA, включающие в себя решение следующих основных задач:

- определение источников входных сообщений, генерирующих информацию о несоответствиях;

- идентификацию и классификацию несоответствий;
- генерацию отчета, включающего в себя список корректирующих и превентивных действий для управления стабильностью системы.

2.1. Источники входных данных. Обработка сообщений о несоответствиях

В первой главе (см. рисунок 1.2.) мы показали, что основным источником информации о ключевых показателях контроля стабильности являются различные учетные системы и генерирующие ими корпоративные данные. Это могут быть разрозненные или консолидированные базы данных, которые необходимо привести к некоторому единому формату для дальнейшей их обработки и идентификации. Ключевыми показателями являются также заключение аудитора, сообщение от персонала о наблюдаемых несоответствиях и отклонениях, информация от потребителей об обнаруженных недостатках в продукции и предоставляемых услугах. Заключение аудитора представляет собой набор электронных форм, содержащих в основном формализованные поля для заполнения в строгом соответствии с требованиями определенных нормативных документов. Во многих отраслях также действуют строгие правила к оформлению сообщений (заполнению форм) о проблемах качества продукции. Так, в фармацевтической отрасли жалобы пациентов на побочные явления, наблюдаемые в период приема лекарственных препаратов должны быть оформлены надлежащим образом согласно требованиям фармаконадзора. Источниками информации о наблюдаемых несоответствиях могут быть также сотрудники, которые, согласно утвержденным регламентам вводят данные о ключевых производственно-управленческих показателях как в процессах планового мониторинга так и во время случайного обнаружения каких-либо несоответствий, проблем или инцидентов. Ниже представлена структурная схема классификации несоответствий, которая содержит указанные источники сообщений:



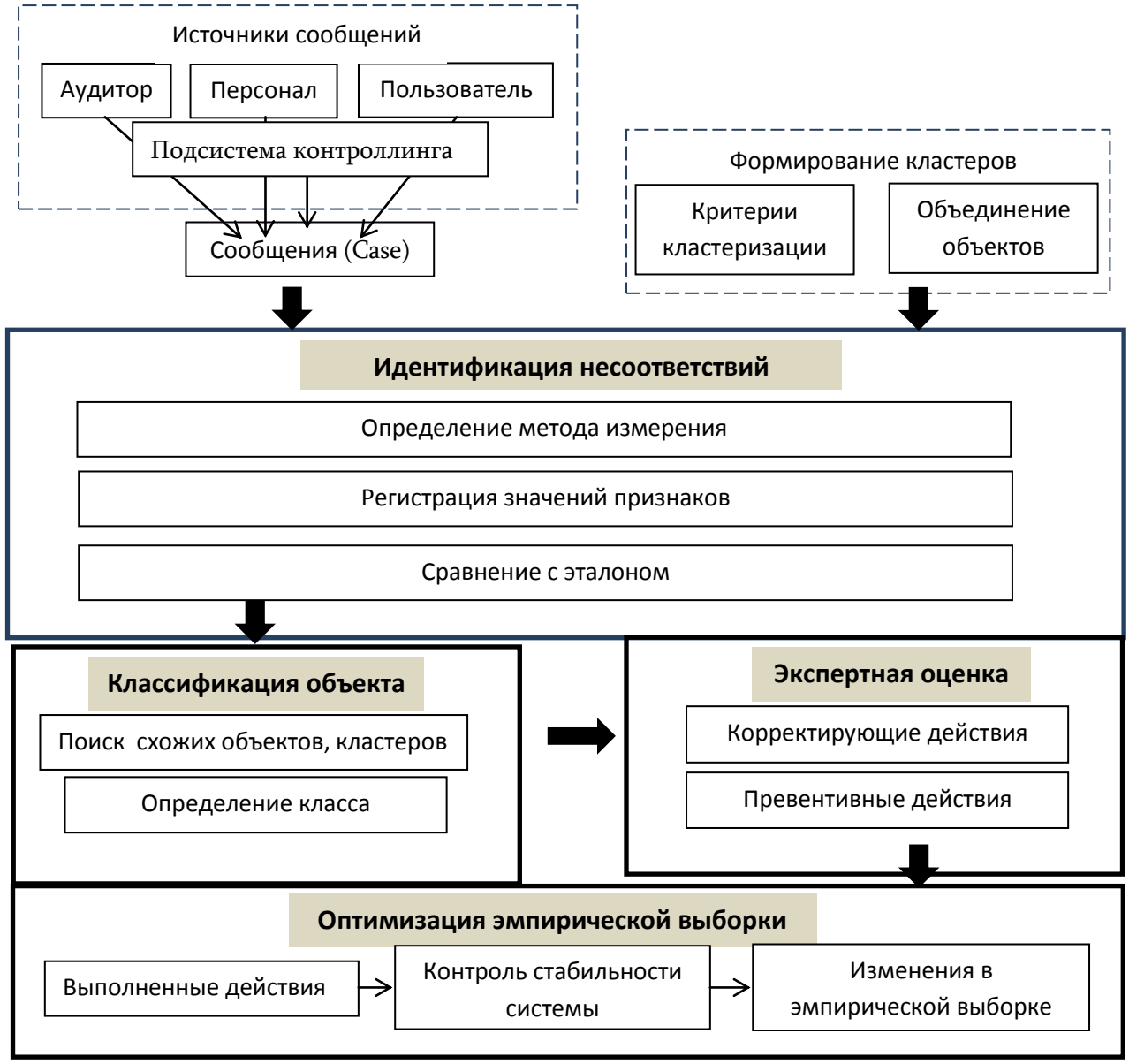


Рисунок 2.1. Структурная схема модели обработки сообщений, идентификации несоответствий, определения класса объектов и соответствующих процедур CAPA.

Актуальной задачей в приведенной модели является сбор и обработка разрозненных данных. Решение данной задачи возможно путем реализации методов интеграции процессов, систем (приложений) и данных. Используя системные службы и серверные приложения можно обеспечить файловый доступ посредством удаленного вызова процедур и тем самым обеспечить **интеграцию на уровне приложений и процессов**. Для этого необходим общий протокол некоторого унифицированного формата передачи информации. **Интеграция на уровне данных** может быть достигнута за счет создания базы данных совместного доступа, при котором все функции обработки данных осуществляются как клиентскими

программами, так и серверными приложениями. В тех случаях, когда не представляется возможным интеграция без изменения кода в существующих системах и проведение этих изменений сопряжено с определенными сложностями, то можно прибегнуть к процедуре репликации данных. В таком случае, дальнейшая обработка информации будет проводиться на базе данных, полученной в результате репликации. С учетом приведенных методов, интеграционная шина (см. рис. 1.2) будет представлять приложения на основе сервисов SOAP/REST с форматами данных XML, JSON, которые будут переводить получаемые с гетерогенных источников информацию в единый унифицированный формат Dynamic Objects:

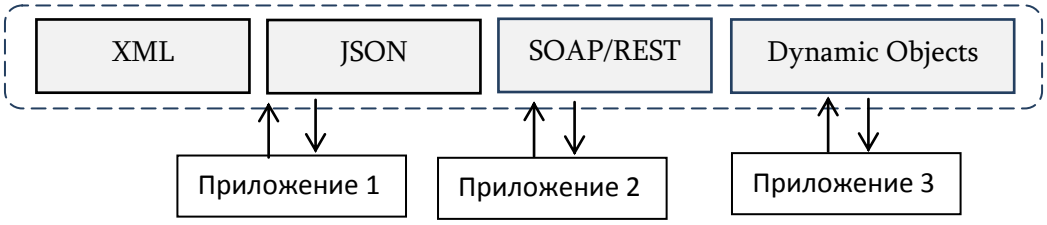


Рисунок 2.2. Структура интеграционной шины.

2.2. Метод идентификации несоответствий в сложных стохастических системах

Задача идентификации несоответствий является следующим этапом реализации методологии CAPA как основы для проведения исследования и полной оценки обнаруженных проблем с целью определения необходимых действий, направленных на обеспечение стабильности рассматриваемых систем. Как было отмечено в предыдущих разделах источниками сообщений о несоответствиях могут быть самыми различными: запрос персонала на обслуживание рабочего оборудования, датчики устройств, внутренний/внешний аудит, результат лабораторного теста ингредиента, жалоба потребителя и др.

Вышеприведенные источники условно можно разбить на две группы:

- источники в запланированных процессах мониторинга (анализ режимов функционирования, контроль производительности процесса, данные о тенденциях, контроль/аудит качества);
- источники случайного обнаружения несоответствий, инцидентов и отклонений.

Рассмотрим процесс исследования и идентификации несоответствий путем оценки признаков контролируемых компонент системы. Для указанных компонент

необходимо определить ключевые параметры и диапазон допустимых значений. Также необходимо для критически важных количественных показателей задать допустимое число последовательно превышающих или последовательно недостающих (хотя и в пределах допустимого диапазона) значений по отношению к нормативному (эталонному) показателю. При этом, если выполняется условие $f_t(x) < f_{t+1}(x)$, где $f_t(x)$ $f_{t+1}(x)$ – значения признака f объекта x в моменты t и $(t+1)$, то возможно это свидетельствует о некоторой негативной тенденции.

Метод идентификации предполагает выполнение следующих 4-х шагов:

Шаг 1. Для каждой компоненты по каждому характеристическому параметру с количественными показателями должны быть заданы допустимые значения или диапазон допустимых значений.

Например, ниже приведена гистограмма (контрольная карта) для компоненты “Машина дозировки и упаковки” (группа “Оборудование”), в которой по горизонтали представлено количество производственных процессов, а по вертикали – показатель давления воздуха. UCL, LCL – верхние и нижние, соответственно, допустимые пределы значений параметра.

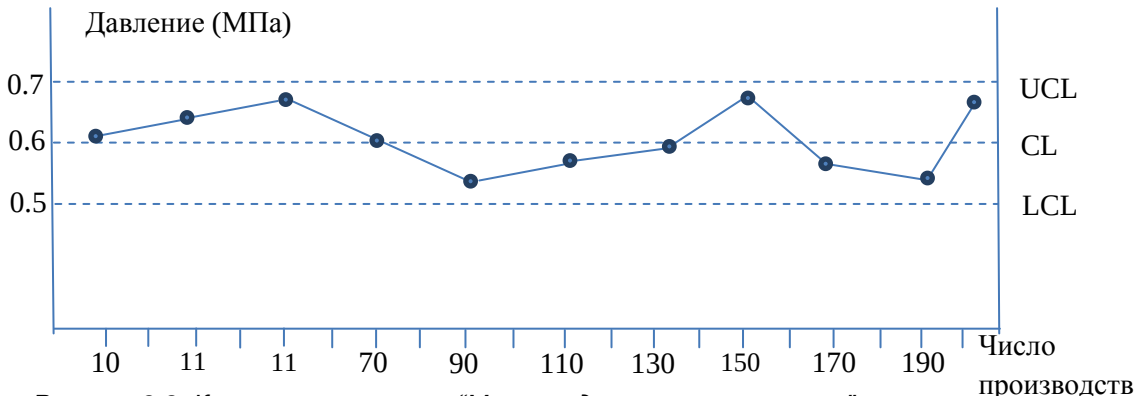


Рисунок 2.3. Контрольная карта “Машина дозировки и упаковки”.

Шаг2. Необходимо определить множество сочетаний значений нескольких параметров при котором можно утверждать о имеющемся несоответствии и отклонениях.

Шаг3. Для определения тенденций необходимо для каждой компоненты задать признаки нарушения стабильности (управляемости) показателей. Например, можно установить предельное число N как количество допустимых точек, подряд

находящихся по одну сторону от средней линии (CL) и одновременно с убывающим или возрастающим значением. Тогда превышение этого количества для заданной компоненты свидетельствует о тенденции к возможному нарушению стабильности в будущем.

Шаг 4. Необходимо задать для каждого контролируемого параметра период и способ проведения мониторинга.

Для нечисловых признаков в качестве допустимых значений рассматривается заданное множество термов (слов) и контроль ключевого показателя будет осуществляться на проверку его совпадения с заданным множеством термов.

Для формального описания процесса обнаружения и идентификации несоответствий зададим следующие процедуры:

1. Формирование структуры исходных данных:

- определение объектов и их ключевых параметров;

- назначение диапазона допустимых значений параметров и периодов мониторинга.

2. Сбор данных и регистрация источников поступления сообщений о несоответствии

- идентификация источников информации и регистрация показателей ключевых параметров;

- анализ данных и классификация результатов регистрации.

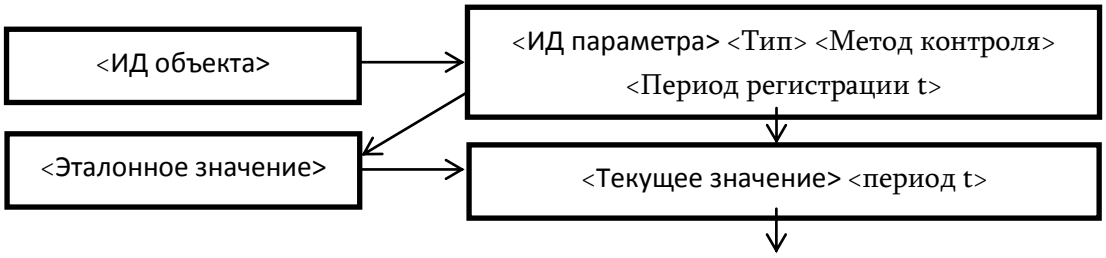
3. Идентификация несоответствий

- протоколирование информации о несоответствии и нарушениях стабильности, источнике сообщения, даты наблюдения.

4. Идентификация (прогнозирование) тенденций

- протоколирование значений наблюдаемых параметров.

Ниже представлены упрощенная и развернутая схемы процессов идентификации несоответствий**:**



<ID: эталон> <ID: текущее значение> <Результат>

Рис. 2.4. Схема идентификации несоответствий.

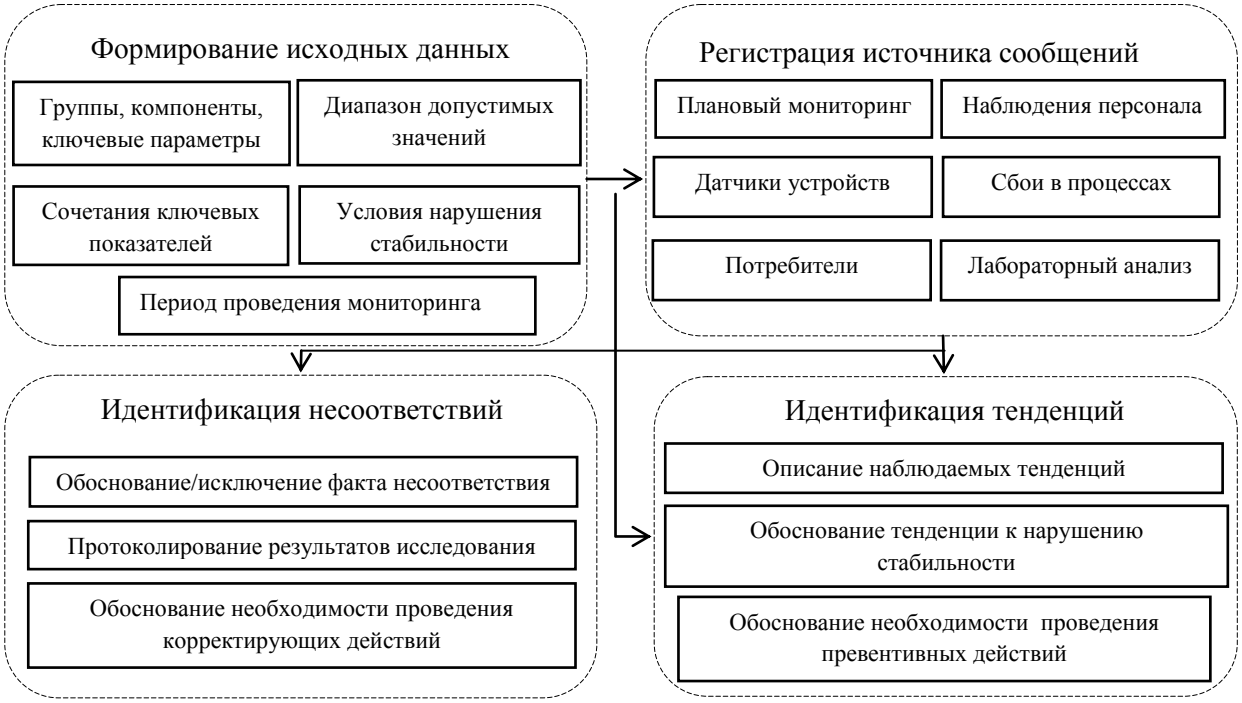


Рисунок 2.5. Структурно-функциональная схема обнаружения и идентификации несоответствий.

Для заданного контролируемого объекта определяется множество признаков (ключевых параметров) и их атрибутов: тип, метод и период регистрации. Далее, на основе заданных атрибутов определяется и регистрируется значение выбранного признака, сравнивается с эталоном и фиксируется результат: подтвержден/опровергнут, факт несоответствия.

2.3. Оценка модели классификации. Метод кластеризации объектов по качественным признакам на основе предикатных выражений

Оценка вероятностных моделей - это одна из основных задач машинного обучения и интеллектуального анализа данных. Перекрестные проверки часто используют как метод оценки качества модели классификации [68-74]. Основным подходом является использование случайно выбранного множества, которое разбивают на непересекающие эмпирическую (обучающую) выборку, на которой будем настраивать алгоритм и контрольную выборку, на которой будем проверять нашу модель и оценивать точность классификации. Однако, если мы имеем

некоторый ограниченный набор данных и хотим на этом наборе оценить качество модели классификации, то к сожалению, из-за неполноты данных не будет обеспечена точность этой оценки. В этом случае изучение модели на имеющемся наборе исходных данных проводится путем многократного использования повторной перекрестной проверки. Так как сами перекрестные проверки содержат достаточно большую дисперсию, то в результате такой многократной проверки можно наблюдать усреднение данных и уменьшение дисперсии результатов. Однако, как показывают результаты статистических расчетов, основанных на повторной подвыборке, многократные перекрестные проверки не всегда гарантируют уменьшения дисперсии.

Проблема моделирования взаимосвязей параметров, характеризующих процессы в системах, обусловлены прежде всего сложностью их структуры, неполнотой информации и стохастичностью их поведения. Исследование подобных систем возможно путем построения моделей многомерного статистического анализа. Рассмотрим кластерный метод статистического анализа ключевых параметров сложных систем и условий нарушения их стабильности путем идентификации несоответствий и моделирования процессов обеспечения устойчивости систем. В основе кластерного метода используется способ разбиения исследуемых объектов на подмножества по схожим признакам и взаимосвязям. Однако, проблема такого построения классов заключается в сложном анализе большого числа признаков, что приводит к необходимости определения наиболее информативных данных и снижению размерности множества признаков, используя такие методы как факторный анализ, метод главных компонент и другие. Рассмотрим задачу моделирования многомерных структур минимальным набором информативных признаков, представленных в виде ключевых показателей (свойств). Необходимо определить набор этих показателей, согласно которому можно будет произвести разбиение объектов на непересекающиеся по определенным условиям множества. При этом, необходимо учесть, что признаки, входящие в этот набор, могут быть произвольной природы и принимать значения различных типов – числовые, текстовые, бинарные, номинальные и др. Это особенно важно при изучении не только одного конкретного объекта, но и сложного процесса, характеризующегося разнообразными признаками и различными единицами измерений, которые сложно объединить в обычной эконометрической функции. Таким образом, актуальной является задача интеллектуального анализа данных, позволяющего достаточно

большой объем информации представить в виде множества меньшей размерности, в наглядном и доступном для восприятия представлении. Основными математическими характеристиками, описывающими кластерную модель , являются [75] :

- Центр кластера - это среднее значение расстояния между объектами кластера.
- Радиус кластера - максимальное расстояние точек от центра кластера.
- Размер кластера – количество объектов и длина радиуса кластера.

Цели кластеризации ставятся в зависимости от особенностей конкретной прикладной задачи. Например, кластеризация может быть использована в качестве предварительной обработки данных для последующего глубокого изучения кластерных групп схожих объектов по отдельности, используя методы классификации и прогнозирования.

Метод кластеризации разбивает некоторое множество объектов $a_i \in A$ на m кластерных групп $Q_1, Q_2, \dots, Q_m$ по следующим правилам:

- объект a_i принадлежит только одному кластеру;

- два объекта a_i и a_j принадлежат одному и тому же кластеру, если у них наблюдается сходство по определенным признакам, то есть они однородные;

- два объекта a_i и a_j принадлежат различным кластерам, если у них имеются различные признаки, они разнородные.

Разбиение на кластеры должно осуществляться согласно некоторому критерию. Этот критерий может представлять собой некоторую целевую функцию, в качестве которой, например, можно выбрать квадратичную сумму отклонений [58], [59] :

$$W = \delta_n = \sum_{j=1}^{n} \left(x_j - \bar{x}\right)^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n} \left(\sum_{j=1}^{n} x_j\right)^2$$

где x_j – представляет собой j-ое признаковое значение x_j-го объекта.

Заметим, при кластерном моделировании для расчета величины расстояния между точками(объектами) во избежании искажений необходимо учитывать возможную неоднородность единиц измерения свойств этих точек (объектов) и привести расчет этих свойств к единому масштабу измерения и сравнимости шкал. Для преодоления этой проблемы производится преобразование переменных к некоторому общему диапазону значений с помощью заранее установленных весовых коэффициентов нормирования. Эту процедуру называют стандартизацией или нормализацией. Таким коэффициентом, в частности может быть отношение показателя отклонения от среднего значения к среднеквадратичному:

$z = (x - \bar{x})/\delta$, где $\bar{x}$ и δ – среднее и среднеквадратичное отклонения переменной x.

В результате нормирования с помощью весовых коэффициентов, используя аналитические расчеты, например эвклидово выражение, можно рассчитать расстояние между точками в многомерном пространстве, имеющими изначально различные шкалы и единицы измерений.

Рассмотрим методы оценки степени близости между объектами одного кластера и определим принадлежность объекта к конкретному кластеру на основе этих оценок. Отметим, что основными трудностями, возникающими при этом являются выбор методов нормирования измеряемых значений объектов и определение расстояния между самими объектами.

Пусть задана эмпирическая выборка $A^l = \{a_1, a_2, \dots, a_l\}$, $A^l \subset A$, каждый элемент которой представляет собой заданную пару отображений объекта в свой класс $a_i \to y_i$, где $a_i \in A$ - множество объектов обучающей выборки , $y_i \in Y$ - множество кластеров (классов). Требуется решение следующих двух задач:

- определить функцию оценки расстояния $p(a_i, a_j)$ между двумя произвольными элементами a_i и a_j;

- определить кластеры $y_1, y_2, \dots$ в виде непересекающихся подмножеств, таких, что близкие по метрике ρ объекты объединены в один кластер, а объекты максимально удаленные рассредоточены в разных кластерах.

Одной из самых популярных метрик вычисления расстояния между двумя объектами с числовыми признаками – это вычисление евклидовых расстояний, представляющих собой расстояние между объектами в n-мерном пространстве [68]:

$$d(a_i, a_j) = (\sum_{l=1}^{m} (a_i^l - a_j^l)^2)^{\frac{1}{2}}$$

Существует простой, итеративный метод кластеризации, основанный на евклидовом расстоянии. Его можно описать следующим образом. Предположим, что мы имеем R случайным образом выбранных кластеров $C_1, C_2 \dots C_R$. Нахождение точек (объектов) в n-мерном пространстве необходимо настроить таким образом, чтобы каждая из них по возможности была бы смещена в направлении центра своего кластера. Представим немаркированные (неклассифицированные) объекты обучающей выборки. Для каждого объекта a_i этой выборки, необходимо найти тот кластер Cj, который максимально приближал бы к себе объект a_i:

$(1 - \beta_j)C_j + \beta_j a_i \to C_j$, где β_j – некоторый параметр, определяющий скорость перемещения в направлении j-го кластера.

Данная процедура перемещает в направлении поиска кластера с шагами меньшими, чем продолжается самообучение алгоритма. Предположим, что каждый алгоритм поиска кластера C_j имеет некоторую весовую переменную m_j, равную числу проведенных перемещений. По мере увеличения массы кластера он движется медленнее к центру. Например, мы можем установить $\beta_j = 1/(1 + m_j)$ и использовать приведенное выше выражение приближения объекта к кластеру . С помощью этого правила настройки, кластер, который мы ищем, всегда находится в центре тяжести (образец среднего) набора объектов, к которому эти объекты приближаются. Интуитивно, если процедура поиска кластеров когда-либо попадает в какой-то достаточно хорошо сгруппированный набор объектов (и если этот алгоритм поиска кластеров является единственным, который находится таким образом), он будет сходиться к центру тяжести этого кластера. С помощью этого правила регулировки поиск кластеров всегда осуществляется по центру массы группы объектов.

Как только поиск кластера максимально приближен, процедура классификации помечает объекты a_i' и на этой основе может быть проведено разбиение пространства путем вычисления расстояний до различных кластерных поисков. Такая классификация может быть реализована некоторой линейной функцией приближения. При базировании разбиения на расстояниях, мы ищем кластеры, объекты которых максимально близки друг другу. Мы можем измерить показатель неудовлетворенности V кластера параметров a_i путем вычисления его выборочной дисперсии, определяемой следующим выражением:

$V = (1/K)\sum_i (a_i - M)^2$, где M - среднее значение выборки кластера, которое определяется как:

$M = \left(\frac{1}{K}\right)\sum_i X_i$, здесь K- количество точек в кластере.

Можно разбить набор объектов на кластеры таким образом, чтобы сумма выборочных дисперсий (погрешностей) этих кластеров была бы минимальной. Конечно, если у нас есть один кластер для каждого объекта, то выборка будет равны нулю, поэтому мы должны учесть, что наша мера погрешности разделения должна увеличиваться с количеством кластеров. Таким образом, мы можем искать компромисс между дисперсиями кластеров и достаточным количеством объектов в них. Разработка основной процедуры поиска кластеров позволяет варьировать количество кластеров в зависимости от расстояний между ними и в зависимости от выборочных дисперсий кластеров. Например, если расстояние, $d(a_{ij})$, между двумя

кластерами C_i и C_j, будет меньше некоторого порога ε, то мы можем их объединить в один кластер, с усредненным центром тяжести этих двух кластеров (с учетом их соответствующих масс). Таким образом, мы можем уменьшить общую возможную погрешность в процедуре разбиения объектов на группы, уменьшив количество кластеров. Можем ввести в рассмотрение функцию штрафа, как разницу между количеством кластеров в текущем и предыдущем шагах итерации. Тогда можно исследовать данное объединение кластеров с минимальным значением штрафной функции. С другой стороны, если какой-либо алгоритм поиска кластера определил кластер C_i, в котором количество объектов превышает некоторую сумму δ, то можно провести новый поиск другого кластера C_j, смежного с C_i, и сбросить массы обоих C_i и C_j, сделав их равным нулю. Таким образом, погрешность разделения может в конечном счете уменьшиться, уменьшив и общую дисперсию выборки со сравнительно небольшим штрафом за дополнительный поиск кластера. Значения параметров δ и ε устанавливаются в зависимости от относительных весов, заданных для выборочных дисперсий и количества кластеров. В методах на основе расстояния важно масштабировать компоненты векторов признаков близости объектов. Изменение значений вдоль некоторых размеров вектора объектов может быть значительно большим, чем изменение других измерений. Одним из обычно используемых методов является вычисление стандартного отклонения (квадратного корня от дисперсии) каждой из компонент по всему набору эмпирической выборки и нормализация значений компонент, так чтобы их скорректированные стандартные отклонения были бы равны между собой [36].

Метод, основанный на вероятностных значениях [36]. Предположим, что мы имеем разбиение эмпирического множества θ на R взаимно исключающих кластеров $C_1, C_{2\ldots}C_R$. Мы можем решить, с какого из этих кластеров C_i должен быть выбран произвольный объект из R, для которого вероятность $p(C_i|A)$ будет больше некоторого порогового значения δ. Мы можем выбрать кластер, для которого значение $p(A|C_i)p(C_i)$ является максимальным. Предполагая условную независимость компонент a_i, максимальная величина будет:

$$S(A, C_i) = p(a_1|C_i)p(a_2|C_i), \ldots, p(a_n|C_i)$$

$p(a_j|C_i)$ можно оценить по выборочной статистике моделей в кластерах и затем использовать в приведенном выше выражении. Назовем $S(A, C_i)$ подобием A кластеру объектов C_i. Таким образом, мы назначаем A кластеру, к которому он

наиболее похож, при условии, что сходство больше, чем заданное значение δ. Как и раньше, мы можем определить среднее значение выборки кластера C_i:

$M_i = (1/K_i) \sum_{a_j \in C_i} a_j$, где K_i-номер элемента в кластере C_i.

Тогда, алгоритм кластеризации, основанный на оценке степени сходства объектов можно записать в виде выполнения следующих шагов:

Шаг 1. Инициируем немаркированную выборку объектов Θ и пустой список L кластеров.

Шаг 2. Для следующего объекта a из Θ вычислим $S(a_i, C_i)$, для каждого кластера C_i. (Первоначально эти сходства равны нулю.) Предположим, что наибольшая из этих сходств - $S(a, C_{max})$. Тогда рассмотрим два случая:

- если $S(a, C_{max}) > \delta$, то назначить a кластеру C_{max} . То есть, $C_{max} + a \rightarrow C_{max}$ и обновить значение вероятности (статистику выборки):

$p(a_1|C_{max})p(a_2|C_{max}), \dots, p(a_n|C_{max})$, и полученное новое значение $p(C_{max})$ взять за основу в новой модели для последующих шагов. Переход в пункт 3.

- если $S(a, C_{max}) \leq \delta$, то необходимо создать новый кластер $C_{new} = \{A\}$ и добавить C_{new} к L.

Переход к 3.

Шаг 3. Объедините все существующие кластеры, C_i и C_j , если $(M_i - M_j)^2 < \varepsilon$. Вычислите новую статистику выборок $p(a_1|C_{merge})p(a_2|C_{merge}), \dots, p(a_n|C_{merge})$ для объединенного кластера $C_{merge} = C_i \cup C_j$.

Шаг 4. Если выборочная статистика кластеров не изменилась в течение всей итерации через Θ, то заканчиваем с кластерами в L; в противном случае перейти к шагу 2. Значение параметра δ контролирует количество кластеров. Если δ высока, в каждом кластере будет большое количество кластеров с несколькими объектами. При малых значениях δ, в каждый кластер будет включено небольшое количество кластеров со многими объектами. Точно так же, чем больше значение ε, тем меньше будет число кластеров, которые будут найдены. Таким образом можно построить классификатор на основе разделения объектов по кластерам путем назначения включения объекта a в тот кластер, при котором будет максимальным згачегие $S(a, C_{imax})$.

Для применения иерархических методов кластеризации, можно воспользоваться одним из подходов, в которых используется ряд последовательных слияний/делений объектов на N групп согласно критерию их близости [36]:

1. агломеративная иерархическая кластеризация (снизу вверх):

- назначаем N кластеров (каждый объект является собственным кластером)
- объединяем наиболее похожие объекты
- повторяем об'единение до тех пор, пока все объекты не будут находиться в одном кластере, в соответствие с заданными критериями объединения;

2. разделительная иерархическая кластеризация (сверху вниз):
   - назначаем один кластер в виде множества всех заданных объектов
   - разделяем по принципу максимальной разнородности объектов
   - повторяем разделение до тех пор, пока все объекты не будут находиться в своем кластере и при этом удовлетворять заданным критериям.

Вернемся к методу, основанном на евклидовом пространстве. Предположим, что у нас есть множество немаркированных объектов обучения. Мы можем сформировать иерархическую классификацию этих объектов в Θ. Сначала мы вычисляем евклидово расстояние между всеми парами объектов θ (опять же, предполагается, что признаки объектов нормированы и находятся в единой шкале измерений). Предположим, что наименьшее расстояние находится между объектами a_i и a_j. Мы включаем a_i и a_j в кластер C, исключаем a_i и a_j из θ и заменяем их на вектор кластера C, равный среднему значению a_i и a_j. Затем мы снова вычисляем евклидово расстояние между всеми парами точек θ. Выбирая минимальное расстояние между парами объектов, мы формируем новый кластер C и как и раньше, заменяем пару объектов в θ по их среднему значению. Если расстояние между объектом a_i и вектором кластера C_j наименьшее, то формируем новый кластер C, состоящий из объединения C_j и $\{a_i\}$. В этом случае мы заменим C_j и a_i в θ по их (соответственно взвешенному) среднему значению и продолжим далее. Если кратчайшее расстояние между двумя кластерными векторами есть расстояние между C_i и C_j, то мы формируем новый кластер C, состоящий из объединения C_i и C_j. В этом случае мы заменим C_i и C_j их средним (взвешенным) и продолжим. Поскольку мы уменьшаем число точек в θ по одному каждый раз, мы в конечном счете закончим построение дерева (иерархии) кластеров, при котором будет только одна вершина дерева.

Мы можем разработать меру качества разбиения на основе того, насколько точно мы можем распознать объект, когда задан раздел, в котором он находится. Это будет вероятностной мерой определения кластера. Предположим, нам дано разбиение Θ в R классы $C_1, C_{2\ldots}C_R$. Как и ранее, мы можем вычислить статистику

выборок $p(a_i|\ C_k)$, которые дают значения вероятности для каждой компоненты, заданной классом, назначенным ему разбиением. Допустим, что каждая компонента a_i из A может принимать значения v_{ij} , где индекс j проходит над областью определения этой компоненты. Обозначение $p_i(v_{ij}|\ C_k)$ будет равно вероятности (a_i = $v_{ij}|\ C_k$. Воспользуемся следующим вероятностным предположением о значениях компонент вектора A: допустим он находится в классе k: $a_i = v_{ij}$ с вероятностью $p_i(v_{ij}|\ C_k)$.

Тогда вероятность того, что мы можем правильно распознать i-ю компоненту равна [36]:

$$\sum_j probability\ (guess\ is\ v_{ij}) p_i(v_{ij}|\ C_k) = \sum_j [p_i(v_{ij}|\ C_k)]^2$$

Среднее число (n) компонент, чьи значения правильно угадываются этим методом, определяются суммой этих вероятностей по всем компонентам множества A

$$\sum_i \sum_j [p_i(v_{ij}|\ C_k)]^2$$

Учитывая разбиение объектов на R-классы, мера качества G этого разбиения является средним значением указанного выше выражения для всех классов:

$$G = \sum_k p(C_k) \sum_i \sum_j [p_i(v_{ij}|\ C_k)]^2$$

где $p(C_k)$ - вероятность того, что объект находится в классе C_k. Чтобы установить меру штрафа за наличие большого количества классов, мы разделим ее на R, чтобы получить общую «качественную» меру разбиениях [36] :

$$Z = (1/R) \sum_k p(C_k) \sum_i \sum_j [p_i(v_{ij}|\ C_k)]^2$$

Приведем пример использования этой меры для тривиально простой кластеризации четырех трехмерных схем. Существует несколько разных разделов.

Оценим значения Z:

$$P_1 = \{a, b, c, d\}, P_2\{\{a, b\}, \{c, d\}\}, P_3\{\{a, c\}, \{b, d\}\}, P_4\{\{a\}, \{b\}, \{c\}, \{d\}\}$$

Первый, P1, помещает все объекты в один кластер. Вероятности выборки $P_i(v_{i1} = 1)$ и $P_i(v_{i0} = 0)$ равны 1/2 для каждой из трех составляющих. Суммирование значений компонент (0 и 1) дает значение $(1/2)^2 + (1/2)^2 = 1/2$.
Суммирование по трем компонентам дает $3/2$. Усреднение по всем кластерам (есть только одно) также дает $3/2$. Наконец, деление на число кластеров дает окончательное значение Z этого раздела, $Z(P_1) = 3/2$.

Второй раздел P_2 дает следующие вероятности выборки:
$p_1(v_{11} = 1|C_1) = 1,\ p_2(v_{21} = 1|C_1) = 1/2,\ p_3(v_{31} = 1|C_1) = 1$

Суммирование по значениям компонентов (0 и 1) дает $1^2 + 0^2 = 1$ для компоненты 1, $1/2^2 + 1/2^2 = 1/2$ для компоненты 2 и $1^2 + 0^2 = 1$ для компоненты 3.

Суммирование по трем компонентам дает $2\,{}^1\!/_2$ для класса 1. Аналогичный расчет также дает $2\,{}^1\!/_2$ для класса 2. Усреднение по двум кластерам также дает $2\,{}^1\!/_2$.

Наконец, деление на число кластеров дает окончательное значение Z этого раздела, $Z(P_2) = 1\ 1/4$, не так высоко, как в случае , $Z(P_1)$. Аналогичные вычисления дают $Z(P_3) = 1$ и $Z(P_4) = 3/4$, поэтому этот метод оценки разделов будет способствовать размещению всех объектов в одном кластере.

Оценка всех разбиений на m групп объектов, а затем выбор наилучшего может быть сложно вычислимым. Следующий итерационный метод основан на иерархической процедуре кластеризации. Процедура увеличивает дерево, каждый узел которого помечен набором объектов. В конце процесса корневой узел содержит все объекты θ. Концы дерева будут содержать одноэлементные наборы. Метод использует значения Z для размещения объектов в узлах; выборочная статистика используется для обновления значений Z всякий раз, когда объект помещается в узел. Алгоритм представляет собой выполнение следующих шагов:

1. Инициируем дерево, корневой узел которого содержит все объекты θ и единственный пустой узел-преемник. Примем, что каждый непустой узел в дереве имеет (помимо любых других преемников) ровно один пустой преемник.
2. Выбираем объект a_i (если объектов больше нет, прекращаем действие).
3. Присваиваем значение μ корневому узлу.
4. Для каждого из преемников μ (включая пустой преемник), вычисляем значение Z для a_i. Выбираем оптимальное (максимальное) значение Z.

5. Если лучший узел является пустым узлом η , то мы помещаем a_i в η . Генерируем пустой узел-преемник η, генерируем пустой родственный узел из η и переходим к 2.
6. Если лучший узел является непустым родственным узлом η мы помещаем a_i в η. Создаем один узел-преемник η, содержащий родственный объект, который был в η, создаем другой узел-преемник η, содержащий a_i, создаем пустой узел-преемник η, и переходим к 2.

Если лучший хост является непустым узлом, не являющимся одиночным η, то мы размещаем a_i в η , группу μ в η и переходим к пункту 4.

Этот процесс довольно чувствителен к порядку, в котором представлены объекты. Чтобы сделать окончательное классификационное дерево менее зависимым от порядка, целесообразно провести процедуру слияния и разделения узлов.

Слияние узлов: может случиться так, что два узла, имеющих один и тот же родительский узел, могут быть объединены, обеспечивая общее повышение качества результирующей классификации, выполняемой наследниками этого родителя. Вместо того, чтобы попытаться объединить все пары, хорошей эвристикой является попытка объединить узлы двух лучших узлов-хозяев. Когда такое слияние улучшает значение Z , новый узел, содержащий объединение объектов в объединенных узлах, заменяет объединенные узлы, а два узла, которые были объединены, устанавливаются как преемники нового узла.

Разделение узла: эвристика для разделения узлов заключается в том, чтобы рассмотреть возможность замены лучшим узлом несколько узлов-наследников. Эта операция выполняется только в том случае, если она увеличивает значение Z класса, выполняемого группой родственных объектов. Для оценки расстояния между случайными объектами (величинами) можно воспользоваться коэффициентом Пирсона [76] или какой-нибудь другой известной метрикой кластерного анализа, используемой для определения близости между объектами:

$$d(a_i, a_j) = \sum_{l=1}^{m} \left| x_i^l - x_j^l \right|$$ - линейное расстояние

$$d(a_i, a_j) = \sum_{l=1}^{m} (a_i^l - a_j^l)^2$$- квадрат евклидова расстояния

$d(a_i, a_j) = (\sum_{l=1}^{m} (a_i^l - a_j^l)^p)^{\frac{1}{p}}$ – обобщенное p-степенное расстояние Минковского.

Существует еще целый ряд метрик, представляющих собой геометрическое расстояние в многомерном пространстве. Возведение в квадрат стандартного

евклидова расстояния позволяет получить большее значение весов для визуального подчеркивания отдаленности друг от друга объектов из разных кластерных групп.

Выводы к главе 2

1. Рассмотрены метрические методы классификации, основанные на гипотезе компактности и методе нахождения k “ближайших соседей”.
2. Поставлена задача оценки эмпирического риска и нахождения решающей функции путем применения функционала наименьших квадратов. Рассмотрена задача выбора критериев оценки качества классификации с помощью решающей функции.
3. Показано, что по мере роста сложности модели, описываемой некоторой полиномиальной функцией, также увеличивается количество наблюдаемых расхождений. Представлен метод оценки адекватности модели с помощью функционала перекрестного (скользящего) контроля.
4. Исследованы итерационные методы кластеризации объектов на основе многомерного статистического анализа данных. Рассмотрены способы разбиения объектов на подмножества по схожим признакам и взаимосвязям.

ГЛАВА 3: ИССЛЕДОВАНИЕ И РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ УПРАВЛЕНИЯ ПРОЦЕДУРАМИ САРА

Как было уже отмечено в предыдущих главах основными требованиями CAPA являются [77-82] :

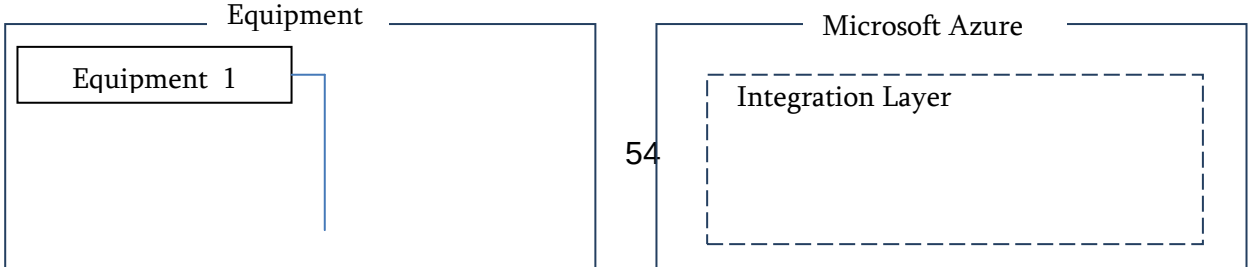
- своевременное выявление несоответствий ключевых параметров заданным нормативным стандартам;
- идентификация источников отклонений и несоответствий;
- анализ причинно-следственных связей между различными несоответствиями;
- планирование и исполнение корректирующих и превентивных действий, направленных на устранение обнаруженных отклонений и несоответствий;
- оценка степени влияния на валидированные компоненты системы: процедуры, оборудование, продукты, процессы и системы;
- тщательное протоколирование и документирование всех этапов CAPA, от обнаружения отклонений и несоответствий до планирования и выполнения корректирующих действий и оценки их эффективности.

В первой главе диссертации было отмечено, что реализация процедур CAPA сопряжена рядом сложностей, среди которых можно выделить следующие:

- неэффективность и трудоемкость выполнения процессов ручными методами;
- длительность процедур согласования и одобрения для проведения изменений;
- сложность организации маршрутизации процессов управления изменениями и действиями сотрудников на различных этапах реализации CAPA;
- невозможность проведения анализа и оценки эффективности выполненных процедур CAPA.

3.1. Архитектура информационной системы BVR QMS управления процедурами CAPA

На основе представленных в главе 2 методов и алгоритмов разработана информационная система управления процедурами CAPA (BVR QMS), логическая структура которой представлена на следующем рисунке:



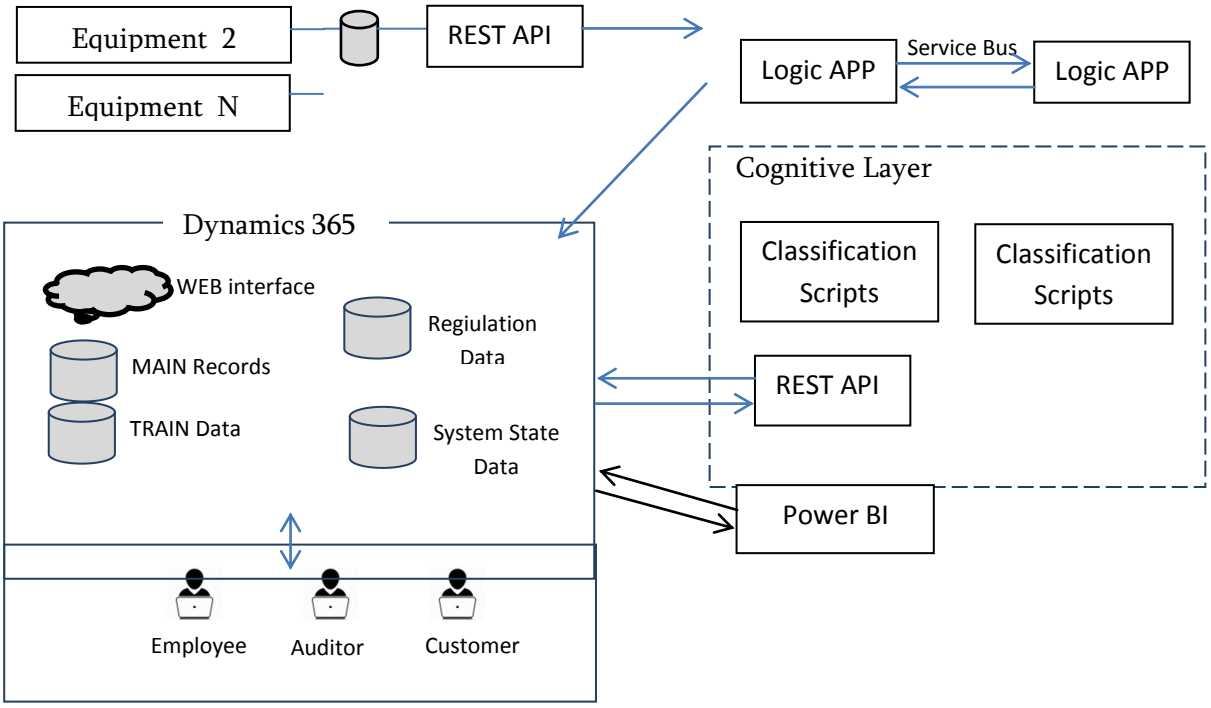


Рисунок 3.1. Структура BVR QMS и ее взаимодействие с другими компонентами.

Моделирование многих процессов основано на разработке и внедрении инструментов интеллектуального анализа данных [83-85]. Процесс обнаружения и идентификации несоответствий возможен путем создания самоорганизующей системы на основе обработки больших данных, анализа ключевых параметров и прогнозирования ожидаемых результатов [86,87]. Поскольку эти модели описывают компьютерную обработку данных и будут генерироваться программными системами, то необходимо “заставить” их развиваться вместе с процессами, которые создают эти данные. Основное внимание при проектировании информационной системы должно быть уделено программному прогнозированию и принятию решений на основе этих прогнозов. Ниже будет представлено описание фукнциональных компонент системы BVR QMS, реализующих функции идентификации несоответствий и обнаружения их корневых причин.

3.1.1. Контроль ключевых показателей и идентификация несоответствий заданным допустимым значениям

Реализовав вышесказанное можно обеспечить полноценный автоматизированный процесс обнаружения и идентификации несоответствий. Современные исследования в области теории информационных систем управления

качеством широко используют статистические методы анализа данных и оценки ключевых параметров. Среди этих методов можно выделить контрольные карты средних, средне-квадратичных и экстремальных значений, оценки экстремальных разбросов, стандартных отклонений и меридиан [88-92]. С помощью контрольных карт (диаграмм Шухарта) задаются диапазоны (верхние и нижние границы, *UCL*, *LCL*) допустимых значений для ключевых показателей, эталонные значения (*CL*), регистрируются фактические данные и рассчитываются методами статистического анализа степени отклонения от допустимых значений.

Значения CL, UCL и LCL вычисляются следующим образом [88]:

$CL=\mu,\quad UCL=\mu+3\sigma,\ LCL=\mu-3\sigma,$ (8)

где μ - эталонное значение (математическое ожидание),

σ – стандартное значение допуска (дисперсия).

Карта размаха (R-карта) рассчитывается следующим образом [88]:

$CL=d_2*\sigma\ ,\ UCL=D_2*\sigma,\ LCL=D_1*\sigma$ (9)

где d_2 , D_2 и D_1 – коэффициенты для вычисления эталонной линии, верхней и нижней границы значений размаха.

Если задать величину среднего скользящего размаха R_{cp} , то можно рассчитать величину уменьшенного допуска

$$\sigma_2=R_{cp}*\ d_2 \qquad (10)$$

Одной из основных проблем обнаружения несоответствий и/или тенденций к нарушению стабильности является необходимость определения следующих параметров для каждого ключевого показателя и последующих статистических расчетов:

- частоту (период) наблюдений и регистрации значений *W* (шаг диагональной оси карты Шухарта);
- число наблюдений *Q* (количество точек в диаграмме Шухарта).

Таким образом, по каждому параметру необходимо определить после какого количества наблюдений и регистраций значений и с какой частотой их проведения можно считать, что расчеты приведенных выше показателей указывают на имеющиеся несоответствия или тенденции к несоответствию. Решение данной проблемы возможно путем ввода дополнительных показателей (критериев), взаимосвязанных с заданным параметром, расчета их значений и построения соответствующих контрольных списков (чек-листов). На основе этих расчетов и

проведения анализа значений можно окончательно рассчитать частоту и количество регистраций.

Введем следующие формальные обозначения:

пусть $A = \{ a_1, a_2, \dots a_n \}$ – множество ключевых параметров. Зададим для A множество $P(A) = \{ P(a_1), P(a_2) \dots P(a_n)\}$, элементами которого являются множества показателей $P(a_i) = \{P_{t1}(a_i), P_{t2}(a_i) \dots P_{tj}(a_i)\}$ для параметра $a_i \in A$, регистрируемых в определенные периоды времени T, t_j – порядковый номер наблюдения (заданный период наблюдения) в цепочке регистрации значений, j- количество наблюдений. Введем понятие количественных характеристических критериев , связанных с параметром $a_i \in A$ и обозначим их в виде следующего множества:

$C(a_i)=\{C_1(a_i), C_2(a_i),\dots, C_m(a_i)\}$, где m – число характеристических критериев для параметра a_i.

Для каждого параметра a_i будет рассчитываться значение коэффициента $V(a_i)$ как средняя величина от множества значений характеристических критериев:

$$V(a_i)=\frac{(\sum_{k=1}^{m} C_k(a_i))}{m*l} \qquad (11)$$

где l- коэффициент нормализации (обычно он равен максимальному значению, которое может принимать параметр $C_k(a_i)$.

Значения характеристических параметров $C_k(a_i)$ могут определяться информационной системой как автоматически, путем выборки входных данных, поступающих от различных источников так и путем заполнения сотрудником динамически генерируемого контрольного списка (чек листа). Этот список может быть заполнен специалистом по качеству, технологом производства, сотрудником лаборатории и другими квалифицированными работниками предприятия.

Далее, составив таблицу ранжирования значений характеристических параметров с назначенными значениями частоты и количества наблюдений для параметра a_i можно однозначно определить необходимые данные для последующего контроля и регистрации ключевых показателей. Таблица может иметь следующий вид:

Заданные пределы допустимых значений, частота и количество наблюдений

Нижний предел $V(a_i)$	Верхний предел $V(a_i)$	Частота наблюдений	Количество наблюдений
$V_{11}(a_i)$	$V_{21}(a_i)$	$W_1(a_i)$	$Q_1(a_i)$

$V_{12}(a_i)$	$V_{22}(a_i)$	$W_2(a_i)$	$Q_2(a_i)$
$V_{13}(a_i)$	$V_{23}(a_i)$	$W_3(a_i)$	$Q_3(a_i)$
$V_{14}(a_i)$	$V_{24}(a_i)$	$W_4(a_i)$	$Q_4(a_i)$

В зависимости от того в какой заданный диапазон(нижний и верхний пределы) попадает вычисляемое значение, с соответствующей строки будут выбраны столбцы “Частота наблюдений” и “Количество наблюдений”. Таким образом, кортеж, состоящий из четверки множеств *{C,V,W,Q}*, – задает первоначальные параметры настройки для последующего построения контрольных карт. Тройка множеств *{A,R,T}* – представляет собой наблюдаемые и регистрируемые динамические значения ключевых параметров в определенные заданные периоды времени. На основе полученных значений можно построить контрольную карту и провести необходимые статистические расчеты.

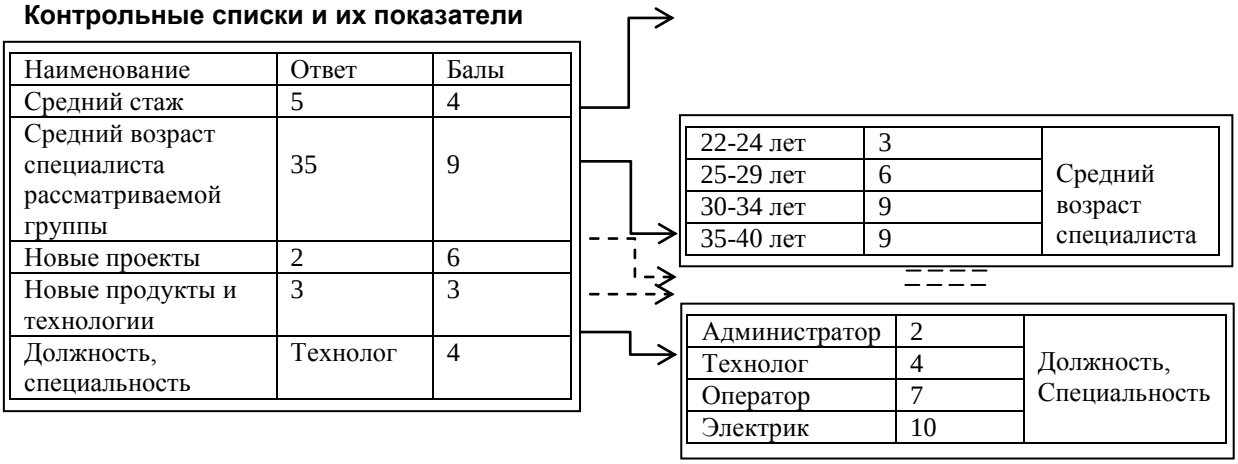
Приведем пример из практики: предположим для параметра “Аттестация” были установлены соответственно, частота проведения аттестации сотрудника раз в месяц с диапазоном балов *UCL=15* , *LCL=20* по двадцатибальной шкале. С данным параметром, допустим, связаны следующие показатели, средние значения которых должны быть вычислены в системе: 1. Средний возраст сотрудников, 2. Стаж сотрудников, 3. Новые проекты в текущем периоде, 4. Новые продукты и технологии в текущем периоде, 5. Рост зарплаты за текущий период, 6. Отпуска за текущий период. Пункты 1,2,5,6 могут быть рассчитаны, например, на основе данных, полученных с компьютерной системы “управления кадрами”, а пункты 3,4 должны быть введены сотрудником.

Ниже представлена контрольная таблица (чек-лист) для параметра “Аттестация сотрудника”, включающая в себя критерии, ответы и таблицу значений соответствующих балов.

Балы значений показателей

< 2 лет	1	Средний стаж работы
3-5 лет	6	
6-7 лет	7	
> 7 лет	10	

Контрольные списки и их показатели



Наименование	Ответ	Балы
Средний стаж	5	4
Средний возраст специалиста рассматриваемой группы	35	9
Новые проекты	2	6
Новые продукты и технологии	3	3
Должность, специальность	Технолог	4

22-24 лет	3	Средний возраст специалиста
25-29 лет	6	
30-34 лет	9	
35-40 лет	9	

Администратор	2	Должность, Специальность
Технолог	4	
Оператор	7	
Электрик	10	

Расчет по формуле (11) с заданным коэффициентом нормирования 10 (максимально возможное значение бала) будет:

*(4+9+6+3+4)/5*10= 0.52*

Далее по заданной таблице ключевых показателей выбираем строку с диапазоном, в котором находится рассчитанный коэффициент (для рассматриваемого примера это строка N3) и выбираются значения *W* и *Q* соответственно : частота наблюдений = *10*, количество наблюдений = *5*. Таким образом, для последующего построения контрольной карты и оценки специалистов рассматриваемой группы необходимо будет провести статистические расчеты, в частности, по формулам (8), (9), (10) после 4-х аттестаций, проводимых через каждые две недели, при этом аттестация будет считаться успешной, если сотрудники в среднем наберут количество балов в пределах от 17 до 20

Ключевые показатели для контрольной карты

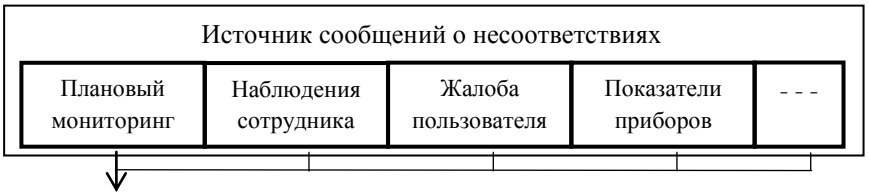
No	Нижний предел $V(A_i)$	Верхний предел $V(A_i)$	Частота наблюдений	Количество наблюдений	UCL	LCL
1	0	0,3	Каждые 5 дней	10	15	20
2	0,3	0,5	Каждые 8 дней	9	16	20
3	0,5	0,7	Каждые 10 дней	5	17	20
4	0,7	1	Каждые 20 дней	4	17	20

Данный пример из практики может означать следующее: внедрение новых технологий на предприятии и аттестация уровня специалистов по степени осваивания этих технологий предполагает достаточный период (10 дней) для осваивания новых задач, высокие требования знаний (больше 17 балов), выполнение статистических расчетов уже после проведения 4-х аттестаций и построения соответствующих контрольных карт.

На основе представленной модели разработана программная подсистема, входящая в систему BVR QMS, которая позволяет рассчитать контрольные карты и получить экспертную оценку по ключевым показателям с целью оперативного

принятия решений, в частности, по улучшению квалификации сотрудников, например, проведения обучения и/или привлечения новых специалистов.

Для формального описания связей между ключевыми и характеристическими параметрами, а также скрытыми переменными можно воспользоваться методами и моделями интеллектуального анализа данных, представленные в работе [93]. На основе формального описания процессов и анализа данных можно построить систему прогнозирования и принятия решений, которая будет генерировать динамические параметры для составления контрольных карт и соответствующих статистических расчетов. На рис. 3.2 представлена структурно-функциональная схема процесса, реализованного в системе BVR QMS, реализующего функции обнаружения и идентификации несоответствий ключевых показателей и обработки данных обратной связи для динамической настройки основных параметров управления контрольными картами и статистическими расчетами.



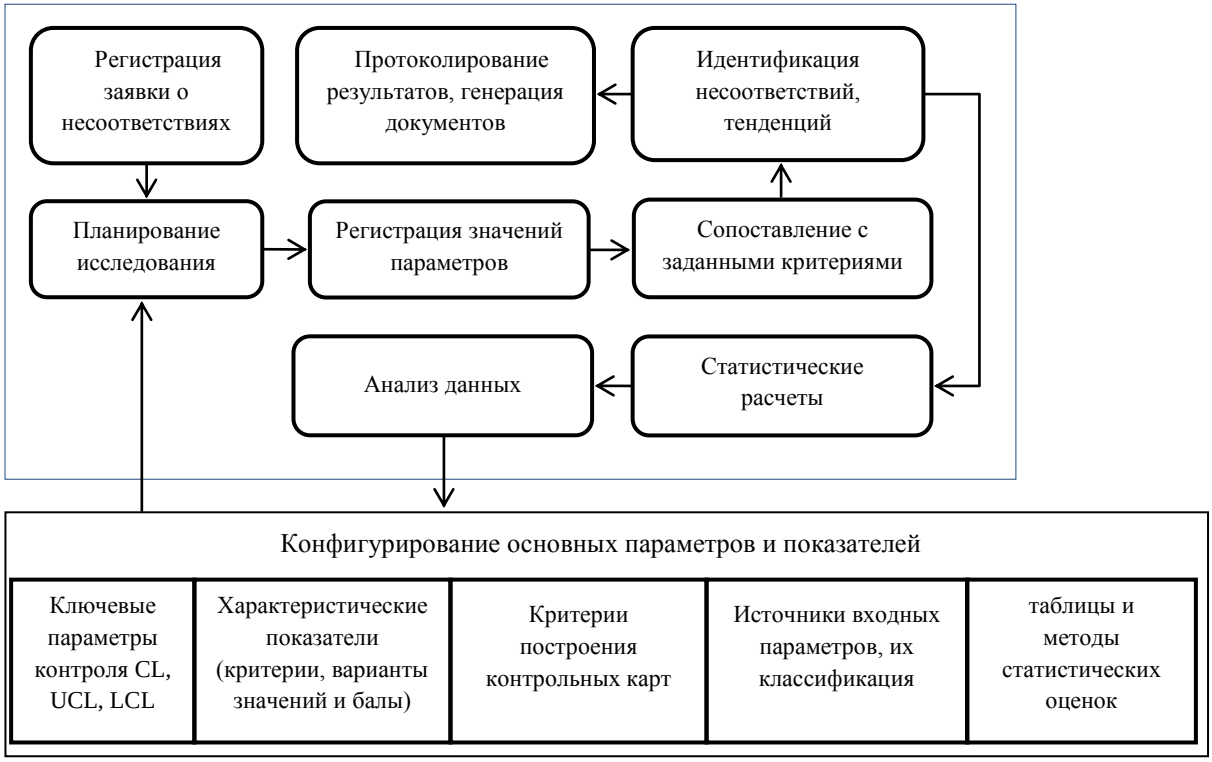


Рисунок.3.2.Функциональная схема процесса идентификации несоответствий ключевых показателей заданным допустимым значениям.

Автоматическое определение несоответствий или тенденций к нарушению стабильности являются первым шагом идентификации и оценки несоответствий для последующего оперативного и точного принятия решений о проведении корректирующих и превентивных действий. В основе предложенного метода - структурированный подход к вопросам автоматизации процессов управления изменениями.

3.1.2. Метод определения корневой причины нарушения стабильности

Идентификация основной причины несоответствий, отклонений и инцидентов в деятельности различных компаний, как было отмечено в предыдущих главах, является неотъемлемой частью всего процесса соответствия определенным

отраслевым стандартам. На сегодня имеется множество различных подходов, методик и процедур определения источников и корневых причин, обнаруженных несоответствий. Однако, зачастую на практике применение этих методик сопряжено рядом сложностей из-за отсутствия инструментов, консолидирующих данные, методы и ресурсы, используемые в процессе идентификации несоответствий и определения причин их порождающих. Основными задачами системы CAPA наряду с обнаружением отклонений ключевых показателей основных процессов деятельности компаний от запланированных значений, являются также устранение причин отклонений и предотвращение их влияния на качество выпускаемой продукции. Анализ обнаруженных несоответствий представляет собой многошаговую процедуру, направленную на изучение взаимосвязанных случайных явлений и процессов в рассматриваемом классе сложных стохастических систем. Процедура анализа всегда зависит от обстоятельств, источников сообщения и методов наблюдения несоответствий, оценки возможных рисков, степени влияний на качество продукции, текущих затрат и многих других параметров. Следующим важным шагом реализации процедур CAPA после идентификации несоответствий являются анализ и определение корневой причины их порождения, включающих в себя выполнение следующие шагов:

- регистрация источников сообщений о имеющихся проблемах (несоответствиях, отклонениях, инцидентах) и идентификация нарушений стабильности;
- поиск вероятностных причин несоответствий;
- определение корневой причины.

Рассмотрим детально приведенные выше три этапа обнаружения несоответствий и их корневых причин.

1. Идентификация нарушения стабильности.

Источниками информации о имеющихся несоответствиях и нарушениях стабильности могут быть наблюдения персонала, жалобы потребителя, заключения аудиторских проверок, показания приборов, результаты лабораторных тестов, регистрация нарушений норм жизнеобеспечения (бесперебойное питание, кондиционирование и прочее). Возможны и другие источники в сообщений, специфичные для конкретной предметной области.

Система менеджмента в организациях требует протоколирования источников сообщений о несоответствиях, детального описания возникшей проблемы и доказательство ее существования. Например, доказательством существования дефектов в оборудовании может быть высокий процент отрицательных результатов лабораторных тестов на физико-химический контроль балк-продуктов (нерасфасованной произведенной продукции).

Контроль стабильности включает в себя последовательное методичное выполнение следующих процессов:

- определение ключевых параметров системы согласно источникам информации о несоответствиях;
- определение нормативных значений ключевых параметров;
- измерение значений ключевых параметров и контроль на соответствие нормативным (эталонным) значениям.

Для идентификации несоответствий и доказательства нарушения стабильности существует множество методов и инструментов оценки данных, среди которых, контрольная карта Шугарта, радарная диаграмма, график Паретто и другие [94-96]. Основной проблемой является определение и выбор для каждого конкретного случая нарушения стабильности соответствующего метода контроля и обнаружения несоответствия. Выбор метода зависит прежде всего от источника сообщения и типа анализируемой информации, который может быть количественным или текстовым.

2. Поиск вероятной причины.

Длительность и сложность расследования вероятной причины, непосредственно вызвавшей несоответствие определяется множеством ключевых показателей, зависящих от источника сообщения и типа несоответствия. Например, причиной такого показателя несоответствия как температура воздуха в производственном помещении может быть вызвана неисправностью кондиционера, что легко может быть определено. В то же время для определения причины обнаруженного побочного влияния на здоровье пациента во время приема лекарственного препарата требует тщательного расследования и привлечения больших ресурсов.

Основной задачей поиска вероятной причины является

- накопление и консолидация “исторических” данных о имеющихся в прошлом подобных отклонениях, значениях ключевых показателей и выявленных причинах;
- структуризация и очистка “исторических” данных для дальнейшей классификации несоответствий;
- выборка наиболее вероятной причины на основе формализованных данных и экспертных оценках.

Исследования, проведенные на этапах идентификации нарушения стабильности, контроля ключевых показателей и определения вероятной причины, дополненные анализом причинно-следственных связей между отдельными процессами системы позволят установить корневую причину наблюдаемых несоответствий. Для визуализации указанной задачи и проведения соответствующих исследований удобно использовать причинно-следственные диаграммы. Решение приведенных выше проблем идентификации несоответствий и обнаружения вероятной и корневой причин возможно путем разработки метода машинной реализации самоорганизующихся систем, который позволил бы на основе классификации несоответствий и накопленных “исторических” данных определить выборку метода анализа и соответствующих ключевых параметров визуализации данных. Программная реализация указанной проблемы на основе формального описания зависимостей между компонентами и их причинно-следственных связей позволяет решить основную задачу, а именно определить корневую причину нарушения стабильности рассматриваемого класса СДС.

Математическое описание процесса идентификации вероятных и корневых причин нарушения стабильности:

пусть $A = \{a_1, a_2, \dots a_n\}$ – заданное конечное множество источников сообщений о имеющихся (наблюдаемых) несоответствиях, относящихся к одной из групп несоответствий $R_i\ (i = 1, \dots m)$.

Будем считать, что паре (a_j, R_i) соответствует упорядоченный вектор методов контроля

$$M(a_j, R_i) = [m_1(a_j, R_i), m_2(a_j, R_i), \dots m_l(a_j, R_i)],\ \text{где } l=1,2,\dots l'\ (a_j, R_i) \qquad (3.1)$$

Зададим для каждого метода $m_l(a_j, R_i)$ множество ключевых параметров и их допустимых (нормативных) значений соответственно

$$Q(m_l(a_j, R_i)) = [q_1(m_1(a_j, R_i)), q_2(m_l(a_j, R_i)), \dots q_k(m_l(a_j, R_i))], \qquad (3.2)$$

где $k=1,2,\ldots k'(m_l(a_j,R_i))$

$$V\left(m_l(a_j,R_i)\right)=[v_1(m_1(a_j,R_i)),v_2(m_l(a_j,R_i)),\ldots v_k(m_l(a_j,R_i))] \quad (3.3)$$

Регистрируемые значения параметров $q_k(m_l(a_j,R_i))$ в заданный момент времени t обозначим следующим образом

$$D\left(t,m_l(a_j,R_i)\right)=[t,d_1(m_1(a_j,R_i)),d_2(m_l(a_j,R_i)),\ldots d_k(m_l(a_j,R_i))], \quad (3.4)$$

Таким образом для каждого метода контроля m_1 тройка множеств

$$Q\left(m_l(a_j,R_i)\right),V\left(m_l(a_j,R_i)\right),D\left(t,m_l(a_j,R_i)\right)$$

задает ключевые параметры, его нормативные и регистрируемые в заданный момент t значения для установленного метода m_l .

Контроль показателей ключевых параметров и их сопоставление с нормативными значениями представим в виде следующей функции

$$G=F\left\{v_1\left(m_1(a_j,R_i)\right),d_1\left(m_1(a_j,R_i)\right),v_2\left(m_l(a_j,R_i)\right),d_2\left(m_l(a_j,R_i)\right),\ldots\right\} \quad (3.5)$$

Значение G функции F равно разнице между эталонными и регистрируемыми количественными показателями для каждого ключевого параметра выбранного метода. Если $G\neq 0$, то имеет место нарушение стабильности.

Каждой паре значений (a_j,R_i) поставим в соответствие множество вероятных причин нарушения стабильности

$Z(a_j,R_i)=[z_1(a_j,R_i),z_2(a_j,R_i),\ldots z_p(a_j,R_i)]$, где $p=1,2,\ldots p'(a_j,R_i)$ (3.6)

По сути причина нарушения стабильности $Z(a_j,R_i)$ представляет собой также возможное несоответствие, которое необходимо будет идентифицировать.

Тогда задачу определения несоответствий и вероятной причины нарушения стабильности можно свести к последовательному выполнению следующих процедур:

- определение цепочки методов $M(a_j,R_i)$, их ключевых параметров $Q(m_l(a_j,R_i))$ и допустимых значений, приведенных соответственно в формулах (3.1), (3.2), (3.3) для поступающих сообщений (a_j,R_i) ;
- измерение и регистрация значений ключевых параметров (3.4) в заданный период t ;

– сопоставление регистрируемых с допустимыми значениями, согласно выражению (3.5) и определение вероятной причины $Z(a_j, R_i)$, представленной в формуле (3.6).

Процесс должен быть выполнен как для идентификации несоответствия, так и для определения вероятных причин, вызвавших это несоответствие.

На рисунке 3.3 представлена структурно-функциональная схема процесса определения вероятной причины обнаруженных несоответствий, реализованного в системе BVR QMS:

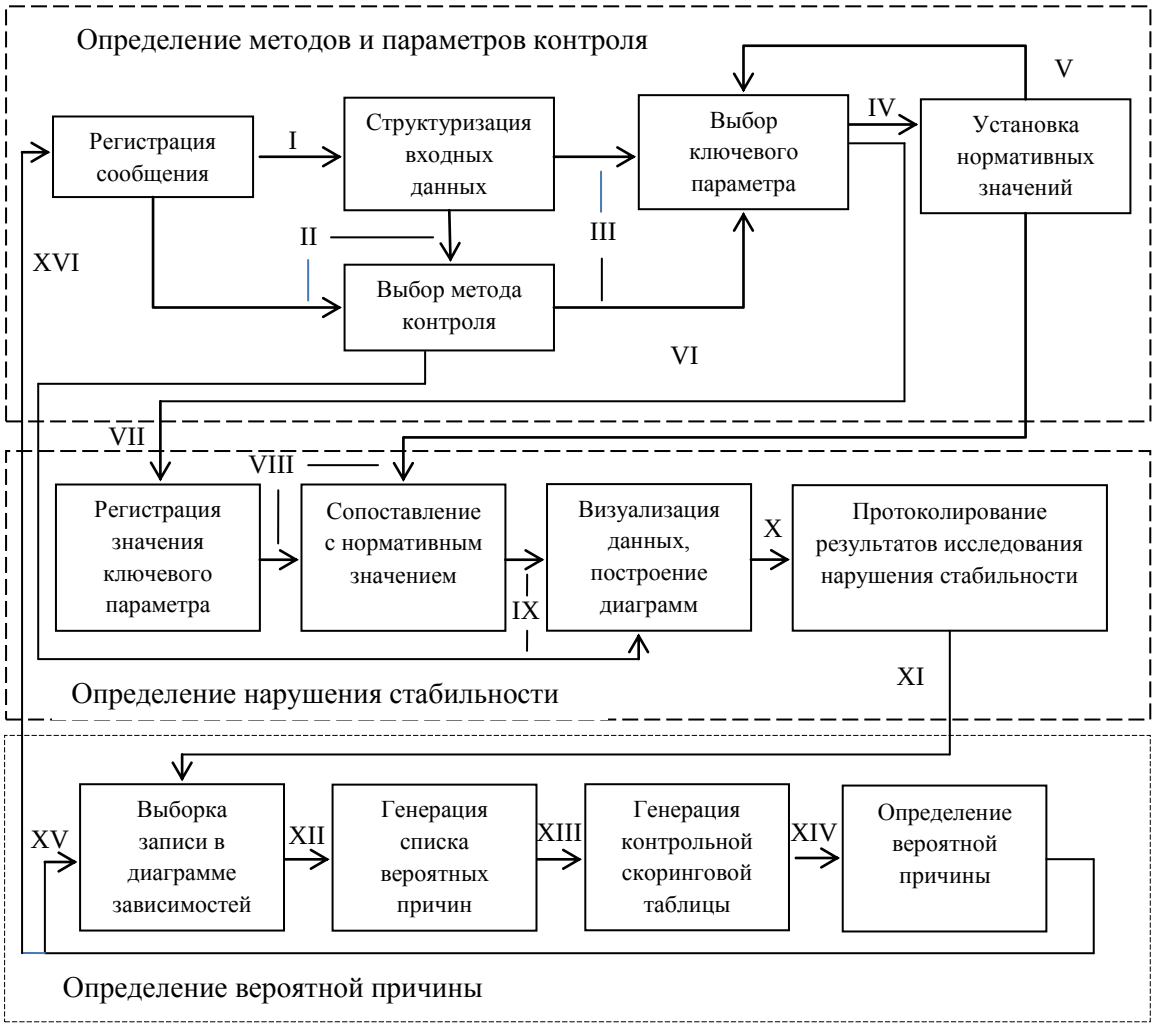


Рисунок 3.3. Схема процесса определения вероятной причины обнаруженных несоответствий.

Описание метода. Поступающие сообщения (I) о несоответствиях приводятся к структурированной форме для унификации и определения группы, которой соответствует данное предполагаемое несоответствие. Далее определяются методы, которые должны быть применимы для данной группы несоответствий и ключевые параметры для исследования и визуализации данных (II, III).

Одновременно, фиксируются для выбранных параметров эталонные (допустимые) значения (IV). Процесс продолжается до тех пор, пока не будет определена вся последовательная цепочка методов и параметров для данной группы несоответствий (V,VI). На этом этапе завершается процесс идентификации поступивших сообщений и определения ключевых параметров контроля и соответствующих методов исследования несоответствий и визуализации данных. На следующем этапе производится регистрация значений ключевых параметров и их сопоставление с нормативными значениями (VII,VIII).

Также производится визуализация полученных результатов с помощью установленных методов контроля (IX) и протоколирование данных о состоянии (возможном нарушении) стабильности системы (X). Если обнаружены несоответствия (факт нарушения стабильности) то на следующем этапе производится анализ вероятной причины путем выборки из базы данных записи о вероятной причине, связанной с уже установленной группой несоответствий (XI). При этом, возможны несколько причин с различными значениями вероятностей, рассчитанных на основе статистических данных по результатам прошлых исследований несоответствий (XII). Для этого дополнительно можно использовать некоторую скоринговую таблицу (XIII) с характеристическими параметрами для каждой отдельной вероятной причины, которую могут заполнять квалифицированные специалисты. По результатам заполнения указанных таблиц может быть определена причина несоответствия с наибольшим вероятностным значением (XIV). Следующим этапом производится исследование достоверности обнаруженной вероятной причины. С этой целью осуществляется возврат к первому этапу (XVI), в котором указанная причина рассматривается как новое сообщение для которой будут выполнены все шаги, описанные выше. Данный процесс будет повторяться для всех остальных предполагаемых вероятных причин до тех пор, пока не будет обнаружена действительная причина наблюдаемого несоответствия (XV). Далее, для выявления корневой причины, необходимо использовать заранее составленную многоуровневую причинно-следственную диаграмму (матрицу) или диаграмму взаимозависимостей, в которой, для каждой вероятной причины верхнего уровня могут быть несколько связанных причин нижнего уровня. Тогда, согласно указанной диаграмме необходимо последовательно выбрать данные об очередной вероятной причине и опять повторить приведенную выше процедуру до тех пор пока не будут проанализированы данные для последнего уровня диаграммы. Реализация

указанных задач в системе BVR QMS позволяет определить корневую причину на основе матрицы зависимостей, отображающей цепочку многоуровневой взаимосвязи вероятных и корневых причин.

На рисунке 3.4 приведен пример матрицы для фармацевтического производства, представляющей собой цепочку взаимосвязанных элементов. Как видно из рисунка, для двух групп несоответствий (Производство, Стандартные Операционные Процедуры, СОП) заданы многоуровневые связанные таблицы взаимозависимостей с вероятностными значениями, полученными на основе статистических данных.

Рисунок 3.4. Пример фрагмента многоуровневой матрицы зависимостей несоответствий и их вероятностных значений в фармацевтическом производстве.

Запись в таблице взаимосвязей каждого уровня содержит следующую информацию:

ГРУППА ОТКЛОНЕНИЙ	ТИП ОТКЛОНЕНИЯ – обнаруженный - запланированный	ИСТОЧНИК – датчики устройств - наблюдение сотрудника - заявка от потребителя - результаты аудита - лабораторные тесты	ТИП ВАЛИДАЦИИ – помещение - устройство - готовая продукция - система, процесс - метод, процедура
Указатели на связанные объекты			

Рисунок 3.5. Структура записи в матрице зависимостей для фармацевтического производства.

Одной из основных задач для машинной реализации представленного метода является идентификация неструктурированных сообщений о несоответствиях, которые могут носить случайный характер как по времени поступления (наблюдения) так и по описанию. Поэтому и пару значений (a_j , R_i) можно считать случайной, хотя источники сообщений представляют собой конечное множество {A } детерминированных значений. Для идентификации неструктурированных сообщений в системе BVR QMS реализован метод анализа данных, включающий в себя выполнение следующих этапов:

- семантический разбор предложения, поиск по ключевым словам, содержащимся в базе данных и по похожим сообщениям, которые были зарегистрированы и обработаны в прошлом
- идентификация сообщения и соответствующей группы, если определены более 80% терминов
- генерация интерактивной контрольной (скоринговой) таблицы для сообщений с менее чем 80% распознанных терминов. Данная таблица заполняется компетентным специалистом для дальнейшего семантического разбора сообщения. Идентификация сообщения и соответствующей группы.

На рис 3.6 представлена структурно-функциональная схема программной системы обнаружения несоответствий, вероятных и корневых причин:

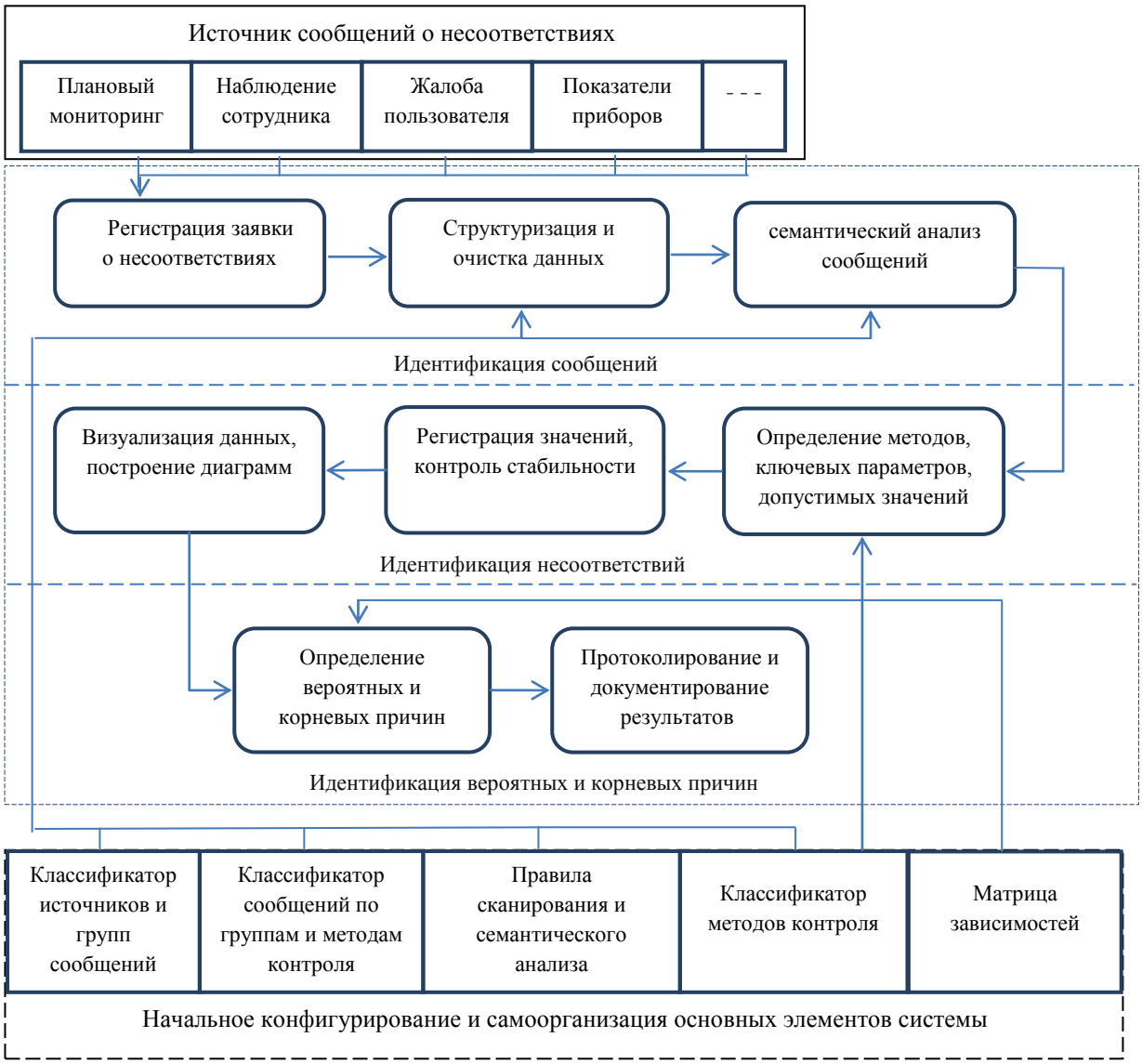


Рисунок 3.6. Структура программной системы контроля стабильности.

Предложенный метод применим как для рассмотренных количественных ключевых параметров, так и для контроля параметров с бинарными и качественными (текстовыми) значениями. В этом случае идентификация отклонений осуществляется путем проверки на соответствие фактического значения контролируемого параметра допустимому значению или одному из элементов заданного множества допустимых значений.

3.2. Метод построения подмножества эталонных признаков и объектов по эмпирической выборке

В предыдущих разделах третьей главы мы рассмотрели основные компоненты и методы, реализованные в системе BVR QMS, позволяющие идентифицировать несоответствия и обнаружить корневые причины их возникновения. В данном разделе мы рассмотрим задачу определения подмножества эталонных объектов для классификации новых сообщений контрольной выборки путем нахождения близких по признакам сообщений из эмпирической выборки. В основе рассматриваемой задачи находится гипотеза отображения схожих сообщений в один и тот же вектор тематических классов (гипотеза компактности) [97-99] .

Введем следующее понятие:

$\beta^l(m'_j) = (\beta(m'_j, m_1), \beta(m'_j, m_2), \dots \beta(m'_j, m_l))$ – множество весовых коэффициентов, показывающих степень близости (схожести) нового сообщения m'_j с сообщениями m_i из выборки M_l. Чем выше значение $\beta(m'_j, m_i)$, тем ближе по признакам сообщение m'_j к m_i. Тогда, алгоритм $a: M \to Y$ классификации m'_j можно записать как нахождение вектора классов $Y(m'_j) = Y(m_i)$ для m'_j, соответствующего тому сообщению $m_i \in M_l$, при котором обеспечивается максимальное значение коэффициента $\beta(m'_j, m_i)$

$$a(m'_j, M_l) = (g(m_i) | max_{(m_i \in M_l)} \beta(m'_j, m_i)) .$$

Определим значения $\beta(m'_j, m_i)$. Для этого рассмотрим два возможных случая:

1. Признаки в сообщениях m'_j и m_i представлены действительными значениями

$(u_1(m'_j), u_2(m'_j), \dots u_v(m'_j))$, $(u_1(m_i), u_2(m_i), \dots u_v(m_i))$, тогда используя формулу евклидова расстояния получим:

$$\beta^l(m'_j) = (\sum_{d=1}^{v} | u_d(m'_j) - u_d(m_i)|)^{1/2}$$

Таким образом можем рассчитать степень близости нового сообщения m'_j к каждому сообщению m_i из выборки M_l.

2. Признаки в сообщениях m'_j и m_i представлены смысловыми (текстовыми) значениями $(w_1(m'_j), w_2(m'_j), \dots w_s(m'_j))$, $(w_1(m_i), w_2(m_i), \dots w_s(m_i))$, тогда в качестве значений $\beta^l(m'_j)$ будем рассматривать следующее выражение

$\beta^l(m'_j) = 1 - \frac{\sum_{k=1}^{s}(w_k(m'_j) \neq w_k(m_i))}{s}$, где s – кол-во смысловых признаков в сообщении m'_j,

Будем считать, что выражение $[w_k(m'_j) \neq w_k(m_i)]$ равно “1”, если оно истинно и равно “0” в противном случае.

На практике часто можно наблюдать признаки, представленные как смысловыми значениями так и в виде действительных чисел. В таких случаях можно рассматривать весовой коэффициент в виде некоторой функции (суммы) от двух рассмотренных выше типов весов:

$\beta^l(m'_j) = \lambda_1(\sum_{d=1}^{l} | u_d(m'_j) - u_d(m_i)|)^{1/2} + \lambda_2 \left(1 - \frac{\sum_{k=1}^{s}(w_k(m'_j) \neq w_k(m_i))}{s}\right)^{1/2}$, здесь λ_1 и λ_2 – коэффициенты нормирования, которые можно подобрать таким образом, чтобы сумма $\beta^l(m'_j)$ находилась бы в некоторой заданной области, например в пределах (0÷1). На рис. 3.7 представлена структурная схема процесса отображения нового сообщения m'_j на вектор тематических классов $Y(m'_j) \subset Y$ путем нахождения ближайших k “соседей” m_i из исходной выборки M_l и соответствующего вектора $Y(m_i) \subset Y$. Ниже представлена структурная схема компонент, входящих в процесс классификации несоответствий, реализованный в системе BVR QMS.

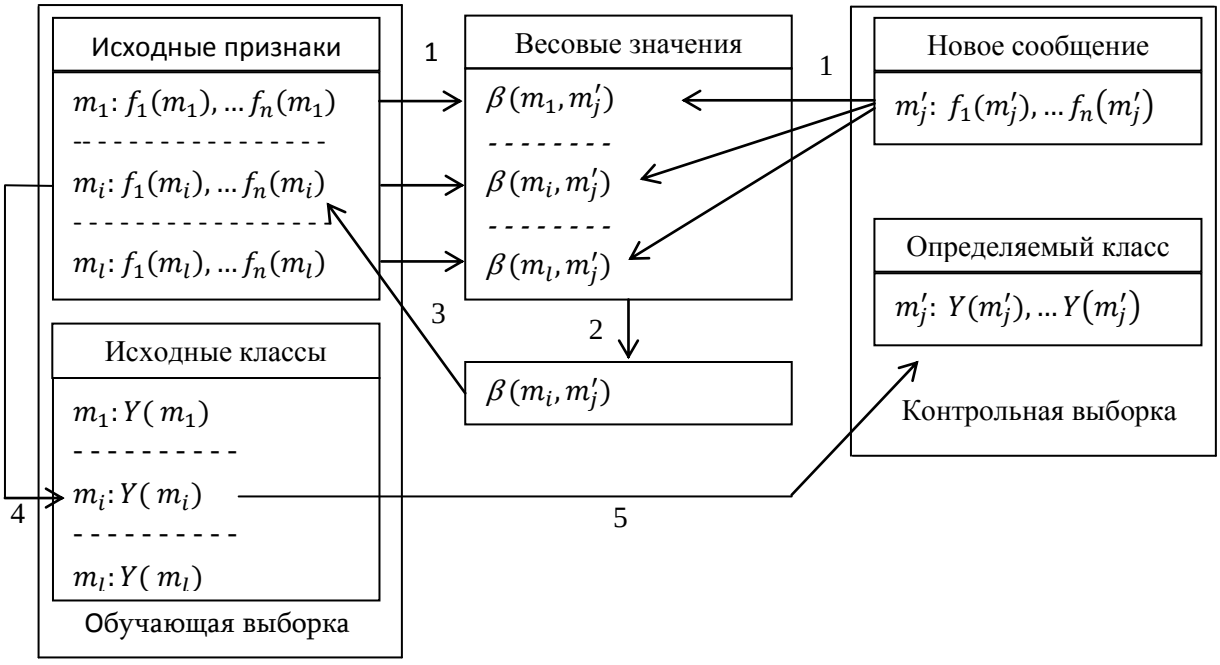


Рисунок 3.7. Структурная схема классификации сообщения на основе обучающей выборки.

Однако, данный подход, основанный на нахождении близкого по признакам сообщения является неустойчивым к возможным погрешностям [100]. Это обусловлено следующими двумя возможными причинами:

- двум близким сообщениям могут соответствовать различные классы;
- класс, определенный по обучающей выборке для сообщения с контрольной выборки не соответствует данным, полученным эмпирическим путем, которые будем считать эталонными.

Одним из подходов к преодолению данной проблемы заключается в решении задачи нахождения такого k ближайших по признакам сообщений при котором алгоритм классификации обеспечит выходные результаты с минимальными погрешностями. Это можно сделать с помощью метода скользящего контроля [101], основанного на эмпирической оценке рассматриваемого нами алгоритма классификации путем сопоставления вычисляемых данных на контрольной выборке с данными, имеющими на обучающей выборке. Как и прежде будем считать, что выражение $[\,Y(m_i) \neq Y(m_j')]$ равно “1”, если оно истинно и равно “0” в противном случае.

Сумму ошибок представим в виде выражения

$$\theta(m_i, m_j') = \sum_{i=1}^{l} [\,Y(m_i) \neq Y(m_j')]$$

Таким образом выбор оптимального алгоритма заключается в последовательном выполнении следующих шагов итерационного процесса:

- нахождение близких сообщений из обучающей выборки для новых поступающих сообщений из контрольной выборки,
- оценка эмпирических данных с вычисляемыми значениями путем назначения весовых коэффициентов и сравнения результатов с допустимым уровнем погрешности,
- динамическая настройка ключевых параметров скользящего контроля.

Ниже представлены шаги итерационного процесса классификации сообщений о несоответствиях, реализованного в системе BVR QMS:

1. Определение функции $g(M_l, Y_l)$ на обучающей выборке M_l.

2. Составление алгоритма, реализующего решающую функцию отображения $\mathrm{a}: \mathrm{M} \to \mathrm{Y}$. Номер первого близкого сообщения примем равным $k \coloneqq 1$.
3. Определение класса Y посредством алгоритма a для сообщения из контрольной выборки путем анализа данных по очередному близкому сообщению из обучающей выборки с номером k.
4. Сопоставление результатов с эмпирическими данными и подсчет коэффициента ошибки $\theta(m_i, m_j')$.
5. Увеличение размерности выборки $k \coloneqq k + 1$, если $\theta(m_i, m_j') > \varepsilon$, где ε - заданный порог (целое число). Повторение пунктов 3,4.

Представленный алгоритм позволяет реализовать основную задачу классификации сообщений о несоответствиях за счет итерационного процесса пошаговой оптимизации ключевых параметров нахождения близких сообщений из обучающей выборки и динамической настройки весовых коэффициентов для оценки и уменьшения погрешностей в определении тематических классов для рассматриваемых текущих сообщений из контрольной выборки.

Приведем пример из практики предприятия фармацевтической отрасли:

Обучающая выборка состоит из 130 сообщений об отклонениях (несоответствиях) с 4-мя признаками:

$$M_l = M_{130},\ f(m_i) = \{f_1(m_i),\ f_2(m_i), f_3(m_i), f_4(m_i)\} \quad \text{где } i = (1 \div 130),$$

и тремя группами классов Y_4, Y_5, Y_3.

Каждому сообщению $m_i,\ i = 130$ ставится в соответствие вектор классов:

$Y(m_i) = (Y_1(m_i), Y_2(m_i), Y_3(m_i))$, где $(Y_1(m_i) \subset Y_4$, $Y_2(m_i) \subset Y_5$, $Y_3(m_i) \subset Y_3$).

Заданы признаки:

признак 1: [Тип отклонения] = {<обнаруженный>, <запланированный>}

признак 2: [Источник обнаруженного несоотвествия] = {<готовая продукция>, <сырье>, <балк-продукт>, <процесс>, <оборудование>, <процедура>, <персонал>, <система жизнеобеспечения>}

признак 3: [Предполагаемая причина] = {<ошибка отбора проб>, <неправильный метод>, <процедурная ошибка>, <ошибка микробиологического анализа>, <ошибка сотрудника>, <неопределенная ошибка>}

признак 4: [Подразделение]= {<Отдел качества><Лаборатория><Технический отдел>}

Заданы группы классов:

Группа 1: <Оценка несоответствия>

Классы: <Несущественный>, <Значительный>, <Критический>, <Неопределенный>

Группа 2: <Действия>

Классы: <Действий не требуется>, <Пригласить экспертов>, <Провести дополнительное исследование>, <Выполнить корректирующие и превентивные действия>, <Перевести в процесс “управления изменениями”>

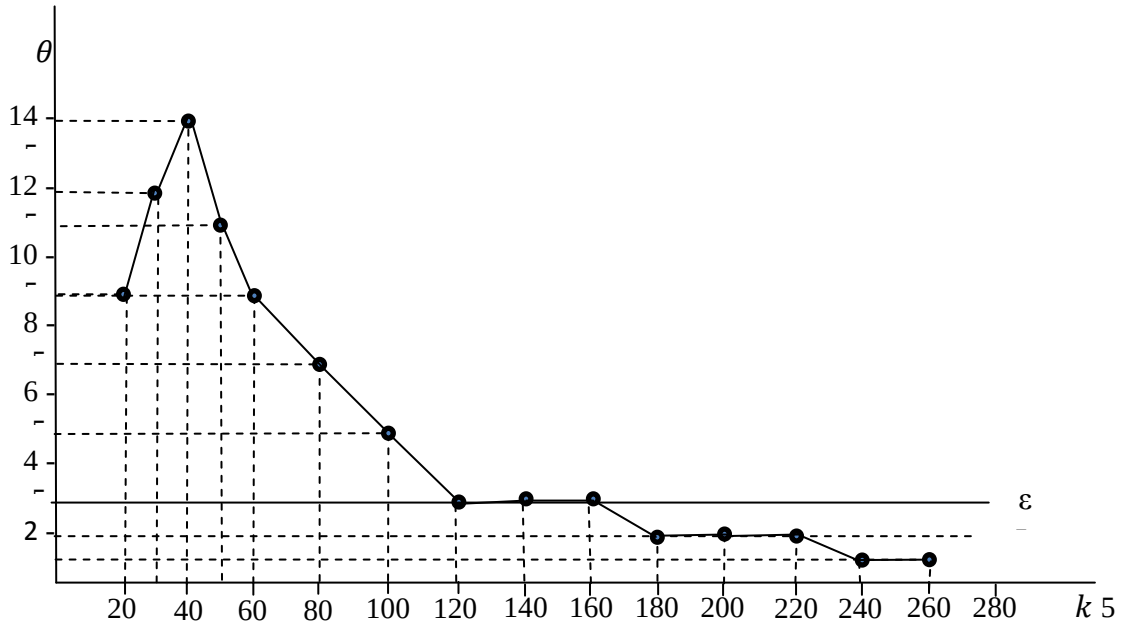
Группа 3: <Стандартные Операционные Процедуры>

Классы: <Изменения в процедуре>, <Повторный анализ ингредиентов>, <Запрос сертификата качества у производителя>.

На рисунке 3.8 приведена структурная схема рассматриваемого примера.

Рисунок 3.8. Структурная схема фрагмента классификации несоответствий в фармацевтической отрасли.

Заслуживает внимания реализованный в системе BVR QMS подход к динамическому назначению каждому элементу обучающей выборки некоторого весового коэффициента, задающего степень важности данного обучающего сообщения. Для этого введем переменную $\gamma(m_i) = \frac{\sum_{j=1}^{g}[Y(m_i)=Y(m_j')]}{n}$, где g - количество сообщений из контрольной выборки, для которых близким по признакам было сообщение m_j' из обучающей выборки, n- количество элементов в контрольной выборке. Очевидно, что коэффициент $\gamma(m_i)$ следует рассматривать при значении k, при котором минимален показатель ошибки при сравнении обучающей и контрольной выборок, иными словами, когда выполняется условие $\theta < \varepsilon$. В таком случае, в представленном алгоритме классификации на шаге итерации, при котором будет достигнуто условие $\theta < \varepsilon$ можно добавить функцию вычисления нового значения $\gamma(m_i)$ для элемента m_i из обучающей выборки. Ниже представлен график зависимости коэффициента ошибки θ от рассматриваемого количества k близких по признакам сообщений из обучающей выборки к текущему сообщению m_i из контрольной выборки. В качестве информации выбраны результаты анализа данных фармаконадзора, проведенного на предприятии фармацевтической отрасли:



Как видно из графика для выбранного значения ε начиная с $\mathrm{k} = 180$ можно говорить об относительной точности алгоритма классификации.

Дополнительно, уменьшение погрешности достигается путем динамической корректировки коэффициента важности γ для сообщений из обучающей выборки.

Данная корректировка производится автоматически после каждой классификации сообщений из контрольной выборки, начиная с $k = 180$. Система регистрирует ключевые параметры, такие как тип контроля “плановый контроль” или “обнаруженный” , наблюдаемый объект: оборудование, система, ингредиент и др. Используя метод скользящего контроля предложенный алгоритм определяет число близких по заданным признакам сообщений, при которых погрешность, допущенная машинным методом классификации меньше заданного допустимого порога. Для сравнения результатов машинной классификации используются данные , получаемые эмпирическим путем. На основе обнаруженных близких по признакам сообщений из обучающей выборки легко определить релевантные классы, которые содержат информацию о степени риска (“Критический”, “Значительный”, “Несущественный”, “Неопределенный”…), вероятной причине отклонения (“Ошибка устройства”, “Неправильный метод”, “Процедурная ошибка”…), и данных по проведению корректирующих и превентивных действий (“Замена датчика автоклава”, “Профилактика упаковочной машины”, “Изменения в процедуре подготовки комнаты”, “Обучение сотрудника” ...). Таким образом, вычислительная система, построенная на предложенном алгоритме сопоставления обучающей и контрольной выборках будет генерировать экспертное заключение по проведению процедур CAPA. При этом, основные ключевые показатели классификации динамически изменяются (оптимизируются) каждый раз после обработки очередного нового сообщения из контрольной выборки. Таким образом, рассматриваемый алгоритм реализует принцип саморазвития и самонастраиваемости системы классификации несоответствий и минимизирует погрешности результатов машинной обработки данных. Использование предложенного метода позволяет легко адаптировать систему BVR QMS для автоматизации задач классификации несоответствий. Дальнейшим развитием предложенного метода является исследование вопросов классификации сообщений, содержащих неструктурированные тексты. [102-109]

3.3. Классификация объектов на основе решающих функций

Задана обучающая выборка $\mathrm{A}_{\mathrm{i}}^{\mathrm{l}}$ объектов из A_{i}, для которых определены значения ключевых признаков F и поставлены в соответствие определенные классы Y:

$A_i^l = (a_{ij}, y_j)_{j=1}^l \subset F \times Y$ – обучающая выборка, Y – конечное множество заданных классов, состоящее из двух элементов $Y = \{y_1, y_2\}$. Логической закономерностью или правилом будем называть такой предикат (решающую функцию), которая каждому объекту в результате оценки признаков ставит в соответствие ноль либо единицу. Необходимо составить семейство предикатов, включающих в себя логические закономерности с признаками $f(a_{ij}) \subset F$ и обладающих достаточной информативностью (предсказательностью), которое позволило бы отнести данный объект к тому или иному классу, другими словами реализовать правило отображения (идентификации): $r: F \to Y$. Далее, для новых значений ключевых показателей $f(a_{ij})$ подобрать оптимальный алгоритм, который на основе выведенных и хорошо интерпретируемых предикатов r позволил бы путем оценки признаков $f(a_{ij})$ классифицировать объект a_{ij} для последующего определения стабильности контролируемой системы. Под оптимальным алгоритмом будем понимать такой алгоритм, который обеспечит минимальный уровень погрешности при определении классов для заданного объекта. Необходимо определить метод оценки этой погрешности. На рисунке 3.9. приведен пример классификации с допущенными погрешностями. Здесь объекты отмечены “0” и “1” согласно значениям соответствующих предикатов. Объекты находятся в пространствах C_p и C_n – представляющих собой классы соответственно, из области правильной классификации для объектов “1” и ошибочной классификации для объектов “0” и наоборот, C_n и C_p – есть области правильной и ошибочной классификации соответственно для объектов “0” и “1” .

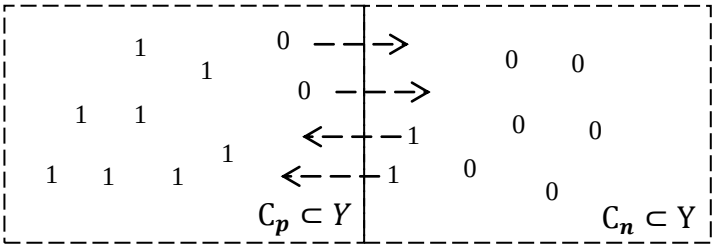


Рисунок 3.9. Распределение объектов по двум классам.

Задача оптимизации алгоритма заключается в том, чтобы из заданной выборки как можно большее число объектов с значением предиката = “1” было бы расположено в области C_p, и как можно большее число объектов с значением предиката = “0” – в области C_n. Мы исследуем вопросы закономерностей путем

выбора некоторого параметрического семейства функций (предикатов), включающих операции над признаками объектов и определим пространство объектов и критерии поиска для этих функций. Необходимо решить задачу объединения обнаруженных закономерностей в определенный алгоритм, который классифицирует рассматриваемые нами объекты согласно значениям вычисляемых предикатов, содержащих в качестве аргумента значения ключевых признаков этих объектов. При поступлении признаков на вход составленных алгоритмов мы должны получить надежную классификацию объектов на основе реализации обнаруженных закономерностей отображения в заданные тематические классы. Для поиска закономерностей в представленном классе задач в системе BVR QMS реализованы следующие три конструкции параметрических функций:

1. Предикат(конъюкция) контроля признаков с вещественными значениями

$r(a_{ij}) = \wedge_{l\in\{1,k(A_i)\}}[\beta_{l1} \leqslant f_l(a_{ij}) \leqslant \beta_{l2}]$,

β_{l1}, β_{l2} - пороговые значения верхних и нижних границ. Допустимы, также частные случаи, когда $\beta_{l1} = -\infty$ или $\beta_{l2} = +\infty$

Приведем пример из фармацевтики для предиката, содержащего конъюнкцию их двух анализируемых признаков:

Объект a_{ij}=<Балк-продукт хлорида натрия 0,9 %>

1-ый признак: $f_1(\mathrm{a_{ij}})$ =< Объем инъекционной воды>

2-ой признак: $f_2(\mathrm{a_{ij}})$ =< Объем хлорида натрия>

Предикат:

$r(\text{Хлорид натрия } 0{,}9\%) = (1600\text{л} \leqslant \text{Объем инъекционной воды} \leqslant 1700\text{л}) \wedge (14800\text{гр} \leqslant \text{Объем хлорида натрия} \leqslant 14900\text{гр})$;

2. **Предикат (конъюнкция с числом признаков = λ) контроля признаков с вещественными числами**

$r(a_{ij}) = [\sum_{l\in\{1,k(A_i)\}}[\beta_{l1} \leqslant f_l(a_{ij}) \leqslant \beta_{l2}] \geqslant \lambda]$,

если $\lambda = l$, то имеет место строгая конъюнкция,

если $\lambda = 1$, то имеет место условие дизъюкции

Например, предикат, содержащий три признака по фармаконадзору, из которых как минимум два произвольных объединены условием конъюнкции:

Объект a_{ij}=<Карта сообщения фармаконадзора>

1-ый признак: $f_1(\mathrm{a_{ij}})$ =< Арт. давление(сист.)>

2-ой признак: $f_2(\mathrm{a_{ij}})$ =< Концентрация сахара в крови>

3-й признак: $f_3(\mathrm{a_{ij}})$ =< Концентрация гемоглобина в крови>

Предикат:

$r(\text{Карта сообщения фармаконадзора}) =$

$\left[(160\text{мм рт. ст} \leqslant \text{Арт. давл. (сист.)})\right] + [8\ \text{ммоль} \leqslant \text{Концентрация сахара в крови}] +$

$\left[\left(\frac{150\text{г}}{1\text{л}} \leqslant \text{Концентрация гемоглобина}\right)\right]) \geqslant 2,$

3. Предикат с текстовыми (тематическими) признаками

$r(a_{ij}) = [\sum_{l\in\{1,k(A_l)\}}[f_l(a_{ij}) \neq f_l'] \geqslant \lambda]$**,**

где f_l' **-** некоторое заданное эталонное значение

или

$r(a_{ij}) = [\sum_{l\in\{1,k(A_l)\}}[f_l(a_{ij}) \notin F_l'] \geqslant \lambda]$,

где $F_l' = \left\{f_{l,1}', f_{l,2}', \dots, f_{l,k(F_l')}'\right\}$, - есть множество эталонных значений.

Ниже приведен пример с данным предикатом:

предикат, содержащий три тематических признака, объединенных условием конъюнкции:

Объект a_{ij}=<Производственный этап>

1-ый признак: $f_1(\mathrm{a_{ij}})$ =< Тип образца(объект)>

2-ой признак: $f_2(\mathrm{a_{ij}})$ =< Тип отклонения>

3-й признак: $f_3(\mathrm{a_{ij}})$ =< Этап производственного процесса>

Предикат:

$r(\text{Производственный этап}) =$

$(\text{Тип образца(объект)} = \text{"Процесс"}] + [\text{Тип отклонения} = \text{"Обнаруженный"}] +$

$[\text{Этап производственного процесса} = \text{"Розлив"}]) \geqslant 3$

Для всех трех приведенных выше примеров алгоритм классификации при значении предиката, равного “1” должен поставить в соответствие класс y_2 (“Стабильность не нарушена”) из заданного множества двух классов $Y = \{y_1, y_2\}$.

Нам необходимо построить алгоритм классификации, согласно которому предикаты будут выдавать значение =1 для того множества объектов, которым в обучающей выборке соответствует определенный класс из Y . Алгоритм тем оптимальнее, чем больше объектов из исходной выборки он может правильно отнести к “своему” классу.

Для решения приведенных задач, в системе BVR QMS реализована процедура стохастического локального поиска, суть которого заключается в следующем:

эмпирическая выборка разбивается на два подмножества: $A_i^M(y^M)$ и $A_i^V(y_j^V)$, включающих в себя M и V объектов, таким образом, чтобы они относились соответственно классам $y^M \in Y$ и $y_j^V \in Y$. Берем начальное множество правил $R(\mathrm{A}_i)$, для объектов класса A_i с назначенными значениями пороговых значений $\beta_{l1}, \beta_{l2}, \lambda$. Далее, на вход алгоритма подаем обучающую выборку $\mathrm{A}_\mathrm{i}^\mathrm{l}$ с предопределенной парой значений $(\mathrm{a_{ij}}, \mathrm{y_j})$. Производится вычисление признаков $\mathrm{f}(\mathrm{a_{ij}})$, значения которых в качестве аргументов используются в формулах-предикатах семейства $R(\mathrm{A}_i)$. Выделим подмножество объектов $a_{ij}^m \in A_i^m(y_j^m) \subset A_i^l$, для которых $R(\mathrm{A}_i) = 1$, а следовательно поставлен в соответствие определенный класс $y_j^m \in \mathrm{Y}$. Выделим также подмножество $a_{ij}^v \in A_i^v(y_j^v) \subset A_i^l$, для которых $R(\mathrm{A}_i) = 0$ и поставлен в соответствие класс $y_j^v \in \mathrm{Y}$.

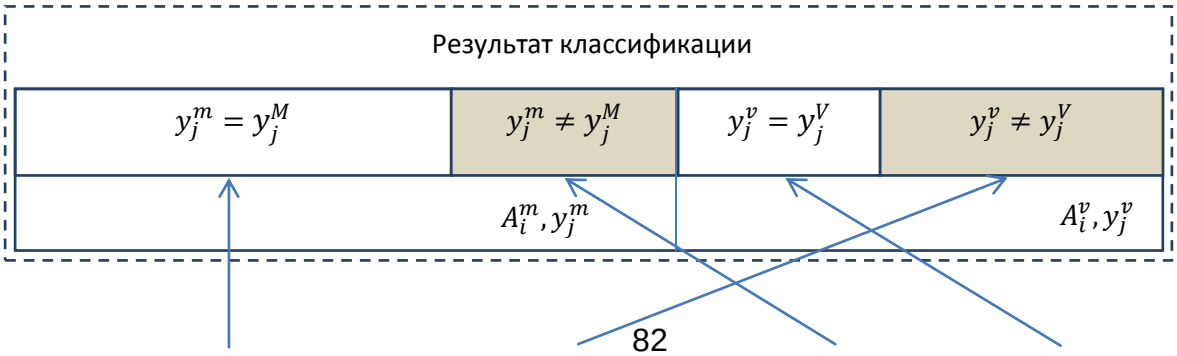
Для оценки корректности классификации выделим количество корректно $P(A_i^m)$ и $P(A_i^v)$ и ошибочно $N/(A_i^m)m$, $N(A_i^v)$ покрытых классификатором объектов:

$$P(A_i^m) = \left(a_{ij}^m \in A_i^m \middle| y_j^m = y_j^M\right)$$
$$N(A_i^m) = (a_{ij}^m \in A_i^m | y_j^m \neq y_j^M)$$
$$P(A_i^v) = \left(a_{ij}^v \in A_i^v \middle| y_j^v = y_j^V\right)$$
$$N(A_i^v) = (a_{ij}^v \in A_i^v | y_j^v \neq y_j^V)$$

Тогда, для достижения оптимальности алгоритма классификации, необходимо добиться того, чтобы количество объектов, правильно покрытых классификатором было бы максимальным, а ошибочно определенных объектов – минимальным, т.е.

$$P(A_i) = (P(A_i^m) + P(A_i^v)) \to max$$
$$N(A_i) = (N(A_i^m) + N(A_i^v) \to min.$$

На рисунке 3.10 приведена иллюстрация классификации объектов по эмпирической выборке:



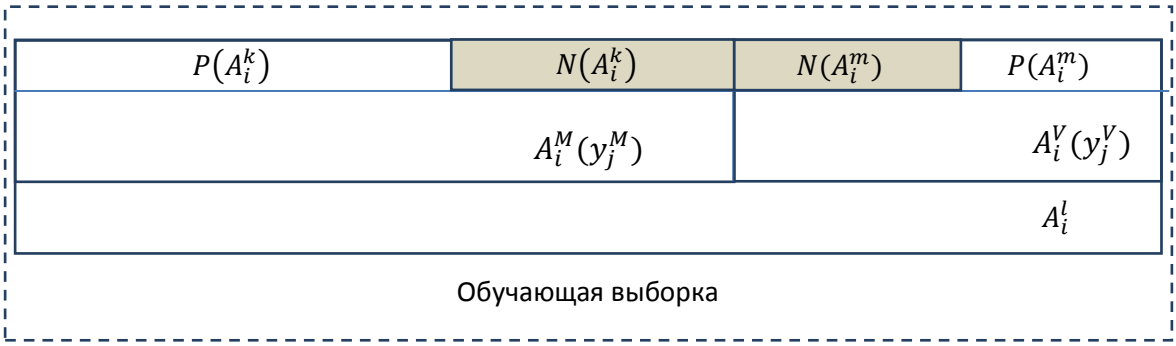


Рисунок 3.10. Классификация предикатными функциями на эмпирической выборке.

Как видно из рисунка, в результатах классификации имеются погрешности при анализе признаков и расчете предикатов для объектов эмпирической выборки. Необходимо оценить степень этих погрешностей. Значимость результатов классификации (информативности) может быть оценена одним из существующих энтропийно-статистических методов.

Допустим, справедлива гипотеза о независимости событий $\{\mathrm{A_i^l} = (\mathrm{a_{ij}}, \mathrm{y_j})_{\mathrm{j=1}}^{\mathrm{l}} \subset \mathrm{F} \times \mathrm{Y}\}$ и $\{\mathrm{r(F)} = 1\}$. Тогда вероятность реализации пары (P,N) подчиняется энтропийному распределению [110,111,112]:

$$\mathrm{F(P,N)} = \mathrm{h}\left(\frac{\mathrm{M}}{\mathrm{l}}\right) - \frac{\mathrm{P(A_i)+N(A_i)}}{\mathrm{l}}\,\mathrm{h}\left(\frac{\mathrm{P(A_i)}}{\mathrm{P(A_i)+N(A_i)}}\right) - \frac{\mathrm{l-P(A_i)-N(A_i)}}{\mathrm{l}}\mathrm{h}\left(\frac{\mathrm{M-P(A_i)}}{\mathrm{l-P(A_i)-N(A_i)}}\right)$$

где, $\mathrm{h(q)} = -\mathrm{g}\log_2 \mathrm{q} - (1-\mathrm{q})\log_2(1-\mathrm{q})$

Задача состоит в построении максимального значения оценок $\mathrm{F(P,N)} \to \max$ для наблюдаемых значений, т.е. мы можем установить некоторый уровень значимости Ω и проверить степень информативности

$$\mathrm{F(P,N)} > \Omega.$$

Тогда, задача будет сведена к выбору таких пороговых значений β_{l1}, β_{l2} в предикатах $\mathrm{R(A_i)}$ при которых будет достигнут требуемый уровень информативности.

В системе BVR QMS для решения этой задачи выполняется определенное количество циклов, на основе которых проводится автоматически оценка информативности. Процесс выполняется до тех пор, пока новые пороговые значения не перестанут улучшать правила. Далее, производится исключение дублирующих друг друга предикатов и остаются только те предикаты, которые несут наибольшую информативность.

Приведем описание метода и алгоритма, основанных на стохастическом локальном поиске:

Входные данные:

1. Исходная эталонная выборка $A_i^l = (a_{ij}, y_j)_{j=1}^l \subset F \times Y$, где
   $F = \{f_1(a_{ij}), f_2(a_{ij}), \dots f_{k(A_i)}(a_{ij})\}$- множество признаков
   $Y = \{y_1, y_2\}$ – классы, $y_j \in Y$.
2. Эталонные подвыборки, разделенные по классам $A_i^M(y_j^M) \subset A_i^l$, $A_i^V(y_j^V) \subset A_i^l$, где $A_i^M \cup A_i^V = A_i^l, M + V = l$, $y_j^M \neq y_j^V$, $y_j^M \in Y$, $y_j^V \in Y$.
3. Семейство предикатов $r(r_1, r_2, \dots, r_{n(A_i)}): F \to Y$.
4. Показатель Ω для оценки энтропии.

Выходные данные:

множество оптимальных пороговых значений β_{l1} и β_{l2} для каждого признака $f_i \in F$

Алгоритм:

1. Выбираем a_{ij} из A_i^l , $j = \{1,2, \dots\}$.
2. Регистрируем значения признаков $\{f_1(a_{ij}), f_2(a_{ij}), \dots f_{k(A_i)}(a_{ij})\} \subset F$.
3. Вычисляем предикаты

$$r_1(f_1(a_{ij}), f_2(a_{ij}), \dots f_{k(A_i)}(a_{ij})),$$
$$r_2(f_1(a_{ij}), f_2(a_{ij}), \dots f_{k(A_i)}(a_{ij})),$$
$$- - - - - - - - - - - - - - -$$
$$r_{n(A_i)}(f_1(a_{ij}), f_2(a_{ij}), \dots f_{k(A_i)}(a_{ij}))$$

4. Вычисляем значение решающего предиката $r(a_{ij})$, содержащего в качестве термов $r_1, r_2, \dots, r_{n(A_i)}$, объединенных операциями конъюнкции и дизъюнкции.
5. Формируем подмножества A_i^m и A_i^v из объектов a_{ij} , согласно значениям решающих предикатов. Если $r(a_{ij}) = 1$, то a_{ij} добавляется к множеству A_i^m, при значении $r(a_{ij}) = 0$, объект a_{ij} будет отнесен к множеству A_i^v.
6. Проверяем достоверность классификации путем сравнения алгоритмически полученного класса объекта a_{ij} с заданным в обучающей выборке. При этом, формируем следующие подмножества объектов при выполнении соответствующих условий:

- если $y_j^m = y_j^M$, то имеется корректная классификация и подсчитывается количество правильно покрытых предикатом объекты $P(A_i^k) := P(A_i^k) + 1$;

- если $y_j^m \neq y_j^M$, то имеется ошибочная классификация и подсчитывается количество неправильно покрытых предикатом объекты $N(A_i^m) := N(A_i^m) + 1$;

- если $y_j^v = y_j^V$, то имеется корректная классификация и подсчитывается количество объектов, соответствующих классу y_j^V : $P(A_i^m) := P(A_i^m) + 1$;

- если $y_j^v \neq y_j^V$, то имеется ошибочная классификация и подсчитывается количество неправильно классифицированных объектов $N(A_i^k) := N(A_i^k) + 1$;

- если классификация была ошибочной, то корректируются пороговые значения $\beta_{l1} := \beta_{l1}(a_{ij})$ и $\beta_{l2} := \beta_{l2}(a_{ij})$;

- если не все объекты исходной выборки рассмотрены, то повторяются пункты 1-7 для следующего объекта a_{ij}.

7. Подсчитывается энтропийное распределение $F(P, N)$ на основе полученных значений $P(A_i), N(A_i), M$. Если $F(P, N) < \Omega$, то повторить пункты 1-8 для значений β_{l1}и β_{l2}, полученных на предыдущей итерации.

Данный алгоритм позволяет построить предикаты с оптимальными пороговыми значениями для выполнения классификации с допустимым уровнем погрешности.

На рисунке 3.11 представлены графики зависимости коэффициента информативности предикатов $F(P, N)$ от количества итераций N для выборочных двух подмножеств объектов $A = \{\text{"Оборудование"}, \text{"Процедура"}\}$ звеньев системы идентификации несоответствий в фармацевтической отрасли.

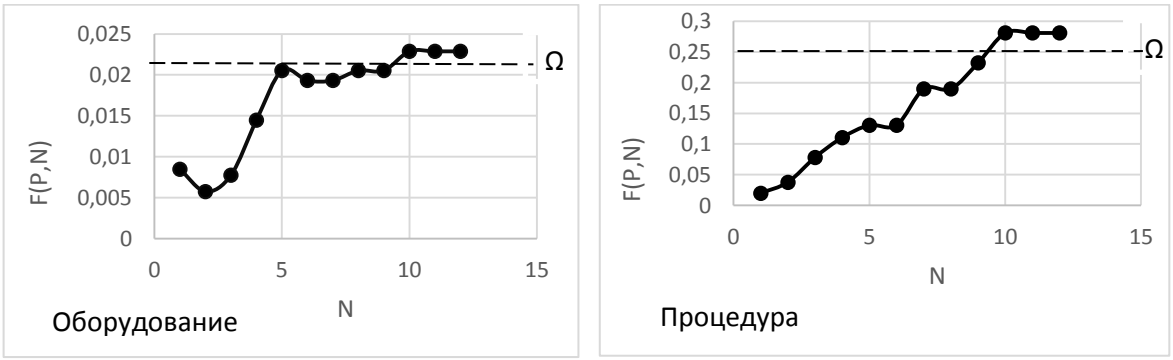


Рисунок 3.11.Энтропийная оценка информативности при различных итерациях.

Как видно из графиков после 10-й итерации прогона алгоритма был достигнут требуемый уровень значения информативности, превышающий порог Ω, после которого наблюдалась стабилизация результатов классификации.

Таким образом, пороговые значения β_{l1}, β_{l2} в предикатах классификации для рассмотренного примера, полученные после 10-й итерации достигают оптимального значения.

Представленный метод классификации позволил эффективно идентифицировать такие несоответствия как инциденты в производственных этапах, сообщения фармаконадзора, отклонения в процессах, в стандартных операционных процедурах, показателях приборов систем и оборудования и др.

3.4. Особенности и основные характеристики информационной системы BVR QMS

В системе BVR QMS использованы все необходимые инновационные технологические инструменты разработки, среди которых, в частности, среда .NET , сервер приложений Microsoft Dynamics, система анализа данных и построения аналитической отчетности Power BI, облачная платформа Azure, языки программирования C#, Python, Javascript. При этом система BVR QMS обеспечивает:

- целостность информации рассматриваемой предметной области в реляционной системе управления базой данных (СУБД) MS SQL;
- протоколирование всех производственных процессов, процедур, этапов производства, включая всевозможные отклонения, инциденты, изменения и др;
- достоверность корпоративных данных обеспечивается за счет контроля вводимых данных и автоматической обработки, согласно действующим алгоритмам и расчетам;
- качество корпоративного управления обеспечивается за счет внедрения технологических инструментов управления, визуализации данных, автоматизации процессов, что обеспечивает быстрое и точное принятия управленческих решений на всех уровнях структуры организации;
- корректное функционирование системы путем использования инструментов напоминания, автоматической генерации заданий, контроля действий

операторов, подлежащих выполнению, выделения полей, обязательных для заполнения (невозможность продолжения процесса, в случае их незаполнения).

На рисунке 3.12 представлена трехуровневая логическая структура информационной системы BVR QMS:

Рисунок 3.12. Трехуровневая логическая структура информационной системы BVR QMS.

Все этапы настройки, разработки, поддержки и администрирования информационной системы были выполнены на основе стандартных промышленных технологий Microsoft.

Система разработана на сервере приложений Microsoft Dynamics, база данных создана на реляционной системе управления базами данных (РСУБД) Microsoft SQL, в качестве платформы для получения аналитической отчетности использована среда разработки Microsoft Visual Studio (MVS) и Microsoft Business Intelligence (BI).

Платформа M**icrosoft Dynamics** позволяет решать весь спектр управленческих задач, связанных с корпоративной и функциональными стратегиями :

- создание единой консолидированной базы корпоративных данных;
- управление бизнес-процессами;
- аналитическая обработка информации, создание отчетов.

Мощные аналитические возможности платформы Microsoft Dynamics, в том числе панели ключевых показателей деятельности для руководителей, предоставляют возможности эффективного планирования, управленческого контроля, измерения основных ключевых показателей и оперативного принятия решений.

Инфраструктура. Все функциональные компоненты BVR QMS работают на платформе Microsoft .NET framework и размещены на корпоративном локальном сервере, функционирующем под управлением Microsoft Windows Server 2012. Структура Microsoft Dynamics CRM позволяет реализовать автоматизацию задач по модульному принципу, что обеспечивает гибкость и простоту внедрения дополнительных функциональных задач и модификацию существующих решений.

Доступ к системе осуществляется с любого устройства (компьютера, ноутбука, планшета, мобильного устройства) с помощью браузера (Internet Explorer, Google Chrome, Mozilla, Forefox, и др) в защищенном режиме путем набора соответствующей строки в браузере и ввода персональных учетных данных: логина и пароля.

Платформа Dynamics представляет собой клиент-серверное приложение. Данная платформа поддерживает все основные интерфейсы веб-сервисов. Клиенты получают доступ к Dynamics CRM через браузер, клиентский плагин к Microsoft Outlook или через планшеты и мобильные устройства.

Права доступа пользователей. Совместная работа сотрудников организации (предприятия) и разграничение доступа к данным обеспечивается согласно ролевым назначениям и политике безопасности, настраиваемыми в системе Microsoft Dynamics. Настройки ролей безопасности производятся уполномоченным сотрудником (администратором) по безопасности.

В системе также, предусмотрена автоматизация процессов, в которых генерируются задания и направляются конкретному сотруднику с инструкциями и напоминаниями для выполнения определенных действий. При этом все произведенные действия в системе протоколируются с указанием данных о

выполнении заданий: исполнитель, дата выполнения, наименование процесса, результаты валидирования и т д .

Основные преимущества:

- консолидированы задачи автоматизации управления и контроля всеми производственно-хозяйственными процессами и регистрации значений соответствующих ключевых показателей;
- представлены в распоряжение персонала широкий спектр инструментов визуализации данных, среди которых списковые представления данных, панели мониторинга, аналитические отчеты, динамические графики и диаграммы, расширенный поиск;
- обеспечен доступ с мобильного устройства в защищенном режиме.

Информационная система BVR QMS в полной мере реализована с требованиями по качеству и надежности, как по вопросу функционирования прикладных задач так и по части выполнения системных заданий. Качество работы также обусловлено путем достижения следующих требований:

- легкость в обучении и работе за счет дружественного пользовательского интерфейса, возможности гибкой настройки под нужды каждого пользователя или группы пользователей;
- оперативность принятия решений и анализа текущего состояния;
- доступность к актуальным данным согласно ролевым назначениям и политики безопасности;
- возможность быстрой и гибкой модификации процессов и дополнения новых задач, которые появляются в силу тех или иных обстоятельств, например по требованию регулятора, при изменении процедур, процессов и т д.

В системе BVR QMS минимизированы возможные риски, связанные с ошибками оператора при вводе данных путем использования широкого набора инструментов, среди которых

- полнофункциональный контроль вводимых данных;
- настройка полей для обязательного ввода;
- автоматическое заполнение полей согласно заданному алгоритму;
- набор подсказок и меню, направляемых действия оператора;
- автоматическое уведомление о необходимости выполнения заданий;

– автоматическая генерация документов, в которых переменные поля заполняются не оператором , а программно с использованием базы данных.

В системе BVR QMS реализована концепция электронных подписей, при которых автоматически проставляются инициалы, дата проставления подписи и назначение подписи.

Модель данных. Данные хранятся в системе СУБД SQL в виде реляционных двумерных таблиц, работа с данными ведется на уровне метаданных, что обеспечивает безопасность и целостность базы данных. С помощью инструментов управления метаданными можно создавать структуру данных в защищенном режиме. В Microsoft Dynamics используется интерфейс программирования приложений (API), который позволяет добавлять или обновлять метаданные. Доступ к данным осуществляется с помощью богатого набора инструментов, предоставляемых сервером приложений Dynamics, а также разработкой отдельных программных модулей на технологиях Java Script, CSS, HTML. Разработка плагинов, работающих на сервере приложений производится на языке программирования C#. Системные задания запускаются с помощью сервисных служб, представляемых операционной системой Windows Server. Резервное копирование осуществляется периодически, согласно временному графику, составленному системным администратором, при необходимости возможно восстановление всей системы и базы данных. Время восстановления системы после сбоя составляет 2-3 мин, время восстановления после отказа, может составлять от 5 до 10 мин.

Данные хранятся на внешних носителях, объединенных в RAID массивы, что делает хранение информации защищенным и безопасным. Доступ к ним осуществляется на протяжении всего периода функционирования системы, согласно политики безопасности, определяющей ролевые назначения сотрудников организации. В системе производится регулярное автоматическое резервное копирование данных согласно заданному периоду копирования. В системе имеются средства аудита, показывающие всю историю изменения тех или иных полей в форме, при этом фиксируются – автор изменения, измененное поле, старое и новое значение поля. Система хранит историю модификации данных, визуализируя их для легкости просмотра, при этом все изменения связанные с процессами, процедурами тщательно регистрируются и протоколируются. События входа в систему формируются контроллерами домена путем проверки учетных записей домена и учетных записей хост компьютеров. При этом производится полный аудит входа

пользователей в систему. Реализован защищенный механизм визирования процедур и процессов, при этом система автоматически проставляет дату электронной подписи, инициалы, назначение подписи (например, составлено, утверждено, согласовано и др), при этом после сохранения данных невозможно эл. подпись удалить или изменить. Разработанная информационная система, построенная на платформах Microsoft® Windows Server, Dynamics Server Application, реляционной системы управления базой данных (СУБД) MS SQL и веб-технологиях характеризуется высокой надежностью и отказоустойчивостью. Одновременно, в системе предусмотрен целый ряд организационно-технических мер и инструкций пользователя, по которым сотрудники в критических ситуациях будут выполнять определенные действия для поддержки непрерывности процессов. Например, в случае останова сервера или в случае длительного сбоя электропитания, данные могут вводиться оператором с хост компьютера в электронную таблицу Excel или другое локальное приложение, затем при введении в строй электропитания и сервера, путем автоматической миграции данных восстанавливается актуальность базы данных. При отказе хост компьютера, сотрудник может легко продолжить работу на другом любом свободном компьютере, предварительно получив доступ к системе с помощью логина и пароля. Вся база данных архивируется автоматически согласно заданному временному периоду архивирования. При необходимости, данные легко могут быть восстановлены сервисами и службами, которые обеспечивает серверная операционная система Windows Server 2008/2017. Все указанные процедуры выполняются согласно разработанным алгоритмам и методам управления несоответствиями и отклонениями, включающими в себя процедуры идентификации проблемы, определения корневых причин и генерации экспертного заключения для персонала с рекомендациями по проведению корректирующих действий и/или изменений.

Система BVR QMS отличается гибкостью, настраиваемостью, масштабируемостью, дружественным пользовательским интерфейсом, доступностью в обучении и использовании в повседневной жизни. При необходимости информационная система может легко масштабироваться с учетом роста компании, формулирования новых требований, модификации существующих и добавления новых процессов. Разработанная система отличается защищенностью, способностью быстро восстанавливаться при сбоях и возможностью удаленного

доступа сотрудников с учетом их ролевых назначений и санкционированности действий.

Выводы к главе 3

1. Представлена структурно-функциональная схема системы BVR QMS.
2. Рассмотрен метод автоматизации процессов идентификации несоответствий и определения корректирующих действий.
3. Представлен алгоритм и вычислительный процесс определения корневой причины случайных нарушений стабильности, рпеализованный в системе BVR QMS.
4. Дано математическое описание стохастических вычислений в задачах управления процедурами CAPA.
5. Разработан метод организации стохастических вычислений в сложных системах рассматриваемого класса на основе предикатных выражений.
6. Представлено описание логической структуры, модели данных и основных характеристик системы BVR QMS.

ГЛАВА 4: РЕЗУЛЬТАТЫ ВНЕДРЕНИЯ СИСТЕМЫ BVR QMS В ФАРМАЦЕВТИЧЕСКОМ ПРОИЗВОДСТВЕ

В предыдущей главе было представлено описание информационной системы BVR QMS, спроектированной на основе исследований, разработанных методов и алгоритмов идентификации отклонений и их классификаций для управления стабильностью систем. В данной главе мы представим результаты внедрения системы BVR QMS, в частности, в фармацевтическом производстве (версия: BVR QMS Pharm), основной задачей которой была автоматизация процессов CAPA посредством контроля соответствия ключевых показателей отраслевым нормативным стандартам. На основе анализа и обработки эмпирических данных автоматизированы процессы расчетов вероятностных значений взаимосвязи различных групп отклонений. Были реализованы задачи автоматизации процесса получения необходимой аналитической отчетности, протоколов CAPA и электронного досье производства, что позволило специалистам управления качеством значительно сократить время, затрачиваемое на внутренние и внешние аудиторские проверки и оптимально определить ресурсы для проведения изменений или корректирующих действий. В основе внедрения системы лежит процессный подход, включающий в себя детальное описание взаимодействия всех процессов между собой. Внедрение системы позволило автоматизировать все основные производственные задачи, в соответствии GMP стандарту.

Все функциональные модули в системе BVR QMS Pharm выполнены с учетом полноты, целостности, безопасности хранения и обработки данных и исключают возможные несоответствия, противоречия, двусмысленность и др.

В организационно-технологической базе BVR QMS Pharm предусмотрены регулярное резервное копирование и архивация всей системы в автоматическом режиме и восстановление данных по запросу пользователя с соответствующими правами доступа.

4.1. Этапы фармацевтического производства. Метод генерации электронного досье производства

Комплекс программ **«Производство»** системы BVR QMS Pharm позволяет проводить протоколирование всех производственных процессов, а так же автоматическую генерацию электронного досье. Система хранит всю историю производства, включая имеющиеся отклонения и инциденты. На рисунке 4.1 представлена структурно-функциональная схема управления производственными этапами и контроля ключевых параметров. База данных включает в себя информацию о следующих группах объектов:

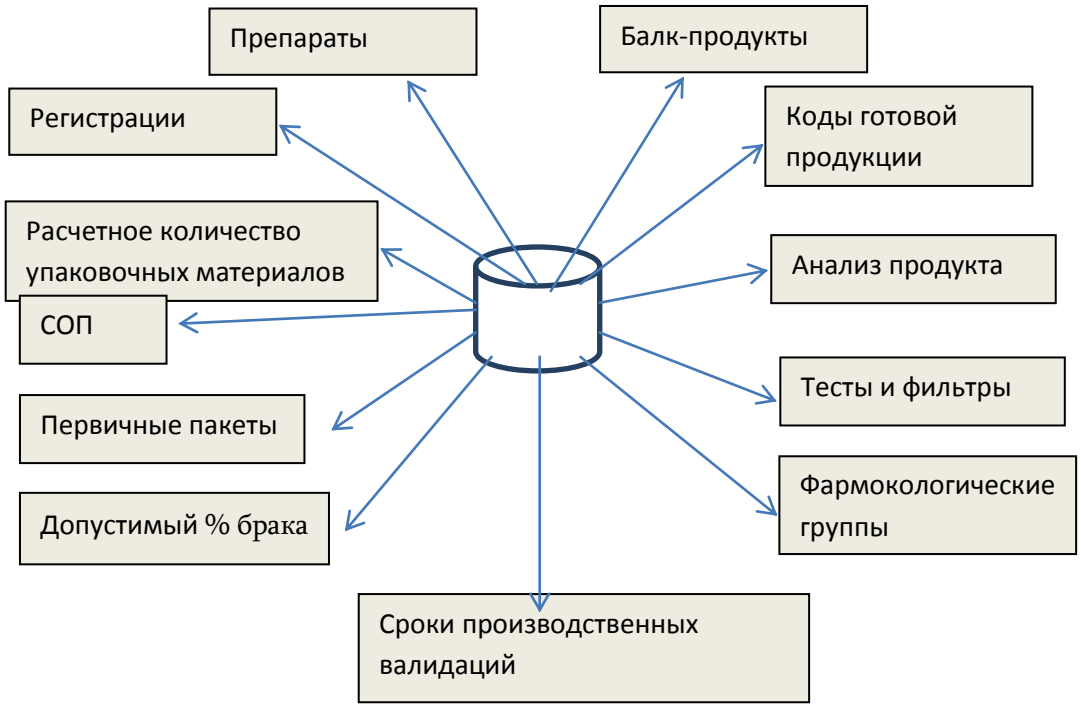
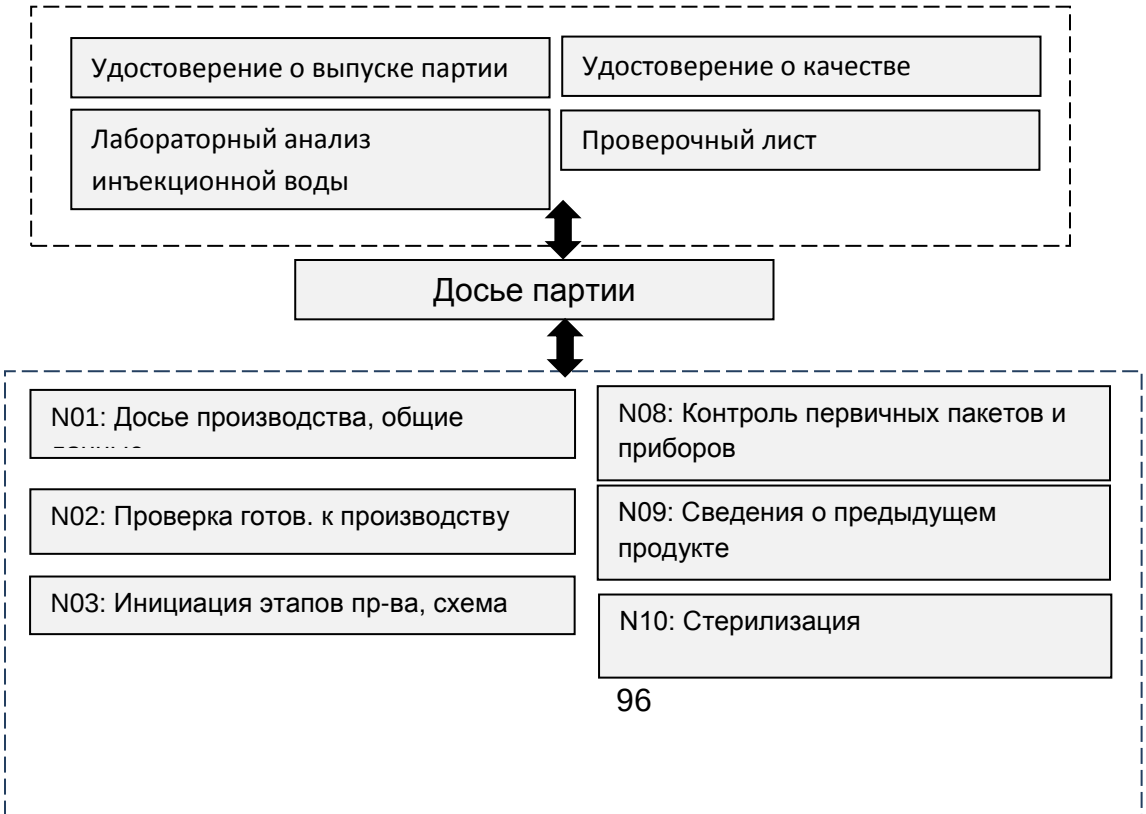


Рисунок 4.1. Основные данные, хранимые в системе фармацевтического производства.

В базе данных хранится вся информация о произведенных балк-продуктах (произведенная продукция до расфасовки) и конечных препаратах, а также их основное описание, включая коды продуктов, их фармокологические группы и состав. С данной подсистемой взаимодействует подсистема о регистрациях лекарств, в которой хранится информация: название лекарства, страна регистрации, срок действия, регистрационный документ. Для каждого производства система генерирует на основе заданных формул расчета необходимое количество упаковочных материалов. В базе данных хранится вся информация об упаковочных материалах, серийных номерах, производителях и поставщиках, характеристики для каждой упаковки. При появлении отклонений система автоматически определяет поставщика и производителя упаковочного материала. Для каждой упаковки есть

основное описание. Подсистема “Досье производства партии” предназначен для протоколирования всех этапов производства. В системе реализована автоматическая генерация электронного досье. Подсистема “Стандартные операционные процедуры”(СОП) предназначена для прикрепления СОП-ов (по темам), связанных с обеспечением качества в производственных процессах. Подсистема “Анализ продукта” предназначен для хранения базы необходимых тестов для каждого продукта. Информация используется в досье производства. Подсистема “Тесты” предназначен для хранения всей базы тестов, проводимых во время производства. Подсистема “Допустимый процент брака” предназначен для протоколирования имеющегося брака и сравнения количества фактических браков с допустимым значениям брака. Используется в досье производства (При выходе за допустимые пределы, система не позволяет продолжать процесс). Подсистема “Сроки производственных валидаций и формул” предназначен для хранения данных о производственной валидации и формулах, связанных с каждым производством (номер и срок действия). На рисунке 4.2 представлена структурная схема “Электронного досье производства” согласно принятому и утвержденному стандарту



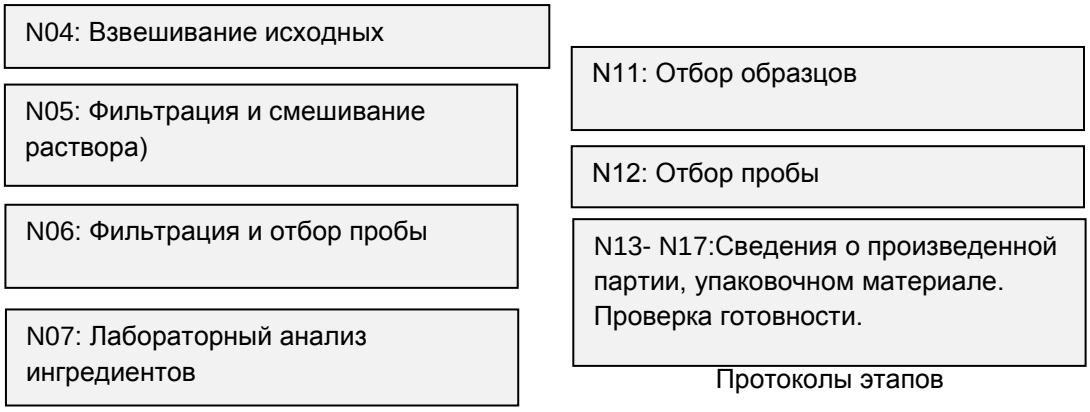


Рисунок 4.2. Структура этапов производства, управляемых информационной системой.

Во всех производственных этапах реализована автоматическая функция электронной подписи. Например, при установлении “флажка” в полях “ Начальник производства”, “КК работник”, “ОК/КК менеджер” в разделе “Подписи” система автоматически генерирует инициалы ответственного лица (пользователя, установившего “флажок”) и дату утверждения. Все поля для модификации доступны только соответствующим ответственным лицам. В разделе **“Общие сведения”** электронного досье вводится следующая информация:

- наименование производимого препарата и единица производства, которая автоматически выдается с помощью динамического меню, согласно выбранному значению препарата;

- фактическая дата производства.

Далее, система согласно указанному пользователем препарату автоматически рассчитывает и генерирует “порядковый номер партии” и «Срок годности»

Протокол N01 (досье производства, общие данные)

Наряду с значениями атрибутов: “Досье N”, “No Партии”, “Дата”, “Продукт” система автоматически определяет и проставляет данные об “Упаковке”, “Дозе (мл)” и “Номере Реактора”. Также в полях “Завершения процесса (Время)” и “Начало процесса (Время)” проставляются соответственно время начала и окончания процессов, предусмотренных в данном протоколе. В системе реализована концепция электронной подписи: при установлении “флажка” в полях “ Технолог производства”, “Начальник производства”, “Менеджер производства”, “ОК/КК менеджер” в разделе “Подписи” система автоматически генерирует инициалы ответственного лица (пользователя, установившего “флажок”) и дату утверждения.

Все поля для модификации доступны только соответствующим ответственным лицам.

Протокол N02 (Проверка готовности к производству)

Система на основе анализа исходных данных о начале производства (наименование препарата, объем и др.) автоматически определяет перечень стандартных процедур (СОП), необходимых для подготовки производства. В разделе “Проверка готовности производства (до производства)” в зависимости от типа препарата система генерирует окно с соответствующими названиями объектов и производственных территорий (например, комната взвешивания (N217/216), комната подготовки раствора (N226) и др) . В соответствующих полях оператор может проставить одно из значений “Да” или ”Нет”.

Протокол N03 (инициация этапов производства и их схема)

Значения полей “Досье N”, “No Партии”, “Продукт”, “Дата” выбираются и проставляются автоматически согласно данным из раздела “Общие сведения”. Оператор задает стандартную операцию, выбирая необходимое значение из списка, генерируемого системой. Далее в разделе “Исходные материалы” в соответствие с производимым препаратом система автоматически генерирует следующие поля: исходный материал, N партии, код поставщика, срок годности.

Протокол N04 (Взвешивание исходных материалов)

Система контролирует фактические результаты взвешивания на соответствие нормативным значениям для каждого препарата. Система также автоматически генерирует электронную схему этапов и последовательности производственного процесса для соответствующего препарата после взвешивания.

Протокол N05 (фильтрация и смешивание раствора)

В поле “СОП” пользователь проставляет соответствующую стандартную операцию, выбирая необходимое значение из динамического списка, генерируемого системой и время начала и окончания процессов взвешивания. В разделе “Весы для взвещивания” представлен список используемых весов, для которых пользователь помечает в поле “Проверено” Да/Нет и утверждает, проставляя электронную подпись. В разделе “Комната No” пользователь вводит данные о наличие или отсутствие имеющихся весов в соответствующих комнатах

взвешивания. Далее, система автоматически вычисляет и выдает в разделе “Количество сырья для взвешивания” таблицу со следующими значениями: название сырья, No партии, количество, единица измерения, технолог, ответственный, состав препарата, согласно производственной формуле

Протокол N06 (фильтрация и отбор пробы)

Система автоматически генерирует информацию о необходимой “Температуре иньекционной воды” и “Скорости смешивания раствора”. Пользователь вводит в систему данные о “Времени (Начало)” и “Времени (Завершения)”, а также данные о “Фильтрации раствора” и “Отборе пробы”, которые система автоматически проверяет на соответствие заданным нормативным показателям. Также система контролирует, чтобы значение поля “Время” в разделе “Проба 2” было бы позже значения поля “Время” в разделе “Проба 1”. Информация о “Температуре” в подразделе “Начало” раздела “Температура иньекционной воды” не может быть меньше 80 (система в данном случае выдает соответствующее сообщение об ошибке) . Система определяет параметры фильтра и выдает соответствующую информацию “Фильтр” в подразделе “0.5 мкм/ 0.22 мкм”. Система контролирует на корректность ввода данных в поле “Давление”, при этом допустимые значения меняются в зависимости от типа препарата. Система автоматически рассчитывает значение поля “Отправляемое количество (л.)” ,по формуле : Qnt(1) – Qnt(p1)-Qnt(p2), где

Qnt(1) – количество, указанное в протоколе N01

Qnt(p1), Qnt(p2) –значения количества образцов, соответственно в подразделах “Проба1” и “Проба2”

Протокол N07 (лабораторный анализ ингредиентов)

Система автоматически генерирует следующую информацию:

- очередной № анализа в текущем месяце
- методы анализа для заданного препарата
- экземпляры для “Отбора пробы”
- Время (значение времени выбирается автоматически на основе данных протокола N06)
- Объем (мл) выбирается автоматически на основе данных протокола N06)

Для заданного препарата и указанного метода анализа система автоматически генерирует необходимые тесты.

Протокол N08 (контроль первичных пакетов и приборов)

Система на основе данных, получаемых от корпоративной учетной системы автоматически генерирует значения для следующих полей из раздела “Первичные Пакеты”: внутренний код, партия, размер.

В подсистеме “Приборы” приведен список оборудования, устройств и систем, задействованных в производственном процессе. Данные “Начало розлива” и “Завершение розлива” заполняются пользователем, при этом система осуществляет контроль набора на корректность даты и времени. Система контролирует, чтобы значения “Фактическое количество”, вводимое пользователем не отличалось бы от значения “Ожидаемое количество” более чем на ±1%, согласно заданным нормативным требованиям.

Протокол N09 (сведения о предыдущем продукте)

Технолог задает стандартную операцию (СОП) из генерируемого системой перечня СОП. В разделе “Общие сведения о предыдущем продукте” автоматически система осуществляет поиск в базе данных и заполняет поля в протоколе N01 о предыдущем препарате, который был приготовленном в данном реакторе. В разделе “Промывание”, “Уборка”, “Стерилизация системы” в зависимости от типа препарата информационная система генерирует список технологических систем, задействованных в производстве и подлежащих стерилизации, промывки и уборки. Отмечая в соответствующих полях знаком “галочка” пользователь протоколирует выполнение соответствующих работ, указывая также время начала и завершения. Значение поля, поступаемое от оборудования (или вводимое оператором) “Результат (мбар)” не должно быть меньше чем 3450, в противном случае система выдает сообщение об ошибке

Протокол N10 (соответствующая форма автоматически создается только для определенных типах препаратов)

Оператор задает СОП для данного этапа и вводит значение поля “количество”, которое не может превышать значения “Фактическое количество” протокола N08 подраздела “Розлив1” . Система контролирует корректность набора значений полей “Завершение” и “Начало”.

Протокол N11, N12 (отбор образцов и проб)

В данных протоколах система автоматически генерирует количество образцов и проб, необходимых для проведения лабораторного анализа. При этом автоматически определяются Методы и тесты анализа и соответствующие допустимые значения для контроля результатов лабораторного анализа.

Протокол N13 (данные о произведенной партии и имеющемся браке)

Система автоматически на основе анализа данных по предыдущим протоколам генерирует значение поля “Брак(кол.) и текущую дату

Протокол N14 (выбор упаковочного материала и оборудования)

На основе списка серийных номеров, не прошедших стерилизацию, система выдает информацию, согласно которой оператор выбирает серийный номер препарата для запуска процесса стерилизации. При этом автоматически проставляется серийный номер и сроки годности продукта. В разделе “Оборудование” представляется список оборудования и упаковочного материала для упаковки и маркировки конкретного препарата. Система автоматически запрашивает информацию о типе и номере упаковочного материала от учетной корпоративной системы.

Протокол N15 (протоколирование архивных образцов и готовой продукции)

Пользователь вводит данные о процессе стериализации, согласно автоматически генерируемому списку номеров серий стерилизации, приведенных в данном досье. Программно производится расчет количества в зависимости от страны экспорта и других исзодных данных, введенных в систему.

Протокол N16 (данные о завершении производственного цикла)

Пользователь выбирает значение серийного номера упаковки, согласно списку, значений, генерируемому системой и проставляет данные о завершении производственного процесса (дата начала и завершения, подписи ответственных лиц).

Протокол N17/ Проверка готовности производства

Система генерирует список наименований производственных территорий, оборудования и стандартных процедур, включающих в себя процедуры стерилизации.

4.2. Производительность, безопасность и масштабируемость системы BVR QMS Pharm

Оптимизация и конфигурирование системы BVR QMS осуществляется с помощью богатого набора административных инструментов, имеющихся в среде Microsoft Dynamics. Тестирование производительности системы Microsoft Dynamics показало, что потенциал масштабирования системы позволяет ей эффективно работать на крупнейших предприятиях, обслуживая до 150 тысяч одновременно работающих пользователей с периодом отклика в доли секунды. Microsoft в сотрудничестве с Intel выполнила тестирование производительности Microsoft Dynamics на серверах Dell R910, основанных на процессорах Intel Xeon серии 7500 и использующих твердотельные накопители Pliant Technology. В ходе тестов применялась только стандартная оптимизация решения в соответствии с рекомендациями, опубликованными в технических документах “Оптимизация и обслуживание Microsoft Dynamics” и “Улучшение производительности Microsoft Dynamics и защита данных Microsoft SQL Server 2008/2015”

Тестирование производилось на последних версиях Microsoft Dynamics 365, включающей последние версии Microsoft Windows Server и Microsoft SQL Server. В этой тестовой среде Microsoft Dynamics показал следующие характеристики производительности:

- кол-во одновременно работающих пользователей: 150000 (каждый из них выполняет транзакцию в системе раз в 8 мин)
- среднее время ответа: 0.4 сек
- веб-запросы: 5.5 млн в час
- бизнес-транзакции: 703080 в час.

При этом были получены следующие результаты:

- среднее использование сервера БД 39,6%
- среднее использование сервера CRM 42%

Защищенность системы и безопасность данных BVR QMS Pharm. Модель безопасности в Microsoft Dynamics CRM обеспечивает целостность и конфиденциальность корпоративных данных. Модель безопасности также способствует эффективному доступу к данным и организации совместной работы сотрудников в защищенном режиме. Microsoft Dynamics CRM обеспечивает:

- многоуровневую методологию лицензирования для пользователей;

- защищенный доступ пользователей к той информации, которая необходима ему для выполнения работ;
- классификацию пользователей и групп пользователей по ролям безопасности и ограничениям доступа к системе согласно этим ролям;
- обмен данными для совместного доступа к объектам, сущностям и отдельным полям.

Модель безопасности позволяет объединить бизнес-единицы(подразделения) , пользователей и правил безопасности на основе заданных ролей для обеспечения политики безопасности в рамках всей корпорации. Среди технологических инструментов управления безопасностью можно выделить:

- санкционированность входа в систему с помощью системы паролей и учетных записей;
- ограниченность доступа к данным путем соответствующих настроек в системе Dynamics CRM;
- отказоустойчивость путем архивирования базы данных и ее восстановления после сбоя на двух уровнях: на уровне SQL базы данных и на аппаратном уровне путем использования технологии RAID массивов;
- работа по протоколу VPN (virtual private network)_для защищенности сети передачи данных и предотвращения несанкционированного внешнего доступа;
- катастрофоустойчивость обеспечивается путем удаленной репликации базы данных в шифрованном виде и восстановление с удаленного устройства хранения.

4.3. Конфигурирование системы BVR QMS Pharm

В системе имеется широкий спектр инструментов конфигурирования, который позволяет:

- добавлять, удалять, модифицировать основные функциональные модули;
- модифицировать пользовательский интерфейс по мере необходимости;

- модифицировать организационно-штатную структуру, в том числе добавляя новые подразделения, изменяя или исключая существующие;
- добавлять , удалять, перемещать по подразделениям пользователей (сотрудников);
- изменять права доступа сотрудников;
- модифицировать аналитическую отчетность путем добавления новых отчетов или изменения существующих;
- изменять при необходимости процессы, правила обработки данных и документов, регламенты коммуникаций сотрудников между собой;
- регламентировать права доступа удаленных сотрудников;
- программировать модели и инструменты для быстрого построения рабочих процессов.

Система BVR QMS Pharm легко может быть модифицирована согласно специфическим требованиям GMP. Одним из приоритетов версий Microsoft Dynamics стало расширение возможностей доступа к данным. Вход в систему в нем возможен не только через веб-браузеры или Outlook, но и с планшетов и мобильных устройств на базе Windows, iOS и Android. В пользовательском интерфейсе акцент сделан на содержании форм и дизайне, который сокращает количество навигационных настроек и оставляет больше места для отображения отдельных веб-ресурсов (дайджеста) и статуса взаимодействия с клиентами.

Приведем пример настраиваемости системы к изменениям в требованиях. В частности, к вопросу оценки поставщиков – при добавлении нового типа сертификата , который должен быть представлен поставщиком конкретного ингредиента, пользователь легко может добавить это требование в перечень документов и связать его с конкретным ингредиентом и далее прослеживать наличие/отсутствие данного документа. Если соответствующий документ отсутствует, то ответственный за работу с поставщиками при размещении заказа получит уведомление об отсутствие соответствующего сертификата у данного поставщика и произойдет блокировка размещения заказа для данного ингредиента через конкретного поставщика.

Пользовательский интерфейс. Взаимодействие пользователя с системой Dynamics осуществляется с помощью широкого набора инструментов, обеспечивающих интуитивно-понятный, дружественный интерфейс. Взаимодействие

Microsoft Dynamics с другими программами, системами, аппаратно-программными комплексами обеспечивается на уровне имеющихся открытых стандартов и широкого спектра программных коннекторов, позволяющих упростить процесс необходимой интеграции. Microsoft Dynamics содержит в себе широкий спектр инструментов для легкой и быстрой настройки пользовательского интерфейса под конкретную должностную роль, включая формы, поля, меню, подсказки, представления, расширенный поиск и др.

4.4. Испытания и внедрение системы BVR QMS Pharm

С целью валидации информационной системы была разработана документированная процедура оценки качества и эксплуатационных характеристик системы на всех этапах ее жизненного цикла с оформлением соответствующих отчетов.

Разработка, внедрение, тестирование, обучение, ввод в зксплуатацию и сопровождение были произведены и продолжаются выполняться согласно стандарту Microsoft Dynamics Sure Step Methodology . Данная методология представляет собой комплексную систему управления проектами на всех этапах, включая рекомендации, стратегии управления, инструменты и шаблоны для разработки, тестирования и внедрения. Методология Sure Step позволила приступить к эксплуатации почти сразу после развертывания системы с дальнейшим поэтапным дополнением специфичных отраслевых задач. Внедрение было проведено по модульному принципу, согласно которому по каждой функциональной компоненте были выполнены следующие этапы:

формирование и согласование спецификаций – демонстрация – устранение замечаний – обучение – тестирование – эксплуатация – создание документации.

Обновление платформы осуществляется согласно методологии обеспечения целостности и сохранности данных. Надежность системы также обусловлена тем , что она позволяет разработчикам работать с данной платформой на уровне метаданных, что обеспечивает целостность и сохранность базы данных. Обновления прикладных задач обеспечиваются встроенными в сервер приложений гибкими инструментами, для модификации и добавления нового функционала. Все функциональные модули в системе выполнены с учетом полноты, целостности, безопасности хранения и обработки данных и исключают возможные несоответствия, противоречия, двусмысленность и др. В рассматриваемой

информационной системе соблюдены все необходимые функциональные требования путем использования широкого набора организационно-технических инструментов, среди которых настраиваемый доступ к данным согласно ролевым назначениям пользователя, автоматический контроль ввода данных, уведомление о наступивших событиях или выходе значения того или иного показателя за пределы допустимых.

В период тестирования и демонстрации системы были представлены результаты соответствия заданным требованиям, в частности, рассмотрены значения ключевых показателей при функционировании системы на допустимых, предельных и запредельных значениях платформа Dynamics обеспечила надежную миграцию данных путем использования развитых инструментов импорта /экспорта данных . Данная платформа не только надлежащим образом обеспечивает корректность экспорта данных, например в офисные приложения – Excel/ Word, но также при импорте данных с этих приложений в систему за счет широкого набора средств визуализации может обнаружить ошибки/некорректные значения в отдельных подсистемах.

Одной из основных отраслей, в которой была внедрена система BVR QMS является фармацевтическая отрасль, которая, как было уже отмечено в предыдущих главах, отличается от других наличием особых требований на соответствие стандарту GMP - «Надлежащей производственной практике». Данный стандарт направлен на гарантирование качества продукции и здоровья пациента на всех этапах от разработки до производства и отгрузки готовой продукции. Сегодня мы наблюдаем частые модификации в требованиях регуляторно-инспекционных органов (постановления правительства, министерства здравоохранения, требования служб надзора и контроля, требования международных стандартов, регулирующих органов и др). Основная цель этих модификаций – обеспечение качества выпускаемых лекарственных препаратов в соответствие с быстро меняющимися потребностями и вызовами в современном мире . Все это вынуждает фармацевтические компании быстро реагировать на данные изменения, т. е. оперативно проводить внедрение соответствующих изменений для соблюдения постоянно меняющихся требований качества: модификация существующих стандартных операционных процедур, внесения изменений в методологию, переобучение сотрудников, добавление новых ранее не существующих процессов, процедур, методов и прочее. Поэтому первостепенной задачей программного

обеспечения управления качеством должно быть способность быстро и гибко перенастраиваться на все новые требования стандартов GMP и управления качеством. Добавим к этому меняющиеся потребности и потребителей – врачей, пациентов, что вынуждает быстро реагировать также и на эти потребности. В разработанной информационной системе реализованы указанные задачи, охватывающие все этапы деятельности фармацевтического производства – от закупа ингредиентов, упаковочных материалов и их лабораторных исследований до производства и поставки продукции конечному потребителю. Разработка, внедрение, тестирование, обучение, ввод в зксплуатацию и сопровождение были произведены и продолжаются выполняться согласно стандарту Microsoft Dynamics Sure Step Methodology. Данная методология представляет собой комплексную систему управления проектами на всех этапах, включая рекомендации, стратегии управления, инструменты и шаблоны для разработки, тестирования и внедрения. Методология Sure Step позволила приступить к эксплуатации почти сразу после развертывания системы с дальнейшим поэтапным дополнением специфичных отраслевых задач. Внедрение было проведено по модульному принципу.

Обновление платформы осуществляется согласно методологии обеспечения целостности и сохранности данных. Надежность и масштабируемость платформы Microsoft® Dynamics подтверждена различными сертификатами и исследованиями международных независимых компаний. Надежность системы также обусловлена тем , что она позволяет разработчикам работать с данной платформой на уровне метаданных, что обеспечивает целостность и сохранность базы данных. Обновления прикладных задач обеспечиваются встроенными в сервер приложений гибкими инструментами, для модификации и добавления нового функционала. Все функциональные модули в системе выполнены с учетом полноты, целостности, безопасности хранения и обработки данных и исключают возможные несоответствия, противоречия, двусмысленность и др. В рассматриваемой информационной системе соблюдены все необходимые функциональные требования путем использования широкого набора организационно-технических инструментов, среди которых настраиваемый доступ к данным согласно ролевым назначениям пользователя, автоматический контроль ввода данных, уведомление о наступивших событиях или выходе значения того или иного показателя за пределы допустимых.

С учетом вышесказанного разработанная система позволила решить следующие целевые задачи**:**

- повышение качества выпускаемой продукции согласно требованиям стандарта GMP;
- обеспечение эффективного взаимодействия с потребителями, оперативный учет их потребностей;
- сокращение текущих расходов;
- управление рисками;
- хранение истории взаимодействия со всеми контрагентами – производителями, поставщиками, партнерами (дистрибьюторами) и конечными потребителями (врачами/клиниками и др.);
- совместная продуктивная работа сотрудников подразделений компании;

Объекты и методы исследования на практике. Одним из объектов исследования явилось фармацевтическое производство дженерик препаратов, включающее в себя все этапы: от контроля ингредиентов до производства конечных продуктов (лекарственных средств). Исследование проводилось на двух ведущих фармацевтических предприятиях Армении – Ликвор и Фарматек, которые являются производителями и экспортерами лекарственных препаратов в страны СНГ, Восточной Европы и Ближнего Востока и имеют в наличие сертификаты GMP. Задача непрерывного соответствия качества GMP стандарту весьма актуальна для этих производителей (как и для других) с целью расширения географии экспорта.

Исследование включало в себя

- классификацию стандартных операционных процедур в деятельности фармацевтических организаций;
- классификацию возможных инцидентов и отклонений от норм в процессе производства и лабораторного контроля , оценку возможных рисков при отклонениях и инцидентах;
- классификацию последовательности действий персонала при имеющихся инцидентах и отклонениях;
- методологию фармаконадзора, включая протоколирование выявленных побочных явлений, исследование возможных причин, заключение и обоснование результатов исследования;

- классификацию проведения изменений и их влияния на производственные и другие процессы.

Все функциональные компоненты реализованы согласно принятым требованиям GMP, а также всем внутренним нормативным документам, регламентирующих процессы управления в фармацевтической компании.

Визуализация данных и аналитическая отчетность. Данная подсистема включает широкий спектр инструментов для визуализации и анализа данных, легкого и быстрого поиска необходимой информации и получения необходимой аналитической отчетности:

- Панели мониторинга, диаграммы, графики и списки
- Таблицы, сгруппированные по заданным критериям
- Отчеты по заранее представленным шаблонам и по запросу оператора

Ниже приведены статистические данные по количеству отклонений до и после внедрения информационной системы, полученные за два года:

Объекты	Число отклонений (До внедрения), 1610 циклов пр-ва, 2012-2014г	Число отклонений (После внедрения), 1580 циклов пр-ва, 2015-2017
Оборудование	15	7
Процессы	25	11
Персонал	80	8
Ингредиенты	30	20
Продукты	40	5

Как видно из таблицы число несоответствий по отношению к числу циклов производства упало с 12% до 4%

Ниже приведена таблица расходов на подготовку досье и аудиторские проверки в чел/час

Процессы	Чел/часы (До внедрения)	Чел/часы (После внедрения)
Подготовка досье производства	10	1
Устранение неисправностей	5	0,5
Аудиторские проверки	127	43

Как видно из таблицы трудоемкость выполнения работ сократилась приблизительно в три раза. Не сложно подсчитать сокращение текущих расходов в денежном выражении из расчета 3-х производств в день и 1/ 3 мес. аудиторской проверки. Возврат инвестиций от внедрения системы составил 13 месяцев.

Внедрение программной системы в фармацевтических компаниях, в лечебных клиниках и организациях инженерного обслуживания сложных систем и оборудования потенциально улучшило качество управленческих решений и создало целый ряд стратегических преимуществ, а именно

- повысило эффективность управления и качество продукции
- сократило текущие расходы
- оптимизировало процессы управления рисками и корректирующими действиями

Указанные преимущества были достигнуты за счет

- повышения скорости и точности исполнения операций
- исключения субъективного человеческого фактора и контроля исполнения работ
- повышения качества и скорости работы с контрагентами
- повышения эффективности совместной работы сотрудников компании
- быстрой подготовки отчетов, в том числе для регулирующих органов.

Внедрение программной системы управления позволило сделать эффективными проведение аудиторских проверок и увеличило инвестиционную привлекательность компаний. В ходе внедрения была проведена окончательная оценка влияния информационной системы на ключевые показатели процессов.

Создание единой интегрированной базы данных и легкий доступ к ним обеспечил быстрое и точное принятие необходимых решений в процессе управления предприятием на всех уровнях – от линейного сотрудника до руководителей подразделений и топ-менеджеров.

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований была построена модель управления процессом стабильности в сложных стохастических системах, включающая в себя функции обнаружения и классификации несоответствий ключевых показателей заданным нормативным значениям для определения необходимых процедур CAPA. На основе построенной модели, разработанных методов и алгоритмов была спроектирована информационная система управления процедурами CAPA и контроля стабильности ключевых показателей в сложных стохастических системах [103 -109].

Основные результаты работы

В ходе выполнения диссертационной работы получены следующие основные результаты:

1. Предложена модель управления стабильностью сложных систем, путем идентификации несоответствий и определения корректирующих действий для их устранения.

2. Разработаны алгоритмы определения источников сообщений о несоответствиях и контроля значений ключевых показателей системы.

3. Исследованы и разработаны методы динамического построения эмпирической выборки, включающей в себя множество отклонений ключевых параметров от заданных нормативных значений и множества, содержащих информацию о необходимых корректирующих и превентивных действиях.

4. На основе анализа задач классификации объектов предложен и реализован алгоритм контроля стабильности в сложных стохастических системах путем определения принадлежности объекта к кластеру (объектам) эмпирической выборки и нахождения класса, определяющего необходимые действия CAPA.

5. На основе проведенных исследований спроектировано и внедрено программное обеспечение BVR QMS управления стабильностью систем. Проведены испытания и получены характеристики программного обеспечения BVR QMS.

СПИСОК ЛИТЕРАТУРЫ

[1] Правила надлежащей производственной практики (GMP) Евразийского экономического союза Версия 4.0 от 20.02.2015

[2] Grazal J.G., Earl D.S. EU and FDA GMP Regulations: Overview and Comparison. Quality Assurance Journal. V.2 , P. 55-60.1997

[3] Overview of GMPs Nov 15, 2004 By Paula J. Shadle, PhD BioPharm International Volume 2004 Supplement, Issue 5

[4] Michael Hiob, Thomas Peither, Ulrike Reuter. GMP Focus. Principles of Equipment Qualification. A Guide for Drug and Device Manufacturers. 2017 Maas & Peither AG - GMP Publishing. First edition 2017

[5] Regulation (EC) No 1394/2007 jf the European parliament and of the council of 13 November 2007 on advanced therapy medicinal products and amending Directive 2001/83/EC and Regulation (EC).

[6] Alsuliman A., Appel SH., Beers DR., Basar R. and others. A robust, good manufacturing practice–compliant, clinical-scale procedure to generate regulatory T cells from patients with amyotrophic lateral sclerosis for adoptive cell therapy // Cytotheraphy. October 2016, V. 18, No. 10, P. 1312–1324.

[7] Raj A. A review on corrective action and preventive action (CAPA) // African Journal of Pharmacy and Pharmacology. 2016. V. 10(1), P. 1-6.

[8] Van Trieste M. CAPA within the Pharmaceutical Quality System // ICH Q10 Conference. P9: Pharmaceutical Quality System Elements: Continual Improvement of the Process (CAPA). 2011. Brussels, Belgium.

[9] Rodriguez J. CAPA in the Pharmaceutical and Biotech Industries. Woodhead Publishing. 1st Edition. 2015. 248 p.

[10] Chopra V., Kumar A., Aiyyer A. Trivedi P., Nagar M. Investigating Out-of-Specification Results and Development CAPA Program for Pharmaceutical Industries: An Overview // Der Pharmacia Lettre. 2011. V. 3(2). P. 368-382.

[11] Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistic al topic models for multi-label document classification // Machine Learning. 2012. Vol. 88, Pp. 157-208.

[12] Zhou S., Li K., Liu Y. Text categorization based on topic model // International Journal of Computational Intelligence Systems. 2009. Vol. 2, no. 4. Pp. 398409.

[13] Blei D. M. Probabilistic topic models // Communications of the ACM. 2012. Vol. 55. № 4. Pp. 77-84.

[14] Chen H., Chiang R., Storey V. Business intelligence and analytics: From big data to big impact // MIS Quart. 2012. 36(4). P. 1165–1188.

[15] Agarwal R., Dhar V. Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research // 2014. Information Systems. Res. 25(3). P. 443–448

[16] Vasant Dhar. Data science and prediction //Communications of the ACM. 2013. V. 56. P. 64-73.

[17] Волкова В.Н., Денисов А.А. Теория систем и системный анализ

[18] Thomas Hofmann probabilistic latent semantic analysis. // International Computer Science Institute, Berkeley, CA & EECS Department, CS Division, UC Berkeley. 2013.

[19] Panagiotis Mazis, Andrianos Tsekrekos. "Latent semantic analysis of the FOMC statements".// Review of Accounting and Finance. 2017. Vol. 16 Issue: 2, pp.179-217

[20] Gunther F., Dudschig C., Kaup B. Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. // The Quarterly Journal of Experimental Psychology. 2016. Volume 69, Issue 4

[21] Ghanem K. Local and Global Latent Semantic Analysis for Text Categorization.// International Journal of Information Retrieval Research (IJIRR). 2014. V. 4. I. 3 , p 1-13

[22] Джонсон Н. Л. Коц С., Кемп А. Одномерные дискретные распределения. М.: Лаборатория знаний. 2017. 563 с.

[23] Рудаков И.В., Шляева А. В. Моделирование входных потоков данных для стохастических моделей дискретных систем. // Вестник МГТУ. 2008. № 2 . стр. 65-69

[24] Langefors. Information systems theory. // Information Systems. Volume 2, Issue 4, 1977, Pages 207-219.

[25] V. S. Lerner. Introduction to information systems theory: concepts, formalism and applications. // International Journal of Systems Science. Volume 35 Issue 7, 15 June 2004 Pages 405-424.

[26] Месарович М., Такахара И. Общая теория систем: математические основы. / -М.: Мир, 1978. -311с.

[27] Садовский В.Н. , Юдин Э.Г. Исследование по общей теории систем: сб. переводов. // -М.: Прогресс, 1969. – 520с.

[28] Садовский В.Н. Основание общей теории систем: логико-методологический анализ / -М.: Наука, 1974. – 279с.

[29] Воронов А. А. Введение в динамику сложных управляемых систем. - М.: Наука, 1985. - 352с.

[30] Бусленко Н.П. Моделирование сложных систем. - М.: изд-во "Наука",1978.- 400 с

[31] Эшби У.Р. Введение в кибернетику. -М.: Изд-во ИЛ, 1959. -432 с.

[32] Анфилатов В.С. Системный анализ в управлении. –М.: Финансы и Статистика, 2002. – 368с.

[33] Берталанфи Л. История и статус общей теории систем. // Системный исследования. - М.: Наука, 1973. – с. 20-37

[34] Садовский В.Н. Основания общей теории систем: логико-методологический анализ. –М.: Наука, 1974. – 279с.

[35] Форестер Дж. Мировая динамика. – М.: Наука, 1978. – 167с.

[36] Nils J. Nilsson. Introduction to machine learning. Robotics Laboratory Department of Computer Science Stanford University. 1998. 188c. http://ai.stanford.edu/people/nilsson/MLBOOK.pdf

[37] Колмогоров, А.Н. Новый метрический инвариант транзитивных динамических систем и автоморфизмов пространства Лебега / А.Н. Колмогоров // ДАН СССР. – 1985. – Т. 119. – С. 94–98.

[38] Renyi, A. On Measures of Entropy and Information / A. Renyi // Proc. Fourth Berkeley Symposium. V.1. Berkeley, Calif.: University of California Press, 1961. – P. 547–561.

[39] Pena, D. Dimensionless measures of variability and dependence for multivariate continuous distributions / D. Pena, A. Van der Linde // Commun. Stat.: Theor. M., 2007. – Vol. 36. – Issue 10. – P. 1845–1854

[40] Goldstein, E. Bruce. Cognitive psychology: connecting mind, research, and everyday experience.// (3rd ed.). Belmont, CA. 2011.
http://wireframe.vn/books/Psychology/Cognitive%20Psychology%20Connecting%20Mind,%20Research%20and%20Everyday%20Experience,%203rd%20Edition%20by%20E.%20Bruce%20Goldstein%200840033559.pdf

[41] А. А. Кулинич. Семиотические когнитивные карты. Ч. 1. Когнитивный и семиотический подходы в информатике и управлении. // Проблемы управления, 2016, выпуск 1, страницы 2–10

[42] А. А. Кулинич. Семиотические когнитивные карты. Ч. 2 Основные определения и алгоритмы. Когнитивный и семиотический подходы в

информатике и управлении. // Проблемы управления, 2016, выпуск 2, страницы 24–40

[43] Гамазов И.Н., Терехов В.И. Анализ задач, возникающих при создании нечетких когнитивных карт. http://scienceproblems.ru/images/PDF/2016/6/analiz-zadach-voznikajushchih-pri-sozdanii.pdf

[44] Абрамова Н.А. Коврига С.В. Некоторые критерии достоверности моделей на основе когнитивных карт.

[45] Авдеева З.К. Коврига С.В. и др. Когнитивный подход в управлении// Проблемы управления. – 2007.-№3.-с. 2-8

[46] Кузнецов О.П., Кулинич А.А., Марковский А.В. Анализ влияний при управлении слабоструктурированными ситуациями на основе когнитивных карт. // Человеческий фактор в управлении. –М.: Комкнига , 2006. – с. 313-344

[47] Абрамова Н. А. Экспертная верификация при использовании формальных когнитивных карт. Подходы и практика / Управление большими системами. Специальный выпуск 30.1 "Сетевые модели в управлении". М.: ИПУ РАН, 2010. С.371-410.

[48] Моргунов, Е.П. Подходы к построению критерия качества границы эффективности в методе Data Envelopment Analysis. // Актуальные проблемы современной науки и пути их решения: Материалы III межвузовской научной конференции аспирантов / КГТЭИ.– Красноярск, 2003.– С. 86– 88.

[49] Переверзев Е.С. Энтропийные методы в теории самоорганизационных процессов

[50] Переверзев Е. С. Вероятностные распределения и их применение / Е. С. Переверзев, Ю.Ф. Даниев. – Днепропетровск : Институт технической механіки НАН Украины и НКА Украины, 2004. – 418 с.

[51] Langley P. Elements of Machine Learning. San Francisco. Morgan Kaufman. 1996.

[52] Jay Wright Forrester. World Dynamics.//Publisher by Wright-Allen Press. 1973 144p

[53] Андерсон Т. Введение в многомерный статистический анализ. М.: изд-во физ.-мат. литературы, 1963, 501с.

[54] Емельянов А.А. Стохастические сетевые системы массового обслуживания. // Прикладная информатика. № 5 (23), 2009, с. 103-111

[55] Корнейчук Б. В. Максимова Т. Г. Стохастические модели переходных процессов марковского типа. // Научно-технические ведомости СПбГТУ № 2 . 2010, С. 120-126

[56] Е.П. Моргунов, О.Н. Моргунова. Многомерная классификация сложных объектов на основе оценки их эффективности.// // Вестник НИИ СУВПТ, 2001, с. 1-19

[57] Леман Э. Проверка статистических гипотез. М.:Наука, 1979, 408с.

[58] Гмурман В.Е. Теория вероятностей и математическая статистика. М.: “Высшая школа” , 2003, 480 с.

[59] Айвазян С. А., Бухштабер В.М. и др. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989, 608с.

[60] Соколов М.В. Параметрические семейства функций, замкнутые относительно допустимых преобразований шкалы измерения. http://sos-homepage.narod.ru/publications.files/families.pdf

[61] С. Стивенс. О шкалах измерения. В сб. «Экспериментальная психология». Пер. с англ. Изд. иностр. лит. М.1961

[62] Mullin M., Sukthankar R. Complete Cross-Validation for Nearest Neighbor Classifiers. — Proceedings of International Conference on Machine Learning. — 2000. — С. 1137-1145

[63] Lance G.N., Willams W.T. A general theory of classification sorting strategies. 1. Hierarchical systems // Comp. J. 1967. № 9. P. 373—380

[64] С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.

[65] Жамбю М. Иерархический кластер-анализ и соответствия. — М.: Финансы и статистика, 1988. — 345 с

[66] Бураго Д.Ю., Бураго Ю.Д., Иванов С.В. Курс метрической геометрии. - М.; Ижевск: Ин-т компьют., 2004. 512 с.

[67] Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — Р. 320.

[68] Соколов Е. Семинары по метрическим методам классификации. http://www.machinelearning.ru/wiki/images/9/9a/Sem1_knn.pdf

[69] Weber, R., Schek, H. J., Blott, S. (1998). A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. // Proceedings of the 24th VLDB Conference, New York C, 194–205

[70] Загоруйко Н. Г., Ёлкина В. Н., Лбов Г. С.Алгоритмы обнаружения эмпирических закономерностей. Новосибирск: Наука, 1985

[71] Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. — 14th International Joint Conference on Artificial Intelligence, Palais de Congres Montreal, Quebec, Canada. — 1995. — С. 1137-1145

[72] Коврига С.В. О методе верификации когнитивных карт, основанном на частных критериях достоверности. XII Всероссийское совещание по проблемам управления. 2014, с. 4132-4143

[73] Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов. — Математические вопросы кибернетики / Под ред. О. Б. Лупанов. — М.: Физматлит, 2004. — Т. 13. — С. 5–36

[74] Ulisses M. Braga-Neto Edward R. Dougherty. Is cross-validation valid for small-sample microarray classification? // Bioinformatics, Volume 20, Issue 3, 12 February 2004, Pages 374–380

[75] Владислав Тарасенко. Территориальные кластеры: Семь инструментов управления. — М.: Альпина Паблишер, 2015. — 201 с

[76] Харченко М.А. Корреляционный анализ. Учебное пособие Воронежского гос. ун. 2008. – 31 с.

[77] Hamid M. A. Risk management in pharmaceutical production. 1-st ed .: Trans. with English. - M.: VIALEK, 2014. 472 p.

[78] Pyatigorskaya N.V. PAT is the basis for modern pharmaceutical research, production and quality assurance. History of development // Development and registration of drugs. 2013. No. 3 (4). P. 48-52.

[79] Report on the research work "Development of the Concept of Quality Assurance of Medicines in the Russian Federation". Moscow: Public Education. MMA named after. I.M. Sechenov. 2009.

[80] Bjyorn Andersen. Tom Vagerhaug. Root Cause Analysis. Simplified tools and tecgniques. ASQ Quality Press. 2006. Second Edition.

[81] Williams P. Techniques for root cause analysis // PMC. 2001. V. 14(2). P. 154-157.

[82] Wilson P., Dell L., Anderson G. Root Cause Analysis : A Tool for Total Quality Management. ASQC Quality Press. Milwaukee. Wisconsin. 1993.

[83] Agarwal R., Dhar V. Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research // 2014. Information Systems. Res. 25(3). P. 443–448.

[84] Chen H., Chiang R., Storey V. Business intelligence and analytics: From big data to big impact // MIS Quart. 2012. 36(4). P. 1165–1188.

[85] Mochen Y., Gediminas A., Gordon B. Mind the Gap: Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining // Information Systems. Research, January. 2018. URL: http://gkmc.utah.edu/winter2016/sites/default/files/webform/abstracts/WCBI%20submission_final.pdf

[86] Provost F., Fawcett T. Data Science for Business. O'Reilly Media. 2013. 414 p.

[87] Vasant Dhar. Data science and prediction //Communications of the ACM. 2013. V. 56. P. 64-73.

[88] Donald J. W. Advanced Topics in Statistical Process Control: The Power of Shewhart's Charts. SPC Press. 1995.

[89] Wheeler Donald, Chambers David. Statistical process management: Optimization of business using Shewhart control charts. Moscow: Alpina Publisher, 2016. 410 p.

[90] Lapidus V. A. The Shewhart system. N.Novgorod: SIC "Priority" Ltd., 2004

[91] Emelyanov A. A. Ways of adapting Shewhart control cards in controlling // Russian Entrepreneurship. 2011. V. 12. № 11. P. 86-90

[92] Tsarev Yu. V., Trostin A. N. Statistical methods of quality management.
Control cards. Teaching-methodical manual / GOU VPO Ivanovo State University of Chemistry and Technology. 2006. 250 p.

[93] Agarwal R., Dhar V. (2014). Editorial—big data, data science, and analytics: The opportunity and challenge for IS research // Information Systems Research, V. 25(3). P. 443-448.

[94] Максимова О. В., Шпер В. Л., Адлер Ю. П. Контрольные карты Шухарта в России и за рубежом. Часть 1. Стандарты и качество. 2011. № 7. С. 82-87.

[95] Фадеев А.Н., Журавлев А.И. Лепестковая диаграмма как средство отображения результатов математического моделирования/А.Н. Фадеев, А.И. Журавлев. – Образование и наука в современных условиях. Чебоксары. Центр научного сотрудничества». 2016. № 2 С. 72 – 75.

[96] Махонченко Ю. Построение диаграммы Парето. Системы менеджмента – консультации и обучение онлайн. 2015.
http://managementsystemsonline.blogspot.am/2015/08/postroenie-diagrammy-pareto.html

[97] Н. Г. Загоруйко. Гипотезы компактности и λ-компактности в методах анализа данных. // Сиб. журн. индустр. матем., 1998, том 1, номер 1, страницы 114–126.

[98] Николай Игнатьев. Интеллектуальный анализ данных и гипотеза о компактности классов: Меры компактности, критерии оценок. 2016. Palmarium Academic Publishing. 100 p.

[99] Поцыкайло Александр Анатольевич. Использование метода k-ближайших соседей при распознавании полутоновых изображений. // Известия ЮФУ.

[100] Р. К. Стрюков, А. И. Шашкин. О модификации метода ближайших соседей. // Вестник ВГУ, серия: системный анализ и информационные технологии, 2015, № 1 114-120 стр.

[101] Демиров В.В. Специфика и направления машинного представления процессов обучения. // Онтология проектирования. 2014. v.1 78-85 стр.

[102] Козлов П.Ю. Методы автоматизированного анализа коротких неструктурированных текстовых документов. // Программные продукты и системы. 2017. № 1 стр. 100-105.

[103] Aghajanyan R. Modeling of Compliance Processes with Specified Criteria in Complex Dynamic Systems // ISSN 0131-4645 Mathematical Problems of Computer Science v. 48, pp. 89-97, 2017.

[104] Markosyan M.V., Aghajanyan R.B., Baizhanova D.O. Information system designing method for non-conformity identification in the procedures for managing corrective and preventive actions// ISSN 2072-9502. Вестник АГТУ. Сер.: Управление, вычислительная техника и информатика. 2018. N 2, стр. 71-80

[105] Агаджанян Р. Б. Метод автоматизации процессов классификации неструктурированных сообщений в сложных стохастических системах// XIV Международная научно-практическая конференция “Advances in Science and Technology”// стр. 102-106, 2018

[106] Агаджанян Р. Б. Контроль ключевых параметров сложной системы методом вычислений логических предикатов// ISSN 2541-9250. Наука через призму времени, №5 (14) 2018, стр. 38-41

[107] Агаджанян Р. Б. Исследование и автоматизация контроля стохастических отклонений в истемах организационного управления производственным процессом// ISSN 2073-6185 «Вестник Дагестанского государственного технического университета. Технические науки», Том 45, №1, 2018г., стр.88-97.

[108] Агаджанян Р. Б. Метод динамического построения признакового пространства эталонной обучающей выборки на основе бинарной классификации// ISSN

2413-7081 (Print), ISSN 2542-0801 (Online) научно-методический журнал издательства «Проблемы науки», №4 (27), 2018 год, стр. 20-24.

[109] Агаджанян Р.Б., Байжанова Д.О. Метод автоматизации процесса определения корневой причины нарушения стабильности в сложных стохастических системах// ISSN 1829-3336 «Вестник национального политехнического университета Армении. Информационные технологии, электроника, радиотехника», 2018 год, N1, стр 45-56

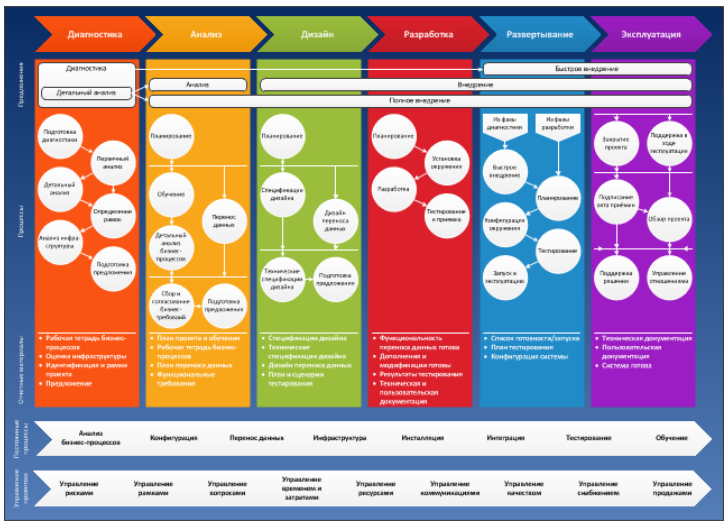
[110] Dubner P. N. Statistical tests for feature selection in KORA recognition algorithms // Pattern Recognition and Image Analysis. — 1994. — Vol. 4, no. 4. — P. 396.

[111] Marchand M., Shawe-Taylor J. Learning with the set covering machine // Proc. 18th International Conf. on Machine Learning. — Morgan Kaufmann, San Francisco, CA, 2001. — Pp. 345–352. http://citeseer.ist.psu.edu/452556.html.

[112] Шаповалов В.И. О фундаментальных закономерностях управления тенденциями. N2. 2005. стр. 1-10

ПРИЛОЖЕНИЕ

Все этапы настройки, разработки, поддержки и администрирования программного обеспечения были выполнены на основе стандартных промышленных технологий Microsoft. Ниже приведена структурная схема методологии Sure Step, включающая в себя этапы разработки и параметры управления жизненным циклом проектирования программного обеспечения:



Информационная система разработана на сервере приложений Microsoft Dynamics, база данных создана на реляционной системе управления базами данных (РСУБД) Microsoft SQL, в качестве платформы для получения аналитической отчетности используется среда разработки Microsoft Visual Studio (MVS) и Microsoft Business Intelligence (BI).

Платформа **Microsoft Dynamics** позволяет решать весь спектр управленческих задач, связанных с корпоративной и функциональными стратегиями :

• создание единой консолидированной базы корпоративных данных;

• управление бизнес-процессами;

• аналитическая обработка информации, создание отчетов.

Аналитические возможности платформы Microsoft Dynamics, в том числе панели ключевых показателей деятельности для руководителей, предоставляют возможности эффективного планирования, управленческого контроля, измерения основных ключевых показателей и оперативного принятия решений.

Интеграция системы с “внешним миром” производится согласно открытым стандартам путем использования различных веб сервисов-коннекторов, которые по необходимости могут постоянно проектироваться и дополняться.

Различные имеющиеся ограничения, касающиеся пользовательского интерфейса, уровня интеграции с другими приложениями преодолеваются в процессе совершенствования технологий а также с выпуском новых версий продуктов Microsoft Dynamics, Microsoft Office, Microsoft SharePoint®, Microsoft SQL Server®, Reporting Services и веб-сервисов.

Все подсистемы связаны друг с другом посредством единого сервера приложений и общей консолидированной базы данных Microsoft SQL. Это позволяет получать всевозможные обобщенные отчеты, в которых могут быть объединены данные, полученные из различных подсистем. При этом доступ к данным осуществляется в защищенном режиме, согласно политике безопасности, принятой в организации и соответствующим ролевым назначениям сотрудников.

Алгоритмы и процедуры обработки данных выполнены в строгом соответствие с основными требованиям, изложенными в проектных документах и спецификациях. Во всех подсистемах проводится контроль на корректность вводимых данных, с учетом их целостности, корректности, и непротиворечивости.

При необходимости все процедуры, функции и алгоритмы обработки легко могут быть модифицированы (изменены, дополнены, удалены), а также могут быть добавлены новые подсистемы без нарушения целостности и правильности работы всех остальных компонент. Надежность функционирования гарантирована программными и системными средствами защиты, реализованными в сервере приложений Microsoft Dynamics CRM и в базе данных Microsoft SQL с одной стороны и технологии поэтапной разработки и модульного внедрения согласно методологии Microsoft Sure Step, с другой стороны.