ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտ

Վարդան Հակոբի Թոփչյան

Ինֆորմացիայի նկատմամբ սեփականության իրավունքը պահպանող հաշվարկներ

Ե.13.05 – «Մաթեմատիկական մոդելավորում, թվային մեթոդներ և ծրագրային համալիրներ» մասնագիտությամբ տեխնիկական գիտությունների թեկնածուի գիտական աստիճանի հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

Երևան — 2014

ИНСТИТУТ ПРОБЛЕМ ИНФОРМАТИКИ И АВТОМАТИЗАЦИИ НАЦИОНАЛЬНОЙ АКАДЕМИИ НАУК РЕСПУБЛИКИ АРМЕНИЯ

Топчян Вардан Акопович

Вычисления сохраняющие права собственности над информацией

АВТОРЕФЕРАТ

Диссертации на соискание ученой степени кандидата технических наук по специальности 05.13.05 - “Математическое моделирование, численные методы и комплексы программ”

ЕРЕВАН – 2014

Ատենախոսության թեման հաստատվել է ՀՀ Գիտությունների Ազգային Ակադեմիայի Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում

Գիտական ղեկավար՝	ֆիզ. մաթ. գիտ. դոկտոր	Լ. Հ. Ասլանյան
Պաշտոնական ընդդիմախոսներ՝	տեխ. գիտ .դոկտոր	Գ. Հ. Խաչատրյան
	տեխ. գիտ. թեկնածու	Ս. Բ. Ալավերդյան

Առաջատար կազմակերպություն՝ Երևանի պետական համալսարան

Պաշտպանությունը տեղի կունենա 2014թ. հունիսի 13-ին ժ. 15:00-ին ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում գործող 037 «Ինֆորմատիկա և հաշվողական համալիրներ» մասնագիտական խորհրդի նիստում, հետևյալ հասցեով՝ Երևան 0014, Պ.Սևակի փ. 1:

Ատենախոսությանը կարելի է ծանոթանալ ՀՀ ԳԱԱ ԻԱՊԻ գրադարանում:
Սեղմագիրը առաքված է 2014թ. մայիսի 12-ին:

Մասնագիտական խորհրդի գիտական քարտուղար, ֆ.մ.գ.դ. Հ. Գ. Սարուխանյան

Тема диссертации утверждена в Институте информатики и проблем автоматизации НАН РА

Научный руководитель:	доктор физ. мат. наук	Л. А. Асланян
Официальные оппоненты:	доктор тех. наук	Г. Г. Хачатрян
	кандидат тех. наук	С. Б. Алавердян

Ведущая организация: Ереванский государственный университет

Защита состоится 13 июня 2014 г. в 15:00 на заседании специализированного совета 037 «Информатика и вычислительные системы» в Институте проблем информатики и автоматизации НАН РА, по адресу 0014, г. Ереван, ул. П. Севака 1.

С диссертацией можно ознакомиться в библиотеке ИПИА НАН РА.
Автореферат разослан 12 мая 2014 г.

Ученый секретарь специализированного совета

доктор физ. мат. наук А. Г. Саруханян

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. По мере интенсивного развития информационных технологий появляются новые подходы и средства автоматизации сбора и анализа персональных и других конфиденциальных данных. Статистические, финансовые и другие информационные структуры помимо выработки агрегированных данных и их предоставления общественности все больше прибегают к распределению данных близких к исходным. Это, во-первых, требование времени, но оно также нацелено на повышение независимого, общественного и научного управления и экспертизы путём целевого анализа этих данных. Социологические данные содержат значительный объем персональных или бизнес ориентированных данных. Публикация данных близких по формату и содержанию к исходным связана с риском раскрытия конфиденциальности этих данных. Противоречие требований прозрачности и конфиденциальности при публикации критических данных явилось основой возникновения нового исследовательского направления известного теперь как технологии ограничения раскрытия данных. Актуальность проблемы обосновывается новыми правовыми актами как например CIPSEA—the Confidential Information Protection and Statistical Efficiency Act of 2002 (Chance, 17(3):2125, 2004), и новыми исследованиями как EUREDIT - the development and evaluation of new methods for editing and imputation (IST-1999-10226 EC FP project, 2000 - 2003).

Сложилось так, что первые подходы решения задачи предопределили две основные и взаимодополняющие технологии. Первый, криптографический подход создал по этому поводу концепцию гомоморфного шифрования, что нацелено на исполнении вычислений над шифрованными данными так, что получаются результаты анализа исходных данных. Второй подход возник на уровне статистической обработки анализа данных, что является подавляющей технологией, применяемой сегодня в данной области. В каждой из указанных двух направлений имеются успехи и нерешенные задачи. Криптография пытается разработать гомоморфные схемы, частичные и полные, однако на сегодняшний день нет эффективной системы для полного объема алгебраических вычислений. Полная система, относительно недавно объявленная компанией IBM, оценочно может стать жизнеспособной десятилетиями позже. Статистическое же направление началось с того, что оно пыталось заменить рассматриваемую задачу схемой оценивания / восстановления отсутствующих данных (missing data). Далее возникли схемы внесения возмущений и схемы генерации синтетических данных. Настоящая работа направлена на исследование иерархических структур анализа данных и пытается лучше использовать дополнительную информацию предметной области для повышения вычислительной скорости, а также для получения более адекватных синтетических данных эксперимента.

Целями диссертационной работы являются дальнейшая разработка моделей, алгоритмов и программ, связанных со сбором и анализом данных в системах предоставления социологической информации исследовательским, общественным,

международным и научным организациям. Основным требованием задачи является сокрытие персональной и иной конфиденциальной информации, и требуется чтобы анализ данных предоставлял результаты необходимые анализирующей стороне. Основной идеей данной работы является алгоритмическая оптимизация генерации данных предоставляемых общественности, а также улучшение системных характеристик, сохранения более глубинных взаимосвязей значений атрибутов задачи, таких как сохранение парных корреляционных связей атрибутов.

Объект исследования. Рассмотрены инфраструктурные, криптографические и эвристические модели и алгоритмы решения задачи. Первые две группы как показывают результаты обзора области, предоставляют ограниченные возможности решения задачи. В связи с этим рассмотрены эвристические модели деревьев классификации и регрессии (CART) и их улучшения, а также иерархический кластерный анализ и проведена системная реализация модели и ее внедрение при замене исходных данных генерированными синтетическими данными. Введены понятия соответствия данных к моделям генерации синтетических данных, и рассмотрены группы парных корреляционных связей стремясь сохранить не только отдельные статистические / частотные характеристики атрибутов, но и их взаимные связи и взаимную коррелированность. Обработаны три группы тестовых задач – Minnesota Population Center (IPUMS), Национальная Статистическая Служба Р.А., Центральный Банк Р.А. Эти эксперименты призваны оценить качество и быстродействие предоставленных моделей и алгоритмов.

Методы исследований. В диссертационной работе использованы теоретические разработки иерархических деревьев решений, иерархического кластерного анализа, и анализа моделей структур данных предоставленных в основном таблицами данных. Используется аппарат теории распознавания образов, который отображает множество обучения задачи на иерархические структуры, используя известные процедуры построения (growing), усечения (pruning), останова (pre-pruning), и бэггинга (bagging, bootstrap aggregation). Проектирование, реализация и тестовое внедрение системы нацелены на выявлении скоростных и качественных характеристик услуг, предоставляемых разработанной системой.

Научная новизна. В сущности, алгоритмическая задача генерации качественных синтетических данных зависит не только от модели генерации, но также от самих данных. Новизна работы заключается именно в предварительном анализе данных задачи, что предоставляет информацию о соответствии данных к модели генерации. Далее, анализ пар атрибутов, определенных / объявленных как взаимосвязанные и последующее использование иерархического дерева генерации позволяет произвести расщепление дерева на части, и эффективный останов алгоритма, оптимизировав этим стандартный процесс построения и отсечения. Таким образом, анализ данных раскрывает

возможности (предел) генерации с сокрытием конфиденциальности, а анализ парных связей ограничивает шаги самой генерации повысив ее производительность.

Выносящиеся на защиту основные положения

- Построение модели анализа атрибутов и их взаимных связей с определением качества данных и последовательности атрибутов в построении иерархии разбиений конфиденциальных данных.
- Улучшение дерева классификации и регрессии в задачах генерации синтетических данных путём совмещения процессов построения и отсечения иерархического дерева.
- Создание программной реализации разработанного альтернативного алгоритма генерации синтетических данных.
- Тестовое внедрение системы в обработке государственных, статистических и финансовых данных.

Особенность и достоверность результатов. Особенностью данной работы является широкий спектр научных задач и областей таких как вычислительные инфраструктуры, криптография, статистика и распознавание образов, рассмотренных в связи с решением основной исследуемой задачи о вычислениях с сохранением конфиденциальности. Достоверность и эффективность полученных оценок подтверждается внедрением работы и проведенными экспериментами. Эксперименты включают три группы данных. Результаты работы публиковались и докладывались на тематических конференциях и семинарах.

Практическая ценность полученных результатов связана с актуальностью задачи предоставления качественных социологических и экономических данных, включающих конфиденциальные персональные данные общественности.

Внедрение. Результаты исследований внедрены в статистическом департаменте Центрального Банка РА.

Апробация работы. Основные результаты и положения диссертационной работы обсуждались на семинарах в ИПИА НАН РА, а также докладывались на конференциях зимней сессии ITA и ITHEA http://www.ithea.org

Публикации. Научные результаты исследований и основные результаты работы отражены в 4 публикациях, список которых приведен в конце автореферата.

Структура и объем диссертации. Диссертационная работа состоит из введения, трех глав, заключения и списка литературы, которая включает в себе 90 работ. Объем работы – 105 страниц, включая рисунки, таблицы и цитируемую литературу. Диссертация написана на русском языке.

СОДЕРЖАНИЕ РАБОТЫ

Во введении формулируется цель работы, обосновывается ее актуальность, очерчивается круг рассматриваемых задач, кратко излагается содержание работы и подчеркивается научная новизна полученных результатов.

В первой главе представлен обзор и анализ возможных прямых подходов/методов для решения поставленной задачи. В параграфе **1.1** рассмотрен возможный подход реконструкции/модификации схемы серверов типичного дата центра/банка с целью ограничения риска раскрытия конфиденциальной информации во время вычислений (Рис. 1).

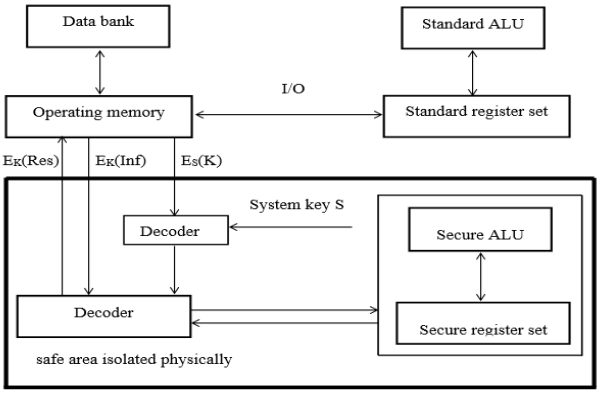

Рисунок 1. Схема инфраструктурных изменений вычислителей дата банка

Далее, в параграфе **1.2** исследуется одна естественная теоритическая модель, а именно, - использование гомоморфных схем криптографии. Основная идея здесь заключается в осуществлении вычислений на зашифрованных данных и дешифровки только после получения результата[1]. С целью демонстрации существующих схем были приведены некоторые примеры. В частности, в одной из приведенных схем в качестве системы оригинальных данных рассматривается

$$\mathrm{P} = < Z_n, +_n, \times_n >,$$

множество целых чисел по модулю n с операциями сложения и умножения по модулю n, где n равен произведению двух больших простых чисел p и q, $n = p \cdot q$. А в качестве системы шифрованных данных -

[1] R. L. Rivest , L. Adleman, M. L. Dertouzous, "On data Banks And Privacy Homomorphisms", *Foundations of Secure Computation*, pp. 169-179, 1978

$$C =< Z_p \times Z_q, (+_p, +_q), (\times_p, \times_q) >,$$

множество $Z_p \times Z_q$ с аналогичными покомпонентными операциями сложения и умножения. Функция шифрования E_k определяется следующим образом:

$$E_k(a) = (a \bmod p, a \bmod q),$$

где k- ключ шифрования, состоящая из чисел p и k, $k = (p, q)$. И наконец, функция дешифрования $D_k((b, c))$ вычисляется с использованием Китайской теоремы об остатках. К сожалению, данная схема неустойчива к вскрытию с использованием атаки открытого текста. Аналогично представленной, рассмотрены и некоторые другие криптосхемы что так же оказываются неустойчивыми к соответствующим типам вскрытия. Отмечается также схема полного гомоморфизма разработанная в фирме IBM, которая станет практически реализуемой лишь десятилетиями спустя.

Ограничения представленных методов явились основой для рассмотрения эвристических моделей (параграф **1.3**). А именно, моделей генерации множеств частично синтетических данных, обеспечивающих одновременно как защиту персональной информации, так и сохранность функциональных связей между соответствующими сегментами множества данных. В параграфе **1.4** вначале представлен сравнительный анализ наиболее распространенных алгоритмов генерации частично синтетических множеств данных[2]. Эти подходы в основном продолжают традиции решений, разработанных для схем восстановления пропущенных значений. Далее работа переходит к изложению известных методов машинного обучения (machine learning) и их расширений, - таких как кластерный анализ, деревья классификации и регрессии (CART), рандомизированные леса (random forests), бэггинг (bagging) и метод опорных векторов (support vector machines).

Во второй главе представлены иерархические модели генерации множеств частично синтетических данных. В параграфе **2.1** представлено описание и анализ наиболее приемлемого алгоритма генерации частично синтетических данных[3]. Генерация синтетических данных осуществляется последовательно, путем наращивания, по каждому конфиденциальному атрибуту. На рисунке 2 представлена структура исходных данных, и система пороговых условий конфиденциальности атрибутов, примененных при работе данного алгоритма.

	A_1	...	A_p	...	A_m	
U_1	a_{11}	...	a_{1p}	...	a_{1m}	$\mathscr{C} = \{C_1, C_2, \dots, C_p\}$

[2] J. Drechsler, J.P. Reiter, "An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets", *Computational Statistics & Data Analysis*, vol. 55(12), pp. 3232-3243, 2011

[3] J.P. Reiter, "Using CART to generate partially synthetic, public use microdata", *Journal of Official Statistics*, vol. 21, pp. 441-462, 2005

$\mathscr{U} =$

U_2	a_{21}	…	a_{2p}	…	a_{2m}
⋮					
U_i	a_{i1}	…	a_{ip}	…	a_{im}
⋮					
U_n	a_{n1}	…	a_{np}	…	a_{nm}

$\mathrm{A_{conf}} = \{\mathrm{A_1}, \mathrm{A_2}, \ldots, \mathrm{A_p}\}$

Рисунок 2. Структура исходных данных и критические интервалы значений атрибутов

Где

$\mathrm{U} = \{U_1, U_2, \ldots, U_n\}$ - множество входных данных (information units set),
$A = \{A_1, A_2, \ldots, A_m\}$ - множество атрибутов, характеризующих элементы данных,
$A_{conf} = \{A_1, A_2, \ldots, A_p\}$ - множество конфиденциальных атрибутов,
$\mathrm{C} = \{C_1, C_2, \ldots, C_p\}$ - множество пороговых условий/интервалов, определяющих степень конфиденциальности атрибутов A_{conf}.

Работа рассматриваемого в параграфе **2.1** алгоритма основывается на деревьях CART. CART используется с целью моделирования и управления условным распределением критических значений конфиденциальных атрибутов. Принцип их построения заключается в рекурсивном разбиении множества рассматриваемых элементов данных на подмножества, однородные относительно зависимой переменной данного шага алгоритма. А именно, на каждом шаге определяется наилучшее условие по некоторому предиктору и производится разбиение текущего множества (growing). Поскольку, полученное при таких итерациях дерево может состоять из неоправданно большого количества узлов и ветвей, то для достижения приемлемого размера этих деревьев производится их отсечение (pruning) на основании некоторого критерия оптимальности. По существу, листья дерева CART представляют условное распределение зависимой переменной для рассматриваемого набора предикторов.

В данном алгоритме для каждого конфиденциального атрибута строится соответствующее дерево CART на основании элементов данных, содержащих критические значения этого атрибута. В качестве набора предикторов рассматриваются все остальные атрибуты множества $\mathscr{A}$, что обеспечивает максимальную информативность в процессе построения. В отличие от традиционного метода построения деревьев CART, в данном алгоритме вместо механизма отсечения используется методика ранней остановки с применением проверки на нетривиальность разбиения, где в качестве критерия рассматриваются минимальное количество элементов и различных значений рассматриваемого конфиденциального атрибута.

Что касается замещений критических значений рассматриваемого атрибута, $A_k (1 \le k \le p)$, то они осуществляются последовательно в листьях соответствующего дерева, с использованием метода Байесовского бутстрапинга. Данный метод генерирует значения на основании некоторого множества возможных значений (donor pool). Для текущего листа L, в качестве этого множества берется множество значений атрибута A_k в данном листе, $A_{(k)}^L = \{a_{(k)1}^L, a_{(k)2}^L, \dots, a_{(k)n_L}^L\}$. В согласии с процедурой Байесовского бутстрапинга, во-первых, генерируются $(n_L - 1)$ равномерно распределенные, произвольные числа в интервале $(0, 1)$ и они упорядочеваются в порядке возрастания: $a_0 = 0, a_1, a_2, \dots, a_{(n_L-1)}, a_{n_L}$. Во-вторых, генерируются n_L таких же чисел в интервале $(0, 1]$, $u_1, u_2, \dots, u_i, \dots, u_{n_L}$, (Рис. 3), и наконец, для каждого u_i $(1 \le i \le n_L)$ определяется интервал $(a_{j-1}, a_j]$, в котором оно содержится, $u_i \in (a_{j-1}, a_j]$, и соответствующее значение $a_{(k)i}^L$ заменяется на $a_{(k)j}^L$.

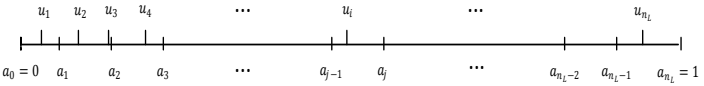

Рисунок 3. Замещение значений по процедуре Байесовского бутстрапинга

В результате последовательных замещений критических значений конфиденциальных атрибутов генерируется соответствующее множество частично синтетических данных.

Согласно анализу проведенному в работе данный алгоритм не лишен недостатков, к которым в частности относится некоторая нерациональность, поскольку для отдельного рассмотрения элементов данных, содержащих некоторую комбинацию критических значений конфиденциальных атрибутов, алгоритму приходится обходить соответствующее дерево CART, даже при их отсутствии, что отрицательным образом сказывается на его производительности. Важно так же, что алгоритм производит замещения критических значений в элементах данных без дополнительной обработки значений, или без группировки элементов по однородным / близким значениям конфиденциальных атрибутов и их соответствующих комбинаций.

Выше приведенное послужило основой для модификации представленного алгоритма с целью улучшения как производительности, так и качества синтетических данных (параграф **2.2**). Процесс модификации был начат с введения системы $\mathscr{R}$, элементы которого констатируют наличие коррелированности между определенными группами атрибутов множества $\mathscr{A}$

$$\mathscr{R} = \{R_1, R_2, \dots, R_t\}.$$

$R_k (1 \le k \le t)$ является подмножеством $\mathscr{A}$, которое указывает на существование коррелированности (или выдвигает требование сохранения формы и степени коррелированности, т.е. факта приема пар значений) между элементами этого множества атрибутов. Дальнейший анализ был основан на том выборе, что в качестве конфиденциальных рассматриваются только количественные атрибуты. Во-вторых, все

атрибуты множества A_{conf} представлены в системе $\mathscr{R}$ где каждый ее элемент R_k, понятно что содержит хотя бы один конфиденциальный атрибут. В-третьих, атрибуты, участвующие в определении коррелированности по $R_1, R_2, \dots, R_t$ трактуются только на уровне пар и они содержат цепные (последовательность пар с общей вершиной) связи в системах R_k. При анализе системы $\mathscr{R}$ было введено понятие *условной коррелированности*, которое подразумевает, что атрибуты A_{j_1} и A_{j_v} условно коррелированы, при условии рассмотрения атрибутов $A_{j_2}, A_{j_3}, \dots, A_{j_{v-1}}$, если существует набор парных коррелированностей $R_{k_1}, \dots, R_{k_{v-1}}$ так, что $R_{k_1} = \{A_{j_1}, A_{j_2}\}$, $R_{k_2} = \{A_{j_2}, A_{j_3}\}, \dots, R_{k_{v-1}} = \{A_{j_{v-1}}, A_{j_v}\}$. Кроме того, было выявлено бинарное отношение α между элементами множества A_{corr} (correlated), представленными в этой системе. Отметим, что атрибут A_{j_1} входит в бинарное отношение α коррелированности с атрибутом A_{j_2} , $A_{j_1} \alpha A_{j_2}$, если A_{j_1} и A_{j_2} удовлетворяют одному из следующих условий:

- Атрибуты A_{j_1} и A_{j_2} совпадают: $A_{j_1} = A_{j_2} \Rightarrow A_{j_1} \alpha A_{j_2}$,
- Атрибуты A_{j_1}, A_{j_2} объявлены коррелированными множеством $\mathscr{R}$: $\exists R_k \in \mathscr{R}$, $R_k = \{A_{j_1}, A_{j_2}\}$ или $R_k = \{A_{j_2}, A_{j_1}\}$,
- Атрибуты A_{j_1}, A_{j_2} условно коррелированы: $\exists A_{j_3}, \dots, A_{v_v} \in A_{corr}$, такие что $R_{A_{j_1}.A_{j_3},\dots,A_{j_v},A_{j_2}} = \{A_{j_1}, A_{j_2}\} \Rightarrow A_{j_1} \alpha A_{j_2}$.

Поскольку, отношение α коррелированности является отношением эквивалентности, то оно разбивает множество A_{corr} на непересекающиеся классы:

$$A_{corr} = A^1_{corr} \cup A^2_{corr} \cup \dots \cup A^s_{corr},$$

$$A^i_{corr} \cap A^j_{corr} = \emptyset,\ 1 \le i \ne j \le s.$$

При этом, любые два атрибута одного и того же класса взаимосвязаны друг с другом, а между атрибутами различных классов коррелированность отсутствует.

Данный анализ позволяет заключит, что подобное разбиение множества A_{corr} на классы эквивалентности дает возможность ограничится рассмотрением возможных определенных комбинаций конфиденциальных атрибутов в пределах взятого одного класса. Кроме того, дальнейшее рассмотрение конфиденциальных атрибутов целесообразней производить последовательно в каждом классе в отдельности.

В процессе построения соответствующего дерева решений для очередного конфиденциального атрибута в текущем классе эквивалентности предлагается в первую очередь произвести разбиения рассматриваемого множества элементов по остальным конфиденциальным атрибутам этого класса и в качестве условий разбиения рассматривать наличие критических значений этих атрибутов в элементах этого множества. Это дает возможность изначально отделить группы элементов с определенными комбинациями критических значений конфиденциальных атрибутов текущего класса. Поскольку, атрибуты различных классов никак не связаны по системе $\mathscr{R}$, то можно ограничится рассмотрением групп по отдельности. На рисунке 4 представлен частный случай класса эквивалентности с тремя конфиденциальными

атрибутами. Дальнейшие разбиения узлов, содержащих элементы этих групп необходимо осуществить таким образом, чтобы с одной стороны сохранить коррелированность между элементами текущего класса, а с другой стороны обеспечить однородность данных по соответствующей комбинации в узлах-потомках. В связи с тем, что мы ограничиваемся рассмотрением только количественных атрибутов в качестве конфиденциальных, то разбиения этих узлов производятся с помощью метода разбивающего иерархического кластерного анализа. Что касается узла, содержащего критические значения только рассматриваемого атрибута, то его разбиения осуществляются так же, как и в ранее представленном алгоритме. По той же причине, в качестве предикторов берутся остальные атрибуты текущего класса эквивалентности, вместо всех остальных атрибутов множества $\mathscr{A}$.

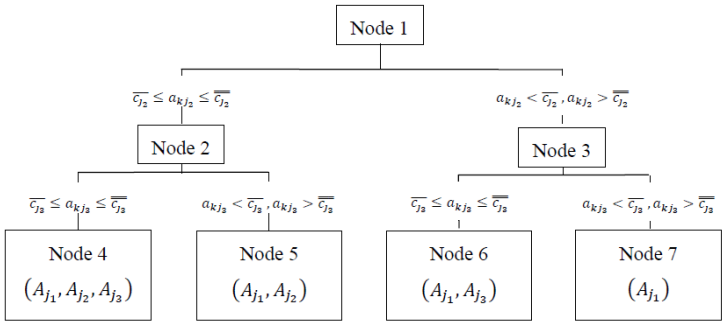

Рисунок 4. Начало дерева по группе критических атрибутов.

По существу, в листьях построенного дерева будут содержатся элементы данных однородные либо по рассматриваемому атрибуту, либо по некоторым комбинациям этого атрибута и остальных атрибутов соответствующего класса эквивалентности. В тех листьях, где содержатся критические значения только рассматриваемого атрибута, замещения осуществляются так же, как и в ранее представленном алгоритме, а в остальных листьях - по наборам значений соответствующих атрибутов комбинации вместо последовательных замещений по каждому из них. Благодаря этому, по возможности сохраняется связи между атрибутами отдельных комбинаций.

Таким образом, представленные данные с очевидностью свидетельствуют, что при наличии дополнительной информации в виде системы $\mathscr{R}$ представляется возможность вместо построения и дальнейшей обработки одного относительно большого дерева изначально детерминировать подмножества элементов, содержащих первостепенные комбинации критических значений конфиденциальных атрибутов, или критические значения лишь одного атрибута, и на их основании работать с несколькими малыми поддеревьями. Причем, основываясь на рассмотрении количественных атрибутах в качестве конфиденциальных, создается возможность разбиения множества элементов, содержащую некоторую комбинацию классов на более однородные, по значениям

атрибутов этой комбинации, а дальнейшие замещения по наборам значений атрибутов позволят сохранить первостепенные корреляции между этими атрибутами по системе $\mathscr{R}$. Что касается последнего подмножества, то его дальнейшая обработка осуществляется на основании относительно малого количества предикторов. По существу, данные изменения ранее представленного алгоритма ведут к повышению его производительности и возможному улучшению качества генерирующихся синтетических данных.

В параграфе **2.3** представлена модель сохранения парных корреляций при генерации синтетических данных. Анализ существующих алгоритмов генерации синтетических данных свидетельствует об их эвристической сущности. А это в свою очередь означает, что состоятельность этих методов обосновывается методами симуляции и нет теоретической обоснованности использования того или иного подхода. В этом параграфе сформулирована модель рассматриваемой задачи для выявления и исследования самой ее сути, в случае сохранения парных связей/корреляций, т.е. исследование структур самих входных данных, естественных ограничений, накладываемых на них методами обработки данных, а также различными требованиями типа конфиденциальности.

При анализе критических областей атрибутов A_{conf}, в первую очередь, очень важно обратить внимание на *предсказуемость значений* в этих областях. В связи с этим, целесообразно оценить информационную энтропию (степень неопределенности) в критических областях для каждого элемента множества A_{conf}. Результаты симуляций показывают, что при значении энтропии в пределах 7.5-8 возможна генерация синтетических данных, обеспечивающих ограничение риска раскрытия конфиденциальной информации.

Далее, для ясности при анализе критических областей конфиденциальных атрибутов были введены такие понятия как *одиночные атрибуты* (single attributes), которые представляют из себя атрибуты, не коррелированные ни с одним другим атрибутом множества $\mathscr{A}$ по системе ограничений $\mathscr{R}$, и *связанные атрибуты* (linked attributes) – остальные атрибуты множества $\mathscr{A}$.

В процессе анализа данных множества одиночных конфиденциальных атрибутов, $A'_{sng} = \{A_1, A_2, \dots, A_l\}$, и дополнительное множество связанных атрибутов $A'_{lnk} = \{A_{l+1}, A_{l+2}, \dots A_p\}$ рассматриваются по отдельности. Изменения критических значений элементов множества A'_{sng}, при построении синтетических данных, осуществляется независимо друг от друга. Мы можем допустить, что критические значения рассматриваемого атрибута, $A_j\ (1 \le j \le l)$, располагаются в верней части соответствующей колонки (Рис. 5) указывающей значения атрибута на элементах входных данных задачи. Выше представленная группировка, полученная путем простой перестановки элементов данных служит основой для рассмотрения изменения критических значений рассматриваемого атрибута в двух отдельных областях:

(1) изменение значений в критической области колонки,
(2) изменение значений по всей колонке.

Конкретное изменение значений атрибута будет зависеть от предполагаемых процедур анализа данных. Если нет других ограничений, то можно рассмотреть перестановки по всей области (2). Перестановки не меняют объективные значения, они меняют их распределения по индивидуумам - строкам. Таким образом, сокрытие критических значений атрибутов A'_{sng} можно осуществить независимо от остальных атрибутов множества $\mathscr{A}$, в пределах соответствующей области.

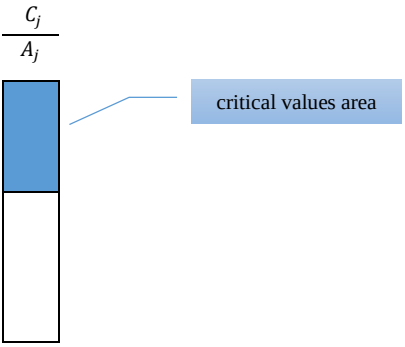

Рисунок 5. Схема расположения значений одиночных атрибутов и области их конфиденциальных значений.

Что касается анализа критических областей атрибутов множества A'_{lnk}, то аналогично анализу, проведенному в параграфе **2.2**, рассматривается система $\mathscr{R}$, элементы которого представляют парные корреляции, которые должны быть сохранены при генерации синтетических данных. Согласно ранее полученным результатам, отношение α коррелированности разбивает множество A_{corr} на непересекающиеся классы эквивалентности. В связи с этим дальнейший анализ критических областей элементов A'_{lnk} производится последовательно по соответствующим классам. При рассмотрения очередного класса эквивалентности, критические значения его конфиденциальных атрибутов могут быть сгруппированы удобным образом. На рисунке 6 представлен пример класса эквивалентности, содержащего r атрибутов, из которых первые два являются конфиденциальными. Благодаря данной группировки строки таблицы данных представляются в виде дискретных областей, содержащих некоторую комбинацию критических значений конфиденциальных атрибутов. Области, содержащие критические значения определенного атрибута или комбинации атрибутов подвергаются дальнейшей обработке. А именно, разбиваются на группы однородные по этим значениям, с учетом сохранения коррелированности между атрибутами рассматриваемого класса эквивалентности. И наконец, в полученных группах осуществляются замещения этих значений. Причем, в группах, содержащих комбинации критических значений, замещения производятся по наборам критических значений соответствующих атрибутов вместо последовательных замещений по каждому из них. В результате обработки

критических значений атрибутов каждого класса эквивалентности генерируется множество синтетических данных, сохраняющее парные связи по заданной системе $\mathscr{R}$.

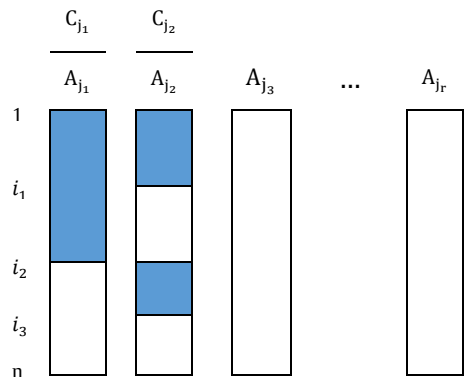

Рисунок 6. Схема расположения значений связанных атрибутов и области их конфиденциальных значений.

В третьей главе дается описание программного обеспечения, реализующего предложенный нами модифицированный алгоритм генерации частично синтетических данных, а также представлены результаты, проведенных экспериментов. В параграфе **3.1** дается описание распределенного (клиент-серверного) программного обеспечения, реализованного на платформе Microsoft .NET Framework 4.5 с использованием технологии WCF (Windows Communication Foundation) 3.5, среды программирования- Microsoft Visual Studio 2012, и в качестве базы данных- Microsoft SQL Server 2012. При помощи технологии Windows Forms реализован внешний интерфейс клиентской части, где устанавливаются необходимые параметры для генерации синтетических данных. А именно, из базы данных выбирается таблица оригинальных данных, отмечаются конфиденциальные атрибуты, устанавливаются их критические области, определяется система коррелированности $\mathscr{R}$ и устанавливается количество необходимых множеств синтетических данных. И наконец, в серверной части, на основании входных параметров, генерируется соответствующее количество синтетических данных, которые предоставляются пользователю (Рис. 7).

Рисунок 7. Общая схема работы распределенного программного обеспечения.

В параграфе **3.2** проведен сравнительный анализ между стандартными алгоритмами области исследования и алгоритмами, предложенным нами. Один из наших экспериментов, наипростейший, был проведен на основании микроданных интегрированного обследования домашних хозяйств РА 2011 года. Из всего множества атрибутов, характеризующих эти данные, были рассмотрены следующие: *Food Purchased, Food Consumed, Non Food Purchased, Expenditure, Monitory Income, Total Income* (таблица 1).

Таблица 1. Описание атрибутов.

Name	*Label*	*Description*
Food Purchased	FP	Food purchased of household per month.
Food Consumed	FC	Food consumed of household per month.
Nonfood Purchased	NFP	Nonfood purchased of household per month.
Expenditure	E	Expenditures of household per month.
Monitory Income	M	Monetary income of household per month.
Total Income	I	Total income of household per month.

Мы предполагаем, что конфиденциальная информация содержится в атрибутах E и I, и в качестве пороговых условий рассматриваются: $E > 200000$ и $I > 250000$. Что касается парных связей, которые необходимо сохранить, то они следующие: (I, M), (I, E), (E, NFP) и (E, FP). Ниже приведены результаты этого эксперимента (таблице 2).

Таблица 2 Результаты эксперимента**.**

Estimands	original data	synthetic data (modified algorithm)	synthetic data (viewed algorithm)
Mean:			
Total income	132862.38	132523.88	132627.714
Expenditure	109818.56	108960.52	109447.164
Standard deviation:			
Total income	105518.35	100335.54	100351.228
Expenditure	95060.95	78792.584	86459.4434
Coefficient in regression of *expenditure* on:			
Constant	-1664.49	1813.41	5643.5738
Total income	0.026	0.041	0.0802
Food purchased	1.017	0.919	0.9916
Food consumed	0.992	0.9262	1.007
Nonfood purchased	1.089	1.110	0.800
R	0.983	0.957	0.7834

Согласно этим результатам значения математического ожидание и среднеквадратичного отклонения, вычисленных на множествах синтетических данных, близки к оригинальным. А это, в свою очередь, свидетельствует о том, что числовые

характеристики атрибутов *Total income* и *Expenditure* сохраняются в синтетических данных, сгенерированными как рассмотренным, так и модифицированным алгоритмами. Однако, существенные отклонения наблюдаются в моделях линейной регрессии. А именно, на множествах синтетических данных, сгенерированных с помощью рассмотренного алгоритма, значение параметра R, показывающего степень правильности интерпретации зависимой переменной независимыми, соответствующей модели на много меньше, чем 0.9, а это указывает на то, что данная модель не столь корректна/правильна. В свою очередь, значение R, вычисленное на множествах, синтезированных с применением модифицированного алгоритма, превышает данное пороговое значение, т.е. эти синтетические данные лучше отражают связь между соответствующими атрибутами. Таким образом, полученные результаты с очевидностью свидетельствуют о том, что синтетические данные, сгенерированные на основании модифицированного алгоритма, качественней, чем данные, синтезируемые с помощью существующего алгоритма. Эксперименты полного объема, правда менее прозрачно, представляют те же интерпретации точности и быстродействия построенных алгоритмов.

ВЫВОДЫ

Основные результаты диссертационной работы заключаются в следующем:

1. Проведен общий анализ существующих технологий ограничения риска раскрытия конфиденциальной информации, в результате которого были выявлены их сильные и слабые стороны [2].

2. Создана модель генерации синтетических данных для обеспечения сохранения парных корреляций атрибутов данных [4].

3. Разработан новый/альтернатичный алгоритм генерации синтетических данных с целью повышения производительности и возможного улучшения качества синтезируемых данных [1, 3].

4. Создано программное обеспечение с целью реализации альтернативного алгоритма генерации синтетических данных [1, 4].

Список публикаций по теме диссертации

1. L. Aslanyan, V. Topchyan, “Hierarchical Cluster Analysis for Partially Synthetic Data Generation”, *Mathematical Problems of Computer Science*, vol. 40, pp. 55-67, Institute for Informatics and Automation Problems of NAS RA, Yerevan, Armenia, 2013.

2. L. Aslanyan, V. Topchyan, H. Danoyan, “Brief Analysis of Technique for Privacy Preserving Computation”, International Journal “Information content and Processing”, issue 1, pp. 4-19, 2014, Sofia.

3. Л Асланян, В. Топчян, «Улучшенные CART технологии генерации частично синтетических данных», International Journal “Information content and Processing”, issue 2, pp. 4-16 , 2014, Sofia.

4. V. Topchyan, “Pair Correlations Preserving Model in the Synthetic Data Generation”, *Mathematical Problems of Computer Science*, vol. 41, pp. 74-86, Institute for Informatics and Automation Problems of NAS RA, Yerevan, Armenia, 2014.

Անփոփում

Վարդան Հակոբի Թոփչյան

Ինֆորմացիայի նկատմամբ սեփականության իրավունքը պահպանող հաշվարկներ

Թեմայի արդիականությունը: Ինֆորմացիոն տեխնոլոգիաների ինտենսիվ զարգացմանը զուգընթաց առաջանում են անձնական և այլ կոնֆիդենցիալ տվյալների հավաքման և վերլուծման ավտոմատացման նոր միջոցներ և մոտեցումներ: Վիճակագրական, ֆինանսական և այլ ինֆորմացիոն կազմակերպությունները ագրեգացված տվյալների մշակումից և դրանց հասարակությանը տրամադրելուց բացի ավելի շատ միտված են տվյալները սկզբնականին մոտ բաշխելուն: Դա, նախ և առաջ, ժամանակի պահանջ է, բայց այն նաև ուղղված է այդ տվյալների նպատակային վերլուծության միջոցով անկախ, հասարակական և գիտական վերահսկման և փորձաքննության բարձրացմանը: Սոցիոլոգիական տվյալները պարունակում են մեծ ծավալի անձնական և բիզնես կողմնորոշում ունեցող ինֆորմացիա: Ֆորմատով և բովանդակությամբ սկզբնական տվյալներին մոտ տվյալների հրապարակումը կապված է այդ տվյալների գաղտնիության բացահայտման ռիսկի հետ: Կրիտիկական տվյալները հրապարակելիս դրանց բաց լինելու և կոնֆիդենցիալության պահանջների հակասությունը հիմք հանդիսացավ նոր հետազոտական ուղղության առաջացմանը, ներկայումս հայտնի ինչպես տվյալների բացահայտման սահմանափակման տեխնոլոգիաներ: Խնդրի արդիականությունը հիմնավորվում է նոր իրավական ակտերով, ինչպես օրինակ CIPSEA—the Confidential Information Protection and Statistical Efficiency Act of 2002 (Chance, 17(3):2125, 2004), և նոր հետազոտություններով ինչպես EUREDIT - the development and evaluation of new methods for editing and imputation (IST-1999-10226 EC FP project, 2000 - 2003).

Գիտական նորույթը: Ըստ էության, որակյալ սինթետիկ տվյալների գեներացման ալգորիթմական խնդիրը կախված է ոչ միայն գեներացման մոդելից, այլ նաև տվյալներից: Աշխատանքի նորույթը կայանում է խնդրի տվյալների նախնական վերլուծության մեջ, ինչը տրամադրում է տվյալների՝ գեներացման մոդելին համապատասխանության մասին ինֆորմացիա: Այնուհետև, ատրիբուտների զույգերի վերլուծությունը, սահմանված ինչպես փոխադարձորեն կապված, և գեներացման հիերարխիկ ծառերի կիրառումը թույլ է տալիս կատարել արդյունավետ ընդհատում դրանով օպտիմիզացնելով կառուցման և հատումների ստանդարտ պրոցեսը:

Պաշտպանությանը ներկայացվող հիմնական դրույթներն են

- Ատրիբուտների և դրանց փոխադարձ կապերի վերլուծության մոդելի կառուցումը որոշելով տվյալների որակը և ատրիբուտների հաջորդականությունը կոնֆիդենցիալ տվյալների հիերարխիայի կառուցման մեջ:
- Դասակարգման և ռեգրեսիայի ծառի լավացումը սինթետիկ տվյալների գեներացման խնդրում հիերարխիկ ծառի հատման և կառուցման պրոցեսների համատեղմամբ:
- Մշակված սինթետիկ տվյալների գեներացման այլընտրանքային ալգորիթմի ծրագրային ապահովման ստեղծումը:
- Համակարգի փորձնական ներդրումը պետական, վիճակագրական և ֆինանսական տվյալների մշակման գործընթացում:

Աշխատանքի հիմնական արդյունքներն են

1. Իրականացվել է կոնֆիդենցիալ տվյալների բացահայտման ռիսկը սահմանափակող գոյություն ունեցող տեխնոլոգիաների ընդհանուր վերլուծություն, որի արդյունքում հայտնաբերվել նրանց ուժեղ և թույլ կողմերը [2]:
2. Ստեղծված է ատրիբուտների զույգային կորրելացիաների պահպանումը ապահովող սինթետիկ տվյալների գեներացման մոդել [4]:
3. Մշակվել է նոր/այլընտրանքային ալգորիթմ արդյունավետության բարձրացման և սինթեզվող տվյալների որակի հնարավոր լավացման նպատակով [1, 3]:
4. Ստեղծված է ծրագրային ապահովում սինթետիկ տվյալների գեներացման այլընտրանքային ալգորիթմն իրականացնելու նպատակով [1,4]:

Resume

Vardan H. Topchyan

Privacy preserving computations

There appear new approaches and tools for automation of acquisition and analysis of personal and other confidential data due to the intensive development of information technologies for state and social needs. Statistical, financial and other information structures today are releasing not only the aggregated data, but also the data that is close to the original raw data. Firstly, this approach is the requirement of the time, but it is also aimed at growing the independent, public and scientific management of information, providing expertise through targeted analyses of the such data. Sociological data contain elements or personal or business oriented data. Publication of data close to the format and content of the original source is associated with the risk of that data aiming the so called privacy disclosure. Contradiction of transparency and confidentiality requirements in publication of critical data has been the basis for new research directions now known as a disclosure limitation technology. The importance and actuality of this problem area is visible by the new legal acts as for example CIPSEA - the Confidential Information Protection and Statistical Efficiency Act of 2002 (Chance, 17 (3): 2125, 2004), and the new Research projects such as EUREDIT-the development and evaluation of new methods for editing and imputation (IST-1999-10226 EC FP project, 2000-2003).

Scientific novelty. In its essence, the algorithmic task of generation of high-quality synthetic data replacing the original raw data depends not only on model of data generation, but also on the raw data itself. The inside innovation of the work is in the preliminary analysis of the input data that provides additional information about the correspondence of input data to the generated synthesizing model. Further, analysis of pair attributes, which defined as correlated (accepting paired than the independent values), and subsequent use of the hierarchical tree with sets of such attributes allows generating an effective stop optimizing of the standard growing and pruning processes. Thus, analysis of original data uncovers opportunity of generating synthesized data replacement concealing privacy, and this analysis of pair correlations limits the algorithmic steps of the generation by raising its productivity.

The following statements presented for defense:

- Design of model of analysis of data characteristic attributes and their interrelationship with the aim of estimating the data quality and consistency of attributes in a hierarchy splits of confidential data.
- Improvement of implementation of classification and regression tree in synthetic data generation by combining the growing and pruning processes in the hierarchical tree model.
- Development of software implementation of alternative algorithm for synthetic data generation.
- Test deployment of system in public, statistical and financial data processing.

The main results of the thesis are:

- A general analysis conducted for existing technologies limiting the risk of disclosure for confidential information, which allowed finding out strong and weak sides of them [2].
- A model of generating synthetic data created for the preservation of pair correlations between data attributes [4].
- A new / alternative algorithm has developed to generate synthetic data in order to increase productivity and improve the quality of synthesized potential data [1, 3].
- Development of application with the aim of implementation of an alternative algorithm for synthetic data generation [1, 4].