УЛУЧШЕННЫЕ CART ТЕХНОЛОГИИ ГЕНЕРАЦИИ ЧАСТИЧНО СИНТЕТИЧЕСКИХ ДАННЫХ

Левон Асланян, Вардан Топчян

Аннотация: *Работа посвящена исследованию вопросов анализа персональных данных обеспечивающих конфиденциальность данных. Предполагается что даны частично критические социологические данные и перед представлением этих данных общественности требуется их модифицировать так, чтобы конфиденциальные данные не раскрывались, и чтобы анализ этих данных не отличался от анализа исходных данных. Работа строит улучшенные алгоритмы класс деревьев классификации и регрессии, которые предоставляют решение задачи генерации так называемых синтетических данных. Новое решение учитывает структуры областей конфиденциальности и проводит оптимизацию дерева замены данных на синтетические.*

Ключевые слова: *классификация, регрессия, раскрытие данных, синтетические данные.*

ACM Classification Keywords: H.1 Information Systems – Models and principles, I.2.0 Artificial intelligence.

Введение

Предоставление экономических, социальных данных общественным структурам является неотъемлемой частью деятельности государственных статистических организаций. Открытый доступ к данным имеет большие преимущества. Прежде всего такие данные могут явится источником осуществления разного рода исследований, в том числе и в учебных целях. Тем не менее ограничение риска раскрытия (disclosure limitation) конфиденциальной информации продолжает оставаться одной из главных задач статистических организаций потому что даже при удалении очевидных идентификаторов таких как имя или адрес не исключают возможности доступа к персональным данным. Ведь как показано многими авторами благодаря сопоставлениям значений общих ключевых атрибутов в нескольких таблицах данных можно выявить определенные персональные данные. Для решения данной задачи/проблемы часто прибегают к модификации (perturbation) исходных данных или к их замене другими, новыми данными. Эти данные генерируются на основе разных моделей и алгоритмов [4]. Часто модифицируются значения отдельных атрибутов/дескрипторов. Таким образом защищая отдельные поля информации, они могут привести к побочному эффекту, к искажению связей между разными сегментами множества данных, что в свою очередь может привести к ошибочным выводам на этапе статистического анализа данных.

Альтернативным подходом решения поставленной задачи, который одновременно пытается сохранить функциональные связи между сегментами множества данных, является подход генерации так называемых полных синтетических данных (synthetic data generation, SDG) [5]. В этом случае, статистическая организация должна, во-первых, произвольно и независимо отмечать общий формат и критическое содержание единиц информации и включать их в соответствующее предполагаемое множество синтетических данных, во-вторых: по выбранной стратегии/алгоритму устанавливать новые,

синтетические, значения в единицах информации, и в-третьих: предоставить общественности некоторое количество множеств, сгенерированных синтетических данных. Известны различные методы [6] генерации полных синтетических данных, обеспечивающих получение значимых результатов с использованием стандартных статистических методов.

Несмотря на отмеченные преимущества института полных синтетических данных, процесс их генерации довольно трудоемкий. Не понятен также подход когда изменяются неконфиденциальная составляющая исходной информации. В связи с этим часто прибегают к использованию схемы генерации частично синтетических данных [6], представляющих из себя сочетание оригинальных и синтетических данных. Потребность в генерации частично синтетических данных возникает в тех случаях, когда статистическое агентство стремится защитить конфиденциальность для определенных записей. С этой целью, генерируются синтетические значения лишь для определенных атрибутов, а значения остальных не изменяются.

Как и в случае полных синтетических данных, частично синтетические данные так же обеспечивают ограничение риска раскрытия информации, позволяя получать значимые результаты с использованием стандартных статистических методов. Отметим, что, в силу своей природы, применение частично синтетических данных, обеспечивает более точные результаты статистических вычислений. По той же причине, риск раскрытия информации выше по сравнению с полными синтетическими данными. Однако, известны различные алгоритмы [5, 6] для их генерации, используемые многими статистическими организациями (U.S. Federal Reserve Board, U.S. Bureau of the Census, Statistical agencies of Germany and New Zealand, etc.), что говорит о перспективах данного метода.

Анализ существующих алгоритмов генерации синтетических данных показывает их эвристическую структуру. Таким образом обоснованием является эксперимент и нет теоретической обоснованности использования того или другого подхода. Вместе с тем область конфиденциальности данных задачи хорошо интерпретируема и она подлежит формальному описанию. Данная работа, впервые, сформулирует формальную модель критических данных и попытается построить улучшенные алгоритмы генерации синтетических данных следя за сохранением как отдельных значений параметров задачи так и за совместными значениями групп параметров и атрибутов. В теоретическом плане, как это замечено отдельными авторами, сформулированные задачи схожи с вероятностными задачами восстановления отсутствующих значений (missing value). В таких схемах возможно получение оценок ошибки однако практические задачи не обладают достаточной информацией для восстановления вероятностных распределений и наша цель не в получении таких оценок а в формализации и использовании дополнительных свойств задачи для сформирования более адекватного прикладного результата.

Выше представленное послужило основой для нашего изучения методов генерации частично синтетических данных [1, 2]. Для осуществления идеи этой работы нам необходимо выбрать и остановиться на одном из подходящих методов генерации синтетических данных. Как показывает анализ литературных данных [5, 6], приемлемых является [6], работа которого основана на использовании деревьев CART (*Classification and Regression Trees*). Прежде чем перейти к более подробному рассмотрению этого алгоритма, дадим краткое описание формата наших данных и предполагаемых синтетических данных задачи.

Формат частично синтетических данных

Процесс генерации частично синтетических данных состоит из двух этапов: (1) предварительная обработка (preprocessing) входных данных, (2) замещение отмеченных, критических значений на синтетические. Формально данный процесс можно описать следующим образом.

Пусть, $\mathcal{U}$ множество отдельных элементов, из которых составлены входные данные, $\mathcal{U} = \{U_1, U_2, \dots, U_N\}$, где каждый элемент характеризуется множеством атрибутов $\mathcal{A} = \{A_1, A_2, \dots, A_p\}$:

$$U_i = \left(a_{i1}, a_{i2}, \dots, a_{ij}, \dots, a_{ip}\right), (1 \le i \le N, 1 \le j \le p).$$

На этапе препроцессинга данных произвольным образом отмечается определенное количество элементов множества $\mathcal{U}$ для текущего наблюдения (observation) и отмечаются конфиденциальные атрибуты (строки и столбцы матрицы $\mathcal{U}$ соответственно), и устанавливаются пороговые условия для атрибутов.

Пусть, $n\ (n \le N)$ есть количество произвольно выбранных элементов множества $\mathcal{U}$, а $d\ (d \le p)$ – количество конфиденциальных атрибутов. Обозначим выбранные элементы и конфиденциальные атрибуты через $\{U_{i_1}, U_{i_2} \dots, U_{i_n}\}$ и $\{A_{j_1}, A_{j_2}, \dots, A_{j_d}\}$ соответственно. С целью определения этих элементов и атрибутов введем вспомогательные наборы индикаторов $I = (I_1, I_2, \dots, I_N)$ и $J = (J_1, J_2, \dots, J_p)$:

$$I_r = \begin{cases} 1, U_i \in \{U_{i_1}, U_{i_2} \dots, U_{i_n}\} \\ 0, U_i \notin \{U_{i_1}, U_{i_2} \dots, U_{i_n}\} \end{cases} 1 \le i \le N,$$

$$J_k = \begin{cases} 1, A_j \in \{A_{j_1}, A_{j_2} \dots, A_{i_d}\} \\ 0, A_j \notin \{A_{j_1}, A_{j_2} \dots, A_{i_d}\} \end{cases} 1 \le j \le p.$$

Схематически данный процесс представлен на Рис. 1.

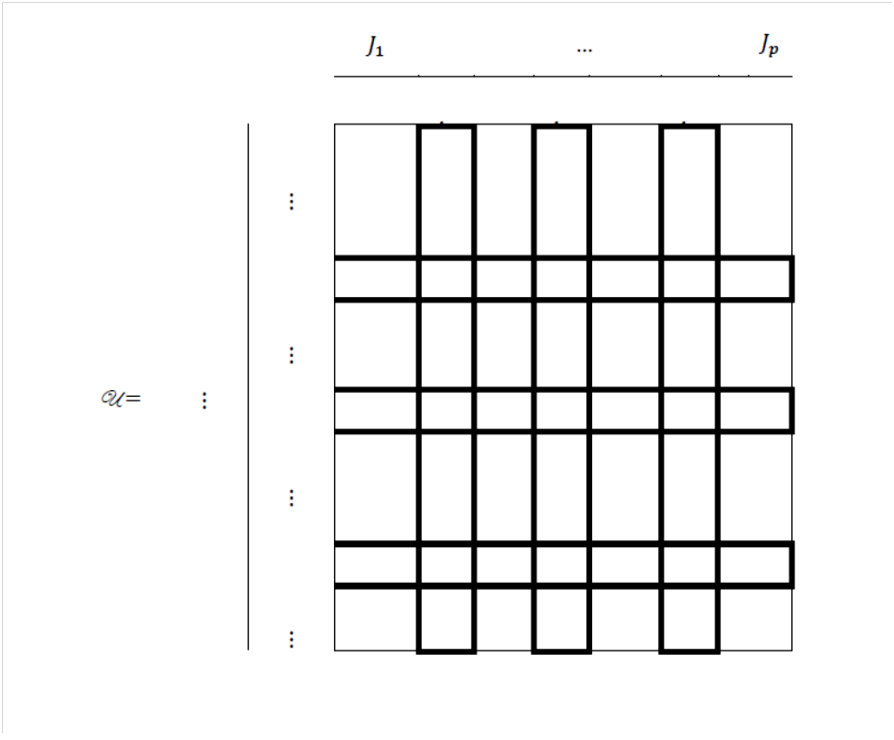


Рис. 1 Исходные данные задачи с выделением критических данных

По существу, в результате определяется расщепление $U_{obs} = (U_{rep}, U_{nrep})$ исходной матрицы наблюдений размерами $[n \times p]$ составленного из рассматриваемых *(observed)* единиц информации. Здесь U_{rep} – представляет собой матрицу размера $[n \times d]$ значений конфеденциальных атрибутов $A_{j_1}, A_{j_2}, \dots, A_{j_d}$ (replaced vs. not replaced). А U_{nrep} – $[n \times (p-d)]$ матрица значений остальных атрибутов (атрибутов значения которых не подвергаются замещению). Для упрощения представления матрицы U_{obs} ее столбцы соответствующим образом можно переставить (Рис. 2).

$U_{obs} =$

	U_{rep}			U_{nrep}		
	A_{j_1}	…	A_{j_d}	$A_{j_{(d+1)}}$	…	A_{j_p}
U_{i_1}	$a_{i_1 j_1}$	…	$a_{i_1 j_d}$	$a_{i_1 j_{(d+1)}}$	…	$a_{i_1 j_p}$
U_{i_2}	$a_{i_2 j_1}$	…	$a_{i_2 j_d}$	$a_{i_2 j_{(d+1)}}$	…	$a_{i_2 j_p}$
⋮						
U_{i_n}	$a_{i_n j_1}$	…	$a_{i_n j_1}$	$a_{i_n j_{(d+1)}}$	…	$a_{i_n j_p}$

Рис. 2 Переставленная исходная матрица

Далее, для определения критических областей значений атрибутов $A_{j_1}, A_{j_2}, \dots, A_{j_d}$ устанавливаются соответствующие пороговые условия (threshold conditions) посредством некоторого множества $\mathcal{C} = \{C_1, \dots, C_d\}$ (Рис. 3).

$U_{rep} =$

	C_1	…	C_d
	A_{j_1}	…	A_{j_d}
U_{i_1}	$y_{i_1 j_1}$	…	$y_{i_1 j_d}$
U_{i_2}	$y_{i_2 j_1}$	…	$y_{i_2 j_d}$
⋮			
U_{i_n}	$y_{i_n j_1}$	…	$y_{i_n j_d}$

$\mathcal{C} = \{C_1, \dots, C_d\}$

Рис. 3 Критическая матрица и пороговые значения атрибутов

На основании этих условий определяется индикаторная матрица $Z\,[n \times d]$, характеризующая необходимость изменения значений конфиденциальных атрибутов (Рис. 4). А именно, значение элемента z_{rk} берется равным единице, если значение соответствующего атрибута A_{j_k} , $a_{i_r j_k}$, удовлетворяет пороговому условию C_k, $z_{rk} = 1\ (1 \le r \le n)$, $(1 \le k \le p)$. В обратном случае $z_{rk} = 0$.

Индикаторная (0,1) матрица в принципе может иметь произвольную форму. Понятно что обычно условия C_k представляются простыми логико-арифметическими выражениями и что этим характеризуется структура самой матрицы Z. Однако в общем случае матрица произвольная и вопрос ее эффективного применения связан с задачами ее оптимального представления. Здесь можно рассмотреть схему представления матрицы в виде суммы малого числа матриц простой структуры, представление при помощи иерархического дерева или другие виды представления. Индикаторная (0,1) матрица таким образом имеет вид

$$Z = \begin{array}{|c|c|c|} \hline z_{1j_1} & \cdots & z_{1j_d} \\ \hline z_{2j_1} & \cdots & z_{2j_d} \\ \hline & & \\ \hline z_{rj_1} & \cdots & z_{zj_d} \\ \hline & & \\ \hline z_{nj_1} & \cdots & z_{nj_d} \\ \hline \end{array}$$

Рис. 4 Индикаторная матрица

и согласно замеченному, с этим связано некоторое адекватное структурное представления данной матрицы.

Суммарным результатом предварительной обработки множества входных данных $\mathcal{U}$и вспомогательных индикаторных наборов I, J являются матрица U_{obs} , соответственно матрицы U_{rep} и U_{nrep} , и индикаторная матрица Z. Полученный набор данных обозначим через $D = (U_{rep}, U_{nrep}, Z)$.

Вторая, в принципе основная часть процесса генерации частично синтетических данных представляет из себя процесс изменения/замещения критических значений. А именно, на основании полученного набора данных $D = (U_{rep}, U_{nrep}, Z)$ и выбранного алгоритма, производятся замещения соответствующих значений матрицы U_{rep} на новые, синтетические значения. Процесс замещения производится независимо некоторое количество m раз, в результате чего генерируются m различных множеств частично синтетических данных:

$$SD_t = \left(U_{syn}^t, U_{nrep}\right),\ 1 \le t \le m\ ,$$

где $U_{syn}^t -$ матрица с установленными синтетическими значениями $t-$го наблюдаемого множества. Отметим, что матрица U_{nrep} одинакова для всех множеств $SD_t, 1 \le t \le m$. Это характерное но не обязательное условие задачи. При желании можно синтезировать и не критические значения и к тому есть своя причина. Первая причина сложностная - она упрощает работу алгоритмов замещения данных. Вторая связана с специфичным раскрытием, когда добывая нужное количество некритических значений атакующий пытается восстановить ее связи к критическим данным.

Таким образом, полученные множества частично синтетических данных $SD_1, SD_2, \dots, SD_m$ являются теми данными, которые предоставляются соответствующим организациям и общественным структурам. На этапе оценки предлагаемой модели и алгоритма замещения данных необходимо подтвердить (validation) что по совокупности предполагаемых методов анализа результаты обработки синтетических данных не будет отличаться от оригинальных результатов по исходным данным.

Описание алгоритма CART в виде удобном для SDG

Теоретико-графовое понятие дерева служит основой ряда моделей поиска, классификации, предоставления привилегий, и др. [8-12]. Примеры известных древовидных моделей принятия решений включают ID3, C4.5, CART, CHAID. В статистике, наряду с иерархическим кластерным анализом используются процедуры типа Bagging, Random Forest, Boosted Trees, которые строят и оптимизируют классы деревьев. Деревья являются существенной компонентой алгоритмов сжатия данных (gif, lzw), систем вычислений, распознавания, семантических анализов данных и др. Наша ближайшая задача в этих терминах характеризуется как задача построения моделей иерархических ресурсов данных и задача иерархической классификации с минимизацией ошибки. Генерирующая идея работы заключается в ограничении решающих правил используемых в вершинах деревьев классом условий и ограничений, характеризующих критические данные задачи. Это возможно и эффективно потому что в основной модели дерево строится по критерию минимизации ошибки и только после этого снижается ее сложность учитывая критические и не критические значения переменных. Построение модели требует некоторую детализацию описания древовидных структур и процедур описания критических данных, к которомы и сейчас мы переходим.

Алгоритм CART является одним из представителей древовидных моделей обработки данных. Для нашего случая алгоритм предназначен для генерации частично синтетических множеств данных, которые

возможно использовать для вычисления простых статистических величин переменных (математического ожидания, дисперсии и т.д.) и построения классификационных и линейных регрессионных моделей. Работа алгоритма основана на бинарных деревьях с условиями на вершинах. Они используются с целью управления условного распределения критических значений конфиденциальных атрибутов.

Деревья CART используются для прогнозирования значений зависимой переменной на основании набора предикторов. Принцип построения CART заключается в рекурсивном разбиении множества рассматриваемых элементов данных на подмножества, однородные относительно зависимой переменной. А именно, на каждом шаге определяется наилучшее условие по некоторому предиктору и производится разбиение текущего множества (growing). В результате, в листьях полученного дерева будут содержатся элементы данных с одинаковым значением зависимой переменной. Поскольку, полученное 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.

Построение дерева. При описании работы алгоритма мы будем придерживаться обозначений, введенных в предыдущей части. Не нарушая общности, допустим, что матрица U_{obs}, полученная в результате предварительной обработки данных, состоит из первых n элементов множества $\mathcal{U}$. Учитывая то, что порядок атрибутов не фиксирован по смыслу нашей задачи, в этапе обработки данных их переставлением мы можем добиться того, чтобы конфиденциальные данные оказались только в первых d столбцах данных, т.е. они определены атрибутами

$$A_{conf} = \{A_1, A_2, \dots, A_d\},\ A_{conf} \subseteq \mathcal{A}$$

В алгоритме, генерация синтетических данных осуществляется последовательно, путем наращивания, по каждому конфиденциальному атрибуту. В связи с этим, на первом этапе работы производится упорядочивание атрибутов $A_1, A_2, \dots, A_d$ по мере убывания количества критических значений по входным данным. С этой целью, для каждого атрибута A_k $(1 \le k \le d)$, на основании индикаторной матрицы Z, вычисляется следующее значение:

$$\sum_{i=1}^{n} z_{ik}\ .$$

Однако, не исключено, что для некоторой группы атрибутов данное значение может быть одинаковым. В этом случае, порядок для этих атрибутов устанавливается по мере важности каждого из них при построении соответствующих деревьев CART для остальных членов этой группы. Для наглядности рассмотрим частный случай с двумя атрибутами. Допустим, что A_b и A_c – атрибуты с одинаковым количеством критических значений,

$$\sum_{i=1}^{n} z_{ib} = \sum_{i=1}^{n} z_{ic}\ .$$

Для этих атрибутов строятся соответствующие им деревья *CART*, T_b и T_c (Рис. 5). Далее, для атрибута A_b определяется величина P_b, равная глубине в дереве T_b, где впервые встречается разбиение по атрибуту A_c (если такое разбиение отсутствует, то P_b условно берется равным $P_b = \infty$). Величина P_b характеризует то насколько сильным предиктором является A_c для атрибута A_b. А именно, чем меньше

значение P_b, тем больше зависимость атрибута A_b от A_c. Аналогичным образом определяется величина P_c для атрибута A_c. На основании полученных данных, порядок для этих атрибутов устанавливается по мере убывания величин P_a и P_b.

Рис. 5. Определение уровня разбиения по данному атрибуту

В результате, упорядоченные атрибуты обозначаются следующим образом: $A_{(1)}, A_{(2)}, \dots, A_{(d)}$.

Пусть, $A_{(k)}$ – текущий атрибут. С целью генерации синтетических значений для атрибута $A_{(k)}$, в первую очередь, строится дерево CART, $T_{(k)}$. Поскольку, $T_{(k)}$ используется с целью определения условного распределения атрибута $A_{(k)}$ в пространстве критических значений, то в качестве объектов для ее построения рассматриваются лишь те элементы U_{obs}, для которых $z_{i(k)} = 1$. Однако, если количество этих элементов не достаточно для построения корректной модели, тогда используются все элементы U_{obs}. А в качестве предикторов берутся остальные $(p-1)$ атрибутов, что обеспечивает максимальную информативность во время построения $T_{(k)}$. 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 атрибута $A_{(k)}$.

Замещение значений. Далее, на основании полученного дерева $T_{(k)}$, осуществляются замещения значений атрибута $A_{(k)}$. В связи с тем, что в листьях $T_{(k)}$ содержатся элементы U_{obs}, однородные относительно значений $A_{(k)}$, то процесс замещения реализуется последовательно по листьям данного дерева. В данном алгоритме, замещения критических значений осуществляется благодаря методам перестановки (relocation) и переоценки (reevaluation) значений. Пусть, L есть текущий лист в дереве $T_{(k)}$, а $A_{(k)}^L = \{a_{(k)1}^L, a_{(k)2}^L, \dots, a_{(k)n_L}^L\}$ – множество значений атрибута $A_{(k)}$ в данном листе. Сначала осуществляется перестановка значений множества $A_{(k)}^L$. С этой целью применяется метод Байесовского бутстрапинга. Данный метод генерирует значения на основании некоторого множества возможных значений (donor pool). Для листа L в качестве данного множества рассматривается $A_{(k)}^L$. 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 числе в интервале $(0,1]$, $u_1, u_2, \dots, u_i, \dots, u_{n_L}$, (Рис. 7), 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 $a^L_{(k)i}$ заменяется на $a^L_{(k)j}$.

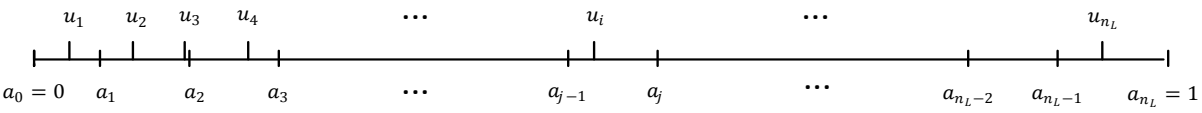


Рис. 6 Замещение значений по процедуре Байесовского бутстрапинга

Не исключено, что после перестановки значений атрибута $A_{(k)}$ некоторые из них могут остаться неизменными. В таких случаях, производится переоценка новых/переставленных значений $A_{(k)}$. С этой целью, в листе L определяется вероятностная плотность этих значений с помощью вычислителя плотности Гауссовского ядра (Gaussian kernel density estimator):

$$\hat{f}(x) = \frac{1}{hn_L}\sum_{i=1}^{n_L}\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(x-a^L_{rep,(k)i}\right)^2}{2}},$$

$$h = \left(\frac{4\hat{\sigma}^5}{3n_L}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}{n_L}^{-\frac{1}{5}},$$

где $a^L_{rep,(k)1}, a^L_{rep,(k)2}, \dots, a^L_{rep,(k)n_L}$ – новые значения атрибута $A_{(k)}$ в листе L, а $\hat{\sigma}$ – среднеквадратическое отклонение этих значений. Затем, для каждого элемента данных произвольным образом выбирается значение вероятностной плотности из множества вычисленных. На основании этого значения вычисляются функция распределения $F(x)$ и ее обратная функция $F^{-1}(y)$:

$$F(x) = \int_{-\infty}^{x}\hat{f}(t)d(t),$$

$$F^{-1}(y) = \inf\{F(x) \geq y\}.$$

В результате, полученное значение функции $F^{-1}(y)$ устанавливается в качестве нового значения атрибута $A_{(k)}$ для рассматриваемого элемента данных.

Не исключено, что некоторые элементы данных, участвующие в построении дерева $T_{(k)}$ $(k > 1)$, могут содержать синтетические значения атрибутов $A_{(1)}, A_{(2)}, \dots, A_{(k-1)}$ построенные на предыдущих шагах алгоритма. Тогда, для обеспечения согласованности в замещениях, элементы данных с одинаковой комбинацией критических значений $A_{(k)}$ и новых значений $A_{(1)}, \dots, A_{(k-1)}$ рассматриваются отдельно от остальных. В тех случаях, когда эти элементы не содержатся в одном из листьев $T_{(k)}$, то алгоритм осуществляет обходит дерева снизу-вверх для определения внутренней вершины, в которой они содержатся. Замещения значений атрибута $A_{(k)}$ для этих элементов осуществляются в полученной вершине.

Таким образом, в результате последовательных замещений критических значений конфиденциальных атрибутов генерируется множество частично синтетических данных. Весь процесс повторяется

независимо m раз и полученные множества $SD_t, 1 \leq t \leq m$ предоставляются соответствующим организациям и общественным структурам.

Анализ алгоритма

Очевидно, что преимуществами приведенного алгоритма являются, во-первых, ее сложностные характеристики. Это связано с тем, что, в алгоритме осуществляется последовательная обработка конфиденциальных атрибутов, и это дает приемлемое приближение, и возможность получения качественных результатов даже при большом количестве конфиденциальных атрибутов и других данных. Во-вторых, использование деревьев CART дает возможность работать как с “качественными”, так и с количественными атрибутами, т.е. работать со смешанными данными (mixed data). Более того, благодаря этим деревьям возможно обнаружение сложных, не линейных функциональных связей между атрибутами данных. Однако, алгоритм не лишен недостатков. Так, ограничением алгоритма является его возможное применение при наличии относительно большого количества элементов данных. Поскольку, при существенно большом количестве последних возможно построение не точных а приближенных (усеченных) моделей CART, то это в свою очередь может явится причиной потери функциональных связей между различными атрибутами, что явно негативный фактор. Кроме того, согласно исследованиям [6] отсутствуют данные подтверждающие, что используемый порядок последовательного рассмотрения критических атрибутов является оптимальным. Не исключено, что при некотором другом порядке могли бы быть получены лучшие результаты. И наконец, алгоритм характеризуется некоторой нерациональностью, поскольку с целью отдельного рассмотрения элементов данных, содержащих некоторую комбинацию критических значений конфиденциальных атрибутов, алгоритм обходит соответствующее дерево CART, даже при их отсутствии. А это отрицательным образом сказывается на его производительности.

Основные причины - недостатки CART побудившие данное исследование тем не менее адресуют другие – нестандартные характеристики. Во первых, речь идет о трудоемкой процедуре отсечения (pruning) деревьев при помощи которой отсекаются часть ветвей при минимальной потере точности классификации или регрессионного приближения. Понятно что более логично выработать процедуры остановки построения дерева чем ее построить и потом отсечь. Это где то напоминает критерии кластеризации которые в иерархическом кластерном анализе управляют процесс завершения роста кластеров. По нашему предположению в задаче SDL по процедуре CART это может быть достигнуто путем применения на вершинах дерева не произвольных а критически характеризованных условий разделения текущей смеси объектов. Во вторых, порой очень сомнительны результаты замещений данных и непонятно что об этом не напоминают существующие публикации по теме. Имеется в виду следующее. Пусть определенное критическое значение оно только одно, и принимается на каком то количестве объектов. Существующие процедуры замещения не имеют пространства для замены значений ь значение одно и оно может быть замещено самым собой. Эли нужно расширить пространство замен, или нужно ввести меру соответствия задачи идее SDL. Мера соответствия для отдельного атрибута зависит от интервала применяемых значений и от количества различных значений, а общая мера для задачи интегрирует отдельные меры атрибутов. Ниже содержится продолжение описания улучшения CART учитывающее указанные недостатки внедрения стандартного метода.

Модификация алгоритма

Рассмотрим основную модель задачи. Пусть, $\mathcal{U} = \{U_1, U_2, \dots, U_N\}$ множество отдельных элементов, из которых составлены рассматриваемые входные данные задачи (*information units set*). Пусть отдельный элемент данных характеризуется множеством значений атрибутов множества $\mathcal{A} = \{A_1, A_2, \dots, A_p\}$.

Не нарушая общности, мы будем предполагать, что в данном наблюдении матрица U_{obs} состоит из первых n элементов множества $\mathcal{U}$. А конфиденциальные данные содержатся в первых d атрибутах:

$$A_{conf} = \{A_1, A_2, \dots, A_d\},\ A_{conf} \subseteq \mathcal{A}$$

В принципе, в качестве конфиденциальных атрибутов могут быть рассмотрены как так называемые "качественные", так и количественные атрибуты. Однако, мы ограничимся рассмотрением только количественных атрибутов.

Для этих атрибутов вводится множество пороговых условий (*threshold conditions set)*, которые и определяют степень их конфиденциальности,

$$\mathcal{C} = \{C_1, C_2, \dots, C_d\}.$$

Условие C_j $(1 \le j \le d)$ определяет критическую область *(critical area)* значений атрибута A_j. Для простоты рассмотрений будем предполагать, что условия C_j задают числовые интервалы в области определения соответствующего атрибута A_j, хотя рассмотрение других структур ограничений может оказаться вполне естественным и полезным. Пусть, C_j определяет интервал $(\bar{c}_j, \bar{\bar{c}}_j)$ критических значений в области $(\bar{a}_j, \bar{\bar{a}}_j)$ определения атрибута A_j, $\bar{a}_j \le \bar{c}_j \le \bar{\bar{c}}_j \le \bar{\bar{a}}_j$.

Далее, мы предполагаем, что в качестве дополнительной информации нам дано множество $\mathcal{R}$, элементы которого констатируют наличие коррелированности между определенными группами атрибутов множества $\mathcal{A}$,

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

R_k $(1 \le k \le t)$ является подмножеством $\mathcal{A}$, $R_k \subseteq \mathcal{A}$, которое указывает на существование коррелированности (или выдвигает требование сохранения формы и степени коррелированности) между элементами этого множества атрибутов. Дальнейший анализ будет основан на предположении, что все атрибуты множества A_{conf} представлены в системе $\mathcal{R}$ и каждый ее элемент содержит хотя бы один конфиденциальный атрибут. Действительно, в обратном случае, если некоторый элемент R_k $(1 \le k \le t$ не содержит ни одного конфиденциального атрибута, то его рассмотрение не имеет смысла, ибо изменения критических значений атрибутов A_{conf} никак не отразятся на связи, представленного этим элементом. Кроме того, если некоторый атрибут A_j $(1 \le j \le d)$ не представлен в системе $\mathcal{R}$, то это означает, что A_j не коррелирован ни с одним из других атрибутов множества A_{conf}, что в свою очередь свидетельствует о том, что нет необходимости в рассмотрении элементов данных с комбинацией A_j и других конфиденциальных атрибутов.

Очевидно, что рассмотрение всех комбинаций конфиденциальных атрибутов не рационально, что выдвигает необходимость анализа системы $\mathcal{R}$ для выявления наиболее характерных комбинаций, которыми можно было ограничится. Пусть $d > 2$. Предположим также, что атрибуты, участвующие в определении коррелированности по $R_1, R_2, \dots, R_t$ только парные и содержат пересечения. Рассмотрим

подмножества $R_{k_1}, R_{k_2} \in \mathscr{R}$, характеризующие коррелированность между атрибутами A_{j_1}, A_{j_2} и A_{j_2}, A_{j_3}, $(1 < j_1 \neq j_2 \neq j_3 \leq m)$ соответственно, $R_{k_1} = \{A_{j_1}, A_{j_2}\}, R_{k_2} = \{A_{j_2}, A_{j_3}\}$. Допустим так же, что дополнительно не задана коррелированность между атрибутами A_{j_1} и A_{j_3}. В целях сохранения связи по R_{k_1} необходимо, что бы изменения значений A_{j_1} и A_{j_2} были согласованы. А именно, значения A_{j_1} должны быть изменены с учетом соответствующих значений атрибута A_{j_2} и наоборот. Аналогичные суждения имеют место для R_{k_2} и атрибутов A_{i_2}, A_{i_3}. Очевидно, что атрибут A_{j_2} зависит как от A_{j_1}, так и от A_{j_3}. Поэтому, для сохранения коррелированности по R_{k_1}, R_{k_2} значения атрибутов A_{j_1} и A_{j_3} так же должны изменены в согласии друг с другом. В результате, между атрибутами A_{j_1} и A_{j_3} возникает взаимосвязь, при условии рассмотрения атрибута A_{j_2}. Выше приведенные данные позволяют ввести следующее естественное определение.

Определение 1.

Скажем, 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 $R_{k_1} = \{A_{j_1}, A_{j_2}\}, R_{k_2} = \{A_{j_2}, A_{j_3}\}, \dots, R_{k_{v-1}} = \{A_{j_{v-1}}, A_{j_v}\}$.

Условную коррелированность атрибутов A_{j_1}, A_{j_v} обозначим через $R_{A_{j_1}, A_{j_2}, \dots, A_{j_v}} = \{A_{j_1}, A_{j_v}\}$.

Дальнейшим анализом системы $\mathscr{R}$ явилось изучение бинарного отношения между атрибутами, представленными в этой системе. Рассмотрим множество этих атрибутов обозначенное через A_{corr} (*correlated*).

Определение 2.

Скажем, 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 $R_{A_{j_1}, A_{j_3}, \dots, A_{j_v}, A_{j_2}} = \{A_{j_1}, A_{j_2}\} \Rightarrow A_{j_1} \alpha A_{j_2}$.

Очевидно, что α удовлетворяет свойствам рефлексивности и симметричности:

$$\forall A_{j_k} \in A_{corr} \Rightarrow A_{j_k} \alpha A_{j_k},$$

$$\forall A_{j_k}, A_{j_r} \in A_{corr}, A_{j_k} \alpha A_{j_r} \Rightarrow A_{j_r} \alpha A_{j_k}.$$

Покажем, что это отношение удовлетворяет также и свойству транзитивности, а именно:

$$\forall A_{j_k}, A_{j_r}, A_{j_s} \in A_{corr}, A_{j_k} \alpha A_{j_r}, A_{j_r} \alpha A_{j_s} \Rightarrow A_{j_k} \alpha A_{j_s}.$$

Так как $A_{j_k} \alpha A_{j_r}$, $A_{j_r} \alpha A_{j_s}$, то из определения отношения α следует, что между атрибутами A_{j_k}, A_{j_r} и A_{j_r}, A_{j_s} существует либо прямая, либо условная коррелированность. Тогда в силу **определения 1** атрибуты A_{j_k} и A_{j_s} будут условно коррелированными. А это в свою очередь означает, что A_{j_k} входит в отношение α с атрибутом A_{j_s}: $A_{j_k} \alpha A_{j_s}$.

Итак, отношение α удовлетворяет свойствам рефлексивности, симметричности и транзитивности, следовательно, оно является отношением эквивалентности. В этом случае, α разбивает множество A_{corr} на непересекающиеся классы эквивалентности:

$$A_{corr} = A_{corr}^1 \cup A_{corr}^2 \cup \ldots \cup A_{corr}^s,$$

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

Причем, 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 одного класса. Кроме того, дальнейшее рассмотрение конфиденциальных атрибутов целесообразней производить последовательно в каждом классе в отдельности.

Пусть, A_{corr}^i – текущий класс эквивалентности. Для удобства интерпретации, рассмотрим частный случай, когда класс A_{corr}^i состоит только из трех конфиденциальных атрибутов: $A_{corr}^i = \{A_{j_1}, A_{j_2}, A_{j_3}\}$; $A_{j_k} \in A_{conf}$ $(k = 1,2,3)$. Не нарушая общности, допустим, что порядок рассмотрения этих атрибутов в построении дерева расщеплений следующий: $A_{j_1} - A_{j_2} - A_{j_3}$. Дерево T_{j_1}, соответствующее атрибуту A_{j_1}, строится на основании множества элементов данных с критическими значениями этого атрибута, $U^{A_{j_1}} = \{U_k,\ \overline{c_{j_1}} \le a_{kj_1} \le \overline{\overline{c_{j_1}}}\}, U^{A_{j_1}} \subseteq U_{obs}$. Поскольку, элементы данных, содержащие комбинации критических значений атрибутов A_{j_1} и A_{j_2}, A_{j_3}, должны быть рассмотрены в отдельности от остальных элементов множества $U^{A_{j_1}}$, тогда целесообразней осуществить процедуру их отделения на начальной стадии/этапе построения дерева T_{j_1}. С этой целью, в первую очередь произвести разбиения множества $U^{A_{j_1}}$ по атрибутам A_{j_2}, A_{j_3} и в качестве условий разбиений рассматривать наличие критических значений этих атрибутов в элементах множества $U^{A_{j_1}}$ (Рис.7).

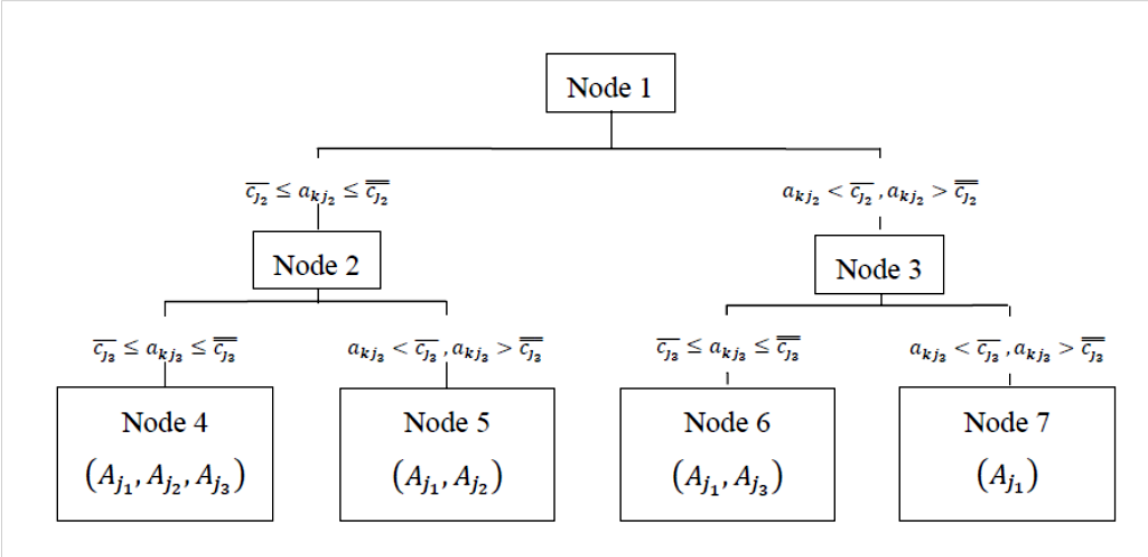


Рис. 7 Начало дерева по группе критических атрибутов

Далее, как видно из рисунка 7 в узлах Node 4, Node 5, Node 6 содержатся критические значения определенных комбинаций атрибутов, поэтому дальнейшие разбиения каждого из этих узлов необходимо осуществить таким образом, чтобы с одной стороны сохранить корреляции между элементами класса A^i_{corr}, а с другой стороны обеспечить однородность данных по соответствующей комбинации в узлах-потомках. Поскольку, в наших исследованиях мы ограничиваемся рассмотрением только количественных атрибутов в качестве конфиденциальных, то для дальнейших разбиений этих узлов мы применяем методику разбивающего иерархического кластерного анализа [1]. В этом случае, элементы данных рассматриваются как 3-мерные векторы, состоящие из значений атрибутов класса A^i_{corr}, что позволяет осуществлять разбиения с учетом корреляций между атрибутами этого класса. В качестве меры расстояния между элементами данных рассматривается Эвклидово расстояние, а в качестве меры однородности полученных подмножеств - мера RMSSTD (Root-Mean-Square Standard Deviation) [14], равная среднеквадратичному отклонению критических значений атрибутов соответствующей комбинации. Что касается узла Node 7, то, поскольку в нем содержатся критические значения только атрибута A_{j_1}, его разбиения будут осуществляться тем же способом, что и в деревьях CART.

По существу, листья дерева T_{j_1} будут содержат элементы данных однородные либо по атрибуту A_{j_1}, либо по некоторым комбинациям A_{j_1} и остальных атрибутов класса A^i_{corr}. В тех листьях, где содержатся критические значения только атрибута A_{j_1}, замещения могут быть осуществлены так же, как и в ранее представленном алгоритме (с использованием Байесовского бутстрапринга и вычислителя плотности Гауссовского ядра), а в остальных листьях - по наборам значений соответствующих атрибутов комбинации вместо последовательных замещений по каждому из них. Благодаря этому, по возможности сохраняется корреляция между атрибутами отдельных комбинаций.

Для атрибутов A_{j_2}, A_{j_3} процесс построения соответствующих деревьев T_{j_2}, T_{j_3} и дальнейшие замещения их критических значений осуществляются аналогичным образом. Очевидно, что с целью построения дерева T_{j_2} будут применены элементы данных множества $U^{A_{j_2}} \setminus U^{A_{j_1}}$. Это предполагает, что в этом дереве выборка данных будет основана на отделении элементов с комбинацией критических значений атрибутов $\left(A_{j_2}, A_{j_3}\right)$ от остальных элементов. Что касается дерева T_{j_3}, то анализ всех возможных комбинаций атрибутов класса A^i_{corr} в деревьях T_{j_1} и T_{j_2}, позволяет рассматривать T_{j_3} как дерево CART со множеством данных $U^{A_{j_3}} \setminus \left(U^{A_{j_1}} \cup U^{A_{j_2}}\right)$.

Приведенный простой пример позволяет заключить, что если класс эквивалентности A^i_{corr} состоит из m_i атрибутов, $A^i_{corr} = \left\{A_{j_1}, A_{j_2}, \dots, A_{j_{m_i}}\right\}$, из которых первые q являются конфиденциальными, тогда для атрибута A_{j_k} $(1 \le k \le q)$ бинарное дерево решений T_{j_k} строится на основании множества элементов $U^{A_{j_k}} \setminus \cup_{r=1}^{k-1} U^{A_{j_r}}$. Причем, при построении этого дерева, в первую очередь, рассматриваются разбиения по атрибутам $A_{j_{k+1}}, A_{j_{k+2}}, \dots, A_{j_q}$ с целью определения и отдельного рассмотрения элементов данных, содержащих комбинации критических значений атрибутов A_{j_k} и $A_{j_{k+1}}, A_{j_{k+2}}, \dots, A_{j_{m_i}}$. В результате, в листьях T_{j_k} будут содержатся элементы данных однородные либо по критическим значениям некоторой

комбинации атрибутов A_{j_k} и $A_{j_{k+1}}, A_{j_{k+2}}, \dots, A_{j_{m_i}}$, либо по критическим значениям только атрибута A_{j_k}. А замещения осуществляются либо по наборам значений соответствующих комбинаций, либо по атрибуту A_{j_k}.

Таким образом, представленные данные с очевидностью свидетельствуют, что при наличии дополнительной информации в виде системы $\mathscr{R}$ изначально могут быть детерминированы элементы данных, содержащие наиболее важные/первостепенные комбинации конфиденциальных атрибутов. Кроме того, основываясь на рассмотренных количественных атрибутах в качестве конфиденциальных, становится ясным возможность разбиения множества элементов, содержащую некоторую комбинацию классов на более однородные, по значениям атрибутов этой комбинации, и подмножества и дальнейшие замещения по наборам значений атрибутов позволят сохранить первостепенные корреляции между этими атрибутами по системе $\mathscr{R}$. Данные модельные структуры и анализ подтверждаются вычислительными экспериментами.

Заключение

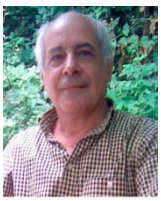
Задачи сохранения конфиденциальности данных при распределенных вычислениях связаны с новыми теоретическими и прикладными исследованиями. С одной стороны, криптография пытается синтезировать схемы кодирования, в которых результат анализа этих данных совпадает с анализом исходных, не кодированных данных, однако известно что такие прикладные системы будут созданы не так скоро. Альтернативные модели анализа данных существуют и имеют эвристический характер. В данной работе исследовались схемы вычислений с сохранением конфиденциальности данных которые основаны на использовании моделей CART и введены дополнительные компоненты модели повышающие скорость вычислений и близость результатов анализа данных к исходным. Результат достигается путем исключения этапа урезания деревьев из процесса анализа данных, что основано на том, что в качестве условий разветвления деревьев используются условия определяющие конфиденциальность данных задачи.

Литература

1. L. Aslanyan, V. Topchyan, Hierarchical Cluster Analysis For Partially Synthetic Data Generation, Transactions of IIAP of NAS of RA, Mathematical Problems of Computer Science, submitted, 2013.
2. Vardan Topchyan, Statistical Disclosure Limitation of Public Use Data by Syntheses with Clustering, ITA 2013 – ITHEA ISS Joint International Events on Informatics, Winter Session, December 18 – 19, 2013, Sofia, Bulgaria, pp. 17.
3. Willenborg, L. and de Waal, T. (2001). Elements of Statistical Disclosure Control. New York: Springer-Verlag.
4. Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 462–468.
5. Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, 1–16.
6. Reiter, J.P. (2005). Significance Tests for Multi-component Estimands from Multiply-imputed, Synthetic Microdata. Journal of Statistical Planning and Inference, 131, 365 - 377.

7. Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.
8. Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.
9. Barros, Rodrigo C., Basgalupp, M. P., Carvalho, A. C. P. L. F., Freitas, Alex A. (2011). A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 42, n. 3, p. 291-312, May 2012.
10. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, Foundations of Machine Learning, MIT Press, 2012.
11. Donald E. Knuth, The Art of Computer Programming, Volume 3, Sorting and Searching, Second Edition (Reading, Massachusetts: Addison-Wesley, 1998), 780pp, ISBN 0-201-89685-0
12. A. V. Aho, J. E. Hopcroft, J. D. Ullman, Data Structures and Algorithms. Addison-Wesley, 1983. ISBN 0-201-00023-7.
13. M. Yu. Moshkov, Conditional tests, Problemi Kibernetiki (in Russian), issue 40, pp. 131-170, Moscow, Nauka, 1983.
14. Subhash Sharma: Applied multivariate techniques, John Wiley & Sons, Inc., 1996.
15. Little, R.J.A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.
16. Kennickell, A.B. (1997). Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, (eds), Record Linkage Techniques, 1997, 248–267. Washington, D.C.: National Academy Press.
17. Abowd, J. M. and Woodcock, S. D. (2001). Disclosure Limitation in Longitudinal Linked Data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds). Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, 215–277. Amsterdam: North-Holland.
18. Liu, F. and Little, R.J.A. (2002). Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata. In Proceedings of the Joint Statistical Meetings of the American Statistical Association, 2133–2138.
19. Jorg Drechsler (2011). Synthetic Datasets for Statistical Disclosure Control. Theory and Implementation.
20. Reiter, J.P. (2005). Using CART to Generate Partially Synthetic, Public Use Microdata. Journal of Official Statistics, Vol. 21
21. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworh, Inc.
22. Reiter, J.P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, 181–189.
23. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, (2008) Introduction to Information Retrieval,378-382.
24. Rubin, D.B. (1981). The Bayesian Bootstrap. The Annals of Statistics, 9, 130–134.
25. Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis, (2002) Clustering Valifity Checking Methods: Part II, (19-27).
26. L. Aslanyan, H. Danoyan, P. Zelenko, Applied Best-Match Type Algorithms, Proceedings of the 7th annual scientific conference of RAU, Yerevan, pp. 56-62, 2013.
27. L. H. Aslanyan, H. E. Danoyan, Complexity of Elias algorithm based on codes with covering radius three, Proceedings of the Yerevan state university, 2013 №1, pp. 44-50.
28. L. H. Aslanyan, H. E. Danoyan, Complexity of Elias algorithm based on Hamming and extended Hamming codes, Reports of NAS RA, vol. 113, no. 2, pp. 151-158, 2013.
29. L. H. Aslanyan, H. E. Danoyan, “On the optimality of a hash-coding type search algorithm”, Proceedings of the 9th conference CSIT, Yerevan, Armenia, pp. 55-57, 2013.

Информация об авторах



Levon Aslanyan – ITHEA ISS, Sofia, Bulgaria; EC Horizon2020 ICT NCP Armenia; Institute for informatics and automation problems of NAS RA, 1, P. Sevak street, Yerevan 0014, Armenia, e-mail: lasl@sci.am

Major Fields of Scientific Research: Discrete optimization, Artificial intelligence, NLP, WSN, Privacy preserved computation

Vardan Topchyan – Institute for informatics and automation problems of NAS RA, 1, P. Sevak street, Yerevan 0014, Armenia, e-mail: vardan.topchyan@gmail.com

Major Fields of Scientific Research: Decision models, Homomorphic encryption, Privacy preserved computation

Enhanced Cart Technologies in Partial Synthetic Data Generation

Levon Aslanyan, Vardan Topchyan

Abstract: This work aims at studying personal data analysis area, when confidentiality property of data is ensured. It is supposed that we are given partially critical social science data and prior to the submission of data to the public it is required to modify them so that confidential information is not disclosed, and that the analysis of these data did not differ from the analysis of raw data. Our work builds improved algorithms of class of classification and regression trees, which provide solution to the problem of generation of the so-called synthetic data. The new solution of generation takes into account the structure of the areas of privacy and is providing optimized tree replacement for synthetic data sets.

Keywords: classification, regression, data disclosure, synthetic data.