Общество с ограниченной ответственностью «Системы информационной безопасности»
(ООО «СИБ»)

УДК _____
Рег. № НИОКТР
Рег. № ИКРБС

СОГЛАСОВАНО
Проректор по научной работе и инновациям ФГБОУ ВО «ТУСУР», канд. техн. наук, доц.

_____А.Г. Лощилов
«__» _____ 2024 г.

УТВЕРЖДАЮ
Директор ООО «СИБ»

_____А.А. Помешкин
«__» _____ 2024 г.

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ РАБОТЕ
ИССЛЕДОВАНИЯ, РАЗРАБОТКА И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ
по теме:
РАЗРАБОТКА МЕТОДИКИ АВТОМАТИЗИРОВАННОГО ОБЕЗЛИЧИВАНИЯ ДАННЫХ
(заключительный)

Грант на государственную поддержку центров Национальной технологической инициативы на базе образовательных организаций высшего образования и научных организаций

Соглашение о предоставлении субсидии от 14.12.2021 г. № 70-2021-00246

Руководитель НИР,
Руководитель центра ГосСОПКА _____И.В. Коротких

Новосибирск 2024

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель НИР, Руководитель центра ГосСОПКА	_____ подпись, дата	И.В. Коротких (введение, заключение)
Отв. исполнитель, Специалист отдела внедрения и сопровождения информационных систем	_____ подпись, дата	А.А. Балабанов (разделы 1, 2, 3, заключение)
Исполнители: Начальник отдела внедрения и сопровождения информационных систем	_____ подпись, дата	А.А. Вяткин (раздел 1, 2)
Ведущий специалист отдела внедрения и сопровождения информационных систем	_____ подпись, дата	А.Г. Марковин (раздел 2)
Специалист отдела внедрения и сопровождения информационных систем	_____ подпись, дата	А.И. Болдин (раздел 1)
Специалист отдела внедрения и сопровождения информационных систем	_____ подпись, дата	Н.И. Калабин (раздел 2)
Специалист отдела внедрения и сопровождения информационных систем	_____ подпись, дата	Н.А. Синигин (раздел 2)
Главный специалист отдела внедрения и сопровождения информационных систем	_____ подпись, дата	Д.В. Соколов (раздел 2)
Заместитель руководителя центра ГосСОПКА	_____ подпись, дата	В.В. Тюменцев (разделы 2,3)
Специалист центра ГосСОПКА	_____ подпись, дата	И.А. Федотов (раздел 3)
Инженер по внедрению и эксплуатации программно-технических средств центра ГосСОПКА	_____ подпись, дата	М.А. Киселев (раздел 3)
Нормоконтроль	_____ подпись, дата	И.В. Коротких

РЕФЕРАТ

Отчет 165 с., 31 рис., 8 табл., 41 источн., 0 прил.

ПЕРСОНАЛЬНЫЕ ДАННЫЕ, ОБЕЗЛИЧИВАНИЕ ДАННЫХ, МЕТОДИКА ОБЕЗЛИЧИВАНИЯ, МЕТОД ОБЕЗЛИЧИВАНИЯ, КРИТЕРИЙ СРАВНЕНИЯ, КРИТЕРИЙ ЭФФЕКТИВНОСТИ, РИСК РАСКРЫТИЯ ИНФОРМАЦИИ, КОНТЕКСТНЫЙ РИСК, КОМПЛЕКСНЫЙ РИСК, ИНФОРМАЦИОННАЯ ПОТЕРЯ, ВРЕМЯ ОБЕЗЛИЧИВАНИЯ

Объектом исследования является методика автоматизированного обезличивания данных.

Цель работы: разработка методики автоматизированного обезличивания данных.

В ходе исследования разработан укрупненный алгоритм, реализующий методику автоматизированного обезличивания данных, в виде последовательности взаимосвязанных этапов. Разработаны алгоритмы реализации каждого из этапов методики. Разработаны критерии и методика сравнения эффективности разных вариантов обезличивания данных в соответствии с предложенными критериями.

В результате исследования была достигнута поставленная цель и сформулированы выводы по результатам проведенного анализа.

Результаты исследований применимы для дальнейшего использования при реализации проектов: 2.5.2.10 «Разработка математических моделей на основании разработанных методик», «Разработка модулей подсистемы интеллектуального управления, обогащения и обезличивания данных».

СОДЕРЖАНИЕ

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящем отчете о НИР применяют следующие термины с соответствующими определениями:

Автоматизированная обработка персональных данных – обработка персональных данных с помощью средств вычислительной техники [1]

Алгоритм – формальное представление, точное предписание (в виде конечного набора правил), однозначно определяющее содержание и последовательность чисто механически выполняемых действий (алгоритмических операций), переводящих исходные данные задачи в искомый результат [2]

Атрибут (признак) обезличенных данных субъекта – элемент структуры обезличенных данных (параметр обезличенных данных). Атрибут имеет название и может иметь множество возможных количественных и качественных значений [1]

Атрибут (признак) персональных данных субъекта – элемент структуры персональных данных (параметр персональных данных). Атрибут имеет название и может иметь множество возможных количественных и качественных значений применительно к конкретным субъектам персональных данных [1]

Деобезличивание – действия, в результате которых обезличенные данные принимают вид, позволяющий определить их принадлежность конкретному субъекту персональных данных, то есть становятся персональными данными [1]

Критерий сравнения	– это свойство или характеристика (признак, показатель, мера), по которой сопоставляют исследуемые объекты между собой [2]
Методика	– алгоритм (процедура), описывающий (описывающая) последовательность взаимосвязанных этапов, реализуемых для решения поставленной научно-исследовательской или прикладной задачи. Методика также включает описание методов, подходов, моделей, приемов, средств и других способов достижения поставленных целей, используемых на каждом этапе. Подобное определение методики с точки зрения алгоритмизации применяется в проектах, нацеленных на разработку программного обеспечения в качестве конечного результата, поэтому данное определение подходит под задачи текущего мероприятия [3]
Обезличенные (анонимизированные) данные	– это данные, хранимые в информационных системах в электронном виде, принадлежность которых конкретному субъекту персональных данных невозможно определить без дополнительной информации [1]
Обезличенные данные (ОД) субъекта	– это данные, представляемые в виде записи, которая является самостоятельной единицей данных, имеет определенную структуру и содержит множество значений атрибутов (признаков) обезличенных данных [1]

Обезличивание (анонимизация) персональных данных	– действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность персональных данных конкретному субъекту персональных данных [1]
Обработка обезличенных данных	– любое действие (операция) или совокупность действий (операций), совершаемых с использованием средств автоматизации, с обезличенными данными, без применения их предварительного деобезличивания [1]
Обработка персональных данных	– любое действие (операция) или совокупность действий (операций), совершаемых с использованием средств автоматизации или без использования таких средств с персональными данными, включая сбор, запись, систематизацию, накопление, хранение, уточнение (обновление, изменение), извлечение, использование, передачу (распространение, предоставление, доступ), обезличивание, блокирование, удаление, уничтожение [1]
Персональные данные (ПД)	– любая информация, относящаяся к прямо или косвенно определенному, или определяемому физическому лицу (субъекту персональных данных) и хранимая в информационных системах в электронном виде [1]
Персональные данные субъекта	– это данные, представляемые в виде записи, которая является самостоятельной еди-

	ницей данных, имеет определенную структуру и содержит множество значений атрибутов (признаков) персональных данных субъекта [1]
Раскрытие информации	– повторная идентификация субъекта, которая выполняется злоумышленником [4]
Информационные потери	– потери, связанные с искажением значений исходного атрибута (признака), возникающие вследствие применения методов обезличивания [4]
ISO/IEC 27018:2019	– международный стандарт, которые устанавливает требования и определяет методики по защите идентифицирующей информации в информационной среде [5]
ISO/IEC 20889:2018	– международный стандарт «Терминология, касающаяся способов повышения конфиденциальности, и классификация методов обезличивания» [6]

ВВЕДЕНИЕ

Работа проводится за счет собственных средств ООО «СИБ» из внебюджетных источников в рамках выполнения мероприятия «Исследования, разработка и математическое моделирование. Разработка методики автоматизированного обезличивания данных» ключевого проекта «Технология интеллектуального управления данными для платформы «Доверенная среда обмена информацией», включающая в себя автоматизированные системы обезличивания и обогащения данных» Программы создания и развития Центра компетенций Национальной технологической инициативы «Технологии доверенного взаимодействия» в редакции от 26.04.2023 г. (пп. 8 пункт 2 таблицы 5.4.1 Приложения №5 Программы создания и развития ЦК НТИ ТДВ); результат исполнения проекта – «Разработана методика, пояснительная записка, отчет о НИР» (пп. 8 пункт 2 таблицы 5.4.1 Приложения №5 Программы создания и развития ЦК НТИ ТДВ).

Обезличивание персональных данных используется для обеспечения конфиденциальности персональной информации о субъекте ПД и является ключевым этапом всего процесса управления и защиты информации. Основная идея состоит в преобразовании данных таким образом, чтобы снизить риск раскрытия информации, но при этом сохранить полезность данных на приемлемом уровне. Решение этой задачи требует рассмотрения целого ряда вопросов, связанных с подготовкой данных, выбором и реализацией методов обезличивания, статистической оценкой риска раскрытия информации и информационных потерь, созданием программно-инструментальных средств обезличивания данных. Должна быть разработана общая методика автоматизированного обезличивания данных, охватывающая все этапы процесса: от определения исходных данных, условий и целей использования обезличенной БД, выбора и обоснования методов и инструментальных средств до оценки риска раскрытия информации после проведения обезличивания ПД и оценки информационных потерь, возникающих вследствие обезличивания ПД.

Цель данного исследования заключалась в разработке методики автоматизированного обезличивания данных, которая, в свою очередь, может стать осно-

вой для разработки модуля «Обезличивание данных» для системы «Интеллектуальное управление, обогащение и обезличивание данных» в рамках платформы «Доверенная среда обмена информацией».

Методика разработана в соответствии с требованиями, сформулированными в рамках проекта 2.5.2.5 «Формирование требований к методике автоматизированного обезличивания данных».

В первой главе приведена общая характеристика методики автоматизированного обезличивания данных и описан укрупненный алгоритм реализации методики в виде последовательности взаимосвязанных этапов.

Во второй главе описаны алгоритмы реализации каждого из этапов методики автоматизированного обезличивания данных. Рассмотрены основные подходы и методы, показатели и меры, которые используются на каждом из этапов методики.

В третьей главе приведены критерии сравнения эффективности разных вариантов обезличивания данных и описана методика сравнения разных вариантов обезличивания данных в соответствии с принятыми критериями и выбора оптимального варианта.

1 Методика автоматизированного обезличивания данных

1.1 Общая характеристика

1. Методика автоматизированного обезличивания данных предназначена для реализации процедуры обезличивания данных с целью расширения возможностей их использования и предоставления широкому кругу потребителей (внешних пользователей) при обеспечении конфиденциальности персональных данных субъектов.

2. Методика является автоматизированной, что предполагает использование программно-технических средств при реализации этапов методики, но не является автоматической и не может быть реализована без участия специалиста.

3. Методика разработана с учетом требований российского законодательства в сфере защиты информации [1,4,7-9] и международных рекомендаций и стандартов в области обезличивания, использования и предоставления обезличенных данных внешним пользователям [5,6, 10-12].

4. Методика представлена в виде укрупненного алгоритма, включающего последовательность взаимосвязанных этапов, направленных на реализацию процедуры обезличивания данных и оценки ее эффективности на основе расчета набора количественных показателей и мер. Каждый из этапов методики также представлен в виде реализующего его алгоритма. Алгоритмы описаны в словесной форме и в виде блок-схем, оформленных по ГОСТ 19.701-90.

5. Методика включает, как один из этапов, оценку риска раскрытия информации на обезличенных данных. В рамках этапа математически оценивается эффективность процедуры обезличивания на основе расчета набора количественных показателей риска, значения которых сравниваются с установленными пороговыми значениями. Таким образом, выполняется математически гарантированное обезличивание данных на базе теории риска в соответствии с мировой практикой и международными стандартами в сфере защиты информации.

6. Методика разработана в соответствии с требованиями, предложенными в проекте 2.5.2.5 «Формирование требований к методике автоматизированного обезличивания данных».

7. Методика может быть взята за основу при разработке модуля «Обезличивание данных» в рамках системы «Интеллектуальное управление, обогащение и обезличивание данных» для платформы «Доверенная среда обмена информацией».

1.2 Укрупненный алгоритм автоматизированного обезличивания данных

На рисунке 1.1 представлена блок-схема укрупненного алгоритма, реализующего методику автоматизированного обезличивания данных. Ниже рассмотрены и описаны основные этапы алгоритма.

1. Постановка и формализация задачи обезличивания данных

На первом этапе выполняется постановка и формализация задачи обезличивания данных. Содержательно определяются конечные цели обезличивания данных, задачи, которые предполагается решать на обезличенной БД, контекст использования/публикации ОБД, участники информационного обмена (поставщик данных, получатель данных), объем ИБД, подлежащий обезличиванию.

На этапе формализации задачи определяются методы обезличивания, которые будут использоваться для обезличивания каждого из признаков ИБД, а также параметры методов. При выборе методов обезличивания, помимо типов признаков ИБД и типов значений признаков ИБД, учитываются также требования к свойствам обезличенных данных, временные (скоростные) ограничения на реализацию процедуры обезличивания.

В ряде случаев скорость и время, необходимое для выполнения операций обезличивания, могут стать серьезным ограничением при выборе методов обезличивания. Например, применение методов микроагрегирования к группе признаков связано со значительными затратами вычислительных ресурсов, и время обезличивания этими методами от объема (количества записей) ИБД зависит квадратично. Поэтому методы микроагрегирования ограниченно применяются при работе с большими данными.

Определяется набор (множество) временных (скоростных) показателей S для оценки быстроты реализации процедуры обезличивания данных, в частном случае, могут быть также заданы пороговые значения показателей S_{max}:

$$S = \{s_1, s_2, \dots, s_m\} = \{s_i\},\ i = \overline{1, v}, \quad (1.1)$$

где s_i – i-ый временной (скоростной) показатель; v – количество показателей.

$$S_{max} = \{s_{max1}, s_{max2}, \dots, s_{maxv}\} = \{s_{maxi}\},\ i = \overline{1, v}, \quad (1.2)$$

где s_{maxi} – пороговое значение i-ого временного (скоростного) показателя.

Задание пороговых значений имеет смысл в условиях обработки ИБД большого объема, либо при условии, что операции обезличивания ИБД будут выполняться постоянно или с некоторой периодичностью, например, при добавлении (обновлении) информации в ИБД и при жестких временных ограничениях. В любом случае, оценка показателей времени (скорости) обезличивания апостериорна и выполняется после завершения процедуры обезличивания.

Определяется набор (множество) показателей риска раскрытия информации R и их пороговые значения R_{max}, которые оценивают эффективность реализации процедуры обезличивания данных:

$$R = \{r_1, r_2, \dots, r_p\} = \{r_i\},\ i = \overline{1, p}, \quad (1.3)$$

где r_i – i-ый показатель риска раскрытия информации; p – количество показателей риска раскрытия информации.

$$R_{max} = \{r_{max1}, r_{max2}, \dots, r_{maxp}\} = \{r_{maxi}\},\ i = \overline{1, p}, \quad (1.4)$$

где r_{maxi} – пороговое значение i-го показателя риска раскрытия информации.

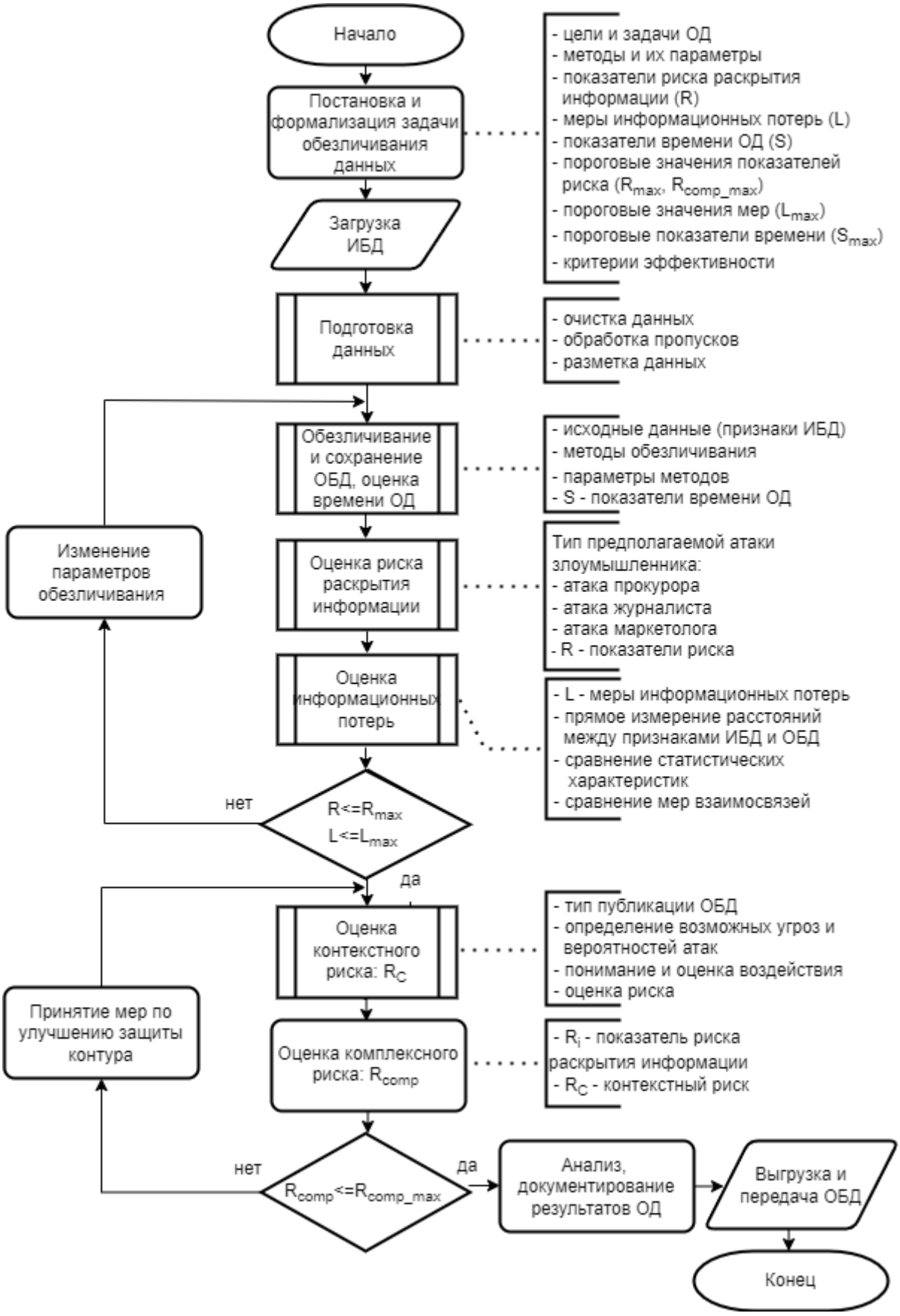


Рисунок 1.1 – Блок-схема укрупненного алгоритма автоматизированного обезличивания данных

Также определяется пороговое значение показателя комплексного риска R_{comp_max}, учитывающего не только риск раскрытия информации, но и контекстный риск – степень защиты контура обработки данных.

Определяется набор (множество) мер информационных потерь L и их пороговые значения L_{max}, которые оценивают эффективность обезличенной БД для ее дальнейшего использования и решения поставленных на ОБД задач:

$$L = \{l_1, l_2, \dots, l_m\} = \{l_i\},\ i = \overline{1,m}, \quad (1.5)$$

где l_i – i-ая мера информационных потерь; m – количество мер информационных потерь.

$$L_{max} = \{l_{max1}, l_{max2}, \dots, l_{maxm}\} = \{l_{maxi}\},\ i = \overline{1,m}, \quad (1.6)$$

где l_{maxi} – пороговое значение i-ой меры информационных потерь.

Формулируются критерии эффективного решения задачи обезличивания:

– достижение минимальных информационных потерь при выполнении ограничений на пороговые значения показателей риска раскрытия информации;

– достижение минимальных показателей риска раскрытия информации при выполнении ограничений на пороговые значения мер информационных потерь;

– выполнение временных (скоростных) ограничений на реализацию процедуры обезличивания данных;

– сохранение свойств обезличенных данных и т.п.

Определяются программные средства решения задачи обезличивания данных: программные среды, библиотеки, пакеты программ и т.п. В рамках проекта «Технология интеллектуального управления данными для платформы «Доверенная среда обмена информацией»», разрабатывается собственная программная система автоматизированного обезличивания данных. Программная система будет поддерживать основные этапы методики автоматизированного обезличивания данных, требующие программной реализации. Для разработки функционала и пользовательского интерфейса программной системы выбрана связка тех-

нологий реляционной СУБД *PostgreSQL* и языка программирования *Java*, позволяющая создавать и поддерживать сложные программные решения.

2. Загрузка исходных данных (ИБД)

На втором этапе загружаются исходные данные для обезличивания – исходная база данных (ИБД). ИБД представляет собой реляционную БД, основанную на реляционной табличной модели представления данных. Каждая строка, содержащаяся в таблице реляционной базы данных, представляет собой запись с уникальным идентификатором, который называют ключом. Столбцы таблицы соответствуют атрибутам (признакам) данных, а каждая запись содержит значение для каждого атрибута, что дает возможность легко устанавливать взаимосвязь между элементами данных. В контексте задачи обезличивания строки соответствуют субъектам персональных данных, а столбцы – атрибутам (признакам) персональных данных субъекта, которые могут принимать как количественные, так и качественные значения.

Выбор реляционной модели представления данных обусловлен преимуществами реляционного подхода в организации данных, а именно: обеспечение безопасного и надежного управления данными на основе правил целостности; обеспечение стандартного способа представления данных и отправки запросов; обеспечение табличного способа хранения структурированной информации и доступа к ней, который интуитивно понятен, гибок и эффективен; обеспечение высокой производительности выполнения операций чтения/записи, поиска данных.

В качестве СУБД выбрана *PostgreSQL*, учитывая ее основные достоинства и преимущества в сравнении с другими СУБД. К основным достоинствам *PostgreSQL* относятся: использование объектно-реляционных БД; высокая расширяемость; поддержка широкого набора типов данных; поддержка БД неограниченных размеров по количеству записей и возможность хранения таблиц размером в 32 Тбайта; многоверсионный контроль параллелизма; открытый и бесплатный исходный код.

Также на вход модуля (подсистемы) обезличивания данных могут подаваться данные в виде файла формата csv.

Математически ИБД представляется в виде матрицы:

$X = \{x_{ij}\}_{i,j=1}^{N,P}$, где x_{ij} – значение j-го атрибута (признака) i-го субъекта ПД в ИБД; N – количество строк, записей (субъектов персональных данных); P – количество признаков.

Обезличенная БД (ОБД) также представляется в виде матрицы:

$X_{\text{об.}} = \{x_{ij\text{об}}\}_{i,j=1}^{n,p}$, где x_{ij} – значение j-го атрибута (признака) i-го субъекта ПД в ОБД; n – количество строк, записей (субъектов персональных данных); p – количество признаков.

В общем случае не все признаки и/или записи ИБД включаются в ОБД в обезличенном виде, т.е. может быть $P \neq p$ и $N \neq n$.

3. Подготовка данных перед проведением процедуры обезличивания

На третьем этапе выполняются операции, связанные с подготовкой исходных данных перед проведением процедуры обезличивания: очистка данных; обработка пропусков; разметка данных.

Очистка данных предполагает обнаружение и удаление ошибок и несоответствий в данных в целях повышения их качества. В ходе очистки данных выявляются информационные дубли, ошибки регистрации, несоответствия в значениях атрибутов (признаков).

В ходе обработки пропусков выполняется либо разметка пропусков (пропуски сохраняются и обезличенная БД сдержит пропуски в тех же позициях, что и исходная БД), либо пропуски заполняются с использованием одного из статистических методов заполнения пропусков.

Операция разметки данных предполагает определение типов признаков и типов значений признаков, подлежащих обезличиванию. Признаки соответствуют одному из четырех типов: прямые идентификаторы; квазиидентификаторы (косвенные идентификаторы); чувствительные и нечувствительные признаки. Значения признаков измеряются либо в количественной, либо в качественной

шкалах, соответственно признаки относятся либо к количественному, либо к качественному типам. Качественные признаки подразделяются, в свою очередь, на номинальные (классификационные) и ранговые (ординальные). Принадлежность признака к одной из шкал измерений зависит от вида операций, которые допустимо выполнять со значениями признака (арифметические операции, операции сравнения). Отдельно выделяется тип признака – дата/время. Типы признаков и их значений являются основным критериями при выборе методов и процедур, которые будут использоваться на этапе обезличивания данных.

4. Обезличивание и сохранение обезличенной БД, оценка времени ОД

На четвертом этапе выполняется обезличивание ИБД и последующее сохранение обезличенных данных. Обезличивание реализуется с помощью выбранных на предыдущем этапе методов обезличивания данных. Выполняется также оценка временных (скоростных) показателей реализации процедуры обезличивания данных. В частном случае, если предполагается многократное выполнение процедуры обезличивания данных в условиях обработки больших данных и/или жестких временных ограничений, и скоростные показатели превышают заданные пороговые значения S_{max}, выполняется возврат на этап постановки задачи (этап 1) и выбираются более эффективные методы обезличивания с точки зрения временных затрат.

На блок-схеме (рисунок 1.1) частный случай не рассмотрен.

5. Оценка риска раскрытия информации

На пятом этапе выполняется оценка риска раскрытия информации путем расчета набора показателей риска раскрытия информации R, определенных на этапе постановки задачи (этап 1).

Формулы расчета показателей риска раскрытия информации зависят от предполагаемой модели атаки злоумышленника (атака прокурора, атака журналиста, атака маркетолога). Как правило, невозможно предположить какому именно типу атаки будет подвергаться обезличенная БД, поэтому имеет смысл рассчитать весь набор показателей риска раскрытия информации для обеспече-

ния комплексной оценки эффективности процедуры обезличивания данных с учетом разных внешних условий и угроз.

6. Оценка информационных потерь

На шестом этапе выполняется оценка информационных потерь, возникающих вследствие обезличивания данных. Выполняется расчет набора мер информационных потерь, определенных на этапе постановки задачи (этап 1). Используются разные типы мер информационных потерь для обеспечения комплексной оценки дальнейшей применимости ОБД для решения поставленных на ОБД задач:

– прямое измерение расстояний между значениями признаков в ИБД и ОБД;

– сравнение статистических характеристик признаков в ИБД и ОБД (расчет разностей статистических характеристик);

– сравнение мер взаимосвязей между признаками в ИБД и ОБД (расчет разностей мер взаимосвязей).

7. Сравнение расчетных и пороговых значений показателей риска раскрытия информации, мер информационных потерь

Выполняется сравнение расчетных значений показателей риска раскрытия информации с установленными пороговыми значениями показателей: $R \leq R_{max}$. Выполняется сравнение расчетных значений мер информационных потерь с установленными пороговыми значениями мер: $L \leq L_{max}$. Если неравенства выполняются, то переходим к следующему этапу, иначе изменяются параметры обезличивания и реализуется возврат на этап обезличивания (этап 4). Изменение параметров обезличивания сводится к выбору других параметров методов обезличивания или даже к выбору других методов ОД, либо к установлению менее жестких пороговых (предельных) значений для показателей риска раскрытия информации и/или мер информационных потерь.

8. Оценка контекстного риска

На восьмом этапе выполняется оценка контекстного риска, т.е. оценивается насколько защищен контур, в котором обрабатываются ИБД и ОБД. Оценка

выполняется с использованием методов экспертного оценивания, т.е. поставщик данных отвечает на ряд вопросов, касающихся организации среды обработки данных (внешней или внутренней), особенностей организации процесса обработки данных, угроз информационной безопасности, возникающих на разных уровнях (сетевые и технические ресурсы, аппаратное и программное обеспечение, стороны и лица, участвующие в обработке, масштабы обработки и т.п.). В зависимости от ответа на вопрос выставляется оценка в баллах, отражающая степень защиты контура, затем формируется комплексная оценка по совокупности ответов на вопросы, и на заключительном этапе рассчитывается оценка контекстного риска с учетом возможных угроз и их вероятностей – R_c.

Отметим, что совокупность вопросов, оценивающих защиту контура, должна определяться для действующей платформы «Доверенная среда обмена информацией» с учетом ее специфики (архитектура, аппаратное и программное обеспечение, организация процессов обработки, хранения и передачи данных и т.д.). Учитывая, что платформа находится в процессе разработки, на данный момент можно предложить только примерный набор общих вопросов, оценивающих контекстный риск.

9. Оценка комплексного риска

На девятом этапе выполняется расчет показателя комплексного риска R_{comp}, представляющего собой произведение риска раскрытия информации (данных) и контекстного риска:

$$R_{comp=}R_c \cdot R_i, \tag{1.7}$$

где R_i – показатель риска раскрытия информации; в качестве показателя риска раскрытия информации может использоваться один из показателей в наборе R.

10. Сравнение расчетного и порогового значений показателя комплексного риска

Выполняется сравнение расчетного значения комплексного риска R_{comp} с установленным пороговым значением $R_{comp} \leq R_{comp\_max}$. Если неравенство вы-

полняется, то переходим к заключительным этапам (этап 11, 12), иначе – принимаются технические и организационные меры, направленные на улучшение защиты контура и снижение контекстного риска при обработке информации. Далее происходит возврат на этап оценки контекстного риска (этап 8) с учетом изменившихся условий.

11. Анализ, документирование результатов ОД

Проводится анализ полученных результатов и их интерпретация, описываются и документируются ход и результаты выполнения процедуры автоматизированного обезличивания данных. В описании результатов обезличивания данных представляется следующая информация:

– признаки (атрибуты) субъектов ПД из ИБД, значения которых обезличивались;

– методы обезличивания данных, показатели риска раскрытия информации, меры информационных потерь, которые использовались в процессе обработки данных;

– пороговые (предельные) значения показателей риска раскрытия информации и комплексного риска, мер информационных потерь, которые были заданы;

– расчетные значения показателей риска раскрытия информации для ИБД и ОБД;

– расчетные значения показателей контекстного и комплексного рисков;

– расчетные значения мер информационных потерь на ОБД;

– расчетные значения временных показателей обезличивания данных;

– критерии оценки эффективности процедуры обезличивания данных и выбора оптимального варианта обезличивания;

– выводы о соответствии расчетных показателей риска раскрытия информации и комплексного риска установленным пороговым значениям, о соответствии расчетных мер информационных потерь установленным пороговым зна-

чениям, о степени обезличивания данных и степени защищенности конфиденциальной информации о субъектах ПД, хранящейся в ОБД.

12. Выгрузка и передача ОБД

На последнем этапе реализуется выгрузка и передача ОБД получателям для дальнейшего использования (первичное или вторичное внутреннее исследование, внешнее исследование) или открытая публикация (публичный доступ к данным).

1.3 Выводы по главе 1

В главе 1 приведена общая характеристика и рассмотрены особенности разрабатываемой методики автоматизированного обезличивания данных. Предложен и описан укрупненный алгоритм, реализующий методику автоматизированного обезличивания данных. Алгоритм представляет собой последовательность взаимосвязанных этапов (рисунок 1.1), которые итерационно взаимодействуют друг с другом: на любом шаге возможен возврат на предыдущие этапы для их корректировки с учетом анализа промежуточных результатов. Также возможно многократное повторение этапов алгоритма для выполнения установленных требований к процедуре автоматизированного обезличивания данных и достижения конечных целевых показателей.

Дано краткое описание каждого из этапов алгоритма, которые далее детально представлены во второй главе отчета.

2 Основные этапы методики автоматизированного обезличивания данных

2.1 Этап постановки задачи обезличивания данных

На этапе постановки задачи должны быть четко определены и сформулированы ключевые положения (пункты), связанные с процессом обезличивания данных, а именно:

– цели и задачи обезличивания данных;

– исходный набор или исходная БД, подлежащая обезличиванию;

– подходы, методы, модели, алгоритмы, применяемые на каждом из этапов обезличивания информации;

– условия, ограничения и предполагаемый вид публикации обезличенной БД;

– программное обеспечение и т.д.

В таблице 2.1 представлены основные пункты, которые должны быть описаны на этапе постановки задачи обезличивания ПД, и краткие к ним пояснения. Представленный набор пунктов можно рассматривать как структуру паспорта постановки задачи обезличивания ПД. На рисунке 2.1 представлен паспорт постановки задачи обезличивания данных в виде структурированного графа.

Далее рассмотрены и описаны основные пункты постановки задачи.

1. Цели и задачи обезличивания данных

Основная цель обезличивания персональных данных – обеспечение конфиденциальности информации о субъекте (субъектах) персональных данных и гарантирование безопасности любых операций, связанных с ПД (сбор, хранение, обработка, передача) в отношении угроз и атак злоумышленников, направленных на раскрытие личных данных.

Рисунок 2.1 – Структура паспорта постановки задачи обезличивания данных

Таблица 2.1 – Пункты постановки задачи обезличивания ПД

№ п/п	Пункт	Пояснение
1.	***Цели и задачи обезличивания данных***	
1.1.	Цель обезличивания ПД	Обеспечение конфиденциальности информации о субъекте (субъектах) ПД и гарантирование безопасности выполнения задач обработки ПД: – **статистическая обработка и статистические исследования ПД;** – сбор и хранение ПД; – обработка поисковых запросов (поиск данных о субъектах и поиск субъектов по известным данным); – актуализация ПД; – интеграция данных различных Операторов; – ведение учета субъектов ПД.
1.2.	Задачи обезличивания ПД	– выделение и описание признаков (атрибутов) субъектов ПД в ИБД, содержащих конфиденциальную информацию; – выбор методов ОД и программно-инструментальных средств ОД; – предварительная обработка ИБД; – обезличивание ИБД на основе выбранных методов и средств, формирование ОБД; – оценка эффективности обезличивания данных (оценка рисков повторной идентификации ОБД, оценка информационных потерь); – документирование основных этапов обезличивания данных и полученных результатов, описание ОБД.
2.	***Исходные (входные) данные***	
2.1.	Обладатель данных	– полное наименование; – общие сведения; – контактные данные; ответственное лицо.

Продолжение таблицы 2.1

2.2.	Содержательное описание данных	– предметная область; – словесное (смысловое) описание; – цель сбора данных; – структура данных; – название и смысловое содержание признаков (атрибутов).
2.3.	Ограничение доступа к данным	– наличие признаков, содержащих конфиденциальную (персональную) информацию; – описание признаков.
2.4.	Дата создания/последнего изменения данных	
2.5.	Периодичность обновления, срок актуализации и хранения данных	– оперативные (ежедневно, еженедельно); – долгосрочные (ежемесячно; ежеквартально; ежегодно; по мере поступления и т.п.); – срок актуализации и хранения.
2.6.	Формат представления данных	– табличный (*csv, xls, xlsx, sql, json, xml*); – текстовый (*doc, docx, pdf*); – графический (*jpeg, gif, tiff, png*); – иной.
3.	***Получатель (потребитель) данных, информационное соглашение о передаче данных, требования к данным со стороны получателя***	
3.1.	Получатель (потребитель) данных	– полное наименование; – общие сведения; – контактные данные, – ответственное лицо.
3.2.	Задачи обработки, которые планируется решать на ОБД	для случая статистической обработки данных: – первичный анализ данных; – исследование состава и структуры данных; – исследование и оценка взаимосвязей признаков; – кластеризация субъектов данных; – классификация субъектов данных; – прогнозирование признаков.

Продолжение таблицы 2.1

3.3.	Информационное (пользовательское) соглашение о передаче данных	– стороны обмена; – права и обязанности сторон; – требования к составу, структуре, формату представления передаваемых данных; периодичности (сроков) предоставления данных; – режим использования данных получателем; – технические условия для обеспечения информационного взаимодействия; – средства контроля и защиты для обеспечения безопасности и конфиденциальности данных.
3.4.	Объем и структура передаваемых данных	– объем выборки из исходной БД; – набор атрибутов (признаков) субъектов ПД из ИБД.
3.5.	Контекст публикации (использования) ОБД	– открытая публикация; – закрытая публикация (первичное или вторичное внутреннее исследование; внешнее исследование); – полуоткрытая публикация.
4.	***Операции подготовки данных перед проведением процедуры обезличивания***	
4.1.	Очистка данных	– регламент процедуры очистки данных; – методы и способы очистки данных.
4.2.	Заполнение пропущенных значений	– подход к обработке пропущенных значений; – методы обработки пропущенных значений.
4.3	Разметка данных	классификация признаков (атрибутов) ИБД: – прямые идентификаторы; – косвенные идентификаторы; – чувствительные признаки; – нечувствительные признаки. определение типов значений признаков ИБД: – количественные (числовые); - ранговые (ординальные); - качественные (номинальные, классификационные); дата/время.

Продолжение таблицы 2.1

5.	***Методы обезличивания данных***	
5.1.	Метод (методы) обезличивания для каждого признака ИБД	– введение идентификаторов; – перемешивание (перестановка); – декомпозиция; – изменение состава или семантики: – глобальное и локальное обобщение; – кодирование сверху и/ил снизу; – локальное подавление; – микроагрегирование; – добавление шума; – округление; – маскирование по шаблону; – случайная выборка; – удаление.
5.2.	Параметры метода ОД	зависят от метода обезличивания
6.	***Оценка риска раскрытия информации***	
6.1.	Типы атаки злоумышленника на уровне данных	– атака прокурора; – атака журналиста; – атака маркетолога.
6.2.	Показатели риска раскрытия информации с учетом возможных типов атак (R)	– вероятность повторной идентификации i-ой записи; – доля записей, вероятность повторной идентификации которых выше заданного порога; – максимальный риск раскрытия информации; – средний риск раскрытия информации (глобальный риск).
6.3.	Пороговые значения показателей риска раскрытия информации (R_{max})	$R_{max} = \{\alpha, \tau, \gamma\}$; – α – порог максимальной доли записей с высоким риском повторной идентификации; – τ – порог максимального риска раскрытия i-ой записи; – γ – порог среднего риска раскрытия информации.

Продолжение таблицы 2.1

7.	***Оценка контекстного и комплексного рисков повторной идентификации ОБД***	
7.1.	Типы атаки злоумышленника на уровне контура обработки данных	– внутренняя атака; – случайная атака; – внешняя атака.
7.2.	Уровни воздействия на субъекта ПД	– потеря конфиденциальности; – потеря целостности; – потеря доступности.
7.2.	Показатели контекстного и комплексного рисков (R_c, R_{comp})	– R_c – показатель контекстного риска; – R_{comp} – показатель комплексного риска.
7.3.	Пороговое значение комплексного риска повторной идентификации ОБД (R_{comp_max})	– R_{comp_max} – пороговое значение показателя комплексного риска.
8.	***Оценка информационных потерь***	
8.1.	Меры информационных потерь (L)	– меры, основанные на прямом измерении расстояний/частот между исходными и обезличенными данными; – меры, основанные на сравнении статистических характеристик и характеристик взаимосвязей, рассчитанных по исходным и обезличенным данным.
8.2.	Пороговые значения информационных потерь (L_{max})	– L_{max} – пороговые значения мер информационных потерь.
9.	***Оценка скоростных/временных показателей обезличивания данных***	
9.1.	Показатели скорости/времени обезличивания (S)	– показатель скорости обезличивания записей БД; – показатель времени обезличивания БД.
9.2.	Пороговые значения показателей скорости/времени обезличивания (S_{max})	– S_{max} – пороговое значение показателя скорости/времени обезличивания.

Продолжение таблицы 2.1

10.	***Оптимизация процедуры обезличивания данных***	
10.1.	Показатели оптимизации	– показатели риска; – меры информационных потерь; – скоростные/временные показатели.
10.1.	Критерии оптимизации	– минимизация информационных потерь при выполнении ограничений на пороговые значения показателей риска раскрытия информации; – минимизация риска раскрытия информации при выполнении ограничений на пороговые значения показателей информационных потерь; – минимизация информационных потерь и риска раскрытия информации; введение вектора предпочтений; – минимизация скорости/времени обезличивания при соблюдении ограничений на пороговые значения риска раскрытия информации.
11.	***Программные средства обезличивания данных (программные решения, среды, пакеты программ, библиотеки).***	

В [1] представлены типовые классы задач, состоящие из наиболее часто встречающихся задач обработки персональных данных в государственных и муниципальных органах. Согласно методическим рекомендациям к задачам обработки персональных данных относят: статистическую обработку и статистические исследования ПД, сбор и хранение ПД, обработку поисковых запросов (поиск данных о субъектах и поиск субъектов по известным данным), актуализацию ПД, интеграцию данных различных Операторов, ведение учета субъектов ПД [1]. Обезличивание ПД направлено на безопасное выполнение задач обработки ПД с точки зрения обеспечения информационной безопасности и защиты субъекта ПД от раскрытия его личности и/или персональной информации о нем.

Процесс обезличивания включает решение ряда задач, связных как непосредственно с реализацией обезличивания, так и с оценкой его эффективности. Можно выделить следующие основные задачи:

– выделение и описание признаков (атрибутов) субъектов ПД в ИБД, содержащих конфиденциальную информацию;

– выбор методов ОД и программно-инструментальных средств ОД;

– предварительная обработка ИБД;

– обезличивание ИБД на основе выбранных методов и средств, формирование ОБД;

– оценка эффективности обезличивания данных (оценка рисков повторной идентификации ОБД, оценка информационных потерь);

– документирование основных этапов обезличивания данных и полученных результатов, описание ОБД.

2. Исходные (входные) данные

Одним из пунктов постановки задачи является описание исходного набора или исходной БД (ИБД). При описании набора данных указываются следующие ключевые позиции:

– обладатель (хранитель) ИБД (полное наименование, общие сведения, контактные данные, ответственное лицо);

– содержательное описание ИБД (предметная область, словесное смысловое описание; цель сбора данных, структура данных, название и смысловое содержание признаков);

– ограничение доступа (наличие признаков, содержащих конфиденциальную информацию о субъекте персональных данных, описание этих признаков);

– дата создания/последнего изменения ИБД;

– периодичность обновления, срок актуализации и хранения ИБД;

– формат представления данных (текстовый, графический, табличный, иной).

Обладатель (хранитель) данных – лицо, которому принадлежат данные. Обладатель данных несет ответственность за безопасность обработки и хранения данных, обеспечивает контроль процессов передачи и обмена данных путем их обезличивания, а также отвечает за внедрение других средств контроля, которые предотвращают неправомерное (незаконное) использование данных и/или их повторную идентификацию.

Обладателями ИБД могут быть государственные органы и структуры, коммерческие организации, частные компании и физические лица. Согласно Федеральному закону от 27 июля 2006 г. №149-ФЗ (ред. от. 12.12.2023) «Об информации, информационных технологиях и о защите информации» обладателем информации может быть гражданин (физическое лицо), юридическое лицо, Российская Федерация, субъект Российской Федерации, муниципальное образование.

Содержательное описание ИБД включает описание предметной области, к которой принадлежат данные (медицина, демография, экономика, экология и т.п.), что представляют собой данные (например, результаты исследований в определенной сфере), для чего собирались данные, структура данных (набор признаков), название и смысловое содержание признаков ИБД. Выделяются и описываются признаки ИБД, которые содержат персональные данные субъектов. Именно эти признаки ограничивают доступ к данным и подлежат обезличиванию.

Данные могут обновляться с некоторой периодичностью (частотой), обусловленной требованиями пользователей и зависящей от сути данных. Для оперативных данных периодичность обновления может быть ежедневная или еженедельная. Для долговременных данных может быть установлена частота обновлений: ежемесячно; ежеквартально; ежегодно; по мере поступления и т.п. Также может быть установлен срок актуализации и хранения данных, хотя и рекомендуют бессрочное хранение всех версий наборов данных для полного обеспечения информационных потребностей пользователей.

Исходные данные могут быть представлены в разных форматах, в том числе не являющимися машиночитаемыми (например, форматы презентаций – *ppt, pptx*; форматы текстовых документов – *doc, docx, pdf*; форматы изображений – *jpeg, gif, tiff, png* и т.п.). Для последующей обработки данные должны быть преобразованы в один из машиночитаемых форматов (*csv, xls, xlsx, sql, json, xml*) и представлены в структурированном табличном виде. Предполагается, что на вход модуля «Обезличивания данных» подается структурированная табличная информация в формате БД *PostgreSQL* или в виде *csv* файла.

В [13] предложено ввести паспорт набора данных, который содержит классификационные признаки, позволяющие отнести данные к определенным классам. Паспорт описывает ключевые параметры данных в унифицированном виде, позволяет однозначно идентифицировать набор данных при его публикации в открытых источниках. При разработке структуры паспорта набора данных учтены требования международных и национальных стандартов: *RDF, RDFa (Resource Description Framework in attributes)* – модель представления данных (метаданнных), пригодных для машинной обработки, разработанная международным консорциумом сети Интернет *(World Wide Web Consortium – W3C); W3C Recommendation: RDFa Core 1.1* – рекомендации по использованию *RDFa*; ГОСТ Р 7.0.10-2019 – российский стандарт «Набор элементов метаданных «Дублинское ядро»». В [14,15] описан протокол обработки (анонимизации) наборов данных для их публикации в открытых источниках.

3. Получатель (потребитель) данных, информационное соглашение о передаче данных, требования к данным со стороны получателя

Наборы данных могут обрабатываться как их обладателем, так и передаваться для использования внешнему получателю (потребителю), в том числе возможна открытая публикация данных на общедоступных Интернет ресурсах. В случае внешнего потребителя поступает запрос на предоставление набора данных, в котором указывается: для чего нужны данные и как будут использоваться (какие задачи предполагается решать на ОБД); требования к данным со стороны получателя (необходимый объем данных; допустимые информационные потери

и степень обезличивания). Между обладателем и получателем заключается информационное (пользовательское) соглашение, в котором оговариваются права и обязанности сторон. В частности, режим использования данных пользователем, например: возможность загрузки; использование только в научных и образовательных целях; предоставление результатов обработки данных третьим лицам или их открытая публикация; создание программных продуктов с использованием полученных данных и т.п. Также описываются меры и средства контроля безопасности и конфиденциальности данных, защиты от раскрытия персональной информации, которые должны обеспечить стороны информационного соглашения.

Потребитель (получатель) данных – физическое лицо или организация (компания), которая получает данные от обладателя данных в соответствии с условиями соглашения о предоставлении данных. В частном случае информационный обмен может осуществляться внутри компании или организации, владеющей данными, а не только с внешними потребителями.

Важно определить и содержательно описать задачи, которые в дальнейшем предполагается решать на обезличенной БД. От этого зависит выбор признаков (атрибутов) ИБД, подлежащих обезличиванию, необходимый объем обезличенных данных, выбор методов обезличивания и их параметров, а также пороговых значений рисков раскрытия информации и информационных потерь. Наиболее часто обезличивание выполняется для последующего решения на ОБД исследовательских задач, связанных со статистической обработкой и анализом данных (осуществление выборки по заявленным параметрам и проведение исследований по заданным параметрам субъектов). Отметим, что этот класс задач наиболее интересен с позиции широты применяемых методов обезличивания, так как только для этого класса задач рекомендованы к использованию методы изменения состава и/или семантики, наравне с методами других классов (декомпозиция, перемешивание). В текущем исследовании акцент сделан на решение задач статистической обработки и анализа обезличенной БД. Задачи статистического исследования данных условно классифицируются на следующие основные типы:

– первичный (предварительный) анализ данных;

– изучение структуры и взаимосвязей между показателями субъекта ПД;

– кластеризация субъектов ПД;

– классификация субъектов ПД;

– прогнозирование показателей субъекта ПД.

В рамках одного статистического исследования возможно решение задач разных классов. Например, первичный анализ данных выполняется, как правило, в начале любого исследования, и включает расчет числовых характеристик признаков, графический анализ, исследование законов распределения признаков. Взаимосвязи между показателями оцениваются для определения предикторов, которые оказывают статистически значимое влияние на результирующий признак и которые следует учитывать при решении задач классификации, регрессии. После решения задачи кластеризации, как правила, решается задача классификации, заключающаяся в отнесении объекта к одному из выделенных классов (кластеров).

Определение объема и структуры передаваемых данных (объем выборки из исходной БД, набор признаков из исходной совокупности признаков) – важное решение, которое зависит от задач, которые планируется решать на ОБД. Если подробно рассмотрены и учтены возможности будущего использования обезличенных данных, сужена область их применения, то может потребоваться меньший объем обезличивания ПД при соблюдении заданного уровня требований к защите информации.

Еще один важный вопрос, который требует рассмотрения – контекст публикации (выпуска) или использования данных, всего возможно три контекста: открытая публикация; закрытая публикация; полуоткрытая публикация. Каждый из них представляет разные отношения и, следовательно, уровень доверия между поставщиком данных и получателем данных. В свою очередь, они будут представлять разный уровень контроля и риска раскрытия конфиденциальной информации. Рассмотрим подробнее возможные контексты публикации (выпуска) данных.

Открытая публикация – публикация данных в открытом доступе в сети Интернет, данные доступны для скачивания и использования без каких-либо условий и ограничений. В случае открытого доступа данные становятся общедоступными, нет никакого контроля над тем, как они будут использоваться. Поэтому требуются самые жесткие предположения об угрозах конфиденциальности, поскольку невозможно оценить мотивы злоумышленника и уровень знаний и инструментов, которыми он может обладать, и которые будет использовать.

Закрытая публикация – публикация данных в ограниченном доступе для определенной категории лиц/организаций. В случае закрытой публикации возможны два варианта: либо проводится внутреннее первичное или вторичное исследование, либо данные передаются для проведения внешнего исследования.

Внутреннее первичное или вторичное исследование (повторное использование данных). Предполагается, что исследование данных проводится внутри организации, владеющей данными. Например, медицинские учреждения и организации ведут истории болезней пациентов, в которых содержится персональная информация, спонсоры клинических испытаний хранят и поддерживают огромное количество данных, собранных в ходе клинических исследований. Использование данных в исследовательских целях требует их обезличивания для сохранения конфиденциальной информации. В ходе клинических испытаний данные собираются для достижения целей отдельных исследований, но совокупная информация часто может оказаться бесценной при выявлении закономерностей, которые не были в центре внимания первоначального анализа. Чтобы иметь возможность использовать данные для целей, не указанных в первоначальном протоколе, спонсоры должны получить согласие на такое использование данных пациентов. Это не всегда может быть возможно или эффективно, альтернативой является обезличивание данных. Любой тип вторичного исследования, где данные используются для целей, отличных от тех, которые указаны в первоначальном протоколе и не охватываются информированным согласием, может потребовать определенного уровня обезличивания.

Внешнее исследование предполагает передачу данных внешним исследователям на договорной основе с соблюдением заданных ограничений и требований к безопасности процесса передачи и использования ПД. В качестве примеров ограничений можно привести: запрет на попытки повторной идентификации; запрет на попытки связаться с любым из субъектов в наборе данных; требование аудита, позволяющее проводить выборочные проверки для обеспечения соответствия соглашению, или требование о регулярных проверках третьей стороной и т.п. Обмен данными с известными исследователями, в соответствии со строгими договорами, с помощью безопасных средств, гарантирует, что процесс безопасен, а связанные с ним риски очень низки.

Полуоткрытая публикация – публикация данных, сочетающая варианты как открытого, так и закрытого доступа к данным. Набор данных доступен любому пользователю для открытого скачивания, однако условием получения данных является необходимость регистрации в организации, предоставляющей данные, и подтверждение согласия на условия использования, обработки и обмена данными (соглашение об условиях использования). В этом случае дополнительные меры защиты и конфиденциальности данных предусмотрены соглашением об использовании данных, но их трудно обеспечить в силу предоставления открытого доступа к данным.

4. Операции подготовки данных перед проведением процедуры обезличивания

Определяются операции подготовки данных, которые требуется выполнить перед проведением процедуры обезличивания данных, а также способы и методы их реализации. К основным операциям предобработки данных относятся: очистка; заполнение пропущенных значений; разметка. В п. 2.2 приведен алгоритм подготовки данных перед проведением процедуры обезличивания в виде последовательности взаимосвязанных шагов, описаны подходы (методы), которые применяются на каждом шаге.

5. Методы обезличивания данных

В этом пункте для каждого признака ИБД, подлежащего обезличиванию, указывается метод (методы) обезличивания и его (их) параметры.

Для обезличивания ПД могут быть выбраны разные методы обезличивания и их параметры, определяющие степень (уровень) анонимизации данных. Для каждого признака исходной БД используется определенный метод обезличивания или их совокупность в зависимости от типа признака и типа значений признака. Таким образом, для исходной БД можно определить множество моделей (вариантов) преобразования (обезличивания) данных, каждая из которых соответствует определенной комбинации методов ОД с заданными параметрами, примененных для признаков, описывающих субъекта ПД. В п. 2.3 описаны алгоритмы реализации разных классов методов обезличивания, описаны условия и ограничения использования каждого метода.

6. Оценка риска раскрытия информации

Указываются показатели риска раскрытия информации R, которые будут использоваться для оценки эффективности процедуры обезличивания данных. Выбор расчетных формул для оценки показателей риска зависит от предполагаемого типа атаки злоумышленника на уровне данных (атака прокурора, атака журналиста, атака маркетолога). Учитывая, что тип атаки априорно неизвестен, можно предложить полный расчет всех показателей риска в условиях разных типов атак для всесторонней оценки риска. Также должны быть установлены пороговые значения показателей риска раскрытия информации: $R_{max} = \{\alpha, \tau, \gamma\}$. Пороговые значения показателей определяются в зависимости от ряда факторов, в том числе от специфики предметной области, к которой принадлежат данные, типа публикации (выпуска) ОБД, условий информационного соглашения об обмене данными. Алгоритм оценки риска раскрытия информации, показатели риска раскрытия информации приведены в п. 2.4.

7. Оценка контекстного и комплексного рисков повторной идентификации ОБД

В данном пункте указываются подходы и методы оценки контекстного и комплексного рисков (R_c, R_{comp}) повторной идентификации ОБД. Расчет показа-

телей риска зависит от ряда факторов, а именно: предполагаемая атака злоумышленника (внутренняя, случайная, внешняя), уровень воздействия, которое оказывает на субъекта персональных данных раскрытие личной информации о нем в контексте потери конфиденциальности, целостности, доступности; вид публикации (выпуска) ОБД. Также устанавливается пороговое значение комплексного риска повторной идентификации ОБД R_{comp_max} в зависимости от возможных угроз информационной безопасности и требований к уровню защиты ПД. Алгоритм оценки контекстного и комплексного рисков повторной идентификации ОБД, показатели риска описаны в п. 2.6.

8. Оценка информационных потерь

Указываются меры информационных потерь L, которые будут рассчитываться на ОБД и их пороговые значения L_{max}. Пороговые значения мер информационных потерь устанавливаются в соответствии с целями и задачами дальнейшей статистической обработки и анализа ОБД. Оценка информационных потерь обсуждается в п.2.5.

9. Оценка временных (скоростных) показателей обезличивания данных

Указываются показатели времени (скорости) обезличивания S, которые будут использоваться и их пороговые значения S_{max}. Данный пункт важен в условиях обработки больших данных и многократном (постоянном) повторении процедуры обезличивания данных с заданной периодичностью, например, по мере поступления новых данных и при жестких временных ограничениях. Основные временные показатели обезличивания данных приведены в п. 2.3.7.

10. Оптимизация процедуры обезличивания данных

Указывается критерий оптимизации, который позволяет выбрать один из вариантов (моделей) обезличивания данных. Для статистической оценки эффективности и выбора варианта обезличивания данных можно использовать набор показателей риска раскрытия информации, мер информационных потерь, временных (скоростных) показателей с установленными пороговыми значениями. Фактически решается оптимизационная задача выбора модели преобразования

данных в соответствии с установленными критериями при заданных ограничениях в одной из постановок.

Первый вариант постановки оптимизационной задачи: минимизация информационных потерь при выполнении ограничений на пороговые значения показателей риска раскрытия информации. В качестве критерия оптимизации выступает показатель (показатели) информационных потерь, целевая функция минимизирует значение (значения) показателя, в качестве ограничений используется показатель (показатели) риска раскрытия информации.

Второй вариант постановки оптимизационной задачи: минимизация риска раскрытия информации при выполнении ограничений на пороговые значения показателей информационных потерь. В качестве критерия оптимизации выступает показатель (показатели) риска раскрытия информации, целевая функция минимизирует значение (значения) показателя, в качестве ограничений используется показатель (показатели) информационных потерь.

Третий вариант постановки оптимизационной задачи: минимизация информационных потерь и риска раскрытия информации; введение вектора предпочтений.

Четвертый вариант постановки оптимизационной задачи: минимизация времени (скорости) обезличивания при соблюдении ограничений на пороговые значения риска раскрытия информации. Этот вариант актуален только в условиях обработки больших данных и многократного решения задачи обезличивания данных, например, по мере поступления новой информации.

Важно подчеркнуть, что конечной целью обезличивания является снижение риска повторной идентификации до приемлемого уровня при сохранении максимально возможной полезности данных.

11. Программные средства обезличивания данных (программные решения, среды, пакеты программ, библиотеки)

Выбор программного обеспечения зависит от того, какие методы и модели обезличивания данных, какие показатели их эффективности планируется использовать. В идеале программное обеспечение должно поддерживать различ-

ные методы и алгоритмы преобразования данных, статистической оценки показателей риска раскрытия информации и информационных потерь.

Разработано достаточно много зарубежных открытых программных систем обезличивания данных, среди которых можно особо выделить: *ARX - Data Anonymization Tool* [16], *sdcMicro* [17], *μ-ARGUS* [18]. Однако эти программные решения реализуют спектр подходов, методов и моделей, которые рекомендованы к применению органами по защите конфиденциальности персональных данных в зарубежных странах, что не всегда соответствует требованиям российского законодательства в сфере защиты информации.

Поэтому разрабатывается собственное программное обеспечение обезличивания данных в рамках системы «Интеллектуальное управление, обогащение и обезличивание данных» для платформы «Доверенная среда обмена информацией», реализующее широкий набор методов обезличивания данных, оценки рисков раскрытия информации и информационных потерь с учетом требований Роскомнадзора в области защиты данных и специфики российского законодательства. Основные требования к программному обеспечению:

– полная функциональность (реализация методов обезличивания данных разных классов; широкого набора показателей риска раскрытия информации; широкого набора мер информационных потерь);

– удобный русифицированный пользовательский интерфейс;

– высокая надежность: обработка ошибок ввода, некорректного задания входных параметров методов и т.п.;

– легкая расширяемость: возможность добавления новых методов, показателей, мер.

2.2 Этап подготовки данных перед проведением процедуры обезличивания

Данные должны быть определенным образом подготовлены перед началом проведения процедуры обезличивания для обеспечения эффективного решения задач обезличивания и возможности дальнейшего использования обезличенной

БД. Этап подготовки данных перед проведением процедуры обезличивания включает три основных шага (подэтапа), которые будут далее рассмотрены и описаны:

1 шаг. Очистка данных.

2 шаг. Обработка пропущенных значений.

3 шаг. Разметка данных.

Шаги очистки данных и обработки пропущенных значений реализуются в рамках модулей «Интеллектуальное управление данными», «Обогащение данных» системы «Интеллектуальное управление, обогащение и обезличивание данных». На вход модуля «Обезличивание данных» поступает таблица с данными или база данных, содержащая очищенные данные надлежащего качества с обработанными пропущенными значениями. В модуле «Обезличивание данных» реализуются функции разметки данных на основе постановки задачи и содержательного анализа природы признаков, описывающих субъекта персональных данных.

2.2.1 Очистка данных

Под очисткой данных (*data cleaning*) понимают процесс, направленный на выявление и удаление ошибок и несоответствий в данных с целью улучшения их качества [19]. Очистка данных выполняется с помощью специальных приемов и алгоритмов. Необходимость в очистке данных возникает вследствие того, что данные собираются и интегрируются (обогащаются) из множества разных источников и содержат ошибки регистрации, информационные дубли, пустые значения и другие артефакты, которые помешают в дальнейшем решению задач обезличивания.

Ошибки в таблицах БД могут возникать как на уровне записей, так и на уровне конкретных значений признаков. Основные проблемы с записями:

– наличие дублирования;

– неуникальные значения идентификаторов (например, двум разным записям соответствует одно значение уникального идентификатора);

– противоречивые записи (например, для одного объекта приведено несколько записей с разными данными).

Основные проблемы на уровне значений признаков следующие:

– недопустимые значения;

– наличие опечаток или орфографических ошибок;

– аномальные значения;

– многозначность (используются синонимы в описании значений признаков);

– ошибки в типах, форматах, кодировках и т.п.

Выбор метода очистки зависит от конкретной ситуации и типа ошибок, при этом выделяют три базовых способа очистки:

– полностью автоматизированный с помощью встроенных в инструменты разработки средств;

– средствами программного обеспечения (скриптов), написанных аналитиком

– ручной способ.

Учитывая уникальность каждой из БД, а, следовательно, и возникающих ошибок в данных, используют комбинацию способов для достижения наилучших результатов.

Алгоритм очистки данных разрабатывается для конкретной БД с учетом ее структуры, содержания и специфики, поэтому здесь можно предложить только обобщенный алгоритм, определяющий основные шаги (этапы) очистки данных.

Ниже приведен обобщенный алгоритм очистки данных.

1. Загрузка ИБД.

2. Определение регламента процедуры очистки данных. На этом этапе определяется последовательность процесса очистки данных, используемые методы, способы и программный инструментарий; устанавливается периодичность выполнения процедуры очистки данных (если речь идет о хранении и постоянном обновлении/дополнении исходной БД).

3. Очистка данных на уровне записей.

3.1. Идентификация и удаление дубликатов (поиск и удаление повторяющихся записей).

3.2. Идентификация и удаление неуникальных значений идентификаторов.

3.3. Идентификация и удаление противоречивых записей.

4. Очистка данных на уровне значений признаков.

4.1. Коррекция ошибок и пропусков (исправление ошибочных данных, устранение многозначности, заполнение отсутствующей информации из дополнительных информационных источников, если это возможно). Операции по коррекции ошибок выполняются с помощью встроенных средств проверки орфографии; поиска и замены текста; удаления пробелов и непечатных знаков из текста.

4.2. Стандартизация форматов (обеспечение единого формата представления данных для их согласованности и упрощения дальнейшего анализа). Операции стандартизации форматов выполняются с помощью встроенных средств преобразования форматов и переформатирования данных.

4.3. Валидация и верификация данных (подтверждение достоверности данных путем их проверки на соответствие определенным критериям и правилам).

5. Обслуживание БД посредством применения встроенных команд и функций очистки и автоочистки данных.

6. Сохранение ИБД после выполнения процедур очистки.

Базы данных *PostgreSQL* требуют периодического проведения процедуры обслуживания, которая называется очисткой, также в *PostgreSQL* реализованы встроенные средства автоочистки [20]. В процессе очистки для каждой из таблиц БД выполняется высвобождение дискового пространства, занятого удаленными или измененными строками, и обеспечивается возможность его повторного использования [20].

2.2.2 Обработка пропущенных значений

Данные в ИБД могут быть неполными и содержать пропущенные значения. На этапе очистки данных решается задача коррекции пропусков (раздел 2.2.1, п. 4.1 алгоритма), если пропуски связаны с ошибками ввода и могут быть заполнены на основе имеющейся информации или информации из дополнительных внешних источников. Если же коррекция пропусков невозможна, необходимо решить вопрос об обработке пропущенных значений до проведения процедуры обезличивания данных.

Общий алгоритм работы с пропущенными данными состоит из следующих основных этапов (см. рисунок 2.2):

1. Загрузка ИБД.

2. Идентификация пропущенных значений и их описание. Описание включает: структуру пропусков (доля пропущенных значений, где расположены пропущенные значения, насколько широкую область они охватывают); выявление причин наличия пропусков (распределение пропущенных значений по записям и признакам БД, можно ли считать пропуски случайными или пропущены данные в строго определенных записях).

3. Выбор подхода к обработке пропущенных значений (P): 1 – сохранение; 2 – заполнение/удаление.

4. Если выбран подход P=1, то

– сохранение пропусков, обработка пропусков не выполняется перед проведением процедуры обезличивания;

иначе (выбран подход P=2):

– выбор метода (способа) обработки пропущенных значений (S): 1 – удаление записей с пропущенными значениями; 2 – применение методов взвешивания; 3 – применение статистических методов заполнения пропусков.

5. Если выбран метод (способ) S=1, то

– удаление записей или признаков с пропущенными значениями;

иначе: если выбран метод (способ) S=2, то

– выбор метода взвешивания;

– заполнение пропусков с помощью выбранного метода;

иначе (выбран метод (способ) S=3):

– выбор метода заполнения пропусков;

– заполнение пропусков с помощью выбранного метода.

6. Сохранение ИБД.

Для идентификации пропусков и описания их структуры используются разные графические и статистические методы. К графическим методам относятся: тепловая карта; совмещенная диаграмма рассеяния и диаграмма размаха; график структуры пропусков; гистограмма пропущенных значений и т.п.

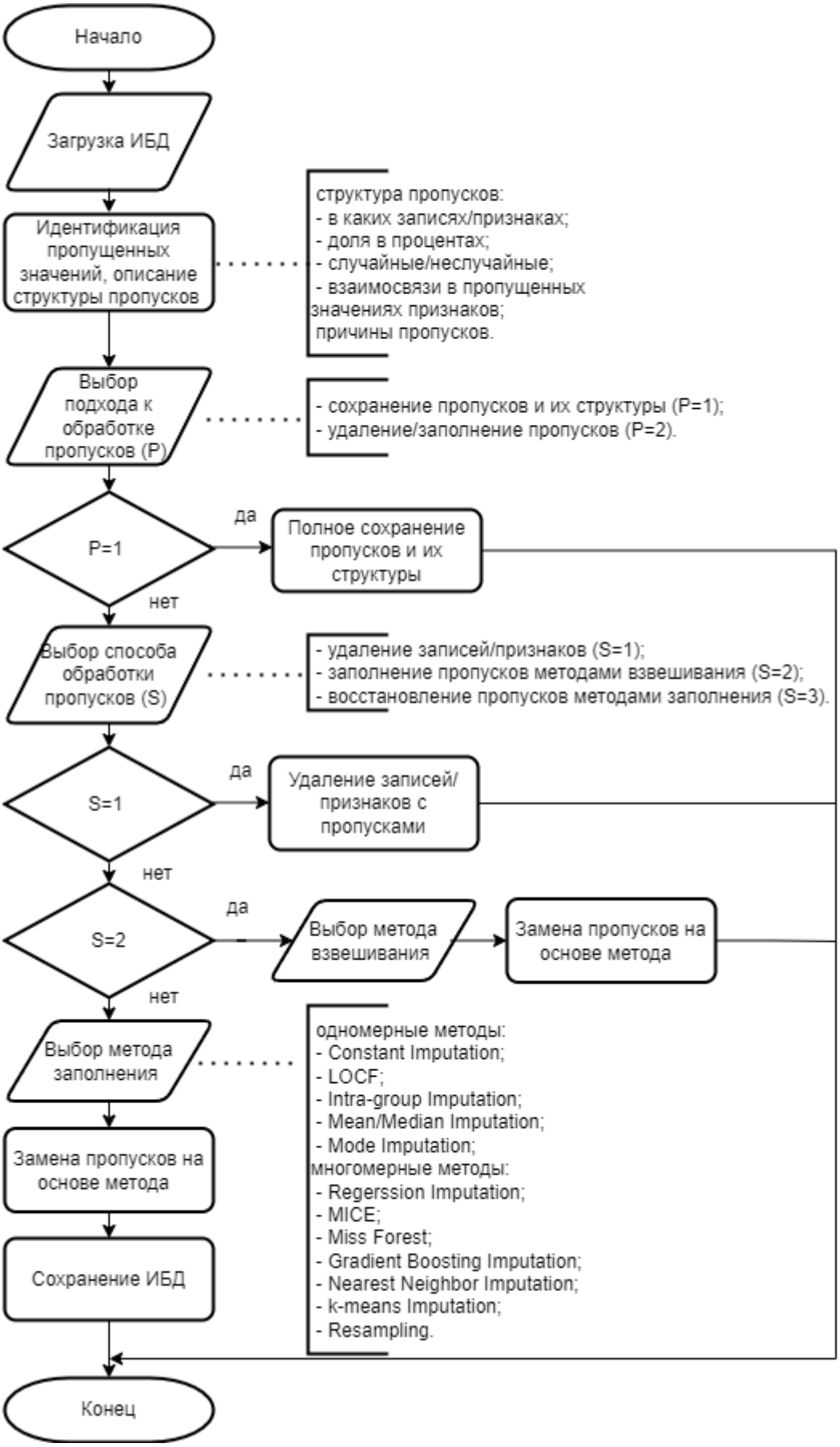


Рисунок 2.2 – Блок-схема алгоритма обработки пропущенных значений

На рисунке 2.3 показан пример тепловой карты. В первой строке приведены названия признаков, слева – количество записей; справа – количество пропусков в значениях признаков в записи; внизу – количество пропусков в значениях признака по всей таблице с данными. Красный цвет обозначает пропуски; синий цвет – заполненные значения. Согласно тепловой карте в таблице с данными: 175 записей не имеют пропусков; 19 записей содержат один пропуск в признаке *Mg*; 19 записей содержат один пропуск в признаке *K*; 1 запись содержит два пропуска; пропущено по 20 значений в признаках *Mg* и *K*.

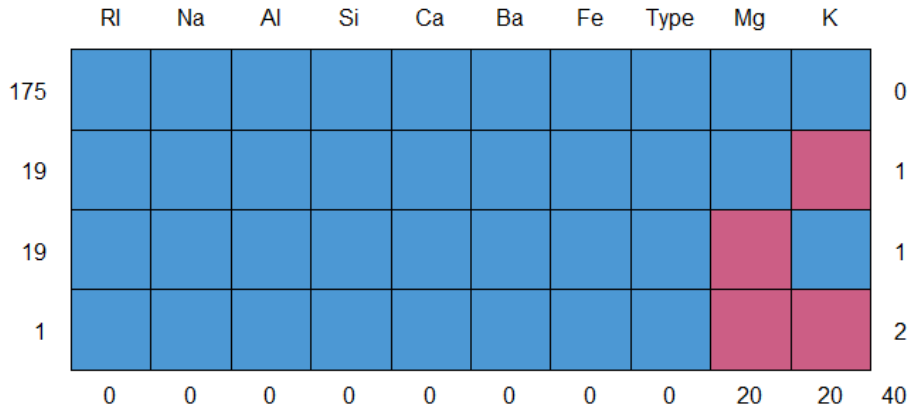


Рисунок 2.3 – Тепловая карта структуры пропусков

На рисунке 2.4 приведена совмещенная диаграмма рассеяния и диаграмма размаха. Синим цветом выделены наблюдаемые значения, красным – пропуски. Число в левом нижнем углу – это количество наблюдений, которые отсутствуют в обоих признаках (равно 1). Диаграмма позволяет проанализировать взаимосвязи между пропущенными значениями в двух признаках, сопоставить статистические характеристики признака при условии наличия/отсутствия пропусков в значениях другого признака, выявить аномальные выбросы в значениях признаков.

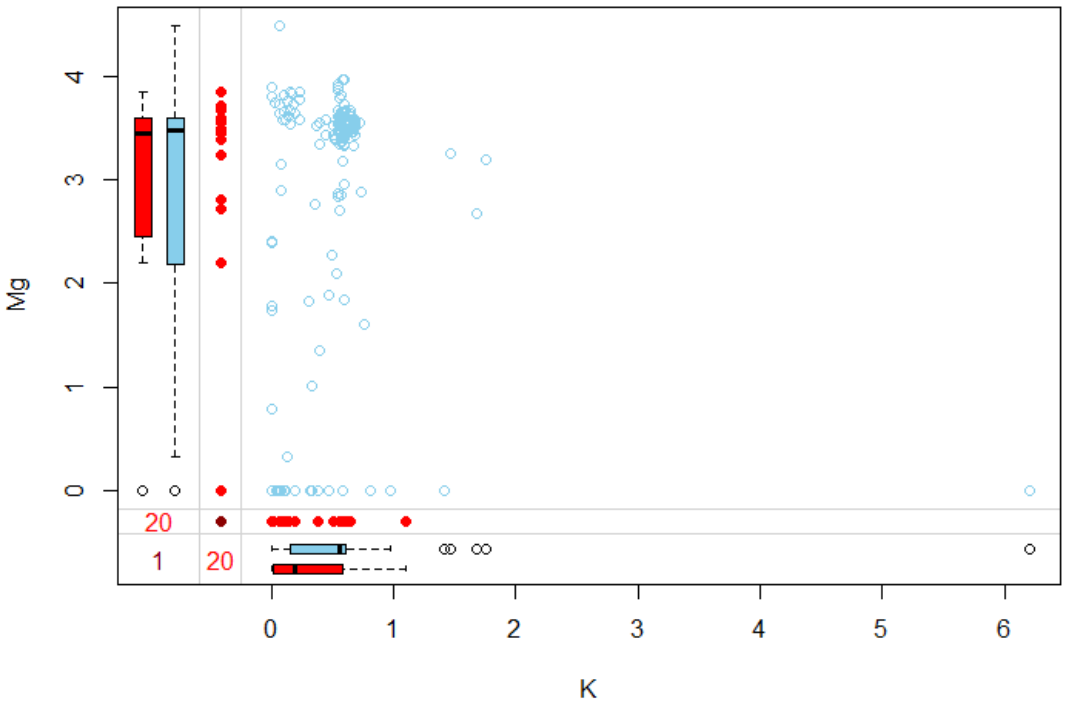


Рисунок 2.4 – Совмещенная диаграмма рассеяния и диаграмма размаха для изучения структуры и взаимосвязей в пропусках данных

На рисунке 2.5 приведен график структуры пропусков, который позволяет выяснить, есть ли в пропусках какая-то закономерность, и связаны ли пропуски между собой. Исходя из полученного графика видно, что распределение пропущенных значений для признаков случайно. Также для каждого признака выводится процент пропусков, и соотношение пропусков к наблюдаемым данным.

Статистически методы изучения структуры пропусков сводятся к расчету процентных долей пропусков в признаках и записях БД. Также используются методы корреляционного и кластерного анализа данных для изучения взаимосвязей между пропусками в нескольких признаках. Фактически результаты статистических расчетов визуализируются на графиках для наглядности и упрощения анализа структуры пропусков.

Отметим, что выявление причин наличия пропусков требует содержательного анализа природы признаков ИБД, рассмотрения особенностей сбора и регистрации данных и не может быть сведено к использованию только формальных методов анализа.

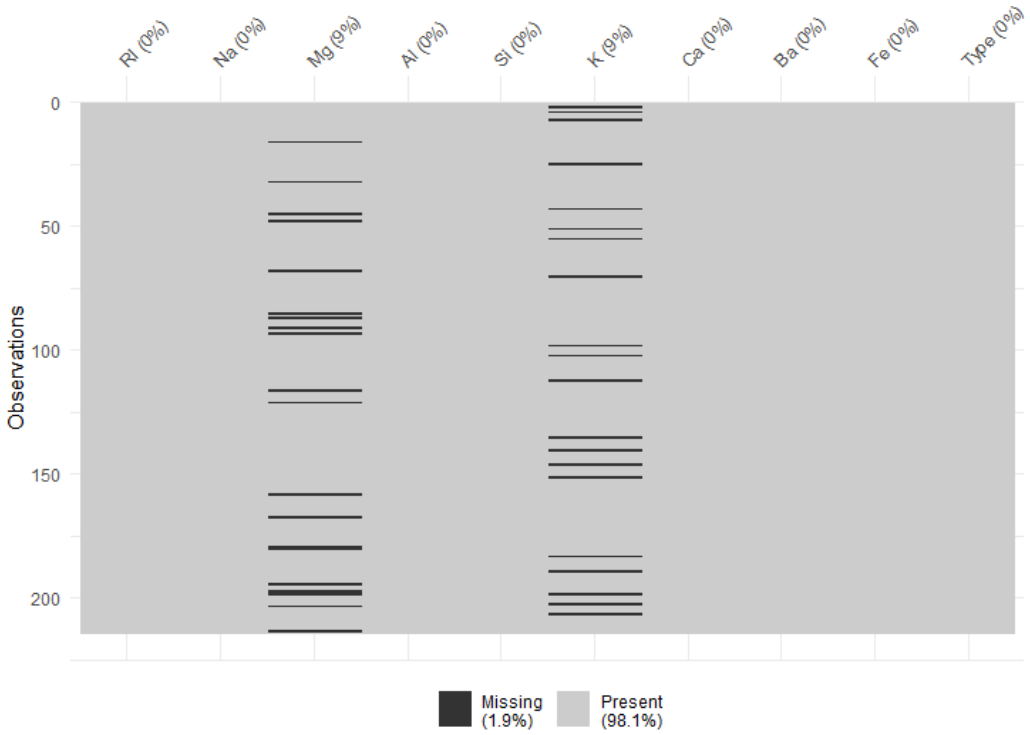


Рисунок 2.5 – График структуры пропусков

В контексте обезличивания данных возможны два подхода к учету пропущенных значений. В первом случае пропуски в данных обрабатываются до проведения процедуры обезличивания: либо заполняются, либо удаляются записи, содержащие пропуски в исходной БД; во втором случае – пропуски сохраняются при проведении процедуры обезличивания, и, в результате, обезличенная БД содержит пропущенные значения в тех же позициях, что и в исходной БД.

Для решения проблем, связанных с наличием пропущенных данных, используют три основных способа: удаление записей или признаков с пропущенными значениями; применение методов взвешивания данных; применение методов заполнения (импутации).

В рамках первого способа исключаются (удаляются) записи, либо признаки, содержащие пропущенные значения. Если доля записей с пропущенными значениями невелика в процентном отношении, то такой подход является хорошим решением. Однако полное удаление записей или признаков всегда приводит к потере информации и статистическая мощность выводов, сделанных на основе

анализа «урезанной» БД, снижается. Поэтому такой способ имеет ограниченное применение, несмотря на его легкость и быстроту использования.

Второй способ предполагает применение специально разработанных статистических методов взвешивания. В процессе взвешивания выполняется два этапа. На первом этапе исследователь удаляет все неполные записи. На втором этапе взвешиваются оставшиеся полные записи, которые используются для замены удаленных неполных записей. Фактически пропуски в данном случае не заполняются, а происходит замена неполных записей на другие, взятые из имеющихся полностью заполненных записей. Но следует помнить, что метод приводит к искажению статистических выводов, так как изначальные пропорции значений по признакам могут нарушаться.

Третий способ сводится к использованию специальных статистических методов заполнения пропусков (импутации). Пропуски в данных заполняются согласно заранее выбранному алгоритму. В результате формируется БД без пропусков значений признаков, которая далее используется для решения статистических задач анализа данных. Недостатком такого подхода является погрешность вычислений и смещение результатов статистического анализа, обусловленные заменой неизвестных истинных значений признаков на приближенные искусственные значения.

Методов заполнения пропусков разработано достаточно много, выбор метода зависит от целей дальнейшей статистической обработки данных, от структуры пропусков, от типа признака, содержащего пропущенные значения, от вычислительных ресурсов. Приведем некоторые наиболее распространенные методы заполнения пропущенных значений [21-24].

Методы заполнения пропусков можно разделить на два больших класса: одномерные методы (*Single Imputation*) и многомерные методы (*Multivariate Imputation*). Одномерный метод предполагает использование данных одного признака в процессе заполнения. Фактически, чтобы заполнить пропуски в значениях признака, берутся данные того же признака. Многомерный метод сводит-

ся к заполнению пропусков в значениях одного признака на основе данных других признаков.

Приведем примеры методов каждого из классов.

Одномерные методы.

Заполнение константой (Constant Imputation). Один из самых простых способов работы с пропусками в количественных данных — заполнить пропуски константой. Например, нулем (подходит для алгоритмов, чувствительных к масштабу признаков). Заполнение константой позволяет не сокращать размер выборки, однако может внести системную ошибку в данные.

Повторение результата последнего наблюдения (*Last observation carried forward* или, сокращенно, *LOCF*). Данный метод применяется, как правило, при заполнении пропущенных значений во временных рядах, когда последующие значения статистически значимо взаимосвязаны с предыдущими значениями.

Заполнение пропуска средним значением или медианой (*Mean/Median Imputation*). Метод сводится к замене пропущенных значений признаков средним (медианой), вычисленным на основе наблюдаемых данных. Наилучшие по точности результаты при замене на среднее наблюдаются в случае нормальной распределенности значений признаков. В случае отклонения от нормального распределения используют заполнение медианой, так как она не чувствительна к выбросам, в отличие от выборочного среднего. Однако данный метод не применим к классификационным признакам и предполагает, что данные отсутствуют случайно. В случае большой доли пропусков, в результате заполнения, появляется много медианных и средних значений, что может привести к смещению результатов статистической обработки.

Заполнение пропуска внутригрупповым значением (Intra-group Imputation). Метод аналогичен предыдущему, но предполагает предварительное разбиение данных на категории (кластеры). Замена на среднее или медиану происходит в зависимости от категории. В результате применения статистическая погрешность результатов обработки данных сравнительно ниже по сравнению с предыдущим методом.

Заполнение пропуска наиболее частым значением (модой) (Mode Imputation). Для заполнения пропусков в классификационных данных подойдет метод заполнения наиболее часто встречающимся значением (модой). Если пропусков немного, то этот метод вполне статистически обоснован. В случае если наблюдается значительное количество пропусков в классификационном признаке, можно ввести отдельную категорию для пропущенных значений.

Многомерные методы.

Заполнение пропусков на основе регрессионных моделей (Regression Imputation). Данный метод заключается в том, что пропущенные значения заполняются с помощью модели линейной регрессии, построенной на известных значениях набора признаков (независимых переменных). В случае классификационного признака используется модель мультиномиальной логистической регрессии. Мультиноминальная логистическая регрессия – общий случай логистической регрессии, который позволяет использовать категориальные (классификационные) признаки с количеством уровней больше, чем 2 [25]. В модели мультиномиальной логистической регрессии для каждого класса (категории) зависимого признака строится уравнение бинарной логистической регрессии [25]. При этом один из классов зависимого признака становится опорным, и все другие классы сравниваются с ним. Уравнение мультиномиальной логистической регрессии прогнозирует вероятность принадлежности к каждому классу зависимого признака на основе значений независимых признаков. Построенное уравнение используется для заполнения пропусков в зависимом признаке.

Методы регрессии хорошо работают, если между зависимым признаком (значения которого заполняются) и предикторными признаками существуют статистически значимые взаимосвязи. Модель регрессии описывает изменение условного среднего, следовательно, пропуски заполняются условным средним, рассчитанным на основе уравнения. Как следствие, в результате заполнения, дисперсия значений признака становится меньше, и усиливается корреляция между признаками.

Множественное замещение с помощью цепных уравнений (*Multiple Imputation by Chained Equations* или, сокращенно, *MICE*). Это один из самых известных многомерных методов. Для иллюстрации подхода *MICE*, рассмотрим набор переменных $X_1, X_2, \ldots, X_n$, где некоторые или все признаки имеют пропущенные значения.

Алгоритм работает следующим образом: для каждого признака заменяется отсутствующее значение с помощью одномерного метода, например, заменой на среднее значение, и это значение называется «заполнителем».

«Заполнители» для первой переменной, X_1, подвергаются регрессии с использованием регрессионной модели, где X_1 – зависимый признак, а остальные – независимые признаки. Далее X_2 рассматривается как зависимый признак, а остальные – как независимые признаки. Процесс продолжается до тех пор, пока все признаки не будут рассмотрены хотя бы один раз как зависимые признаки.

Эти исходные «заполнители» затем заменяются прогнозами регрессионной модели. Процесс замены повторяется в течение нескольких циклов (обычно таких циклов десять). В конце цикла пропущенные значения в идеале заменяются значениями прогноза, которые лучше всего отражают взаимосвязи, выявленные в данных.

MICE эффективно работает с пропущенными значениями в нескольких признаках. Этот подход может дать гораздо лучшие результаты, чем простые методы замены на среднее или медиану. Многие другие алгоритмы, такие как метод *k*-средних, случайный лес и нейронные сети, могут использоваться в качестве основы прогнозирования *MICE* при построении моделей взаимосвязей.

Метод предполагает, что данные отсутствуют случайным образом. Несмотря на все преимущества, этот подход может быть дорогостоящим в вычислительном отношении по сравнению с другими методами, особенно при работе с большими данными.

Метод *Miss Forest*. Метод основан на алгоритме *Random Forest*. Это непараметрический метод замены пропусков (в явном виде функция взаимосвязи между зависимыми и независимыми признаками не строится). Взаимосвязи опи-

сываются с помощью набора логических правил, представленных в виде дерева решений. Метод основан на построении модели случайного леса (ансамблевая модель деревьев решений) для каждого признака, с последующим ее использованием для прогнозирования пропущенных значений. Алгоритм *MissForest* позволяет заменять пропущенные значения практически для признаков любого типа. В частности, он может одновременно обрабатывать многомерные данные, состоящие из количественных и качественных признаков. *MissForest* не требует априорных предположений о типе распределения данных. Однако алгоритм работает хуже многих линейных методов, когда в выборке много разреженных признаков. Случайный лес не экстраполирует данные, в отличие от линейных регрессионных методов, но это можно считать и достоинством метода, так как в случае попадания аномального выброса не будет получено экстремальных значений. К недостатку алгоритма относят склонность к переобучению.

Заполнение пропусков методом градиентного бустинга (Gradient Boosting Imputation). Градиентный бустинг – это алгоритм машинного обучения, который строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений [26]. В основе данного алгоритма лежит итеративное обучение деревьев решений с целью минимизации функции потерь [26]. Достоинством метода является возможность обрабатывать как количественные, так и качественные признаки. Также метод хорошо работает в случае нелинейных взаимосвязей между признаками. В основе алгоритма метода лежит идея аддитивного и последовательного обучения множества моделей. В рамках алгоритма обучается дерево, в котором для каждого наблюдения находится функция потерь, которая предназначена для отслеживания ошибки с каждым примером обучения. С помощью градиентного спуска находится минимальное значение функции потерь, и на результатах этого дерева строится следующее, для выдачи новых прогнозов. Процесс происходит итеративно, количество итераций задано. В результате прогнозы окончательной ансамблевой модели представляют собой взвешенную сумму прогнозов, сделанных предыдущими деревьями [26]. Прогнозы ансамблевой модели используются для заполнения пропусков.

Замена пропусков методом ближайших соседей (Nearest Neighbor Imputation). Метод обеспечивает заполнение пропущенных значений с использованием подхода *k*-ближайших соседей. Для поиска ближайших соседей используется одна из метрик расстоянии (как правило, евклидова мера). Каждый пропуск рассчитывается с использованием значений от *k*-ближайших соседей, у которых есть значение для данного признака. Характеристики соседей усредняются равномерно или взвешиваются по расстоянию до каждого соседа.

Замена пропусков методом k-средних (k-means Imputation). Алгоритм метода аналогичен методу ближайших соседей. Отличие от метода *k*-ближайших соседей заключается не в поиске ближайших соседей для каждого объекта с пропусками, а в использовании значений центра кластера для заполнения пропусков, в который попадает конкретный объект с пропущенными значениями. Метод требует проведения кластеризации объектов, что, в свою очередь, требует начальной инициализации пропусков.

Заполнение пропусков методом Resampling. Resampling – итеративный алгоритм, в котором записи, содержащие пропущенные значения, заменяются случайно подобранными записями из многократной генерации новых выборок, на базе имеющейся БД. Для предсказания отсутствующего значения выполняется построение регрессионного уравнения. Процедура построения регрессионного уравнения повторяется итерационно. После заданного количества итераций, значения полученных регрессионных коэффициентов усредняют и получают окончательные регрессионные коэффициенты. Итоговое регрессионное уравнение обеспечивает максимальную точность прогноза пропущенных значений.

2.2.3 Разметка данных

Непосредственно перед проведением процедуры обезличивания необходимо выполнить разметку данных. Разметка данных состоит из двух шагов:

1. Определение типа каждого признака (атрибута), входящего в исходную БД.

2. Определение типа значений каждого признака (атрибута), входящего в исходную БД.

Выделяют четыре основных типа признаков (атрибутов), описывающих информацию о субъекте персональных данных: прямые идентификаторы; квази-идентификаторы (косвенные признаки), чувствительные признаки, нечувствительные признаки [1].

Прямые идентификаторы – признаки (атрибуты), которые однозначно идентифицируют субъекта персональных данных. В качестве примера прямых идентификаторов можно привести следующие признаки: ФИО, СНИЛС; адрес проживания, номер медицинского полиса и т.п.

Квазиидентификаторы (*косвенные идентификаторы*) – признаки (атрибуты), которые идентифицируют субъекта персональных данных с той или иной степенью неопределенности [1]. В ряде случаев комбинация косвенных идентификаторов может дать однозначную идентификацию субъекта или обеспечить высокий риск раскрытия информации [1]. Например, сочетание косвенных признаков: дата заболевания, пол, возраст, район местожительства при определенных условиях может привести к идентификации личности и увеличивает риск раскрытия персональной информации о субъекте.

Чувствительны признаки – признаки (атрибуты), которые содержат «деликатную» информацию о субъекте персональных данных. Например, медицинская информация, особенности течения заболевания, индивидуальные показатели и т.п. Именно эта информация представляет наибольший интерес для решения задач статистического анализа БД и не может быть изменена.

Один из подходов в обезличивании данных предполагает деление признаков только на прямые и косвенные идентификаторы. Таким образом, предполагается, что чувствительный признак также может выступать квази идентификатором в ряде случаев. Например, код диагноза по МКБ-10 может рассматриваться как чувствительный признак или как косвенный идентификатор в зависимости от конкретной ситуации и типа предполагаемой атаки злоумышленника, от

того какие задачи статистической обработки будут решаться на ОБД и каковы требования к обезличенным данным.

Нечувствительные признаки – признаки (атрибуты), которые не относятся ни к одной из вышеперечисленных категорий и не представляют интереса для дальнейшей обработки. Нечувствительные признаки могут не включаться в обезличенную БД.

Типы признаков ИБД во многом определяют выбор методов и процедур обезличивания, которые будут применяться в ходе обработки ИБД, а также входные данные для этапа оценки риска раскрытии информации. Риск раскрытия информации считается равным 100% для прямых идентификаторов и оценивается по совокупности показателей в случае работы с косвенными идентификаторами (см. п. 2.4).

Значения признаков измеряются в разных шкалах и соответствуют одному из типов: количественные (числовые) признаки (данные); ранговые (ординальные) признаки (данные); качественные (номинальные, классификационные) признаки (данные); признаки типа дата/время.

Количественные признаки (данные) – признаки, которые регистрируются с помощью чисел, имеющих содержательный смысл. Признаки программно задаются типами: *integer, double, float* и т.п. Количественные признаки измеряются в количественной шкале, на которой допустимы все арифметические операции и операции сравнения. К количественным признакам относятся, например, возраст, доход, стаж работы, рост, вес и т.п.

Ранговые (ординальные) признаки (данные) – признаки, которые измеряются в порядковой шкале, на которой допустимы только операции сравнения. Порядковые признаки имеют тип *ordinal* в программных средах. К ранговым признакам относится, например, уровень неврологических нарушений (0 – без нарушений; 1 – незначительные нарушения, 2 – средняя стадия, 3 – тяжелая стадия).

Качественные (номинальные, классификационные) признаки (данные) – признаки, принимающие значения из некоторого ограниченного набора неис-

числяемых категорий. Измеряются в качественной шкале, для которой не допустимы арифметические операции, а из операций сравнения допустимы только «равно» или «не равно» (например, пол, адрес местожительства, ФИО и т.п.). Для качественных признаков задается тип *text, string, factor* в зависимости от используемого программного средства.

К отдельному типу относятся признаки, значения которых соответствуют дате/времени и представляются в разных форматах записи. Во всех программных инструментах реализован тип *data/time* и средства (функции) для преобразования форматов представления.

При выборе методов и процедур обезличивания данных учитывают типы значений признаков, как один из критериев выбора, так как используются разные методы обезличивания для работы с количественными, качественными признаками или с датами/временем. Кроме того, на этапе оценки информационных потерь используются разные меры информационных потерь в зависимости от типа значений признаков.

2.3 Этап обезличивания данных

Работа модуля «Обезличивание данных» системы «Интеллектуального управления, обогащения и обезличивания данных» начинается с выбора БД в формате *PostgreSQL*. Далее *Java* программа подключается к базе данных *PostgreSQL* и создает её копию с именем «mask_имя исходной БД» (mask_ИБД).

После того как была создана копия исходной БД программа отключается от оригинала и начинает применять методы обезличивания к клонированной базе и проводить все нужные преобразование через запросы из *java*. Таким образом, в *PostgreSQL* создаются и хранятся новая обезличенная база данных, таблицы идентификаторов, таблицы соответствий, таблицы перестановок, если метод обезличивания предполагает их создание. В дальнейшем необходимо решить вопрос о раздельном хранении всех БД для обеспечения информационной безопасности и надежной защиты от возможных атак злоумышленников.

Укрупнено алгоритм реализации этапа обезличивания данных состоит из следующих шагов.

1. Выбор ИБД в формате *PostgreSQL*.
2. Создание копии ИБД с именем mask_ИБД.
3. Задание типов атрибутов (признаков) mask_ИБД.
4. Задание типов значений атрибутов (признаков) mask_ИБД.
5. Выбор атрибутов для обезличивания из mask_ИБД.
6. Выбор метода обезличивания для каждого атрибута или совокупности атрибутов.
7. Задание параметров метода обезличивания.
8. Обезличивание данных на основе выбранного метода.
9. Сохранение ОБД.

Далее обезличенная БД передается на следующие этапы обработки, в рамках которых оцениваются риски раскрытия информации, меры информационных потерь, выполняется сравнение с пороговыми показателями, и делается вывод о степени защищенности ОБД и ее пригодности для решения поставленных задач.

Ниже приведены алгоритмы реализации основных типов методов обезличивания данных согласно классификации, приведенной в [8]: метод введения идентификаторов, метод декомпозиции, метод перемешивания, методы изменения состава и/или семантики.

При описании алгоритмов клонированная база данных mask_ИБД, над которой выполняются все преобразования, обозначается ИБД для простоты и удобства.

2.3.1 Обезличивание на основе метода введения идентификаторов

Метод введения идентификаторов предполагает замену значений признаков в ИБД на идентификаторы и введение таблицы соответствий, в которой каждому оригинальному значению признака в ИБД сопоставлен уникальный идентификатор [1]. Устанавливается взаимно-однозначное соответствие: каждому значению идентификатора должно соответствовать одно значение признака и каждому значению признака должно соответствовать одно значение идентификатора. Таблицы соответствия создаются для каждого признака ИБД, значения которых заменяются идентификаторами. Таблица соответствия хранит исходное значение признака и идентификатор. Для вычисления значений идентификаторов могут использоваться криптографические алгоритмы хэширования, основанные на вычислении хэш-функций.

Условия и ограничения использования метода: применяется для обезличивания, как правило, прямых идентификаторов, которые однозначно определяют субъекта персональных данных; для вычисления значений идентификаторов рекомендуют использовать семейство алгоритмов *sha2* (*sha224, sha256, sha384* и *sha512*), которые наиболее эффективны для защиты конфиденциальной информации; длина идентификатора должна быть одинакова для всех записей.

На рисунке 2.6 приведена блок-схема алгоритма метода введения идентификатора. Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

Входные данные и параметры алгоритма:

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$;

– *Table_j* – имя j-ой таблицы соответствия;

– алгоритм вычисления значений идентификаторов (в частности, алгоритм хэширования).

Алгоритм метода введения идентификаторов:

1. Загрузка ИБД.

2. Выбор признаков X^j из ИБД; выбор алгоритма хэширования.

3. Для каждого j-го признака, $j = \overline{1,p}$:

   3.1. Ввод названия таблицы соответствия *Table_j*.

   3.2. Запись значений признака X^j в таблицу соответствия *Table_j*.

   3.3. Создание столбца идентификаторов в ИБД и в таблице соответствия *Table_j*, которые связывают записи в ИБД и *Table_j*.

   3.4. Вычисление значений идентификаторов на основе выбранного алгоритма хэширования.

   3.5. Запись значений идентификаторов в ИБД и *Table_j*.

   3.6. Удаление столбца с исходными значениями X^j из ИБД.

4. Сохранение $X_{\text{об.}}$.

Таблицы соответствий должны быть отделены от ИБД, так как если пользователь (и злоумышленник тоже!) будет иметь доступ к таблицам соответствий, то он может связать персональные данные с конкретным субъектом персональных данных. Термин «отделить» означает либо физически отдельное хранение ИБД и таблиц соответствий, либо установку между двумя базами данных межсетевого экрана (сертифицированного на соответствие нормативным требованиям).

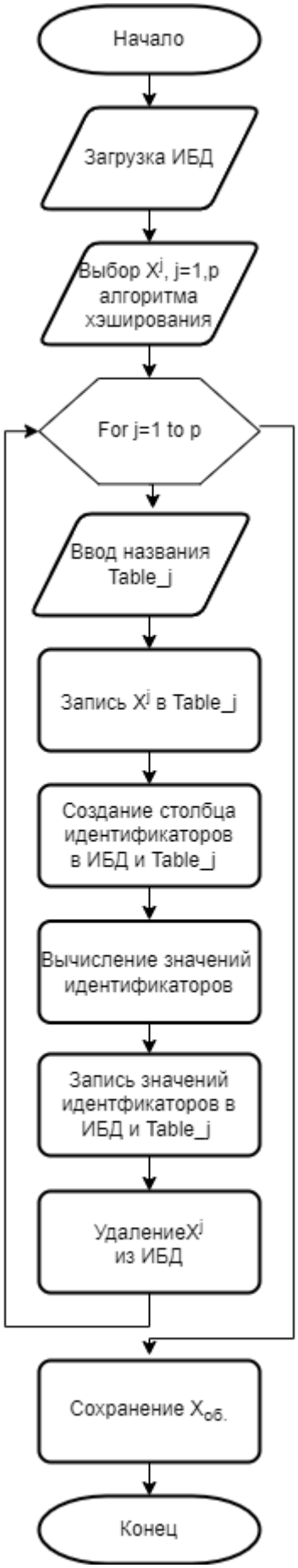


Рисунок 2.6 – Блок-схема алгоритма введения идентификаторов

В первом случае связи между базами не будет совсем, и совместная их обработка возможна только с применением специальных внешних носителей. Во втором случае связь между базами будет односторонняя (со стороны таблицы соответствий), совместная обработка возможна тоже только с одной стороны [27-28].

На рисунке 2.7 приведен пример обезличивания атрибута ФИО, который относится к прямым идентификаторам. Исходные значения ФИО заменяются на идентификаторы, вычисленные с помощью хэш-функции, и создается таблица соответствий.

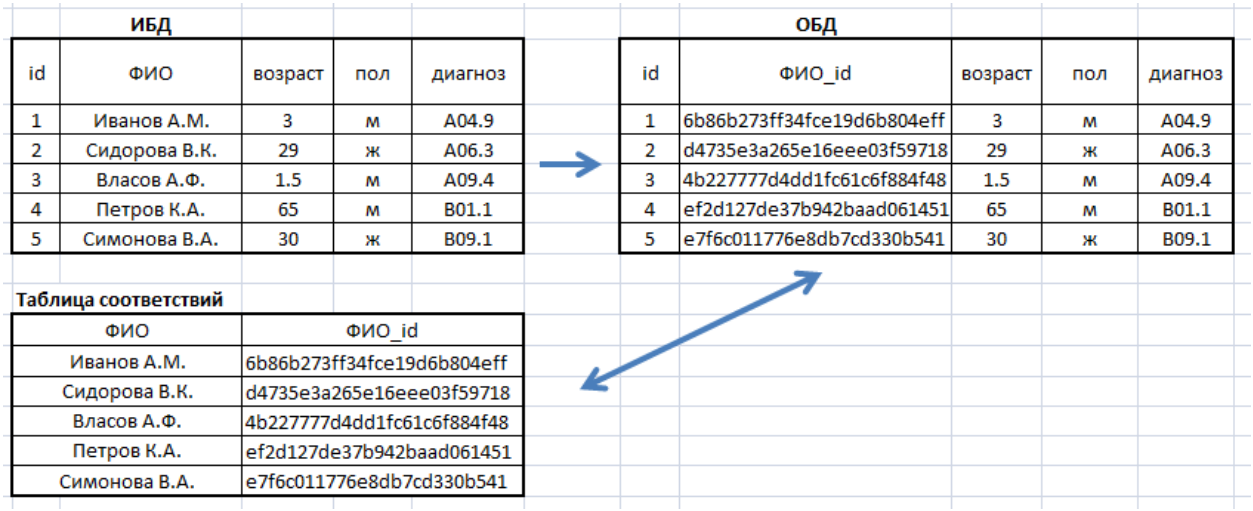


Рисунок 2.7 – Пример применения метода введения идентификаторов

2.3.2 Обезличивание на основе метода декомпозиции

Метод декомпозиции заключается в разделении атрибутов персональных данных, хранящихся в ИБД, на несколько таблиц и подмножеств, с последующим раздельным хранением записей, соответствующих подмножествам этих атрибутов [1]. Метод декомпозиции обеспечивает обезличивание за счет отделения идентифицирующей информации, при этом таблица с идентифицирующей информацией должна быть недоступна злоумышленнику. Между подмножествами (таблицами), на которые декомпозируется ИБД, должна быть обеспечена взаимосвязь, что достигается ведением таблицы соответствий. В таблице соответ-

ствий отражена структура взаимосвязей между идентификаторами строк ИБД и идентификаторами строк в таблицах подмножеств.

На рисунке 2.8 приведена блок-схема алгоритма метода декомпозиции. Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

Входные данные и параметры алгоритма:

– $X = \{x_{ij}\}_{i,j=1}^{n,p}$, где x_{ij} – значение j-го атрибута (признака) -го субъекта ПД в ИБД; n – количество строк, записей (субъектов персональных данных); p – количество признаков; $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ – значения признаков в i-ой записи; $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \ldots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака;

– K – количество частей, на которые декомпозируется ИБД;

– *Table_k* – имя k-ой таблицы для хранения подмножества атрибутов ИБД, $k = \overline{1, K}$;

– *Table_c* – имя таблицы для хранения структуры взаимосвязей между подмножествами ИБД.

Алгоритм метода декомпозиции:

1. Загрузка ИБД – X;
2. Ввод значения параметра K, определяющего количество частей, на которые декомпозируется ИБД.
3. Ввод имени таблицы *Table_c*, в которой хранится структура взаимосвязей между подмножествами ИБД.
4. Создание столбца идентификаторов в ИБД и запись идентификаторов для связей между частями БД.
5. Для каждого k-го подмножества, $k = \overline{1, K}$:
   5.1. Ввод названия таблицы *Table_ k*.
   5.2. Выбор набора атрибутов X^j из ИБД, которые будут храниться в *Table_ k*.
   5.3. Создание столбца идентификаторов в таблицах *Table_ k* и *Table_c* для связей записей.

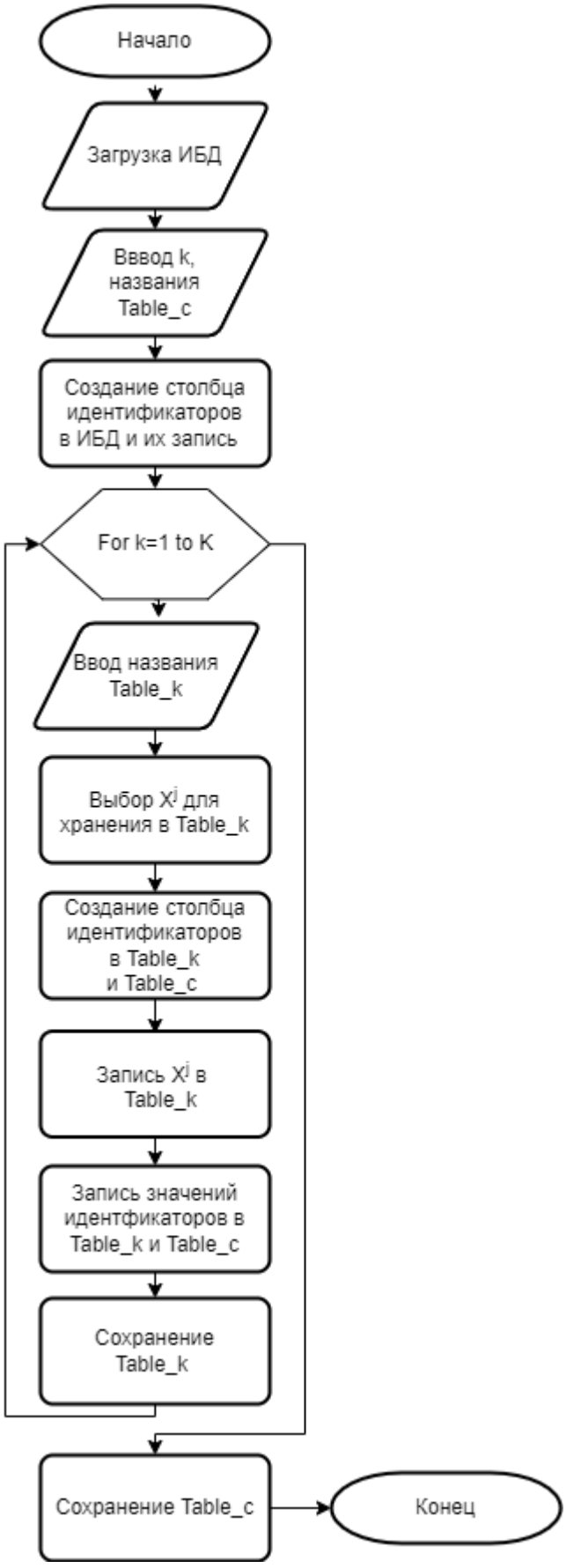


Рисунок 2.8 – Блок-схема алгоритма декомпозиции

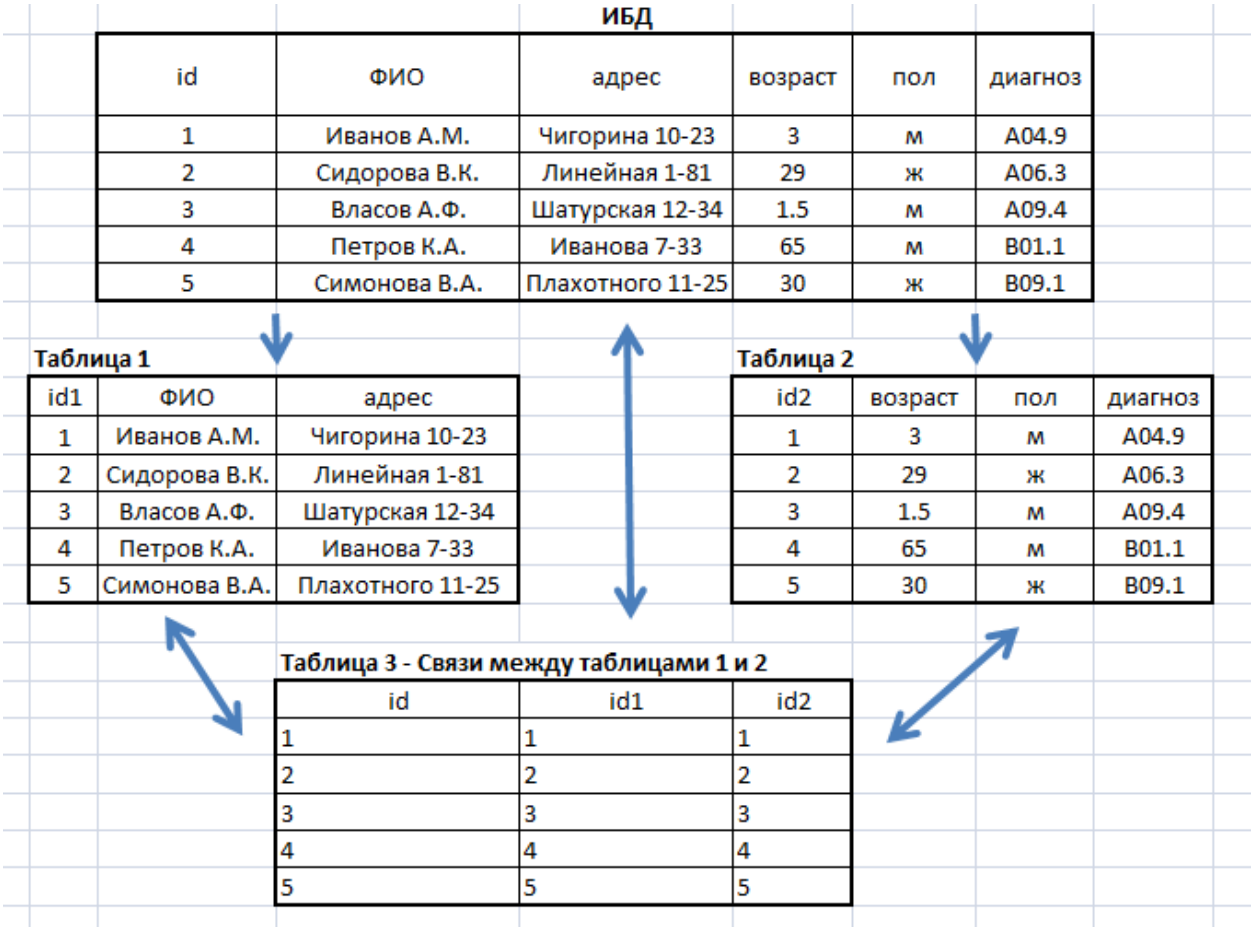
5.4. Запись значений признаков X^j в таблицу *Table_ k*.

5.5. Запись соответствующих значений идентификаторов в *Table_ k* и таблицу взаимосвязей *Table_c*.

5.6. Сохранение *Table_ k*.

6. Сохранение *Table_c*.

На рисунке 2.9 приведен пример работы метода декомпозиции: исходная БД разбивается на два подмножества (таблицы) с данными, которые хранятся раздельно. В первой таблице (таблица 1) хранятся прямые идентификаторы (ФИО, адрес проживания), во второй таблице (таблица 2) – косвенные идентификаторы и чувствительные атрибуты (возраст, пол, диагноз). В таблице 3 содержится информация о связях между ИБД, таблицами 1 и 2.



ИБД

id	ФИО	адрес	возраст	пол	диагноз
1	Иванов А.М.	Чигорина 10-23	3	м	A04.9
2	Сидорова В.К.	Линейная 1-81	29	ж	A06.3
3	Власов А.Ф.	Шатурская 12-34	1.5	м	A09.4
4	Петров К.А.	Иванова 7-33	65	м	B01.1
5	Симонова В.А.	Плахотного 11-25	30	ж	B09.1

Таблица 1

id1	ФИО	адрес
1	Иванов А.М.	Чигорина 10-23
2	Сидорова В.К.	Линейная 1-81
3	Власов А.Ф.	Шатурская 12-34
4	Петров К.А.	Иванова 7-33
5	Симонова В.А.	Плахотного 11-25

Таблица 2

id2	возраст	пол	диагноз
1	3	м	A04.9
2	29	ж	A06.3
3	1.5	м	A09.4
4	65	м	B01.1
5	30	ж	B09.1

Таблица 3 - Связи между таблицами 1 и 2

id	id1	id2
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5

Рисунок 2.9 – Пример применения метода декомпозиции

2.3.3 Обезличивание на основе метода перемешивания

Метод перемешивания заключается в перемешивании отдельных записей или групп записей ИБД между собой на основе использования некоторого правила перемещения [1, 29]. В качестве входных данных и параметров метода используют: набор атрибутов из исходной БД, которые войдут в группу перемешивания, и относительно которых будут перемешиваться записи; объем группы записей ИБД, которые участвуют в перемешивании (как правило, на каждой итерации в перемешивания участвуют две записи, которые обмениваются значениями атрибутов); алгоритм перемешивания. Метод должен обеспечивать возможность обратного преобразования (деобезличивания), поэтому применяются алгоритмы на основе перемещения «один в один». В результате применения метода перемешивания сохраняется структура и смысл атрибутов, но нарушаются связи между атрибутами, входящими в группу перемешивания, и остальными атрибутами ИБД. В частном случая группа перемешивания может включать только один атрибут.

Возможны два варианта реализации метода. В первом случае значения атрибутов, входящих в группу перемешивания, перемещаются для каждого атрибута отдельно, т.е. алгоритм перемешивания применяется к каждому атрибуту (одиночное перемешивание). Во втором случае реализуется сцепленное перемешивание, т.е. выбирается набор атрибутов, значения которых перемещаются относительно других атрибутов, входящих в исходную БД. В этом случае структура взаимосвязей между атрибутами внутри набора сохраняется.

В результате применения метода перемешивания ОБД обладает следующими особенностями:

1. Исходная БД преобразуется в ОБД с тем же количеством записей и идентичной структурой, в которой каждый отдельно взятый атрибут одного субъекта персональных данных имеет прежнюю семантику, но получает значение из другой записи ИБД (второго субъекта), причем другой атрибут первого субъекта может получить значение аналогичного атрибута третьего субъекта и т.д. При этом атрибуты субъекта, не входящие в группу перемешивания, остают-

ся неизменными. Таким образом, для злоумышленника ОБД не выглядит ни обезличенной, ни даже модифицированной.

2. Создаются формализованные алгоритмы перемещения значений атрибутов между записями для автоматизированного использования.

К составу группы атрибутов перемешивания предъявляются следующие требования:

– неизвестность для злоумышленника (т.е. количество и состав группы атрибутов перемешивания является секретом метода);

– объем группы перемешивания (в случае одиночного перемешивания) не должен быть слишком мал, иначе злоумышленник может вычислить его на примере известного ему субъекта персональных данных и сделать вывод о достоверности прочих атрибутов;

– группа перемешивания может включать, как прямые идентификаторы, так и косвенные идентификаторы, рекомендуется включать в группу перемешивания все идентифицирующие атрибуты;

– в группу перемешивания также могут включаться специальные идентификаторы (искусственно введенные атрибуты), которые связывают различные таблицы БД реляционного типа (справочники с одной стороны и изменяемые данные функционального характера – с другой), тем более, если эти идентификаторы являются идентифицирующими атрибутами субъекта персональных данных.

В рамках алгоритма перемешивания могут использоваться алгоритмы создания случайных перестановок конечного множества. Один из наиболее известных – алгоритм перестановки Фишера-Иейтса в современной (модифицированной) версии, адаптированной для эффективной реализации на компьютере. Идея алгоритма заключается в перемещении «вычеркнутых» чисел в конец списка путем замены их последним неотчеркнутым числом на каждой итерации. Временная сложность такого алгоритма равна $O(n)$, т.е. это алгоритм линейного времени.

На рисунке 2.10 приведена схема перестановки 5 элементов (строк) согласно модифицированному алгоритму Фишера-Иейтса. На первой итерации генерируется случайное целое число в диапазоне от 1 до 4, которое определяет элемент (строку) перестановки. Далее переставляются последний элемент (строка номер 5) и выбранный элемент (строка номер 3). На второй итерации в перестановке участвуют предпоследний элемент (строка номер 4) и элемент, определенный генератором случайных чисел в диапазоне от 1 до 3 (строка номер 2). Далее аналогично, пока не дойдем до первого элемента последовательности. Таким образом, количество итераций алгоритма на единицу меньше количества элементов (строк).

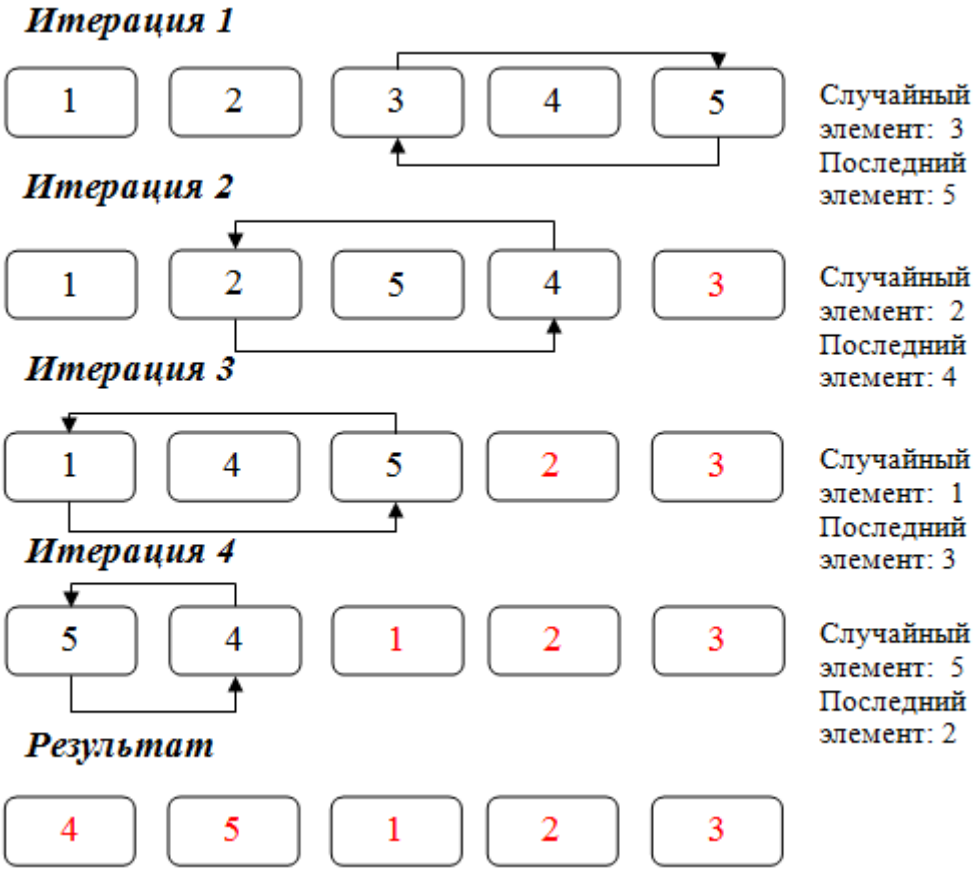


Рисунок 2.10 – Пример работы алгоритма перестановок

Результат работы алгоритма перемещения может быть задан в виде таблицы перестановок (соответствие «идентификатор записи – идентификатор записи») для каждого атрибута, либо в виде формулы, связывающей идентификатор

исходной записи, номер атрибута в этой записи и идентификатор конечной записи в процессе перемещения. Формула может быть любой сложности, но существует достаточно жесткое ограничение: однозначное соответствие [29]. Однозначность соответствия означает, что полный охват заменяемых атрибутов в заданной группе перемешивания должен сопровождаться полным охватом замененных атрибутов в этой же группе, причем акт замены каждого атрибута должен быть однократным в одном цикле перемешивания (все атрибуты группы перемешивания заменяются за один раз) [29]. Для реализации алгоритма был выбран вариант с использованием таблицы перестановок.

Ниже приведен алгоритм перемешивания для случая сцепленного перемешивания, когда значения атрибутов, входящих в группу перемешивания, не перемещаются относительно друг друга. В случае одиночного перемешивания используется этот же алгоритм, применяемый для каждого атрибута (p раз), входящего в группу перемешивания.

На рисунке 2.11 приведена блок-схема алгоритма метода перемешивания. Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

Входные данные и параметры алгоритма перемешивания:

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1, p}$;

– p – количество признаков, входящих в группу перемешивания;

– алгоритм перестановки;

– $Table\_p$ – имя таблицы перестановок.

Алгоритм метода перемешивания:

1. Загрузка ИБД.
2. Выбор признаков X^j из ИБД и алгоритма перестановки.
3. Ввод названия таблицы перестановок *Table_p*.
4. Вычисление $i = \max\_len$ (количество строк в ИБД).

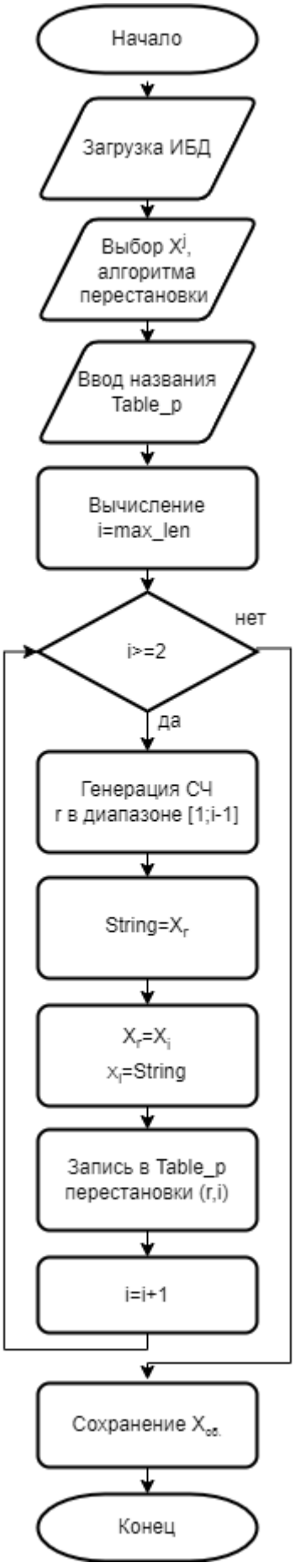


Рисунок 2.11 – Блок-схема алгоритма перемешивания

5. Пока $i \geq 2$

5.1. Генерация случайного целого числа $r \in [1; i-1]$.

5.2. Запись r-ой строки: $string = X_r = (x_{r1}, x_{r2}, \ldots, x_{rp})$.

5.3. Присваивание значениям атрибутов r-ой строки соответствующих значений атрибутов в i-ой строке: $X_r = X_i$.

5.4. Присваивание значениям атрибутов i-ой строки соответствующих значений атрибутов в $string$: $X_i = string$.

5.5. Запись в таблицу $Table\_p$ перестановки: (r, i).

5.6. $i = i - 1$.

6. Сохранение $X_{об.}$.

На рисунке 2.12 приведен пример обезличивания данных с помощью метода перемешивания. Исходная БД включает 10 записей, для которых зарегистрированы значения четырех признаков: ФИО; адрес; возраст; диагноз. В группу перемешивания включены признаки ФИО и адрес, реализуется одиночное перемешивание для каждого атрибута отдельно. В результате работы алгоритма перемешиваются значения атрибутов ФИО и адрес и формируются две таблицы перестановок.

ИБД

id	ФИО	адрес	возраст	диагноз
1	Иванов А.М.	Чигорина 10-23	3	A04.9
2	Сидоров В.К.	Линейная 1-81	29	A06.3
3	Власов А.Ф.	Шатурская 12-34	1.5	A09.4
4	Петров К.А.	Иванова 7-33	65	B01.1
5	Симонов В.А.	Плахотного 11-25	30	B09.1
6	Антонов И.И.	Русская 21-55	4	A04.9
7	Панин П.Р.	Сиреневая 12-22	26	A06.3
8	Уваров Е.К.	Ленина 7-19	68	B01.1
9	Терехин Е.К.	Советская 11-17	55	B09.1
10	Вохмин В.П.	Арбузова 8-11	78	A04.9

ОБД

id	ФИО	адрес	возраст	диагноз
1	Антонов И.И.	Ленина 7-19	3	A04.9
2	Петров К.А.	Иванова 7-33	29	A06.3
3	Симонов В.А.	Сиреневая 12-22	1.5	A09.4
4	Вохмин В.П.	Совесткая 11-17	65	B01.1
5	Панин П.Р.	Арбузова 8-11	30	B09.1
6	Терехин Е.К.	Плахотного 11-25	4	A04.9
7	Сидоров В.К.	Линейная 1-81	26	A06.3
8	Власов А.Ф.	Русская 21-55	68	B01.1
9	Уваров Е.К.	Чигорина 10-23	55	B09.1
10	Иванов А.М.	Шатурская 12-34	78	A04.9

Таблица перестановок (для ФИО)

N	id1	id2
1	1	10
2	8	9
3	3	8
4	2	7
5	3	6
6	2	5
7	1	4
8	2	3
9	1	2

Таблица перестановок (для адрес)

N	id1	id2
1	3	10
2	1	9
3	6	8
4	2	7
5	5	6
6	3	5
7	1	4
8	2	3
9	1	2

Рисунок 2.12 – Пример применения метода перемешивания

2.3.4 Обезличивание на основе методов изменения состава и/или семантики

Методы изменения состава или семантики образуют целый класс (группу) методов, объединенных общим подходом к обезличиванию данных. Эти методы обеспечивают обезличивание данных путем замены исходных значений признаков на результаты их статистической обработки, обобщения, маскирования, добавления шума или удаления части сведений [1, 30].

В рамках группы методов изменения состава и/или семантики представлены разнообразные методы, предполагающие использование разных принципов преобразования данных. Далее буду рассмотрены основные варианты реализации методов этой группы.

2.3.4.1 Обезличивание на основе метода обобщений

Метод обобщения заключается в снижении информативности данных путем уменьшения их детализации, которое может быть достигнуто с помощью увеличения масштаба шкалы измерения признака или сокращением числа категорий, которыми представлен каждый признак [4]. Алгоритм работы метода зависит от типа значений обезличиваемого признака: дата/время; количественный; порядковый (ординальный); классификационный (номинальный).

При обобщении дат/времени переходят к новому формату (более общему) представления даты, либо задают диапазоны представления дат.

Для количественного признака обобщение сводится к замене значений X^j на новые дискретные значения $X^j_{\text{об.}}$ путем перехода от количественной шкалы измерений к ординальной, интервальной или классификационной. Фактически реализуется разбиение значений признака на интервалы. В ОБД значения признаков представлены либо интервалами, либо статистическими характеристиками интервалов (среднее, медиана, среднее по границам интервала и т.п.), либо категориями (номер категории или название категории).

В случае ординального признака исходные значения X^j заменяются интервальными значениями или качественной категорией интервалов (номер интервала, название интервала).

При обобщении классификационного признака X^j объединяются несколько категорий с целью формирования новых (более общих) категорий; в результате получается новый признак $X^j_{\text{об.}}$, для которого $|D(X^j_{\text{об.}})| < |D(X^j)|$, где $|\ |$ – обозначение мощности множества.

Различают методы глобального и локального обобщения. Глобальное обобщение подразумевает, что единое преобразование применяется ко всем записям исходной БД, локальное обобщение – преобразование зависит от выбранного подмножества записей БД.

Условия и ограничения использования метода: применяется для обезличивания косвенных идентификаторов; применяется для обезличивания признаков любого типа.

На рисунке 2.13 приведена блок-схема алгоритма обобщения (для дат/времени). Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

Входные данные и параметры алгоритма обобщения (для дат/времени):

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$; выбирается один признак;

– формат преобразования даты/времени; доступны следующие форматы: секунда/минута/час/день/месяц/год; минута/час/день/месяц/год; час/день/месяц/год; день/месяц/год; неделя/месяц/год; неделя/год; месяц/год; квартал/год; день недели; неделя года; декада года; квартал года; год; век; тысячелетие.

– границы диапазонов: $(x_{k.\text{н}} - x_{k.\text{в}}],\ k = \overline{1,K}$;

– K – количество диапазонов;

Алгоритм метода обобщения (для дат/времени):

1. Загрузка ИБД.

2. Выбор признака X^j из ИБД.
3. Выбор вида преобразования: формат преобразования даты/времени или ввод границ диапазонов для дат/времени.
4. Если выбран формат преобразования даты, то

– преобразование исходных значений признака в соответствии с выбранным форматом; формирование ОБД: $X^j_{\text{об.}}$.

иначе (выбран ввод границ диапазонов):

– ввод границ диапазонов: $(x_{k.\text{н}} - x_{k.\text{в}}], k = \overline{1,K}$;
– создание таблицы соответствий *Table_s*, в которой хранятся идентификаторы диапазонов и диапазоны даты/времени;
– создание столбца $X^j_{\text{об.}}$ в ИБД для хранения обезличенных значений;
– заполнение столбца $X^j_{\text{об.}}$ в соответствии с *Table_s*;
– удаление столбца с исходными значениями X^j из ИБД.

5. Сохранение ОБД.

На рисунке 2.14 приведена блок-схема алгоритма обобщения (для количественного признака). Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

Входные данные и параметры алгоритма обобщения (для количественного признака):

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$; выбирается один признак;
– границы диапазонов: $(x_{k.\text{н}} - x_{k.\text{в}}], k = \overline{1,K}$;
– K – количество диапазонов;

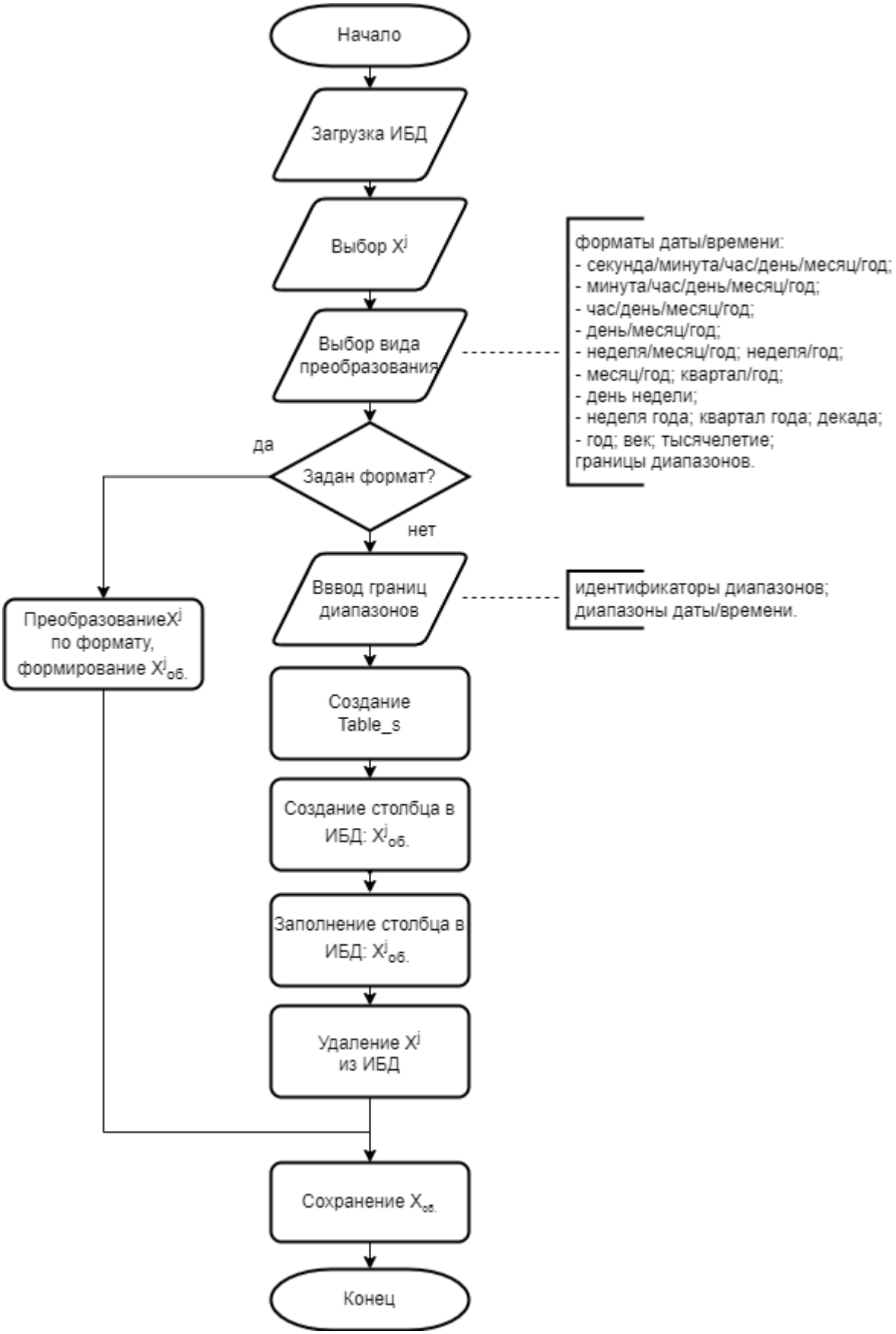


Рисунок 2.13 – Блок-схема алгоритма обобщения (для дат/времени)

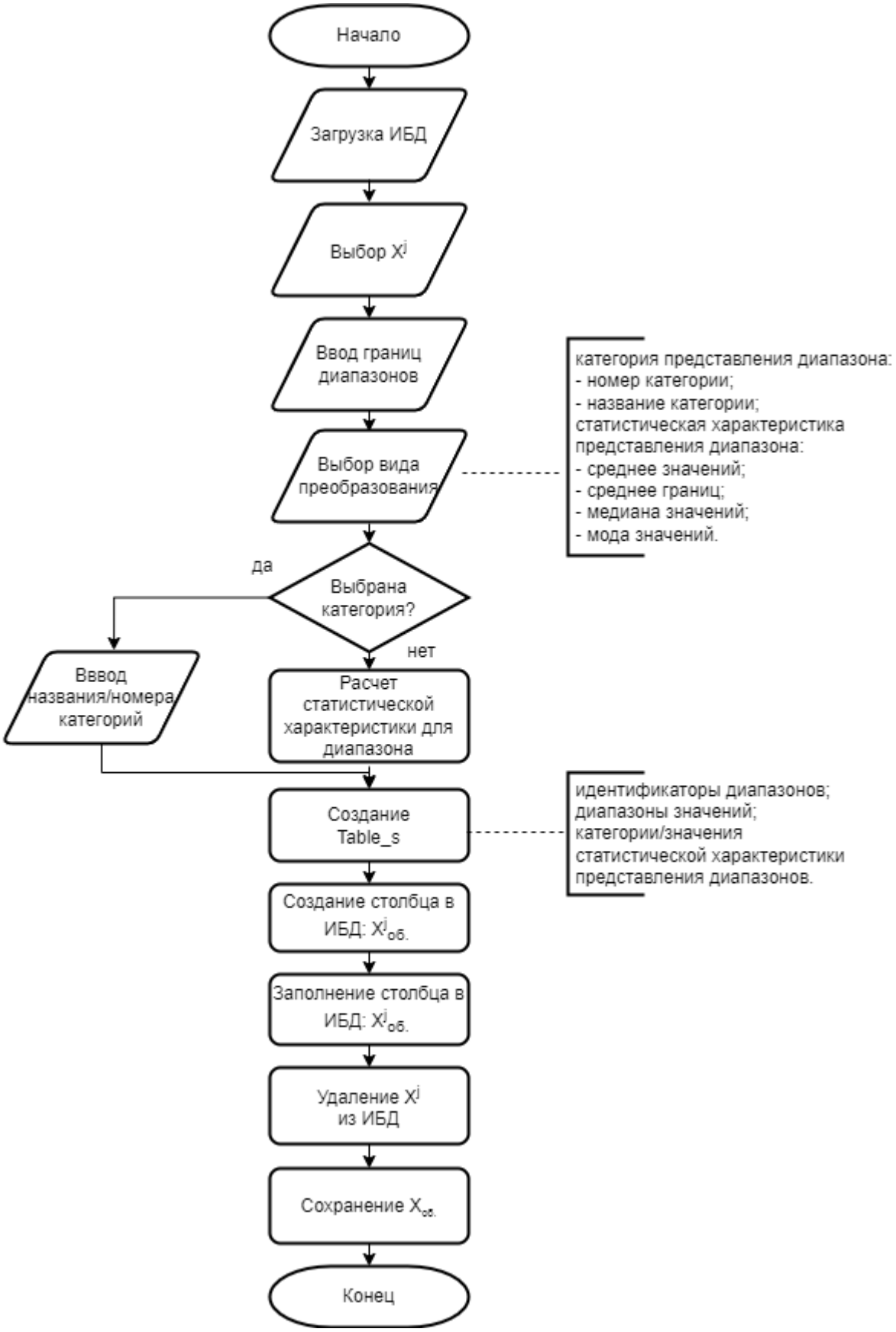


Рисунок 2.14 – Блок-схема алгоритма обобщения
(для количественного признака)

– статистические характеристики и категории представления диапазонов; доступны следующие: среднее значение наблюдений, входящих в интервал; среднее границ интервала; медиана значений, входящих в интервал; мода значений, входящих в интервал; номер категории; название категории.

Алгоритм метода обобщения (для количественного признака):

1. Загрузка ИБД.
2. Выбор признака X^j из ИБД.
3. Ввод границ диапазонов: $(x_{k.\text{н}} - x_{k.\text{в}}],\ k = \overline{1,K}$.
4. Выбор вида преобразования: выбор категории представления диапазона или статистической характеристики представления диапазона.
5. Если выбрана категория представления диапазона, то
   – ввод названия/номера категории для каждого диапазона: $M_k = \overline{1,K}$.

   иначе (выбрана статистическая характеристика):
   – расчет выбранной статистической характеристики для каждого диапазона.
6. Создание таблицы соответствий *Table_s,* в которой хранятся идентификаторы диапазонов, диапазоны значений, категории/значения статистической характеристики представления диапазонов.
7. Создание столбца $X^j_{\text{об.}}$ в ИБД для хранения обезличенных значений.
8. Заполнение столбца $X^j_{\text{об.}}$ в соответствии с *Table_s.*
9. Удаление столбца с исходными значениями X^j из ИБД.
10. Сохранение ОБД.

В случае ординального признака алгоритм метода обобщений аналогичен вышеизложенному алгоритму, но не допустима замена интервального значения признака на статистическую характеристику интервала (среднее, мода, медиана), так как на порядковой шкале не допустимы арифметические операции.

Входные данные и параметры алгоритма обобщения (для классификационного признака):

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$; выбирается один признак;

– множество обобщенных значений признака: $D(X^j_{\text{об.}})$.

Алгоритм метода обобщения (для классификационного признака):

1. Загрузка ИБД.
2. Выбор признака X^j из ИБД.
3. Ввод обобщенного значения признака для каждого оригинального значения признака.
4. Создание таблицы соответствий *Table_s*, в которой хранятся каждое оригинальное исходное значение признака и соответствующее ему обобщенное значение признака.
5. Создание столбца $X^j_{\text{об.}}$ в ИБД для хранения обезличенных значений.
6. Заполнение столбца $X^j_{\text{об.}}$ в соответствии с *Table_s*.
7. Удаление столбца с исходными значениями X^j из ИБД.
8. Сохранение ОБД.

В случае если для исходного значения признака пользователем не введено обобщенное значение, в ОБД будет сохранено исходное значение признака X^j (без изменений).

На рисунке 2.15 приведен пример обезличивания данных с помощью метода обобщения. Исходная БД включает 10 записей, для которых зарегистрированы значения трех признаков: дата заболевания; возраст; диагноз. В ОБД дата заболевания обобщается путем преобразования к формату «месяц/год»; вводятся 5 категорий возраста и обобщаются категории диагноза по МКБ. Для признаков возраст и диагноз формируются таблицы соответствий.

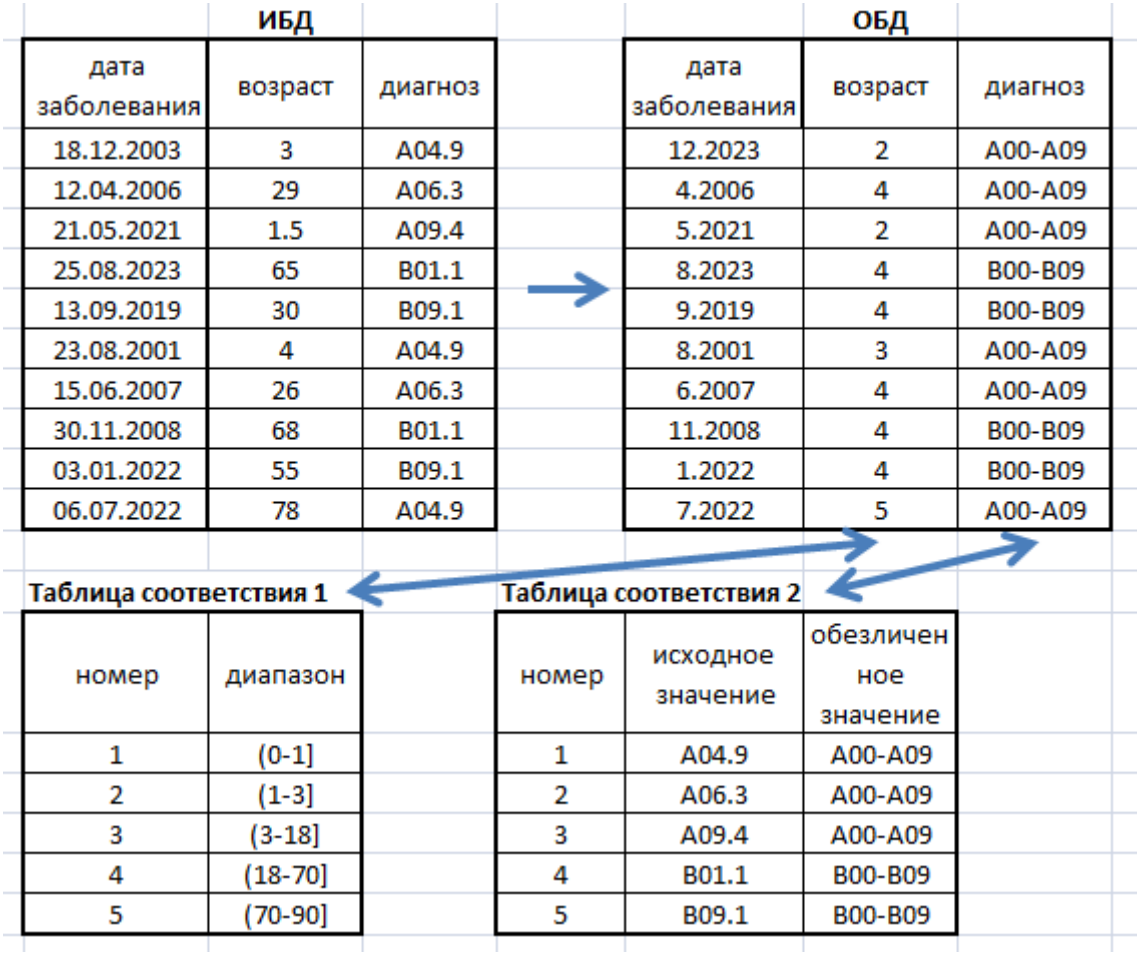


Рисунок 2.15 – Пример применения метода обобщения

2.3.4.2 Обезличивание на основе метода кодирования сверху и/или снизу

Метод кодирования сверху и/или снизу заключается в том, что значения признаков, превышающие или не превышающие некоторый заданный порог соответственно: x_{max}, x_{min} группируются для формирования новой категории. Метод применяется для обезличивания косвенных идентификаторов разных типов: дата/время; порядковые (ординальные); количественные. Этот метод часто реализуется в составе метода обобщений.

На рисунке 2.16 приведена блок-схема алгоритма кодирования сверху и/или снизу. Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

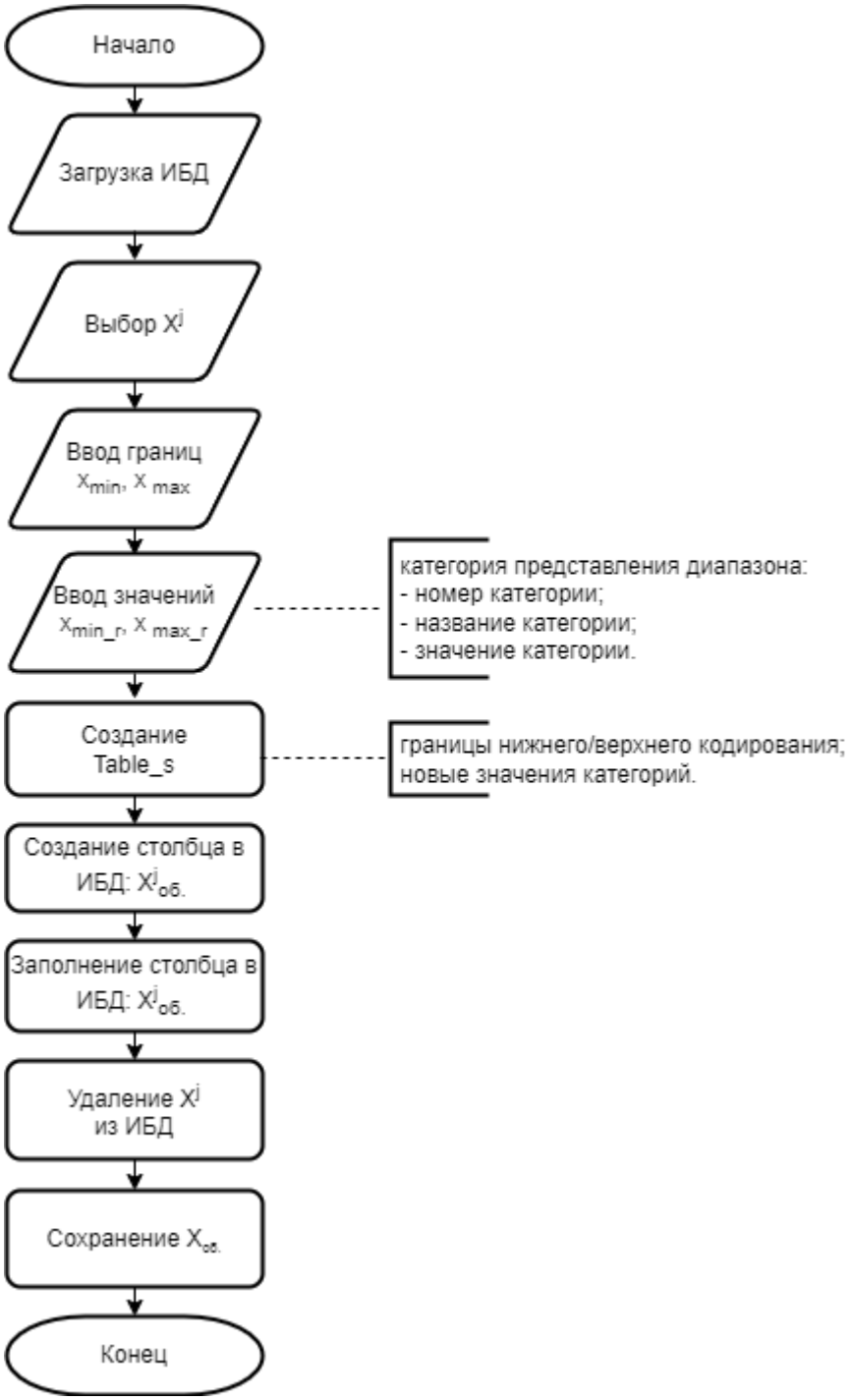


Рисунок 2.16 – Блок-схема алгоритма кодирования сверху и/или снизу

Входные данные и параметры алгоритма кодирования сверху и/или снизу:

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$; выбирается один признак;

– границы нижнего/верхнего кодирования: x_{min}, x_{max};

– значения для новых категорий: $x_{\min\_r},\ x_{\max\_r}$.

Алгоритм метода кодирования сверху и/или снизу:

1. Загрузка ИБД.
2. Выбор признака X^j из ИБД.
3. Ввод границ нижнего/верхнего кодирования: $x_{min},\ x_{max}$.
4. Ввод значений для новых категорий: $x_{\min\_r},\ x_{\max\_r}$. В зависимости от типа значений признака: номер категории; название категории или значение категории. В частном случае, может выполняться замена на пустое значение (фактически реализуется подавление).
5. Создание таблицы соответствий *Table_s,* в которой хранятся границы нижнего/верхнего кодирования и новые значения категорий.
6. Создание столбца $X^j_{\text{об.}}$ в ИБД для хранения обезличенных значений.
7. Заполнение столбца $X^j_{\text{об.}}$ в соответствии с *Table_s.*
8. Удаление столбца с исходными значениями X^j из ИБД.
9. Сохранение ОБД.

На рисунке 2.17 приведен пример обезличивания данных с помощью метода кодирования сверху и/или снизу и метода обобщений. Исходная БД включает 10 записей, для которых зарегистрированы значения признаков: дата заболевания; возраст. В ОБД дата заболевания обобщается путем преобразования к формату «год» и введена нижняя граница кодирования – 2005 (значения ниже этой границы преобразуются в одну категорию: <2005); вводятся 5 категорий возраста, задана верхняя граница кодирования – 70 (если возраст превышает >70, то это соответствует категории 5). Для признака возраст формируется таблица соответствий.

Рисунок 2.17 – Пример применения метода кодирования сверху и/или снизу

2.3.4.3 Обезличивание на основе метода локального подавления

Метод используется в случае наличия в данных выделяющихся наблюдений (экстремальных значений) признаков или экстремальных комбинаций значений признаков. Экстремальные значения или экстремальные комбинации значений подавляются, так как их наличие значительно упрощает процедуру идентификации субъекта, особенно в тех случаях, когда экстремальными являются значения косвенных идентификаторов. Экстремальные значения признаков или экстремальные комбинации значений признаков соответствуют, как правило, уникальным записям в исходной БД. В результате подавления в ИБД всех уникальных записей, ОБД соответствует принципу k-анонимности (k=2).

Используются несколько вариантов метода подавления. Первый вариант сводится к удалению всей записи из исходной БД, содержащей экстремальные значения. Второй вариант предполагает пропуск всех экстремальных значений или комбинаций значений, которые присутствуют в исходных данных (в результате в ОБД появляются пропущенные значения). Третий вариант предполагает их выделение и замену на сглаженные значения. Все варианты метода локального подавления приводят к смещению значений статистических характеристик, рассчитанных по данным. Метод может применяться к любым типам значений

признаков; параметрами метода являются в зависимости от варианта подавления: множество экстремальных значений признака, множество уникальных комбинаций значений признака, либо подавляться могут все уникальные значения признака или все уникальные комбинации значений признака, а также задается способ расчета сглаженных значений (например, замена на среднее значение признака, замена на ближайшее неуникальное значение, замена на введенное пользователем значение).

Рассмотрим модификацию метода подавления, соответствующую подавлению всех уникальных значений или уникальных комбинаций значений признаков ИБД.

На рисунке 2.18 приведена блок-схема алгоритма локального подавления. Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

Входные данные и параметры алгоритма локального подавления:

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$;

– тип подавления: 1,2 или 3;

– способ расчета сглаженных значений: 1 – замена на среднее значение/замена на ближайшее значение – x_m^j; 2 – замена на значение, введенное пользователем – x_r^j.

Алгоритм метода локального подавления:

1. Загрузка ИБД.
2. Выбор признаков X^j из ИБД.
3. Поиск уникальных (экстремальных) значений или уникальных комбинаций значений X^j.

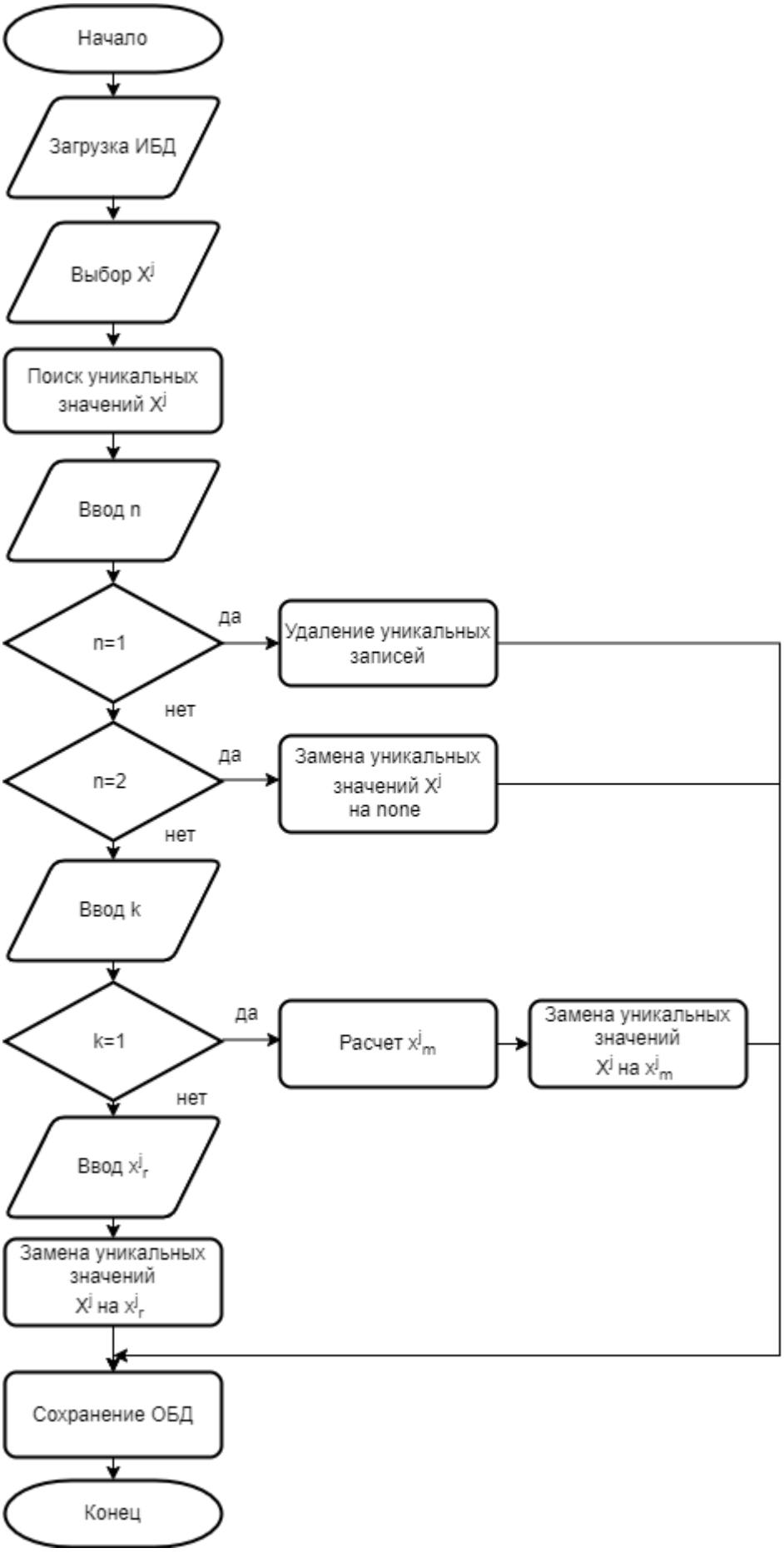


Рисунок 2.18 – Блок-схема алгоритма локального подавления

4. Выбор варианта подавления n:

если n=1: то

– удаление записей, соответствующих уникальным значениям X^j;

если n=2: то

– замена уникальных значений X^j на none;

если n=3: то

– выбор способа расчета сглаженных значений k:

если k=1, то

– расчет среднего значения/ближайшего значения x_m^j;

– замена уникальных значений X^j на x_m^j;

если k=2, то

– ввод значений для замены пользователем x_r^j;

– замена уникальных значений X^j на x_r^j.

5. Сохранение ОБД.

На рисунке 2.19 приведен пример обезличивания данных с помощью локального подавления. Исходная БД включает 10 записей, для которых зарегистрированы значения признаков: дата заболевания; возраст; диагноз. Применен алгоритм локального подавления отдельно для каждого из признаков. Удалена запись, соответствующая уникальной дате заболевания 18.12.2003 (вариант подавления – 1); уникальный диагноз заменен на пустое значение G04.2 (вариант подавления – 2); уникальный возраст 78 заменен на ближайшее неуникальное значение 77 (вариант подавления – 3).

ИБД

дата заболевания	возраст	диагноз
18.12.2003	43	B01.1
12.04.2020	29	A06.3
12.04.2020	43	A06.3
12.04.2020	77	B01.1
13.09.2020	29	B01.1
13.09.2020	77	A04.9
13.09.2020	29	A04.9
13.09.2020	43	A04.9
14.08.2023	77	G04.2
14.08.2023	78	A04.9

ОБД

дата заболевания	возраст	диагноз
12.04.2020	29	A06.3
12.04.2020	43	A06.3
12.04.2020	77	B01.1
13.09.2020	29	B01.1
13.09.2020	77	A04.9
13.09.2020	29	A04.9
13.09.2020	43	A04.9
14.08.2023	77	
14.08.2023	77	A04.9

Рисунок 2.19 – Пример применения метода локального подавления

2.3.4.4 Обезличивание на основе метода микроагрегирования

Метод заключается в объединении ближайших друг к другу по значениям признаков записей ИБД в небольшие группы, размером не менее *k*. Группы формируются по критерию максимального сходства. Метод создает однородные группы записей, рассчитывая попарные расстояния между значениями признаков в двух записях и попарные расстояния между значениями признаков в записи и средними значениями признаков во всех записях в исходном наборе данных. Эта модификация метода микроагрегирования носит название многомерная микроагрегация на основе максимального расстояния до среднего вектора [30]. После того как группы сформированы, для каждой группы рассчитываются средние значения признаков, после чего эти значения используется вместо оригинальных данных для всех единиц данной группы. В результате преобразования обезличенные данные соответствуют принципу *k*-анонимности.

Условия и ограничения использования метода: применяется для обезличивания значений признаков только количественного типа, применяется к косвенным идентификаторам; для реализации алгоритма требуются большие затраты вычислительных ресурсов, наблюдается квадратичный рост объема вычислений с увеличением размера (количества записей) ИБД.

На рисунке 2.20 приведена блок-схема алгоритма метода микроагрегирования. Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

Входные данные и параметры алгоритма:

– $X = \{x_{ij}\}_{i,j=1}^{n,p}$, где x_{ij} – значение j-го атрибута (признака) i-го субъекта ПД в ИБД; n – количество строк, записей (субъектов персональных данных); p – количество признаков; $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ – значения признаков в i-ой записи; $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \ldots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака.

– k – количество элементов в группе микроагрегирования.

Алгоритм метода микроагрегирования (несколько признаков):

1. Загрузка ИБД.

2. Выбор признаков X^j из ИБД, задание параметра k, количества групп микроагрегирования: $g = 0$.

3. Вычисление вектора средних значений признаков $\bar{X}$:

$$\bar{X}=(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p), \tag{2.1}$$

где $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$.

4. Нахождение записи X_r: вычисление квадратов расстояний Евклида между средними значениями признаков $\bar{X}$ и значениями признаков в i-ой записи X_i; нахождение максимального расстояния d_{max}; нахождение r-ой записи X_r, соответствующей максимальному квадратному расстоянию Евклида d_{max}:

$$d(\bar{X}, \bar{X}_l) = \sum_{j=1}^{p}\left(\bar{x}_j - x_{ij}\right)^2, \; i = \overline{1,n}; \tag{2.2}$$

$$d_{max} = \max_i d(\bar{X}, \bar{X}_l);$$

$$d_{max} \to X_r.$$

5. Нахождение записи X_s: вычисление квадратов расстояний Евклида между значениями признаков в r-ой записи X_r и значениями признаков в i-ой записи X_i; нахождение максимального расстояния d_{max}; нахождение s-ой записи X_s, соответствующей максимальному квадратному расстоянию Евклида d_{max}:

$$d(X_r, \bar{X}_l) = \sum_{j=1}^{p} (x_{rj} - x_{ij})^2, i = \overline{1,n}; \quad (2.3)$$

$$d_{max} = \max_i d(X_r, \bar{X}_l);$$

$$d_{max} \to X_s.$$

6. Формирование двух групп микроагрегирования вокруг записей X_r и X_s соответственно: $g = g + 2$. Одна группа включает запись X_r и $k - 1$ ближайших к ней записей по метрике квадрата расстояния Евклида, другая группа включает запись X_s и $k - 1$ ближайших к ней записей по метрике квадрата расстояния Евклида.

7. Формирование нового множества исходных данных X, в котором исключены записи, вошедшие в группы на предыдущем шаге: $n = n - 2k$.

8. Если $n \leq 3k$, то переход на второй шаг алгоритма, иначе – переход на следующий шаг (шаг 9).

9. Если $2k \leq n \leq 3k - 1$, то

- вычисление вектора средних $\bar{X}$;
- нахождение записи X_r, которая наиболее удалена от $\bar{X}$ согласно мере квадрата расстояния Евклида;
- формирование группы микроагрегирования вокруг записи X_r: $g = g + 1$;
- формирование группы микроагрегирования из оставшихся записей: $g = g + 1$;
- переход на шаг 10;

иначе – переход на шаг 10.

10. Если $n < 2k$, то формирование группы из оставшихся записей и переход на следующий шаг алгоритма (шаг 11).

11. Расчет векторов средних значений признаков в каждой из g групп:

$$\bar{X}_i = (\bar{x}_{i1}, \bar{x}_{i2}, \dots, \bar{x}_{ip}), i = \overline{1,g}. \quad (2.4)$$

12. Замена оригинальных значений признаков ИБД на средние значение признака по группе, формирование ОБД: $X_{\text{об.}}$.

13. Сохранение ОБД.

В частном случае алгоритм применяется к одному признаку ИБД, в этом случае алгоритм упрощается и наблюдается линейный рост объема вычислений от размера ИБД.

Входные данные и параметры алгоритма:

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$; выбирается один признак;

– k – количество элементов в группе микроагрегирования.

Алгоритм метода микроагрегирования (один признак):

1. Загрузка ИБД.

2. Выбор признака X^j из ИБД, задание параметра k, количество групп микроагрегирования: $g = 0$.

3. Формирование $X^{j\prime}$, в котором значения X^j отсортированы в порядке возрастания: $x'_{1j} \le x'_{2j} \le \dots \le x'_{nj}$.

4. Формирование группы микроагрегирования их первых k – значений $X^{j\prime}$: $g = g + 1$, удаление первых k – значений из $X^{j\prime}$: $n = n - k$;

5. Если $n > k$, то переход к шагу 3, иначе – формирование группы микроагрегирования из оставшихся элементов: $g = g + 1$ и переход к шагу 6.

6. Вычисление средних значений признака по каждой группе микроагрегирования: $\bar{X}$=($\bar{x}_1, \bar{x}_2, \dots, \bar{x}_g$).

7. Замена оригинальных значений признака X^j на средние значения признака по группе, формирование ОБД: $X^j_{\text{об.}}$.

8. Сохранение ОБД.

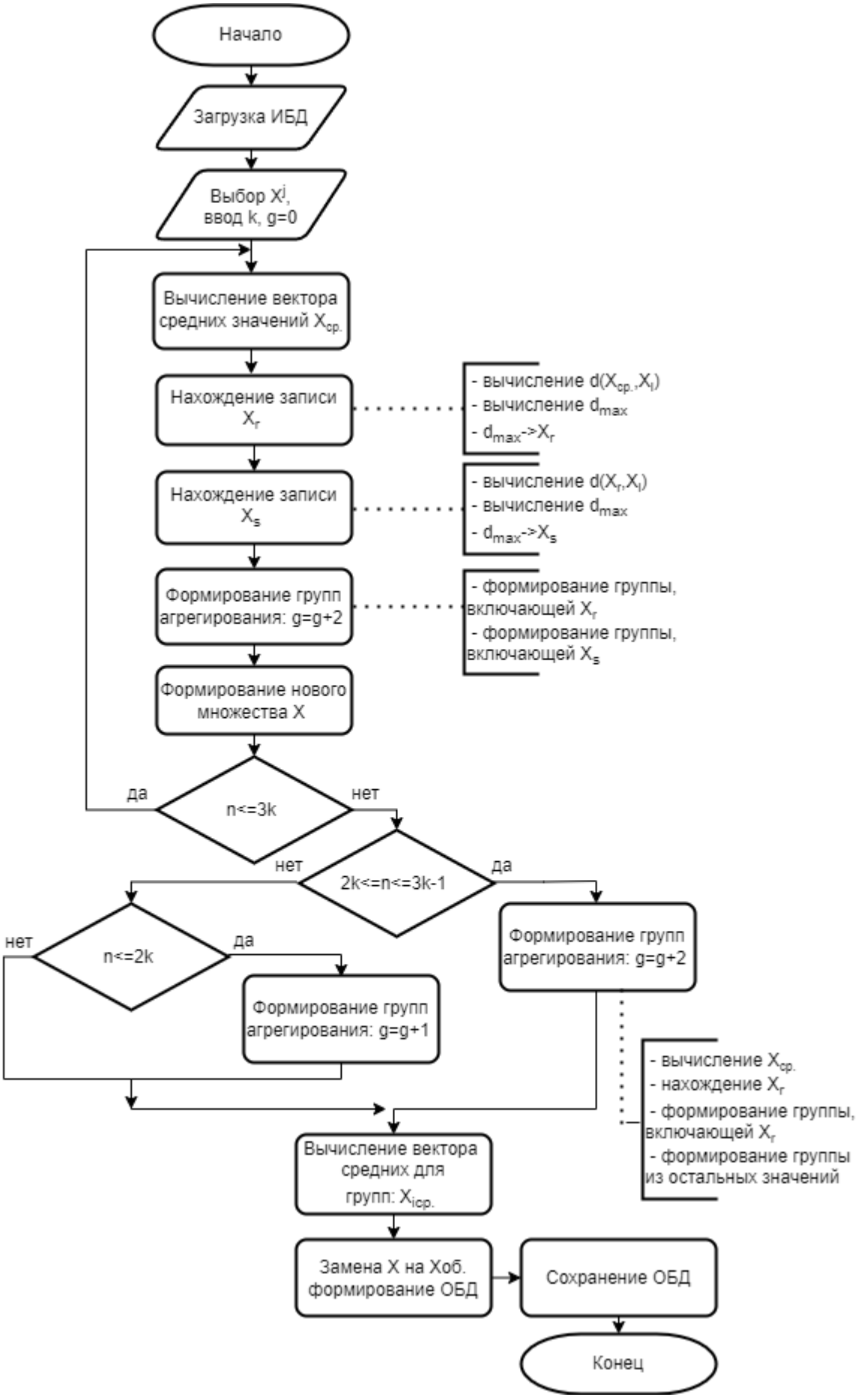


Рисунок 2.20 – Блок-схема алгоритма микроагрегирования

На рисунке 2.21 приведен пример обезличивания данных с помощью метода микроагрегирования. Исходная БД включает 10 записей, для которых зарегистрированы значения трех признаков: возраст; стаж; доход. Задан параметр $k=3$, т.е. каждая группа микроагрегирования включает не менее 3-х записей. В результате работы алгоритма сформировано три группы, состоящие соответственно из 3,3 и 4 записей. Для каждой группы выполнена замена исходных значений признаков на средние значения по группе.

ИБД

возраст	стаж	доход
25	2	50
29	3	120
38	10	100
65	21	70
30	6	90
59	15	60
26	1	30
68	30	80
55	30	60
31	14	150

Группировка

возраст	стаж	доход
25	2	50
29	3	120
38	10	100
65	21	70
30	6	90
59	15	60
26	1	30
68	30	80
55	30	60
31	14	150

ОБД

возраст	стаж	доход
36.67	6	46.67
32.67	9	123.33
32.67	9	123.33
54.5	21.75	75
54.5	21.75	75
36.67	6	46.67
36.67	6	46.67
54.5	21.75	75
54.5	21.75	75
32.67	9	123.33

Рисунок 2.21 – Пример работы алгоритма микроагрегирования

2.3.4.5 Обезличивание на основе метода добавления шума

Метод применяется к количественным признакам, также можно адаптировать метод для модификации значений атрибута типа дата/время; применяется к косвенным идентификаторам. Суть метода сводится к добавлению к истинному значению признака нормально распределенного «шума», и последующей замене истинного значения признака на модифицированное значение. В качестве входных параметров метода используются параметры нормального закона распределения, описывающего «шум»: среднее значение задается всегда равным нулю; дисперсия (или среднеквадратическое отклонение) определяет вклад шумовой составляющей. Чем больше дисперсия, тем больше разброс «шума» и больше степень искажения исходной информации. Возможны и другие законы распределения шумовой составляющей (например, равномерное распределение). В этом случае задаются параметры выбранного закона распределения (например,

для равномерного закона распределения: начальное и конечное значения интервала). Однако, на практике, как правило, используют нормальный закон распределения «шума».

Метод добавления шума к дате (метод старения дат) сводится к изменению исходного значения даты на заданное количество дней. Пользователь задает целое число (положительное или отрицательное), которое определяет, на сколько дней сместится исходная дата. Возможен также вариант, когда число, определяющее смещение даты, генерируется случайным образом в соответствии с заданным законом распределения.

На рисунке 2.22 приведена блок-схема алгоритма метода добавления шума. Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

Входные данные и параметры алгоритма добавления шума:

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$; выбирается один признак;

– law – закон распределения шума и его параметры (σ – среднеквадратическое отклонение для нормального закона распределения; a и b – нижняя и верхняя границы для равномерного закона).

Алгоритм метода добавления шума:

1. Загрузка ИБД.
2. Выбор признака X^j из ИБД.
3. Выбор закона распределения шума – law, ввод параметров закона распределения – σ (для нормального закона); a и b (для равномерного закона).
4. Генерация шумовой составляющей в соответствии с выбранным законом распределения: $\varepsilon_i, i = \overline{1,n}$.
5. Вычисление модифицированных значений: $x_{ij\text{об}} = x_{ij} + \varepsilon_i,\ i = \overline{1,n}, j = \overline{1,p}$.

6. Замена исходных значений x_{ij} признака X^j на модифицированные значения $x_{ijоб}$.
7. Сохранение ОБД.

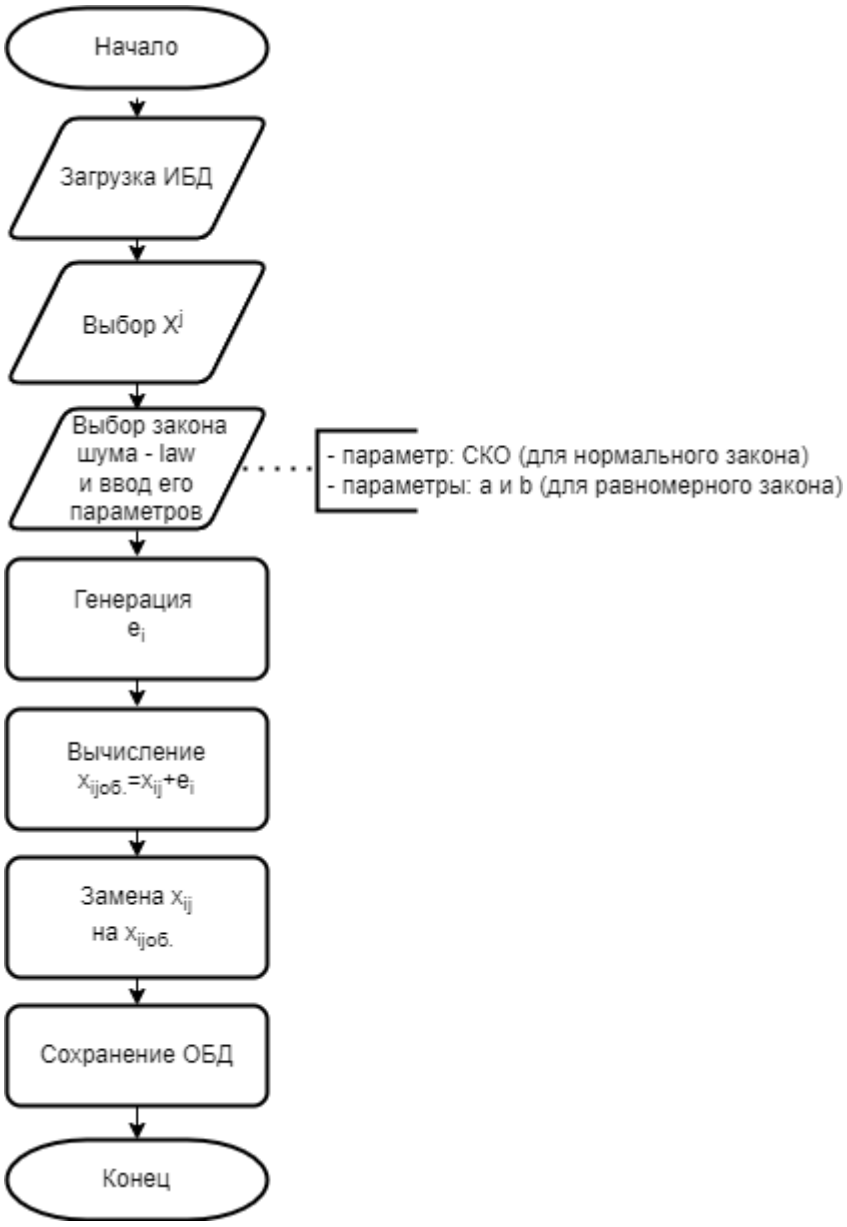


Рисунок 2.22 – Блок-схема алгоритма добавления шума

На рисунке 2.23 приведен пример работы метода добавления шума. Обезличиваются значения признаков возраст и дата заболевания. Шумовая составляющая для признака возраст генерируется по нормальному закону распределения с нулевым средним и дисперсией равной 1. Итоговое модифицированное значе-

ние округляется до целого. Дата заболевания изменяется на равномерно распределенное значение в диапазоне [-2;2].

ИБД

дата заболевания	возраст
18.12.2020	28
12.04.2020	13
15.04.2020	45
17.04.2020	67
13.09.2020	31

ОБД

дата заболевания	возраст
16.12.2020	29
14.04.2020	14
13.04.2020	47
18.04.2020	66
13.09.2020	30

Рисунок 2.23 – Пример работы метода добавления шума

2.3.4.6 Обезличивание на основе метода округления

Метод применяется только к количественным признакам, относящимся к косвенным идентификаторам. Суть метода сводится к округлению истинного значения признака, и последующей замене истинного значения признака на модифицированное значение. В качестве входного параметра метода задается точность округления: количество знаков после запятой в представлении числа.

На рисунке 2.24 приведена блок-схема алгоритма метода округления. Ниже рассмотрены и описаны основные этапы (шаги) алгоритма.

Входные данные и параметры алгоритма округления:

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$; выбирается один признак;

– r – точность округления.

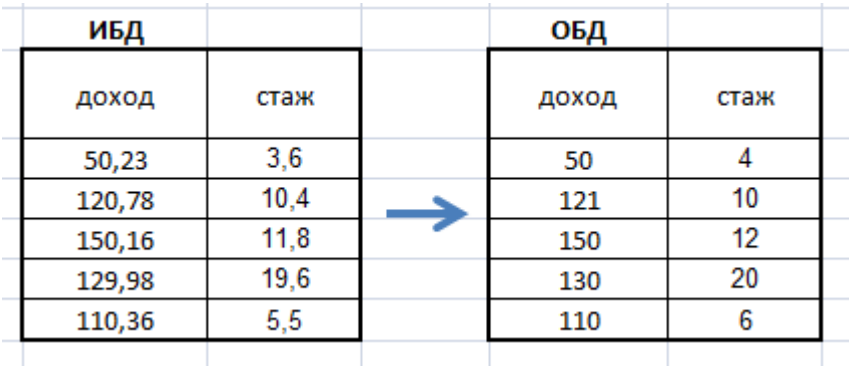
Алгоритм метода добавления шума:

1. Загрузка ИБД.
2. Выбор признака X^j из ИБД.
3. Ввод параметра r – точность округления.

4. Округление исходных значений x_{ij} признака X^j и формирование модифицированных значений: $x_{ij\text{об}}$
5. Замена исходных значений x_{ij} признака X^j на модифицированные значения $x_{ij\text{об}}$.
6. Сохранение ОБД.

Рисунок 2.24 – Блок-схема алгоритма округления

На рисунке 2.25 приведен пример работы метода округления. Обезличиваются значения признаков стаж работы и доход. Применяется округление до целого значения.



ИБД

доход	стаж
50,23	3,6
120,78	10,4
150,16	11,8
129,98	19,6
110,36	5,5

ОБД

доход	стаж
50	4
121	10
150	12
130	20
110	6

Рисунок 2.25 – Пример работы метода округления

2.3.4.7 Обезличивание на основе метода маскирования

Метод маскирования выполняет обезличивание данных путем замены значений исходных признаков на заданные по образцу значения. Можно предложить разные варианты реализации метода маскирования: замена на введенное пользователем значение; замена на значение из файла; замена согласно заданному регулярному выражению. Маскироваться могут часть символов в значении исходного признака, находящихся в определенной позиции, либо может выполняться замена исходных значений признака на новые значения, введенные пользователем. В частном случае, новые значения хранятся в отдельном файле, из которого происходит считывание информации для замены. В качестве параметров метода в зависимости от модификации задаются: позиции символов для замены; новые символы; новые значения признаков, либо название файла с информацией для замены.

Также метод используется для маскирования значений признаков путем нахождения в них регулярных выражений и замены их на символы, введенные пользователем. В этом случае параметрами метода выступают: регулярное выражение (паттерн), по которому будут искаться символы в значениях признака; символы, на которые будут заменяться найденные символы.

Входные данные и параметры алгоритма маскирования (для замены по паттерну):

– $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix}$ – значения j-ого признака, $j = \overline{1,p}$;

– *pattern* – регулярное выражение (паттерн), по которому будут искаться символы;

– *replacement* – символы, на которые будут заменяться найденные символы.

Алгоритм метода маскирования:

1. Загрузка ИБД.
2. Выбор признаков X^j из ИБД.
3. Ввод регулярного выражения для замены – *pattern* и символов для замены – *replacement*;
4. Замена символов в исходных значениях признака X^j на символы *replacement* в соответствии с *pattern*.
5. Сохранение ОБД.

В других модификациях метода алгоритм аналогичен, меняются только входные данные (задаются пользователем новые значения признака, либо задается файл, хранящий информацию для замены).

На рисунке 2.26 приведен пример работы метода маскирования в разных модификациях. Маскируются значения признаков: город; адрес проживания; код диагноза. В коде диагноза маскируются два последних символа, путем их замены на символ «*», в результате реализуется переход к более высокому уровню классификатора болезней. В адресе проживания маскируется цифры согласно регулярному выражению: '[0-9]+', найденные числовые последовательности заменяются на символ *. В признаке город обезличивается информация путем замены исходного значения на none.

ИБД

id	город	адрес	диагноз
1	Новосибирск	Чигорина 10-23	A04.9
2	Новосибирск	Линейная 1-81	A06.3
3	Новосибирск	Шатурская 12-34	A09.4
4	Бердск	Линейная 7-33	B01.1
5	Бердск	Весенняя 11-25	B09.1
6	Бердск	Лесная 21-55	A04.9

ОБД

id	город	адрес	диагноз
1	none	Чигорина *-*	A04**
2	none	Линейная *-*	A06**
3	none	Шатурская *-*	A09**
4	none	Линейная *-*	B01**
5	none	Весенняя *-*	B09**
6	none	Лесная *-*	A04**

Рисунок 2.26 – Пример работы метода маскирования

2.3.4.8 Обезличивание на основе метода выборки и метода удаления

Применение методов выборки и удаления приводит к сокращению объема (размера) обезличенной БД по отношению к исходной БД.

Метод выборки сводится к формированию случайной выборки данных из исходной БД. Использование метода не приводит к обезличиванию данных и требует последующего применения методов обезличивания к вновь сформированной БД. В качестве параметра метода задается объем (доля) выборки относительно ИБД.

Метод удаления заключается в полном удалении признака (признаков) из исходной БД, в обезличенной БД признак (признаки) не представлены. Метод используется для удаления прямых идентификаторов.

2.3.5 Свойства обезличенных данных

Требуемые свойства обезличенных данных перечислены в приказе Роскомнадзора [1,8]. Обезличенные данные должны обладать характеристиками (свойствами), обеспечивающими их дальнейшую обработку.

К свойствам обезличенных данных относятся:

– полнота (сохранение всей информации о конкретных субъектах или группах субъектов, которая имелась до обезличивания);

– структурированность (сохранение структурных связей между обезличенными данными конкретного субъекта или группы субъектов, соответствующих связям, имеющимся до обезличивания);

– релевантность (возможность обработки запросов по обработке персональных данных и получения ответов в одинаковой семантической форме);

– семантическая целостность (сохранение семантики персональных данных при их обезличивании);

– применимость (возможность решения задач обработки персональных данных, стоящих перед Оператором, осуществляющим обезличивание персональных данных, обрабатываемых в информационных системах персональных данных, без предварительного деобезличивания всего объема записей о субъектах);

– анонимность (невозможность однозначной идентификации субъектов данных, полученных в результате обезличивания, без применения дополнительной информации) [1,8].

2.3.6 Свойства методов обезличивания данных

Наличие перечисленных свойств обеспечивается применяемыми методами ОД. В таблице 2.2 приведено соотношение методов обезличивания и свойств данных, которыми обладают ОД в результате применения метода.

Таблица 2.2 – Методы обезличивания и свойства обезличенных данных

Свойства обезличенных данных	**Метод введения идентификаторов**	**Метод изменения состава или семантики**	**Метод декомпозиции**	**Метод перемешивания**
Полнота	Да	Частично	Да	Да
Структурированность	Да	Да	Да	Да
Релевантность	Частично	Да	Да	Да
Семантическая целостность	Да	Частично	Да	Да
Применимость	Да	Да	Да	Да
Анонимность	Частично	Да	Частично	Да

К характеристикам (свойствам) методов обезличивания ПД, определяющими возможность обеспечения заданных свойств ОД, относятся [1,8]:

– обратимость (возможность преобразования, обратного обезличиванию (деобезличивание), которое позволит привести обезличенные данные к исходному виду, позволяющему определить принадлежность персональных данных конкретному субъекту, устранить анонимность);

– вариативность (возможность внесения изменений в параметры метода и его дальнейшего применения без предварительного деобезличивания массива данных); изменяемость (возможность внесения изменений (дополнений) в массив обезличенных данных без предварительного деобезличивания);

– стойкость (стойкость метода к атакам на идентификацию субъекта ПД);

– возможность косвенного деобезличивания (возможность проведения деобезличивания с использованием информации других Операторов);

– совместимость (возможность интеграции персональных данных, обезличенных различными методами);

– параметрический объем (объем дополнительной (служебной) информации, необходимой для реализации метода обезличивания и деобезличивания);

– возможность оценки качества данных (возможность проведения контроля качества обезличенных данных и соответствия применяемых процедур обезличивания установленным для них требованиям).

– практическая реализуемость.

2.3.7 Оценка временных показателей обезличивания данных

Для решения задачи обезличивания данных разработаны методы разных классов (введение идентификаторов, декомпозиция, перемешивание, изменение состава или семантики), причем в рамках каждого класса, в свою очередь, также разработаны различные методы и варианты их реализации. Алгоритм реализации метода может быть оценен с точки зрения временных (скоростных) затрат на его выполнение. Затем, по временным (скоростным) показателям могут быть сравнены разные методы обезличивания и/или варианты реализации одного метода.

Рассмотрим основные временные (скоростные) показатели алгоритма реализации метода обезличивания:

$T_i, i = \overline{1,m}$ – время обезличивания БД объемом n записей с помощью i-го метода; m – количество методов;

$M_i = \frac{T_i}{n}, i = \overline{1,m}$ – время обезличивания одной записи БД i-ым методом;

$V_i = \frac{n}{T_i}, i = \overline{1,m}$ – скорость обезличивания БД (количество записей в единицу времени) i-ым методом.

В процессе обезличивания БД, как правило, применяется комплекс методов, так как для обезличивания разных типов признаков БД требуется использовать соответственно разные методы. Поэтому вводится суммарный показатель времени обезличивания всей БД при использовании m методов обезличивания:

$$T = \sum_{i=1}^{m} T_i . \quad (2.5)$$

Для оценки временной сложности алгоритма также вводят показатель *Big O*, который используется для оценки верхней границы (наихудшего случая) времени работы алгоритма. *Big O* показывает скорость роста времени исполнения алгоритма от размера входящих данных (объема записей БД).

Алгоритмы могут иметь разную сложность: *O(log n)* – логарифмическая сложность; *O(n)* – линейная сложность; *O(n log n)* – линейно-логарифмическая сложность; *O(n^2)* – квадратичная сложность; *O(n^3)* – кубическая сложность; *O(n!)* – факториальная сложность.

Временные (скоростные) показатели алгоритма реализации метода обезличивания зависят не только от самого алгоритма, но и от характеристик компьютера, а именно от производительности процессора и времени записи на жесткий диск. При проведении серии экспериментов показатели будут варьироваться, даже в случае сохранения неизменных условий эксперимента (одна БД, один компьютер). Отклонения (шум) в оценке показателей связаны с загрузкой операционной системы (нагрузка на процессор, загрузка оперативной памяти и жесткого диска). Таким образом, временной (скоростной) показатель работы алгоритма является случайной величиной, что позволяет задать его закон распреде-

ления и ввести статистические характеристики показателя (среднее, дисперсия, минимальное и максимальное значения и т.д.).

2.4 Этап оценки риска раскрытия информации

2.4.1 Показатели риска раскрытия информации

Оценка риска раскрытия информации реализуется для определения степени защиты конфиденциальной информации о субъекте. Риск раскрытия информации может сохраняться после проведения процедуры обезличивания данных, поэтому этот этап включен в методику обезличивания данных, как один из основных. Оценка риска раскрытия информации позволяет сделать выводы об эффективности реализации процедуры обезличивания данных в целом и принять решение о возможности публикации (передачи) и использования ОБД. Исходя из полученных оценок, планируют дальнейшие мероприятия, связанные с защитой данных.

В ИБД могут содержаться как прямые идентификаторы – признаки, которые однозначно идентифицируют субъекта (респондента, пациента, клиента, …), например: ФИО, СНИЛС, ИНН и т.п., так и косвенные идентификаторы – признаки, которые идентифицируют субъекта с той или иной степенью неопределенности. Тем не менее, комбинация косвенных идентификаторов может привести к однозначной идентификации личности. Например, сочетание признаков пол, возраст, район проживания позволяет с большой степенью точности идентифицировать человека. В случае если в БД присутствуют прямые идентификаторы, риск раскрытия информации принимается за 100%. Поэтому прямые идентификаторы любо удаляются, либо обезличиваются с помощью метода введения идентификаторов. После обработки прямых идентификаторов в БД остаются еще косвенные идентификаторы, которые могут использоваться для раскрытия конфиденциальной информации.

Под раскрытием подразумевается повторная идентификация субъекта, которая выполняется злоумышленником. Для каждого субъекта (записи в наборе данных) назначается вероятность успешной идентификации: $\theta_i, i = 1, \dots, n,$ где n

– общее число записей в БД, которая называется индивидуальным риском раскрытия информации [31-34]. На основе этой меры разработан набор производных показателей риска раскрытия информации [31].

В общем случае обезличенная БД рассматривается как случайная выборка *s* объема *n*, выбранная из конечной генеральной совокупности *P*, состоящей из *N* единиц [31-33]. Для любой единицы (записи) *i* из генеральной совокупности, вероятность ее включения в выборку составляет $1/w_i$, где w_i – вес *i*-ой записи. Рассмотрим многомерную таблицу сопряженности, построенную путем сочетания ключевых признаков. Комбинация *j* определяется как *j*-я ячейка в таблице сопряженности. Набор комбинаций *{1, ..., j, ..., J}* определяет разбиение, как генеральной совокупности, так и выборки на ячейки. Значения ключевых признаков для отдельной записи $i \in s$ классифицирует запись в одну из ячеек таблицы сопряженности. Обозначим через *j(i)* индекс ячейки, в которую классифицируется отдельная запись $i \in s$ на основе значений ключевых признаков. Как правило, находится несколько выборочных единиц в пределах одной и той же комбинации *j*. Рассматривается каждая из *j=1, ..., J* ячеек таблицы сопряженности. Пусть f_j и F_j обозначают, соответственно, количество записей в обезличенной БД и количество записей в генеральной совокупности с *j*-й комбинацией значений ключевых признаков; в общем случае F_j неизвестно для каждого *j*. В зависимости от выбранных ключевых признаков общее число *J* комбинаций может быть довольно большим; в обезличенной БД будет наблюдаться только подмножество *J*, и только это подмножество комбинаций, для которых $f_j>0$, представляет интерес для задачи оценки риска раскрытия информации. В дальнейшем под *J* будем понимать количество ненулевых комбинаций сочетаний признаков или количество классов эквивалентности (одинаковых записей по значениям признаков). Все записи в рамках одного класса эквивалентности имеют одинаковую вероятность повторной идентификации, поэтому можно рассматривать вероятность идентификации для каждого класса эквивалентности: $\theta_j, j = 1, \dots, J$.

Частотные характеристики БД, связанные с обезличиванием БД

В качестве общих (частотных) характеристик, описывающих БД с точки зрения степени обезличивания можно выделить следующие:

– J – количество классов эквивалентности;

– минимальный размер (частота) класса эквивалентности;

– максимальный размер (частота) класса эквивалентности;

– средний размер (частота) класса эквивалентности;

– количество уникальных записей;

– процент уникальных записей;

– показатель *k*-анонимности (*k*=2,3,5).

База данных обладает *k*-анонимностью, если для каждого субъекта БД имеется, по меньшей мере, $k-1$ субъект, обладающий такими же значениями признаков. Любая запись БД неотличима, по крайней мере, от ($k-1$) других записей по совокупности значений признаков.

Базовые метрики риска раскрытия информации

Все производные метрики оценки риска, основанные на расчете θ_j, измеряются в диапазоне от 0 до 1 или в процентном соотношении для удобства интерпретации. Рассмотрим подробнее основные метрики [31].

1. Оценка доли записей, вероятность повторной идентификации которых выше порогового значения τ:

$$R_a = \frac{1}{n}\sum_{j=1}^{J} f_j \times I(\theta_j > \tau), \tag{2.6}$$

где $I(.)$ – индикаторная функция, возвращающая значение 0 («ложь») или 1 («истина»). Если значение R_a слишком велико, считается, что ОБД имеет высокий риск повторной идентификации, что неприемлемо. Обозначим порог максимальной доли записей, имеющих высокую вероятность повторной идентификации, через α. Следовательно, решающим правилом для R_a является:

$$D_a = \begin{cases} high,\ R_a > \alpha; \\ low,\ R_a \le \alpha. \end{cases} \tag{2.7}$$

2. Максимальный риск раскрытия информации среди записей БД (оценка худшего сценария):

$$R_b = \max_{j \in J}(\theta_j). \quad (2.8)$$

Правило принятия решения в этом случае:

$$D_b = \begin{cases} high,\ R_b > \tau; \\ low,\ R_b \leq \tau. \end{cases} \quad (2.9)$$

3. Оценка среднего риска повторной идентификации записи (мера глобального риска):

$$R_c = \frac{1}{n}\sum_{j=1}^{J} f_j \theta_j. \quad (2.10)$$

Общее ожидаемое число повторных идентификаций по всей ОБД составляет: $\sum_{j=1}^{J} f_j \theta_j$. Мера (2.10) не зависит от размера ОБД и может быть использована для оценки риска раскрытия информации по всей БД или для сравнения различных типов раскрытия. Процент ожидаемых повторных идентификаций обеспечивает эквивалентную меру глобального риска: $R_c * 100\%$.

Также задается максимальная доля записей γ, которые могут быть успешно идентифицированы, правило принятия решения для метрики:

$$D_{\mathrm{c}} = \begin{cases} high,\ R_{\mathrm{c}} > \gamma; \\ low,\ R_{\mathrm{c}} \leq \gamma. \end{cases} \quad (2.11)$$

Выбираются пороговое значение показателей риска $R_{max} = \{\alpha, \tau, \gamma\}$, и если превышены заданные значения, то делается вывод о высоких рисках раскрытия информации, недостаточной степени защиты конфиденциальных данных в обезличенной БД и неэффективности процедуры обезличивания в целом.

Показатели риска раскрытия информации в зависимости от типа атаки злоумышленника

Разработано два основных подхода к оценке вероятности успешной повторной идентификации *i*-го субъекта. Главный критерий, по которому они различаются, заключается в том, может ли злоумышленник знать, находится ли конкретный субъект (лицо) в обезличенной БД [31,35-36]. В зависимости от этого различаются типы атак, и отличаются формулы расчета базовых показателей риска раскрытия информации (см. таблицу 2.3).

Атака прокурора

Злоумышленник знает, что цель находится в обезличенной БД и мишенью является конкретный субъект. В каких случаях злоумышленник знает, что цель находится в обезличенной БД? В случае если ОБД представляет собой всю генеральную совокупность, либо ОБД является выборкой из внешней БД, но субъект сам сообщает, что он является ее частью (например, сообщает в социальных сетях или знакомым о том, что участвует в исследованиях, результаты которого опубликованы и находятся в открытом доступе).

Атака журналиста

Злоумышленник не знает, находится ли цель в обезличенной БД, однако у журналиста есть доступ к общедоступной (внешней) БД. Существует два основных типа атак, связанных с риском журналиста: первый – злоумышленник нацеливается на конкретный субъект; второй – злоумышленник нацеливается на любой субъект. В первом случае злоумышленник имеет фоновые знания о конкретном субъекте, тогда как во втором – злоумышленнику все равно, какой человек является целью.

В случае оценки журналистского риска под раскрытием подразумевается повторная идентификация субъекта, которая выполняется злоумышленником при сравнении данных целевого субъекта в обезличенной БД с данными субъекта во внешней БД (ВБД), которая содержит прямые идентификаторы, такие как, например, ФИО и СНИЛС. Повторная идентификация происходит, когда запись в ОБД и запись во внешней БД, к которой злоумышленник имеет доступ, принадлежат одному и тому же лицу.

Таблица 2.3 – Формулы для расчета показателей риска раскрытия информации

Вид атаки	Расчетная формула	Пояснение
Прокурора	$R_{p(a)} = \frac{1}{n}\sum_{j=1}^{J} f_j \times I\left(\frac{1}{f_j} > \tau\right)$ $R_{p(b)} = \frac{1}{\min_{j \in J}(f_j)}$ $R_{p(c)} = \frac{J}{n}$	Злоумышленник нацелен на конкретного субъекта и знает, что данные об этом субъекте содержатся в ОБД.
Журналиста	$R_{j(a)} = \frac{1}{n}\sum_{j=1}^{J} f_j \times I\left(\frac{1}{F_j} > \tau\right)$	Злоумышленник нацелен на кон-

	$R_{j(b)} = \frac{1}{\min_{j \in J}(F_j)}$ $R_{j(c)} = max\left(\frac{J}{\sum_{j=1}^{J} F_j}, \sum_{j=1}^{J} \frac{f_j}{F_j}\right)$	кретного субъекта, но не знает, есть ли данные о субъекте в ОБД.
ВБД известна	F_j – реальная частота j-го класса эквивалентности в ВБД	
ВБД не известна	$\widehat{F_j}$ – оценка частоты j-го класса эквивалентности в ВБД	
Маркетолога	$R_{m1(c)} = \frac{J}{N}$ $R_{m2(c)} = \sum_{j=1}^{J} \frac{f_j}{F_j}$	Злоумышленник не нацелен на конкретного субъекта, а стремится повторно идентифицировать как можно большее число субъектов.
ВБД известна	F_j – реальная частота j-го класса эквивалентности в ВБД	
ВБД не известна	$\widehat{F_j}$ – оценка частоты j-го класса эквивалентности в ВБД	

Атака маркетолога

Злоумышленник не нацелен на конкретного человека, а стремится повторно идентифицировать большое количество людей. Таким образом, атаку можно считать успешной только в том случае, если удалось повторно идентифицировать большую часть записей.

На практике, как правило, нет доступа к внешней БД и не известны частоты F_j для расчета показателей риска журналиста и маркетолога. Поэтому риск оцениваться только с использованием информации из ОБД и задается только доля ОБД относительно ВБД. Для этого разработаны различные методы [32,33,37-39]. Более подробно показатели риска раскрытия информации рассмотрены в отчете по проекту 2.5.2.5 «Формирование требований к методике автоматизированного обезличивания данных».

Между разными типами показателей риска существуют взаимосвязи [31]:

$$R_{p(b)} \geq R_{p(c)} \geq R_{j(c)} \geq R_{m1(c)} \geq R_{m2(c)},$$

$$R_{p(b)} \geq R_{j(b)} \geq R_{j(c),} \quad (2.12)$$
$$R_{p(a)} \geq R_{j(a)}.$$

Если принято решение о том, что необходимо учитывать более одного типа риска, то соотношения (2.12) могут помочь решить, какие показатели фактически вычислять и какими рисками управлять.

2.4.2 Алгоритм оценки риска раскрытия информации

На рисунке 2.27 приведена блок-схема алгоритма оценки риска раскрытия информации. Рассмотрим более подробно этапы алгоритма.

1. Загрузка ИБД.

2. Обезличивание, формирование ОБД.

Риск раскрытия информации может быть рассчитан на любом наборе данных, как на исходной БД, так и на обезличенной БД. Рекомендуется расчет показателей риска до проведения процедуры обезличивания и после. В этом случае возможен сравнительный анализ показателей риска, рассчитанных на ИБД и ОБД, что позволит оценить степень улучшения защиты персональных данных субъектов в результате проведения процедуры обезличивания.

3. Оценка частотных характеристик БД.

Оцениваются частотные характеристики БД, непосредственно связанные с оценкой риска раскрытия информации, а именно:

– J – количество классов эквивалентности;

– минимальный размер (частота) класса эквивалентности;

– максимальный размер (частота) класса эквивалентности;

– средний размер (частота) класса эквивалентности;

– количество уникальных записей;

– процент уникальных записей;

– показатель k-анонимности (k=2,3,5).

4. Определение типа атаки злоумышленника A в отношении данных: A=1 – атака прокурора; A=2 – атака журналиста; A=3 – атака маркетолога.

5. Если A=1, то рассчитываются показатели риска раскрытия информации, соответствующие атаке прокурора: $R_{p(a)}$; $R_{p(b)}$; $R_{p(c)}$.

6. Если A=2, то определятся тип внешней БД (известна или не известна).

Если внешняя БД известна, то в формулах расчета показателей риска журналиста $R_{j(a)}$; $R_{j(b)}$; $R_{j(c)}$ используются реальные частоты классов эквивалентности внешней БД – $F_j, j = 1, \dots, J$,

иначе: в формулах расчета показателей риска журналиста $R_{j(a)}$; $R_{j(b)}$; $R_{j(c)}$ используются оценки частот классов эквивалентности внешней БД – $\hat{F}_j, j = 1, \dots, J$.

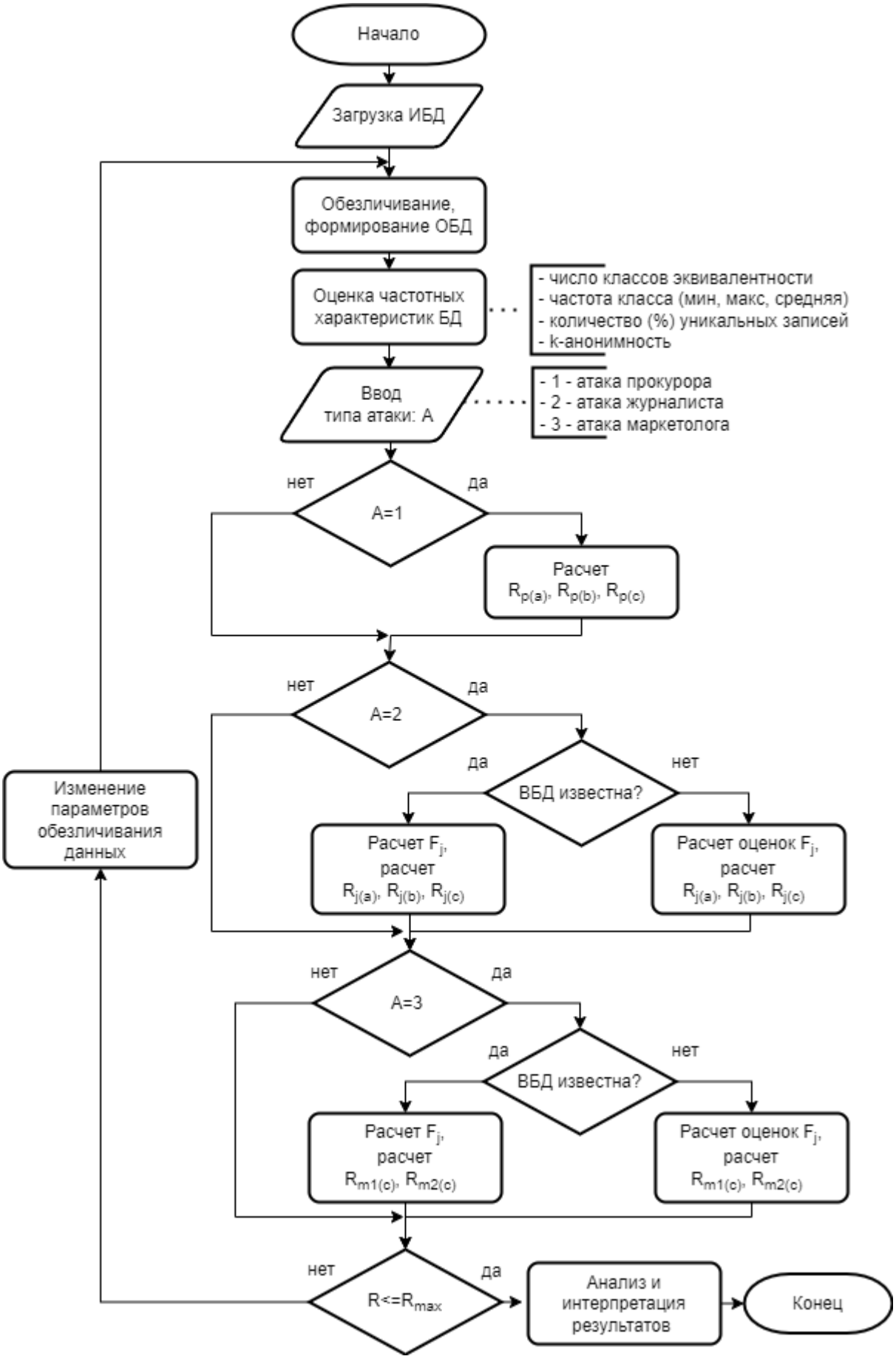


Рисунок 2.27 – Блок-схема алгоритма оценки риска раскрытия информации

7. Если A=3, то определятся тип внешней БД (известна или не известна).

Если внешняя БД известна, то в формулах расчета показателей риска маркетолога $R_{m1(c)}$, $R_{m2(c)}$ используются реальные частоты классов эквивалентности внешней БД – $F_j, j = 1, \dots, J$,

иначе: в формулах расчета показателей риска маркетолога $R_{m1(c)}$, $R_{m2(c)}$ используются оценки частот классов эквивалентности внешней БД – $\hat{F}_j, j = 1, \dots, J$.

8. Сравнение рассчитанных значений показателей риска раскрытия информации R с соответствующими пороговыми значениями $R_{max} = \{\alpha, \tau, \gamma\}$. Пороговые значения рисков раскрытия информации устанавливаются на этапе постановки задачи обезличивания.

Если $R \leq R_{max}$, то выполняется анализ и интерпретация результатов оценки риска раскрытия информации, иначе:

– изменение параметров обезличивания данных;

– возврат на этап обезличивания данных.

2.5 Этап оценки информационных потерь

2.5.1 Меры информационных потерь

В результате обезличивания данных, как правило, снижается их информационная полезность по сравнению с исходными данными.

При выборе методов и решений для обезличивания данных необходимо стремиться к минимизации риска раскрытия информации, обеспечивая при этом максимальную статистическую полезность данных с позиции их последующего анализа. Степень полезности данных определяется требованиями пользователя и планируемыми типами статистического анализа. Для оценки степени полезности были разработаны показатели потери информации, которые оценивают объем потерянной информации применительно к методам обезличивания данных [4].

На настоящий момент не существует единого количественного показателя, который полностью отражал бы структурные различия между исходными и

обезличенными данными, поэтому используют комплекс различных мер информационных потерь.

Для оценки потерь информации существует два разных взаимодополняющих подхода: первый подход – прямое измерение расстояний/частот между исходными и обезличенными данными и второй подход – сравнение статистических характеристик, рассчитанных по исходным и обезличенным данным [4]. В таблице 2.4 приведена сводная информация по используемым мерам информационных потерь.

В рамках первого подхода измеряются потери информации через различия между матрицей X исходных данных и соответствующей матрицей $X_{\text{об.}}$ обезличенных данных. Величину расхождения (ошибки) между матрицами ($X - X_{\text{об.}}$) можно измерять с помощью следующих мер: среднеквадратическая ошибка; средняя абсолютная ошибка; среднее отклонение.

Среднеквадратическая ошибка (*mean square error*) вычисляется по формуле:

$$MSE = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - x_{ij\text{об.}}\right)^2}{np}, \quad (2.13)$$

где p – количество признаков; n – количество записей в БД; x_{ij} – значение j-го признака для i-ой записи в ИБД; $x_{ij\text{об.}}$ – значение j-го признака для i-ой записи в ОБД.

Таблица 2.4 – Формулы для расчета мер информационных потерь

Название меры	**Расчетная формула**	**Тип признака**
Прямое измерение частот/расстояний между ИБД и ОБД		
Среднеквадратическая ошибка	$MSE = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - x_{ij\text{об.}}\right)^2}{np}$	количественный

Продолжение таблицы 2.4

Средняя абсолютная ошибка	$MAE = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij} - x_{ijоб.}\rvert}{np}$	количественный
Среднее отклонение	$MD = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij} - x_{ijоб.}\rvert}{\sqrt{2}s_j}}{np}$	количественный
Сравнение статистических характеристик на ИБД и ОБД		
Потери в оценке среднего значения	$\bar{x}_{j(loss)} = \frac{\bar{x}_j - \bar{x}_{jоб.}}{\bar{x}_j}$	количественный
Потери в оценке среднеквадратического отклонения	$\bar{s}_{j(loss)} = \frac{\bar{s}_j - \bar{s}_{jоб.}}{\bar{s}_j}$	количественный
Мера информационных потерь Шеннона	$I_{loss} = \frac{1}{p}\sum_{j=1}^{p} I_{j(loss)}.$ $I_{j(loss)} = \left(1 - \frac{I_j}{I_{jоб.}}\right) * 100\%$ $I_j = \sum_{s=1}^{m} p_s log_2 p_s$	любой тип признака
Мера потерь информации по признакам, в %	$loss_j$	любой тип признака
Мера средней потери информации по признакам, в %	$L = \frac{1}{p}\sum_{j=1}^{p} loss_j.$	любой тип признака
Мера информационных потерь на основе коэффициента Пирсона	$r_{loss} = \frac{r - r_{об.}}{r}$ $r = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2}}$	количественный
Мера информационных потерь на основе коэффициента Спирмена	$\tau_{loss} = \frac{\tau - \tau_{об.}}{\tau}$ $\tau = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$	порядковый
Мера информационных потерь на основе коэффициента Крамера	$CV_{loss} = \frac{CV - CV_{об.}}{CV}$ $CV = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(c-1)}}}$	классификационный

Средняя абсолютная ошибка (*mean absolute error*) вычисляется по формуле:

$$MAE = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n}|x_{ij}-x_{ij\text{об.}}|}{np}. \quad (2.14)$$

Среднее отклонение (*mean deviation*) вычисляется по формуле:

$$MD = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij}-x_{ij\text{об.}}|}{\sqrt{2}s_j}}{np}, \quad (2.15)$$

где s_j – среднеквадратическое отклонение j-го признака в ИБД.

Преимущество данной меры заключается в нормировке и, как следствие, ее независимости от масштаба измерения признаков.

Вышеперечисленные меры рассчитываются только для количественных признаков.

В рамках второго подхода оценка потери информации для количественных признаков выполняется с использованием следующих описательных статистик, рассчитанных для исходных и обезличенных данных: закон распределения признаков; средние значения; дисперсии (СКО); корреляции; квантили признаков и соотношений признаков.

Потери в оценке среднего значения для j-го признака:

$$\bar{x}_{j(loss)} = \frac{\bar{x}_j-\bar{x}_{j\text{об.}}}{\bar{x}_j}, \quad (2.16)$$

где $\bar{x}_j$ – среднее значение j-го признака в ИБД; $\bar{x}_{j\text{об.}}$– среднее значение j-го признака в ОБД.

Потери в оценке среднеквадратического отклонения для j-го признака:

$$\bar{s}_{j(loss)} = \frac{\bar{s}_j-\bar{s}_{j\text{об.}}}{\bar{s}_j}, \quad (2.17)$$

где $\bar{s}_{j\text{об.}}$– среднеквадратическое отклонение j-го признака в ОБД.

Для качественных и количественных признаков может использоваться информационная энтропия Шеннона, которая отображает меру разнообразия категорий в данных для j-го признака:

$$I_j = \sum_{s=1}^{m} p_s log_2 p_s, \quad (2.18)$$

где p_s – вероятность s-ой категории в данных, m – количество категорий.

Информационные потери (в %) по j-ому признаку составят:

$$I_{j(loss)} = \left(1 - \frac{I_j}{I_{j\text{об.}}}\right) * 100\%, \quad (2.19)$$

где I_j, $I_{j\text{об.}}$ – мера Шеннона для j-го признака соответственно в ИБД и ОБД.

Для оценки степени информационных потерь по всем признакам в ОБД используется соответственно мера (в %):

$$I_{loss} = \frac{1}{p}\sum_{j=1}^{p} I_{j(loss)}. \quad (2.20)$$

Еще одна мера информационных потерь представлена в [39] и определяется как средняя потеря информации по всем признакам БД:

$$L = \frac{1}{p}\sum_{j=1}^{p} loss_j. \quad (2.21)$$

Потери информации для каждого признака $loss_j$ вычисляются в зависимости от его типа (количественный, качественный) и от типа преобразования, которое было применено к признаку для обезличивания информации (подавление, обобщение). Мера не является универсальной, так как не может оценивать информационные потери для любого типа преобразования.

Для оценки информационных потерь с точки зрения искажения статистических взаимосвязей между признаками могут использоваться методы корреляционного и регрессионного анализов данных и сравниваться коэффициенты парной и множественной корреляций между признаками в ИБД и ОБД [40].

Коэффициент парной корреляции Пирсона оценивает степень линейной взаимосвязи между количественными признаками X^j и Y^j:

$$r = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2}}. \quad (2.22)$$

Мера информационных потерь относительно коэффициента корреляции Пирсона:

$$r_{loss} = \frac{r - r_{об.}}{r}, \quad (2.23)$$

где r – коэффициент корреляции Пирсона, рассчитанный на ИБД, $r_{об.}$ – коэффициент корреляции Пирсона, рассчитанный на ОБД.

Коэффициент парной корреляции Спирмена оценивает степень линейной взаимосвязи между порядковыми признаками X^j и Y^j:

$$\tau = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad (2.24)$$

$$d_i = R(x_{ij}) - R(y_{ij}).$$

где d_i – разница между рангами i-го наблюдения для признаков X^j и Y^j; $R(x_{ij})$, $R(y_{ij})$ – ранг i-го наблюдения для признаков X^j и Y^j.

Мера информационных потерь относительно коэффициента корреляции Спирмена:

$$\tau_{loss} = \frac{\tau - \tau_{об.}}{\tau}, \quad (2.25)$$

где $\tau -$ коэффициент корреляции Спирмена, рассчитанный на ИБД, $\tau_{об.} -$ коэффициент корреляции Спирмена, рассчитанный на ОБД.

Оценка информационных потерь для классификационных признаков может быть выполнена на основе расчета критерия χ^2 Пирсона [40].

Критерий χ^2 Пирсона используется для проверки гипотезы о статистической незначимости взаимосвязи между двумя качественными признаками. Критическая статистика критерия подчиняется распределению χ^2 с $(r-1)(\mathrm{c}-1)$ степенями свободы:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}, \quad (2.26)$$

где o_{ij} и e_{ij} – соответственно наблюдаемое и ожидаемое количество значений в $i -$ ой строке и $j -$ ом столбце таблицы сопряженности признаков; r – количество строк в таблице сопряженности; c – количество столбцов в таблице сопряженности.

На основе критерия χ^2 рассчитывается коэффициент корреляции Крамера, который измеряется в диапазоне от 0 до 1 и является мерой связи между двумя классификационными признаками:

$$CV = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(c-1)}}}. \quad (2.27)$$

Мерой потери информации служит относительная разница коэффициентов Крамера, рассчитанных для пар классификационных признаков в ИБД и ОБД:

$$CV_{loss} = \frac{CV - CV_{об.}}{CV}. \quad (2.28)$$

Учитывая, что допускается понижение шкалы измерений значений признака, для количественного признака могут быть рассчитаны также коэффициенты Спирмена и Крамера, для порядкового признака – коэффициент Крамера. Следовательно, мера информационных потерь на основе коэффициента Крамера является универсальной (в определенном смысле) и применима для любого типа признака.

2.5.2 Алгоритм оценки информационных потерь

На рисунке 2.28 приведена блок-схема алгоритма оценки информационных потерь. Рассмотрим более подробно этапы алгоритма.

1. Загрузка ИБД.
2. Обезличивание, формирование ОБД.

Меры информационных потерь оценивают степень уменьшения точности данных в результате проведения процедуры обезличивания. Меры основаны на сравнении частот/расстояний и статистических характеристик данных, рассчитанных до и после проведения процедуры обезличивания, поэтому при оценке информационных потерь используют ИБД и ОБД.

3. Выбор признаков, для которых оцениваются информационные потери, соответственно $X^j, X^j_{об.}$:

X^j – значения j-ого признака, $j = \overline{1,p}$ в ИБД;

$X^j_{об.}$ – значения j-ого признака, $j = \overline{1,p}$ в ОБД.

4. Выбор меры/мер L для оценки информационных потерь.

Выбор мер/меры зависит от типа значений признаков $X^j, X^j_{об.}$ и от того, какие задачи статистической обработки в дальнейшем будут решаться на ОБД. Для количественных признаков доступен весь спектр мер, так как на количественной шкале допустимы все арифметические операции и все операции сравнения. На

порядковой шкале не допустимы арифметические операции, но допустимы все операции сравнения; классификационная шкала допускает только операции сравнения типа: «равно»; «не равно». С учетом этого выбирается мера информационных потерь для признака.

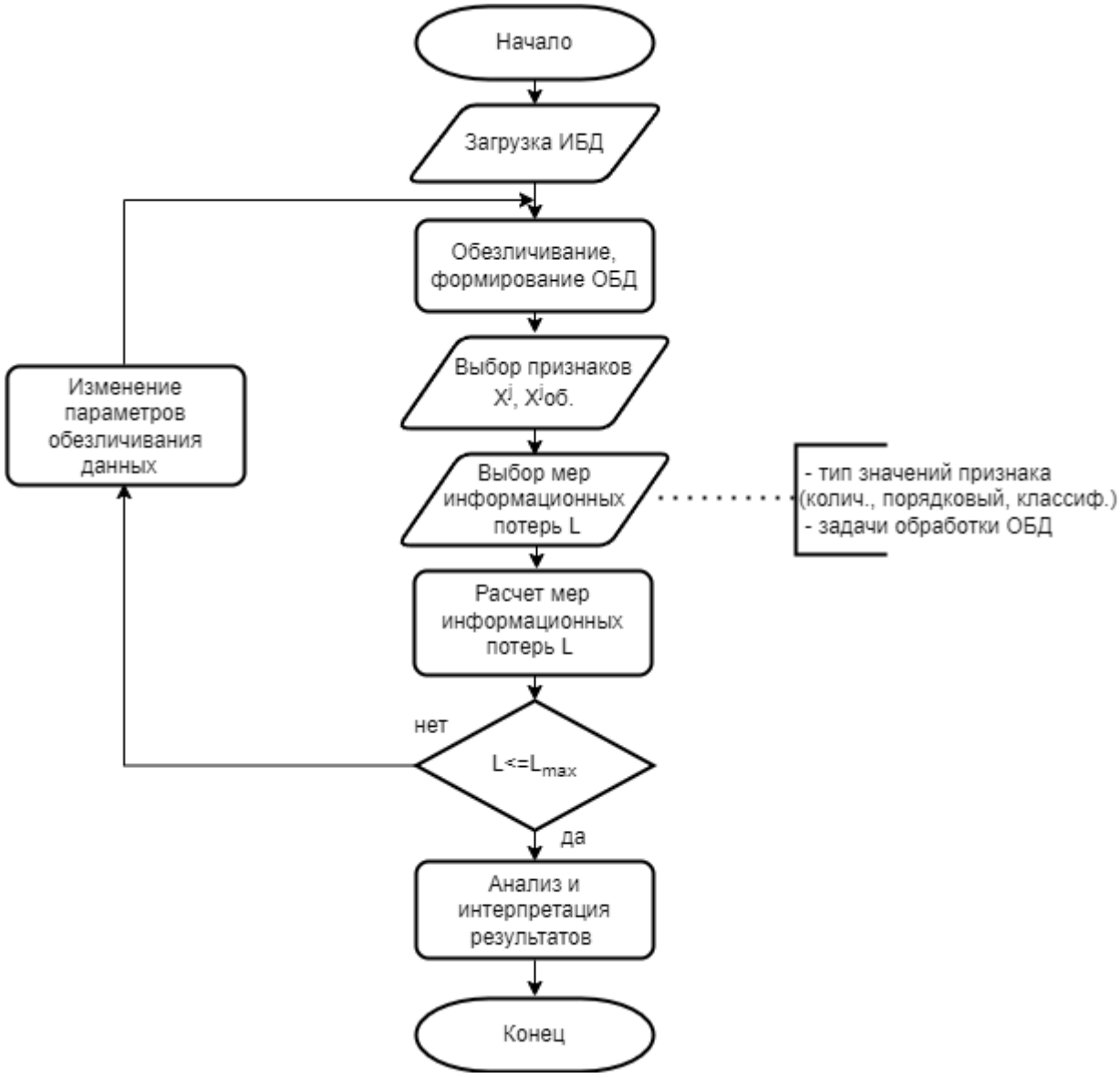


Рисунок 2.28 – Блок-схема алгоритма оценки информационных потерь

На ОБД могут решаться различные задачи, связанные со статистической обработкой данных. Тип задачи выступает одним из факторов выбора меры информационных потерь. Например, если предполагается проведение только первичного анализа данных, достаточно рассчитать меры, основанные на прямом измерении расстояний/частот между исходными и обезличенными данными и сравнить числовые характеристики (среднее, дисперсия), рассчитанные на ИБД

и ОБД. Если же предполагается исследовать структуру взаимосвязей между признаками, то необходим расчет корреляционных характеристик на ИБД и ОБД и анализ степени их искажения в результате обезличивания. Если предполагается решать задачи регрессии или классификации, то рассчитывают показатель точности регрессионного уравнения (например, коэффициент детерминации) или точности классификатора соответственно на ИБД и ОБД, затем анализируют степень потери точности в результате обезличивания.

5. Расчет меры/мер L информационных потерь.

Сравнение рассчитанных значений мер информационных потерь L с соответствующими пороговыми значениями L_{max}. Пороговые значения мер информационных потерь устанавливаются на этапе постановки задачи обезличивания.

6. Если $L \leq L_{max}$, то выполняется анализ и интерпретация результатов оценки информационных потерь, иначе:

– изменение параметров обезличивания данных;

– возврат на этап обезличивания данных.

2.6 Этапы оценки контекстного и комплексного рисков

При оценке риска повторной идентификации записей обезличенной БД рассчитывают набор показателей риска раскрытия информации в зависимости от типа атаки злоумышленника (см. п. 2.4). Показатели риска вычисляют непосредственно на наборе данных.

Однако для расчета комплексного (общего) показателя риска, связанного с выпуском (публикацией) ОБД, необходимо также учитывать вероятность реализации атаки, как одну из составляющих риска. Эта составляющая носит название – контекстный риск.

Под контекстным риском понимают вероятность нарушения защиты контура, в котором обрабатываются данные [12]. Подходы к оценке контекстного риска и методики организации защиты контура рассмотрены в международных стандартах и рекомендациях. В качестве примера можно привести акт Агентства Европейского союза по кибербезопасности (*ENISA*): «Псевдонимизация данных

ENISA: Передовые методы и примеры использования. Технический анализ мер кибербезопасности в области защиты данных и конфиденциальности» (январь 2021) [11]. Также Агентством *ENISA* предложены методические рекомендации и онлайн-калькулятор для оценки контекстных рисков в работе с персональными данными [12]. Широкая нормативная база, связанная с обезличиванием персональных данных (особенно в сфере медицины), разработана в Канаде: например, руководство по публичному раскрытию клинической информации [41]; руководство по обезличиванию структурированных данных [10]. Руководства включают описание методов и подходов к обезличиванию данных и оценке комплексного риска повторной идентификации ОБД.

В России методика оценки угроз информационной безопасности и организации защиты контура описана в требованиях ФСТЭК [9].

Далее рассмотрим этапы (шаги) алгоритма оценки контекстного и комплексного рисков повторной идентификации ОБД (см. рисунок 2.29).

1. Определение типа публикации ОБД – *V*:

V=1 – открытая публикация; *V*=2 – закрытая публикация; *V*=3 – полуоткрытая публикация.

Атаки злоумышленников, направленные на повторную идентификацию записей ОБД, могут быть предприняты для любого набора данных. Однако вероятность атак определенных типов, персона злоумышленника отличаются в зависимости от вида публикации обезличенной БД.

Открытая публикация – публикация данных в открытом доступе в сети Интернет, данные доступны для скачивания и использования без каких-либо условий. В этом случае данные максимально доступны, но наименее защищены от атак злоумышленников.

Закрытая публикация – публикация данных в ограниченном доступе для определенной категории лиц/организаций. Например, хранитель передает данные внешней организации для проведения статистических исследований. Условием получения данных является заключение соглашения об обмене данными, устанавливающего требования соблюдения конфиденциальности и безопасности

данных. Также возможна ситуация, когда проводится первичное или вторичное исследование данных внутри организации, владеющей данными. Закрытый вид выпуска данных предполагает наименьшую доступность, но и обеспечивает высокий уровень защиты от атак злоумышленников.

Полуоткрытая публикация – публикация данных, сочетающая варианты как открытого, так и закрытого доступа к данным. Набор данных доступен любому пользователю для открытого скачивания, однако условием получения данных является необходимость регистрации в организации, предоставляющей данные, и подтверждение согласия на условия использования, обработки и обмена данными (соглашение об условиях использования). В этом случае дополнительные меры защиты и конфиденциальности данных предусмотрены соглашением об использовании данных, но их трудно обеспечить в силу предоставления открытого доступа к данным.

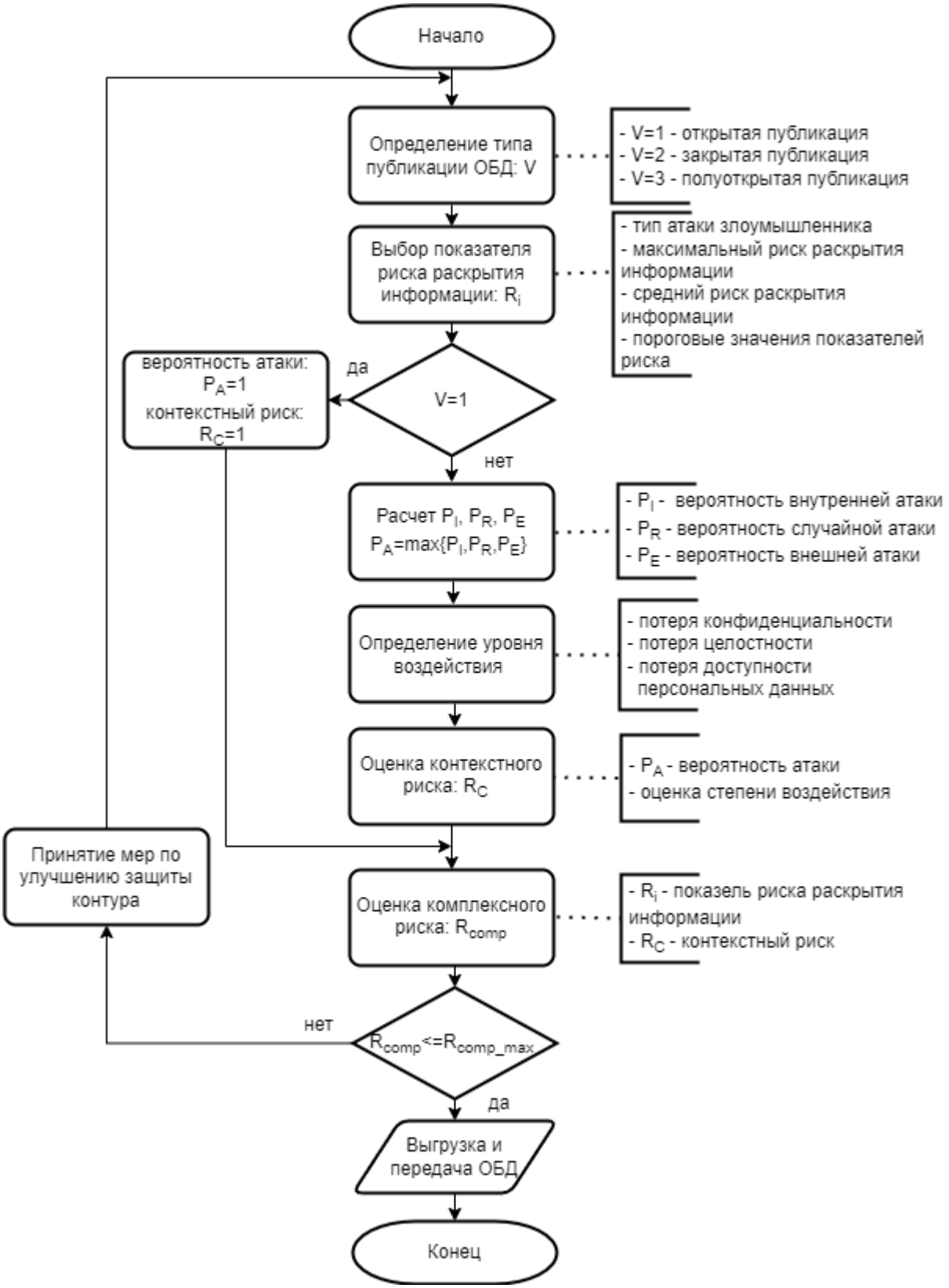


Рисунок 2.29 – Блок-схема алгоритма оценки контекстного и комплексного рисков повторной идентификации ОБД

2. Выбор показателя риска раскрытия информации: R_i.

В п. 2.4 был приведен набор показателей риска раскрытия информации, оценивающих максимальные и средние риски раскрытия персональных данных субъекта в зависимости от условий и типов атак злоумышленника в отношении данных (атака прокурора, журналиста, маркетолога). Для расчета комплексного риска повторной идентификации ОБД необходимо выбрать один ключевой показатель. Выбор показателя зависит от типа публикации данных. Если предполагается открытый доступ к данным, то выбираются максимальные показатели риска, оценивающие максимальную вероятность идентификации отдельной записи ОБД. В случае открытого доступа атаки злоумышленника нацелены на наиболее уязвимые записи в ОБД, соответствующие наименьшему классу эквивалентности с наибольшей вероятностью повторной идентификации. Этим обусловлен выбор максимальных показателей для оценки риска раскрытия информации.

Если публикация ОБД будет закрытого типа, обеспечивающего более высокую степень защиты конфиденциальности и безопасности, то выбираются средние показатели риска раскрытия информации. Однако для защиты редких записей среднее значение должно быть «строгим» средним, т.е. вероятность повторной идентификации отдельной записи не должна превышать заданного порогового значения. В литературе предлагается пороговое значение 0.33, т.е. наименьший размер класса эквивалентности равен 3.14. На практике также может использоваться максимальная вероятность повторной идентификации 0.5, что в случае использования «строгого» среднего гарантирует отсутствие уникальных записей в наборе данных и приемлемый средний риск.

В случае полуоткрытой публикации ОБД используются максимальные показатели риска раскрытия информации, как и для случая открытой публикации, так как данные доступны для скачивания в открытом доступе.

3. Если выбрана публичная публикация (V=1), то

– вероятность атаки злоумышленника устанавливается равной единице: $P_A = 1$;

– контекстный риск равен единице: $R_c = 1$;

– переход на шаг 7.

В случае открытого доступа к данным принимается максимальная вероятность атаки злоумышленника и максимальный контекстный риск, так как предполагается, что всегда найдется человек, который попытается совершить атаку для демонстрации своих возможностей, либо в преступных целях.

4. Если выбрана закрытая публикация (V=2) или полуоткрытая публикация (V=3), то

– расчет вероятности инсайдерской (внутренней атаки): P_I;

– расчет вероятности случайного раскрытия данных: P_R;

– расчет вероятности утечки данных (внешней атаки): P_E;

– выбор максимальной вероятности в качестве вероятности атаки: $P_A = \max\{P_I; P_R; P_E\}$.

5. Определение уровня воздействия на субъект персональных данных.

6. Оценка контекстного риска: R_c.

7. Оценка комплексного риска: R_{comp}.

8. Если $R_{comp} \leq R_{comp\_max}$, то принятие решения о выпуске (передаче) ОБД, иначе:

– принятие мер по улучшению защиты контура;

– возврат на шаг 1.

Поясним более подробно 4 шаг алгоритма.

Расчет вероятности инсайдерской (внутренней атаки)

Вероятность преднамеренной внутренней (*internal*) атаки P_I – вероятность того, что получатель данных предпримет попытку повторной идентификации субъекта/субъектов персональных данных в ОБД.

Оценка вероятности внутренней атаки выполняется путем ответов (да/нет) на вопросы, касающиеся двух аспектов (категорий):

– меры (средства) контроля, предусмотренные информационным соглашением о передаче данных (или об информационном обмене), касающиеся конфиденциальности и безопасности данных;

– технические возможности и/или мотивы получателя данных в отношении проведения атаки, направленной на повторную идентификацию персональных данных субъекта.

Список вопросов приведен в таблице 2.5.

Таблица 2.5 – Вопросы для оценки вероятности реализации атаки

№п/п	**Вопросы**	**Варианты ответа, баллы**
1.	***Меры (средства) контроля, предусмотренные информационным соглашением о передаче данных***	
1.1.	Получатель разрешает только «авторизованным» сотрудникам получать доступ к данным и использовать их для выполнения только должностных обязанностей.	(да/нет) (1/0)
1.2.	Со всеми сотрудниками, включая внешних сотрудников и субподрядчиков, действует соглашение о неразглашении или конфиденциальности (залог конфиденциальности).	(да/нет) (1/0)
1.3.	За раскрытие конфиденциальной информации предусмотрены соответствующие санкции, включая увольнение, которые оговорены в подписанном обязательстве о конфиденциальности.	(да/нет) (1/0)
1.4.	Данные будут удалены по истечении указанного срока использования (хранения).	(да/нет) (1/0)
1.5.	Долгосрочное хранение данных подлежит периодическим проверкам и надзору со стороны независимых органов.	(да/нет) (1/0)

Продолжение таблицы 2.5

1.6.	Получатель публикует только агрегированные данные, которые не позволяют идентифицировать отдельных субъектов персональных данных.	(да/нет) (1/0)
1.7.	Данные не будут разглашаться или передаваться третьим лицам в соответствии с соглашением об обмене данных.	(да/нет) (1/0)

1.8.	Политики и процедуры конфиденциальности и безопасности действуют, контролируются и применяются получателем.	(да/нет) (1/0)
1.9.	Для всех сотрудников проводится обязательное и постоянное обучение по вопросам обеспечения конфиденциальности и информационной безопасности.	(да/нет) (1/0)
1.10.	Установлено антивирусное программное обеспечение.	(да/нет) (1/0)
1.11.	Реализованы меры аутентификации (защита компьютера паролем, уникальная идентификация при входе в систему), доступ к данным имеет только авторизованный персонал.	(да/нет) (1/0)
1.12.	Внедрена и функционирует система мониторинга контрольных журналов для документирования личности, времени и операций, выполняемых с данными.	(да/нет) (1/0)
1.13.	При электронной передаче данных используется зашифрованный протокол.	(да/нет) (1/0)
2.	***Технические возможности и мотивы получателя данных в отношении проведения атаки с повторной идентификацией***	
2.1.	Благонадежность получателя (работал ли получатель в прошлом без инцидентов, связанных с обработкой персональных данных).	(да/нет) (1/0)
2.2.	Существуют ли возможные причины (финансовые или иные), по которым получатель может пытаться повторно идентифицировать персональные данные субъекта/субъектов.	(да/нет) (1/0)
2.3.	Обладает ли получатель техническими возможностями и/или финансовыми ресурсами для попытки любой повторной идентификации.	(да/нет) (1/0)
2.4.	Имеет ли получатель доступ к другим частным БД или наборам данных, которые могут быть связаны с данными для повторной идентификации персональных данных субъекта/субъектов.	(да/нет) (1/0)

Далее выполняется качественная оценка степени контроля конфиденциальности и безопасности данных и технических возможностей (мотивов) получателя для проведения информационной атаки в категориях: «низкий»; «средний»; «высокий».

Если по первой категории набрано 0-4 балла, то оценка «низкий».

Если по первой категории набрано 5-9 баллов, то оценка «средний».

Если по первой категории набрано 10-13 баллов, то оценка «высокий».

Если по второй категории набрано 0-1 баллов, то оценка «низкий».

Если по второй категории набрано 2-3 балла, то оценка «средний».

Если по второй категории набрано 4 балла, то оценка «высокий».

В таблице 2.6 приведена количественная оценка вероятности внутренней атаки P_I, направленной на повторную идентификацию персональных данных субъекта, при сочетании разных уровней мер (средств) контроля и технических возможностей (мотивов) совершения атаки. Наименьшие значения вероятности внутренней атаки соответствуют высокой степени защиты данных и низким техническим возможностям (мотивам) совершения атаки.

Таблица 2.6 – Оценки вероятности реализации атаки

Меры (средства) контроля	**Технически возможности (мотивы)**	**Вероятность атаки P_I**
High (Высокий)	Low (Низкий)	0.05
	Medium (Средний)	0.1
	High (Высокий)	0.2
Medium (Средний)	Low (Низкий)	0.2
	Medium (Средний)	0.3
	High (Высокий)	0.4
Low (Низкий)	Low (Низкий)	0.4
	Medium (Средний)	0.5
	High (Высокий)	0.6

Расчет вероятности случайного раскрытия данных

Получатель обезличенной БД также может непреднамеренно повторно идентифицировать персональные данные субъекта/субъектов. Такая ситуация возможна, если обработчик (получатель) данных узнает коллегу, члена семьи или знакомого.

Вероятность случайного (*random*) раскрытия данных может быть рассчитана по формуле:

$$P_R = 1 - (1 - p)^m, \quad (2.29)$$

где p – доля людей в популяции, у которых есть признак (атрибут), описанный в ОБД; m – среднее число людей, которых знает получатель (обработчик) данных. Значение p определяется по данным официальной статистики о численности населении, а в качестве значения m можно принять «число Данбара» равное 150 (среднее количество социальных связей человека в данный момент времени).

Расчет вероятности внешней атаки

Вероятность внешней (*external*) атаки P_E – вероятность атаки, направленной на повторную идентификацию персональных данных субъекта, внешним злоумышленником на стороне получателя данных. Расчет этой вероятности проводится с учетом информации об утечках данных в отрасли получателя данных (бизнес-сектор и объем обработки данных, сетевые и технические ресурсы). Для оценки вероятности внешней атаки можно использовать подход, разработанный агентством *ENISA* [12]. В отчете по проекту 2.5.2.5 «Формирование требований к методике автоматизированного обезличивания данных» были приведены основные положения (см. п. 2.5). В рамках подхода сформулированы оценочные вопросы о среде обработки данных (которая имеет непосредственное отношение к угрозам информационной безопасности). На основании ответов на вопросы формируется оценка вероятности угрозы (атаки) по четырем категориям среды обработки данных:

– сетевые и технические ресурсы (аппаратное и программное обеспечение);

– бизнес-сектор и объем обработки данных.

– процессы/процедуры, связанные с обработкой ОБД;

– стороны и лица, участвующие в обработке ОБД;

Первые две категории и частично третья категория связаны с внешними угрозами (атаками) и определяют их вероятность. Последняя категория определяет вероятность внутренних угроз (атак). Отметим, что ряд оценочных вопросов *ENISA* (см. таблицу 2.4 из отчета по проекту 2.5.2.5 «Формирование требований к методике автоматизированного обезличивания» данных) пересекаются с оценочными вопросами, приведенными в таблице 2.5.

В любом случае методика оценки контекстного риска предполагает оценку вероятности разных типов атак (внешних, внутренних, случайных), которая опирается на оценочные вопросы. Примеры оценочных вопросов приведены [10,12], но окончательный список вопросов формируется с учетом конкретной ситуации (кто является получателем и хранителем данных, специфика данных, особенности передачи и обработки ОБД, условия информационного соглашения о передаче данных и т.п.). Поэтому список вопросов всегда требует уточнения и/или расширения при решении конкретной задачи в реальных условиях.

Далее выбирается максимальная вероятность из оцененных вероятностей разных типов атак и принимается за вероятность атаки повторной идентификации персональных данных субъекта/субъектов:

$$P_A = \max\{P_I; P_R; P_E\}. \quad (2.30)$$

При оценке вероятности атаки, направленной на повторную идентификацию персональных данных субъекта/субъектов, в случае полуоткрытой публикации ОБД используются те же подходы, что и на закрытых данных. Однако при оценке внутренней угрозы P_I предполагают, что получатель обладает высоким уровнем мотивации и техническими возможностями в сочетании с низким уровнем контроля конфиденциальности и безопасности.

На этапе определения уровня воздействия (5 шаг алгоритма) оценивается степень влияния на субъект персональных данных потери конфиденциальности личных данных. Рассматриваются четыре качественных уровня воздействия: низкий; средний; высокий; очень высокий, как показано в таблице 2.6 ниже.

Оценка уровня воздействия проводится отдельно в трех аспектах:

– потеря конфиденциальности (влияние, которое оказывает на субъект персональных данных несанкционированное раскрытие);

– потеря целостности (влияние, которое оказывает на субъект персональных данных несанкционированное изменение);

– потеря доступности (влияние, которое оказывает на субъект персональных данных несанкционированное уничтожение или потеря персональных данных).

Таблица 2.6 – Описание уровней воздействия

Уровень воздействия	**Описание**
Низкий уровень	Субъект персональных данных может столкнуться с незначительными неудобствами, которые легко преодолеть (например, время, потраченное на повторный ввод информации и т.д.).
Средний уровень	Субъект персональных данных может столкнуться со значительными неудобствами, которые он сможет преодолеть, несмотря на некоторые трудности (дополнительные расходы, отказ в доступе к услугам, непонимание, страх, стресс и т.д.).
Высокий уровень	Субъект персональных данных может столкнуться со значительными последствиями, которые он сможет преодолеть, хотя и с серьезными трудностями (значительный материальный ущерб, потеря работы, ухудшение здоровья и т.д.).
Очень высокий уровень	Субъект персональных данных может столкнуться со значительными или даже необратимыми последствиями, которые он не может преодолеть (нетрудоспособность, длительная болезнь, смерть и т.д.).

При оценке воздействия учитывается ряд факторов, таких как типы персональных данных, критичность операции обработки, объем персональных данных, особые характеристики субъектов данных и т.п.

После этой оценки будут получены три различных уровня воздействия (для потери конфиденциальности, целостности и доступности). Самый высокий из этих уровней рассматривается как окончательный результат оценки воздействия, связанного с общей обработкой персональных данных субъекта.

Оценка контекстного риска (6 шаг алгоритма) проводится с учетом вероятности атаки, связанной с повторной идентификацией персональных данных субъекта/субъектов и уровня воздействия (см. таблицу 2.7). Отметим, что уровень воздействия может не оцениваться. В этом случае за оценку контекстного риска принимается вероятность атаки: $R_c = P_A$.

Таблица 2.7 – Оценка контекстного риска

Вероятность Атаки P_A	**Уровень воздействия**	**Контекстный риск** R_c
Low Низкая вероятность: (0.05-0.2)	Low (Низкий уровень)	0.05
	Medium (Средний уровень)	0.1
	High (Высокий уровень)	0.2
	Very High (Очень высокий)	0.3
Medium Средняя вероятность: (0.2-0.4]	Low (Низкий уровень)	0.3
	Medium (Средний уровень)	0.4
	High (Высокий уровень)	0.5
	Very High (Очень высокий)	0.6
High Высокая вероятность: (0.4-0.6]	Low (Низкий уровень)	0.6
	Medium (Средний уровень)	0.7
	High (Высокий уровень)	0.75
	Very High (Очень высокий)	0.8

На этапе оценки комплексного риска (7 шаг алгоритма) выполняется расчет показателя как произведение риска раскрытия информации (данных) и контекстного риска:

$$R_{comp=}R_c \cdot R_i. \quad (2.31)$$

Выполняется сравнения полученного показателя комплексного риска с установленным пороговым значением R_{comp_max}: $R_{comp} \leq R_{comp\_max}$ (8 шаг алгоритма). Пороговое значение показателя определяется на этапе постановки задачи обезличивания данных. Если неравенство выполняется, то ОБД надежно защищена от раскрытия конфиденциальной информации, как в отношении дан-

ных, так и в отношении контура их обработки. Принимается решение о публикации (передаче) ОБД.

В противном случае должны быть приняты технические и организационные меры по защите контура обработки данных и обеспечения информационной безопасности. Возможные меры по защите среды обработки данных рассмотрены в рекомендациях Агентства *ENISA* [12], ФСТЭК России [9].

Не исключено, что единственным выходом, в случае невозможности обеспечения заданного порогового значения R_{comp_max}, будет изменение типа публикации ОБД.

В заключение подчеркнем, что уменьшать риск повторной идентификации ОБД можно двумя способами: увеличивать степень обезличивания данных, либо улучшать защиту контура. Первый способ не всегда применим, так как связан с информационными потерями в данных вследствие их обезличивания. Чем больше степень обезличивания, тем меньше риски раскрытия информации, но и больше информационные потери. Большие информационные потери могут привести, в свою очередь, к значительной потере точности при решении задач обработки и анализа ОБД, что не всегда приемлемо. Поэтому уменьшать риск повторной идентификации ОБД придется за счет улучшения защиты контура обработки ОБД, либо изменения типа доступа к ОБД (сделать данные закрытыми или частично закрытыми). В случае если невозможно обеспечить заданные показатели рисков раскрытия данных и контекстного риска, принимается решение о невозможности публикации (передачи) ОБД.

2.7 Выводы по главе 2

В главе 2 представлены результаты разработки основных этапов методики автоматизированного обезличивания данных.

Рассмотрены следующие ключевые этапы методики:

– постановка задачи;

– подготовка данных перед проведением процедуры обезличивания;

– обезличивание данных;

– оценка риска раскрытия информации;

– оценка информационных потерь;

– оценка контекстного и комплексного рисков.

Для каждого этапа методики приведено подробное описание реализующего его алгоритма (алгоритмов) в словесной и графической формах (в виде блок-схемы), также описаны методы подходы, которые используются на каждом этапе методики, приведены расчетные формулы показателей и мер, примеры применения методов обезличивания данных.

3 Методика сравнения эффективности разных вариантов обезличивания данных

На этапе постановки задачи выбираются методы обезличивания для каждого из признаков исходной БД и задаются параметры методов, определяющие степень анонимизации данных. Выбор методов зависит от ряда факторов, в частности, от типа признаков ИБД, от того какие задачи планируется в дальнейшем решать на ОБД, от типа публикации ОБД. Однако, выбор методов на этапе постановки задачи не однозначен, и для обезличивания каждого из признаков ИБД могут быть выбраны разные методы и/или значения параметров одного и того же метода. Таким образом, для исходной БД можно определить множество вариантов (моделей) преобразования (обезличивания) данных, каждый из которых соответствует определенной комбинации методов ОД с заданными параметрами, примененных для обезличивания исходных признаков БД. Окончательный выбор варианта преобразования данных может быть только апостериорным, после проведения процедуры обезличивания и оценки ее эффективности. Для сравнения вариантов обезличивания и выбора оптимального варианта должны быть сформулированы критерии сравнения вариантов.

3.1 Критерии сравнения вариантов обезличивания данных

Критерии сравнения основаны на сопоставлении свойств данных до и после проведения процедуры обезличивания и на расчете показателей, оценивающих эффективность проведения процедуры обезличивания данных. Укрупнено критерии можно объединить в следующие группы:

1. Соблюдение требований к данным после проведения процедуры обезличивания.

2. Риски раскрытия информации после проведения процедуры обезличивания, рассчитанные на основе показателей риска раскрытия информации.

3. Информационные потери после проведения процедуры обезличивания данных, рассчитанные на основе мер информационных потерь.

4. Время (скорость) обезличивания данных.

Рассмотрим подробнее критерии каждой группы.

Соблюдение требований к данным

Варианты обезличивания данных могут сравниваться с точки зрения соблюдения требований к данным после проведения процедуры обезличивания. Требования к обезличенным данным рассмотрены в п. 2.3.5, в таблице 2.2 приведена информация о сохранении свойств обезличенных данных при использовании методов обезличивания разных классов.

Первый критерий – сохранение полноты (состав обезличенных данных должен полностью соответствовать составу исходных персональных данных).

Второй критерий – сохранение структурированности обезличиваемых персональных данных (сохранение структуры связей между субъектами ПД).

Третий критерий – сохранение релевантности (возможности обработки запросов по обработке ПД и получения ответов в одинаковой семантической форме).

Четвертый критерий – сохранение семантической целостности ОБД.

Пятый критерий – применимость (возможность решать задачи обработки ПД).

Шестой критерий – анонимность отдельных данных не ниже заданного уровня (уровень анонимности определяется пороговыми значениями показателей риска раскрытия информации).

Риски раскрытия информации

Варианты обезличивания данных могут сравниваться на основе сопоставления показателей риска раскрытия информации, рассчитанных на ИБД и ОБД. Формулы для расчета показателей риска раскрытия информации в условиях разных типов атаки злоумышленника приведены в п. 2.4.1, таблица 2.3.

Первый критерий – доля записей в ОБД, вероятность повторной идентификации которых выше установленного порогового значения τ.

Второй критерий – максимальный риск раскрытия информации среди записей ОБД (оценка худшего сценария).

Третий критерий – средний риск повторной идентификации записи в ОБД (мера глобального риска).

Четвертый критерий – показатель k-анонимности.

Информационные потери

Варианты обезличивания данных могут сравниваться на основе сопоставления мер информационных потерь, возникающих вследствие проведения процедуры обезличивания. Формулы для расчета мер информационных приведены в п. 2.5.1, таблица 2.4.

Первый критерий – среднеквадратическая ошибка (расхождение) между значениями признаков в ИБД и ОБД.

Второй критерий – средняя абсолютная ошибка (расхождение) между значениями признаков в ИБД и ОБД.

Третий критерий – среднее отклонение между значениями признаков в ИБД и ОБД.

Четвертый критерий – потери в оценке среднего значения.

Пятый критерий – потери в оценке СКО.

Шестой критерий – мера потерь Шеннона.

Седьмой критерий – мера потерь информации по признакам ОБД.

Восьмой критерий – мера средней потери информации по признакам ОБД.

Девятый критерий – мера потерь на основе коэффициента Пирсона.

Десятый критерий – мера потерь на основе коэффициента Спирмена.

Одиннадцатый критерий – мера потерь на основе коэффициента Крамера.

Время (скорость) обезличивания данных

Методики обезличивания данных могут сравниваться по скоростному и/или временному показателям выполнения процедуры обезличивания. Формулы для расчета показателей приведены в п. 2.3.7.

Первый критерий – суммарное время обезличивания БД.

Второй критерий – время обезличивания одной записи БД.

Третий критерий – скорость обезличивания БД.

Четвертый критерий – временная сложность алгоритма реализации метода обезличивания данных.

Список критериев сравнения, входящих в каждую группу, может быть расширен. При сравнении вариантов обезличивания данных используется выборка из приведенных критериев, при этом выбор критериев обусловлен спецификой исходной БД, целями и задачами обезличивания данных.

3.2 Алгоритм сравнения вариантов обезличивания данных и выбора оптимального варианта

На рисунке 3.1 приведена блок-схема алгоритма сравнения вариантов обезличивания данных и выбора оптимального варианта. Рассмотрим более подробно этапы алгоритма.

1. Загрузка ИБД.

2. Постановка задачи обезличивания.

На этапе постановки задачи решаются следующие вопросы:

– выбор признаков из ИБД, подлежащих обезличиванию:

– X^j – значения j-ого признака, $j = \overline{1,p}$ в ИБД;

– выбор метода/методов для обезличивания j-го признака и задание параметров методов;

– выбор критериев сравнения вариантов обезличивания данных;

– задание пороговых значений показателей риска раскрытия информации – R_{max}, информационных потерь – L_{max}, времени обезличивания – S_{max};

– задание критерия выбора оптимального варианта обезличивания данных.

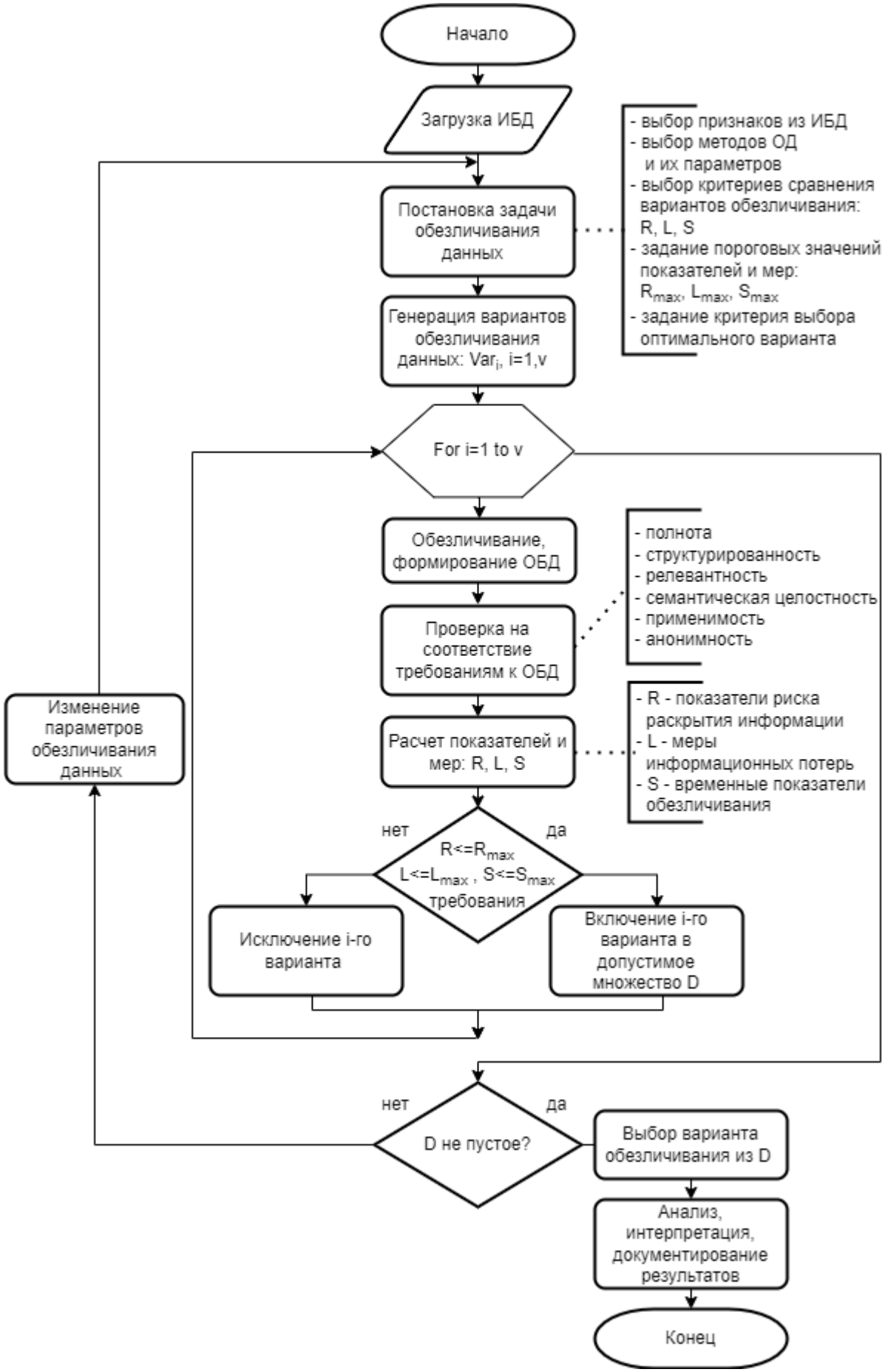


Рисунок 3.1 – Блок-схема алгоритма сравнения и выбора варианта ОД

Для обезличивания j-го признака могут применяться как разные методы обезличивания, так и разные параметры одного метода. Например, в методе обобщения используется параметр, задающий степень (уровень) обобщения. Для признака возраст задаются уровни обобщения: 0-уровень – исходный; 1-уровень – диапазоны возраста шириной 5; 2 уровень – диапазоны возраста шириной 10 и т.д. Каждый из уровней определяет отдельный вариант обезличивания.

Критерии сравнения вариантов обезличивания приведены в п. 3.1. Выбираются критерии из предложенного списка.

Задаются пороговые значения показателей, которые критичны для решаемой задачи обезличивания данных (риск раскрытия, информационные потери, временные показатели). Временные показатели актуальны при обработке больших данных и при постоянном решении задачи обезличивания по мере поступления (изменения) данных, хранящихся в ИБД. Учитывая, что задача обезличивания вряд ли потребует решения в реальном времени, скоростные показатели не являются ключевыми.

Выбор оптимального варианта обезличивания данных сводится к решению оптимизационной задачи, где один из показателей выступает целевой функцией (например, минимум информационных потерь), а для остальных выделенных показателей задаются ограничения (например, глобальный риск раскрытия информации не превышает заданного порогового значения). Возможные постановки оптимизационной задачи приведены в п. 2.1.

3. Генерация вариантов обезличивания данных: $Var_i = \overline{1, v}$, v – количество вариантов обезличивания.

Учитывая, что для каждого признака ИБД могут использоваться разные методы обезличивания и для каждого метода обезличивания могут задаваться разные значения параметров, генерируются разные варианты обезличивания данных. Каждый из вариантов обезличивания соответствует комбинации методов обезличивания и их параметров.

4. Для каждого i-го варианта обезличивания ИБД выполняется последовательность операций:

4.1. Обезличивание, формирование ОБД.

4.2. Проверка на соответствие требованиям к ОБД.

4.3. Расчет показателей риска раскрытия информации – R.

4.4. Расчет мер информационных потерь – L.

4.5. Расчет временных показателей обезличивания – S.

4.6. Сравнение показателей и мер с установленными пороговыми значениями: $R \leq R_{max},\ L \leq L_{max}, S \leq S_{max}$.

4.7. Если условия выполняются и выполняются требования к ОБД, то включение i-го варианта в допустимое множество вариантов обезличивания: D,

иначе: исключение (отбрасывание) i-го варианта.

5. Если множество допустимых вариантов не пустое: $D \neq \emptyset$, то

– выбор варианта обезличивания данных из D в соответствии с заданным критерием оптимальности;

– анализ и интерпретация выбранного варианта обезличивания;

– документирование результатов обезличивания данных;

иначе:

– изменение параметров обезличивания данных;

– возврат на этап постановки задачи.

3.3 Пример сравнения вариантов обезличивания данных

Рассмотрим сравнение вариантов обезличивания данных на примере исследования БД по инфекционной заболеваемости по г. Екатеринбургу, всего 93731 запись. Исследование было подробно описано в отчете по проекту 2.5.2.5 «Формирование требований к методике автоматизированного обезличивания данных». К прямым идентификаторам относились признаки ФИО и адрес проживания. В качестве косвенных идентификаторов, подлежащих обезличиванию, рассматривались следующие показатели: возраст; пол; код диагноза по МКБ-10. Прямые идентификаторы не были включены в ОБД (метод удаления), так как

для решения статистических задач на ОБД прямые идентификаторы не представляют интереса.

Для преобразования косвенных идентификаторов использовались несколько методов ОД, относящихся к классу методов изменения состава и/или семантики, а именно: подавление, обобщение, верхнее кодирование, маскирование. Было выполнено сравнение 4-х вариантов обезличивания. Варианты обезличивания не отличались набором используемых методов, но отличались значениями параметров методов, задающих степень обезличивания:

Вариант 1: код по МКБ-10: обобщенный класс заболеваний (пример – A****), пол и возраст без изменений;

Вариант 2: код по МКБ-10 и пол без изменений, возраст сгруппирован в интервалы шириной 10;

Вариант 3: код по МКБ-10: подкласс заболеваний (пример – A04**), пол без изменений, возраст сгруппирован в интервалы шириной 5;

Вариант 4: код по МКБ-10: класс заболеваний (пример – A0***), пол подавлен (без дифференциации), возраст без изменений.

В качестве критериев сравнения вариантов использовались показатели риска раскрытия информации и меры информационных потерь. В таблице 3.1 приведены результаты оценки эффективности четырех вариантов обезличивания данных по выбранным показателям и мерам. Показатели времени обезличивания в качестве критериев сравнения не рассматривались, так как БД сравнительно небольшого объема, и процедура обезличивания данных проводилась однократно (сбор данных закончен), без жестких временных ограничений.

Варианты обезличивания не отличались по набору используемых методов, поэтому сравнение по свойствам методов также не имеет смысла.

В качестве критичного показателя риска раскрытия информации рассматривался средний риск прокурора – $R_{p(c)}$. Пороговое значение было установлено: 0,35%.

Таблица 3.1 – Оценка эффективности вариантов обезличивания данных

№	Показатель/	Исходная	Вариант	Вариант	Вариант	Вариант

п/п	мера	БД	1	2	3	4
Показатели риска						
1.	$R_{p(c)}$ – средний риск прокурора	1922 2,05%	379 0,4%	368 0,393%	315 0,336%	278 0,297%
2.	$R_{p(b)}$ – максимальный риск прокурора	100%	100%	100%	100%	100%
3.	$R_{j(c)}$ – средний риска журналиста	41,98 (0,04%)	8,99 (0,01%)	7,26 (0,01%)	6,03 (0,01%)	5,62 (0,01%)
4.	$R_{m2(c)}$ – средний риск маркетолога (2)	41,98 (0,04%)	8,99 (0,01%)	7,26 (0,01%)	6,03 (0,01%)	5,62 (0,01%)
5.	$R_{m1(c)}$ – средний риск маркетолога (1)	0,02%	0,004%	0,004%	0,0034%	0,003%
6.	J – количество классов эквивалентности	1922	379	368	315	278
7.	средний размер класса	48,77 (0,052%)	247,31 (0,26%)	254,7 (0,27%)	297,56 (0,317%)	337,16 (0,36%)
8.	максимальный размер класса	5330 (5,67%)	13470 (14,37%)	13582 (14,5%)	21609 (23,05%)	24229 (25,85%)
9.	минимальный размер класса	1 (0,001%)	1 (0,001%)	1 (0,001%)	1 (0,001%)	1 (0,001%)
10.	процент уникальных записей	0,552%	0,026%	0,087%	0,073%	0,076%
11.	k-анонимность 2 3 5	 517 919 1846	 24 50 121	 82 150 264	 68 120 168	 71 99 154

Продолжение таблицы 3.1

Меры информационных потерь						
1.	Мера потерь информации по признакам, в %	0;0;0	75;0;0	0;0;40	25;0;20	50;100;0
2.	Мера средней потери информации, в %	0%	25%	13,33%	15%	50%

3.	Мера потерь Шеннона, в %	0	32,33%	31,55%	41,37%	58,68%

Для выбора оптимального варианта обезличивания данных использовался критерий: минимальные информационные потери при соблюдении ограничений на риск раскрытия информации.

Заданному пороговому значению среднего риска прокурора (0,35%) соответствуют два варианта обезличивания: Вариант 3 и Вариант 4, которые образуют допустимое множество решений. Далее варианты сравниваются по информационным потерям, и выбирается вариант, которому соответствуют минимальные потери. В результате в качестве оптимального выбран Вариант 3 (мера средней потери информации – 15%; мера потерь Шеннона – 41,37%).

Примененные методы обезличивания относились к классу методов изменения состава или семантики, поэтому обезличенные данные удовлетворяли следующим свойствам:

– сохранение полноты – частично (были исключены записи из исходной БД, соответствующие редким заболеваниям и заболеваниям, не относящимся к инфекционным болезням);

– сохранение структурированности обезличиваемых ПД – да (структура связей между субъектами ПД не изменена);

– сохранение релевантности – да;

– сохранение семантической целостности – частично (значения признаков обобщались, что привело к информационным потерям и нарушению семантической целостности данных);

– сохранение применимости – да (возможно решать задачи обработки ОБД);

– сохранение анонимности – да (задан пороговый уровень риска раскрытия информации прокурора, которому удовлетворяют Вариант 3 и Вариант 4 преобразования данных).

Отметим, что если за критичный показатель риска раскрытия информации принимается *k*-анонимность, то ни один из предложенных вариантов не является

допустимым, так как в ОБД присутствуют уникальные записи. Процент уникальных записей во всех вариантах обезличивания данных низок (0,026% – 0,087%), поэтому можно предложить подавить (исключить) уникальные записи из ОБД. В результате исключения ОБД будет удовлетворять принципу анонимности k=2 без значимых изменений в информационных потерях.

3.4 Выводы по главе 3

В главе 3 описаны разработанные критерии сравнения вариантов обезличивания данных. Выделено четыре группы критериев: соблюдение требований к данным; риски раскрытия информации; информационные потери; время (скорость) обезличивания данных и описаны критерии, входящие в каждую группу. Разработана и описана методика сравнения эффективности разных вариантов обезличивания данных и выбора оптимального варианта в виде реализующего ее алгоритма. Сравнение вариантов обезличивание данных выполняется в соответствии с выделенными критериями. Приведен пример сравнения вариантов обезличивания медицинской БД по инфекционной заболеваемости.

ЗАКЛЮЧЕНИЕ

В представленном исследовании выполнен комплекс работ по разработке методики автоматизированного обезличивания данных. Методика разработана на основе анализа и исследования существующих методов и подходов к обезличиванию персональных данных и оценке эффективности обезличивания с учетом требований российского законодательства, международных стандартов и мировой практики в сфере защиты и обеспечения конфиденциальности личной информации субъекта ПД.

Методика автоматизированного обезличивания данных разработана в четком соответствии с требованиями к методике, сформулированными в рамках проекта 2.5.2.5 «Формирование требований к методике автоматизированного обезличивания данных» (см. таблицу 3.2).

Таблица 3.2 – Список требований к методике автоматизированного обезличивания данных

Номер требования	**Наименование требования**	**Критерии выполнения**
2.5.2.5.О.1	Методика должна соответствовать целям, поставленным для решения определенной задачи.	Методика должна содержать в себе раздел «Постановка задачи обезличивания данных» с описанием цели и задач обезличивания данных, исходных данных, участников информационного обмена, если предполагается передача данных третьим лицам, а также методов и средств, используемых в процессе обезличивания информации.
2.5.2.5.О.2	Методика должна быть обоснованной.	Методика должна содержать в себе разделы «Оценка риска раскрытия информации после проведения процедуры обезличивания» и «Оценка информационных потерь, возникающих вследствие обезличивания данных».
2.5.2.5.О.3	Методика должна быть ре-	Выходные данные методики должны

Номер требования	Наименование требования	Критерии выполнения
	зультативной.	включать в себя описание результата выполнения алгоритма автоматизированного обезличивания данных.
2.5.2.5.С.1	Методика должна содержать описание цели обезличивания данных и задач, которые планируется решать на обезличенной БД.	Методика должна включать пункт в разделе «Постановка задачи обезличивания», содержащий цель обезличивания данных и задачи, которые будут решаться на ОБД.
2.5.2.5.С.2	Методика должна содержать описание поставщика и получателя данных, объема исходных данных, подлежащих обезличиванию и передаче, контекста использования/публикации ОБД.	Методика должна включать пункт в разделе «Постановка задачи обезличивания», содержащий информацию о поставщике и получателе данных, объеме обезличиваемых и передаваемых данных, контексте выпуска обезличенных данных.
2.5.2.5.С3	Методика должна содержать описание методов, которые используются для обезличивания данных, набора показателей, которые применяются для оценки риска раскрытия информации на ОБД и набора мер информационных потерь, оценивающих степень потери информации.	Методика должна включать пункт в разделе «Постановка задачи обезличивания», содержащий информацию о выбранных методах обезличивания, о выбранных показателях риска раскрытия информации, о выбранных мерах информационных потерь.
2.5.2.5.С.4	Методика должна содержать описание математических критериев для оценки эффективности процедуры	Методика должна включать пункт в разделе «Постановка задачи обезличивания», содержащий информацию о выбранных критериях оценки эффективности проце-

Номер требования	Наименование требования	Критерии выполнения
	обезличивания данных, пороговых значений показателей риска раскрытия информации и мер информационных потерь.	дуры обезличивания и значениях пороговых показателей риска раскрытия информации и мер информационных потерь.
2.5.2.5.С.5	Методика должна содержать требования к программному обеспечению, используемому для автоматизированного обезличивания данных.	Методика должна включать пункт в разделе «Постановка задачи обезличивания», содержащий информацию о выбранном методе и программном обеспечении для реализации процедуры автоматизированного обезличивания данных.
2.5.2.5.С.6	Методика должна обеспечивать получение обезличенных данных, обладающих заданным набором свойств.	Методика должна содержать раздел «Свойства обезличенных данных», в котором описываются требуемые свойства ОБД.
2.5.2.5.С.7	В рамках методики должны быть использованы методы обезличивания, которые удовлетворяют заданному набору свойств.	Методика должна содержать раздел «Свойства методов обезличенных данных», в котором описываются требуемые свойства используемых методов обезличивания данных.
2.5.2.5.С.8	Методика автоматизированного обезличивания данных должна включать в себя реализацию следующих основных этапов: подготовка данных перед проведением процедуры обезличивания; обезличивание данных; оценка риска раскрытия информации после прове-	Методика должна включать в себя укрупненный алгоритм автоматизированного обезличивания данных.

Номер требования	Наименование требования	Критерии выполнения
	дения процедуры обезличивания; оценка информационных потерь; оценка контекстного риска.	
2.5.2.5.С.9	В рамках этапа методики «Подготовка данных перед проведением процедуры обезличивания» должна быть выполнена очистка данных.	Методика должна включать в себя алгоритм очистки данных.
2.5.2.5.С.10	В рамках этапа методики «Подготовка данных перед проведением процедуры обезличивания» должны быть обработаны пропущенные значения в исходных данных.	Методика должна включать в себя алгоритм обработки пропущенных значений в исходных данных.
2.5.2.5.С.11	В рамках этапа методики «Подготовка данных перед проведением процедуры обезличивания» должна быть выполнена разметка данных.	Методика должна включать в себя алгоритм разметки данных.
2.5.2.5.С.12	В рамках этапа методики «Обезличивание данных» должны быть реализованы следующие классы методов: метод введения идентификаторов, метод декомпозиции, метод перемешивания, метод изменения со-	Методика должна включать в себя алгоритмы применения методов обезличивания разных классов.

Номер требования	Наименование требования	Критерии выполнения
	става или семантики.	
2.5.2.5.С.13	В рамках методики реализация метода введения идентификаторов должна быть выполнена с соблюдением условий применения метода, на входе должны быть заданы параметры метода.	Методика должна включать в себя алгоритм применения метода введения идентификаторов.
2.5.2.5.С.14	В рамках методики реализация метода декомпозиции должна быть выполнена с соблюдением условий применения метода, на входе должны быть заданы параметры метода.	Методика должна включать в себя алгоритм применения метода декомпозиции.
2.5.2.5.С.15	В рамках методики реализация метода перемешивания должна быть выполнена с соблюдением условий применения метода, на входе должны быть заданы параметры метода.	Методика должна включать в себя алгоритм применения метода перемешивания.
2.5.2.5.С.16	В рамках методики реализация метода изменения состава или семантики должна быть выполнена с соблюдением условий применения метода, на входе должны быть заданы параметры метода.	Методика должна включать в себя алгоритм применения метода изменения состава или семантики.
2.5.2.5.С.17	В рамках этапа методики	Методика должна включать в себя набор

Номер требования	Наименование требования	Критерии выполнения
	«Оценка риска раскрытия информации после проведения процедуры обезличивания» должна быть выполнена оценка риска раскрытия информации на основе расчета показателей риска.	показателей риска раскрытия информации и алгоритм их расчета в зависимости от предполагаемой модели атаки злоумышленника.
2.5.2.5.С.18	В рамках этапа методики «Оценка информационных потерь» должна быть выполнена оценка информационных потерь, возникающих вследствие обезличивания данных на основе расчета мер информационных потерь.	Методика должна включать в себя набор мер информационных потерь и алгоритм их расчета.
2.5.2.5.С.19	В рамках этапа методики «Оценка контекстного риска» должна быть выполнена оценка контекстного риска (оценка защиты контура) на основе экспертного оценивания.	Методика должна включать в себя алгоритм оценки контекстного риска на основе экспертного оценивания.
2.5.2.5.С.20	В рамках методики должны быть сформулированы укрупненные критерии сравнения эффективности разных вариантов (процедур) обезличивания данных.	Методика должна включать описание критериев сравнения эффективности разных вариантов (процедур) обезличивания данных, алгоритм сравнения и пример использования.
2.5.2.5.С.21	В рамках методики должны	Методика должна включать описание кри-

Номер требования	Наименование требования	Критерии выполнения
	быть сформулированы критерии сравнения эффективности разных вариантов (процедур) по скоростному и временному показателям.	териев сравнения эффективности разных вариантов (процедур) обезличивания данных по скоростному и временному показателям.
2.5.2.5.С.22	В рамках методики должны быть сформулированы критерии сравнения эффективности разных вариантов (процедур) с точки зрения соблюдения требований к данным после проведения процедуры обезличивания.	Методика должна включать описание критериев сравнения эффективности разных вариантов (процедур) обезличивания данных с точки зрения соблюдения требований к данным после проведения процедуры обезличивания.
2.5.2.5.С.23	В рамках методики должны быть сформулированы критерии сравнения эффективности разных вариантов (процедур) обезличивания на основе показателей риска раскрытия информации, рассчитанных на ОБД.	Методика должна включать описание критериев сравнения эффективности разных вариантов (процедур) обезличивания данных на основе расчета показателей риска раскрытия информации.
2.5.2.5.С.24	В рамках методики должны быть сформулированы критерии сравнения эффективности разных вариантов (процедур) обезличивания на основе мер информационных потерь, возникающих вследствие обезличивания информации.	Методика должна включать описание критериев сравнения эффективности разных вариантов (процедур) обезличивания данных на основе расчета мер информационных потерь.

В ходе выполнения работ по проекту 2.5.2.8 «Разработка методики автома-

тизированного обезличивания данных» получены следующие основные результаты:

1. Разработана методика автоматизированного обезличивания данных, представленная в виде укрупненного алгоритма (словесное описание, блок-схема), состоящего из последовательности взаимосвязанных этапов. Всего выделено 12 ключевых этапов, к которым относятся:

– постановка и формализация задачи обезличивания данных;

– загрузка исходных данных (ИБД);

– подготовка данных перед проведением процедуры обезличивания;

– обезличивание и сохранение обезличенной БД, оценка времени ОД;

– оценка риска раскрытия информации;

– оценка информационных потерь;

– сравнение расчетных и пороговых значений показателей риска раскрытия информации, мер информационных потерь;

– оценка контекстного риска;

– оценка комплексного риска;

– сравнение расчетного и порогового значений показателя комплексного риска;

– анализ, документирование результатов ОД;

– выгрузка и передача ОБД.

Этапы алгоритма итерационно взаимодействуют друг с другом: на любом шаге возможен возврат на предыдущие этапы для их корректировки с учетом анализа промежуточных результатов. Также возможно многократное повторение этапов алгоритма для выполнения установленных требований к процедуре автоматизированного обезличивания данных и достижения конечных целей обезличивания ПД.

2. Разработаны и детально описаны основные этапы методики автоматизированного обезличивания данных. Этапы методики представлены в виде реализующих их алгоритмов (словесное описание, блок-схема).

Этап постановки задачи автоматизированного обезличивания данных

Постановка задачи является начальным и ключевым этапом, определяющим эффективность и результативность всего процесса обезличивания данных в целом. Представлено описание и содержание основных пунктов, которые определяются на этапе постановки задачи, а именно: цели и задачи обезличивания данных; исходные (входные) данные; получатель (потребитель) данных; информационное соглашение о передаче данных; требования к данным со стороны получателя; операции подготовки данных перед проведением процедуры обезличивания; методы обезличивания данных; подходы и методы оценки контекстного и комплексного рисков повторной идентификации ОБД, оценки информационных потерь, оценки временных (скоростных) показателей ОД; критерии оптимизации процедуры ОД; программные средства ОД. Предложенный набор пунктов (см. таблицу 2.1, рисунок 2.1) представляет собой структуру паспорта постановки задачи обезличивания данных.

Этап подготовка данных перед проведением процедуры обезличивания данных

В рамках этапа подготовки данных перед ОД выполняется три основных шага (подэтапа): очистка данных (устранение ошибок регистрации и ввода, информационных дублей, структурных несоответствий и т.п.); обработка пропущенных значений; разметка данных. Разработаны и описаны алгоритмы реализации каждого из подэтапов, описаны методы и подходы, которые применяются для обработки и заполнения пропущенных значений признаков ИБД. Определен состав процедуры разметки данных, включающий два шага:

– определение типа каждого признака (атрибута), входящего в ИБД (прямой идентификатор, косвенный идентификатор, чувствительный признак, нечувствительный признак);

– определение типа значений каждого признака (атрибута), входящего в ИБД (количественный, порядковый, классификационный, дата/время).

Этап обезличивания данных

Описаны алгоритмы реализации следующих методов обезличивания данных (в словесной форме, блок-схемы): метод введения идентификаторов; метод декомпозиции; метод перемешивания; методы изменения состава и/или семантики (обобщение, кодирование сверху и/или снизу, локальное подавление, микроагрегирование, добавление шума, округление, выборка, маскирование, удаление). Приведены примеры использования методов. Описаны свойства методов ОД и свойства, которыми должны обладать обезличенные данные. Также предложены основные временные (скоростные) показатели реализации процедуры обезличивания данных.

Этап оценки риска раскрытия информации

Приведены базовые показатели риска раскрытия информации и их расчетные формулы в зависимости от предполагаемой модели атаки злоумышленника (атака прокурора, журналиста или маркетолога). Описан алгоритм оценки риска раскрытия информации. На обезличенной БД рекомендуется рассчитывать весь набор показателей риска (в условиях разных моделей атак) для всесторонней и комплексной оценки эффективности процедуры обезличивания и обеспечения математически гарантированного обезличивания.

Этап оценки информационных потерь

Приведены базовые меры информационных потерь, используемые на признаках разных типов. Описан алгоритм оценки информационных потерь, возникающих вследствие реализации процедуры обезличивания данных.

Этапы оценки контекстного и комплексного рисков

Разработан алгоритм оценки контекстного риска (степень защиты контура обработки данных) и комплексного риска. Оценка контекстного риска выполняется на основе применения методик экспертного оценивания. В рамках алгоритма:

– определяется вид публикации ОБД (открытая, закрытая, полуоткрытая);

– оценивается вероятность разных типов атак (внешних, внутренних, случайных) с помощью ответов на оценочные вопросы;

– выбирается максимальная вероятность из оцененных вероятностей разных типов атак и принимается за вероятность атаки повторной идентификации ПД субъекта;

– определяется уровень воздействия (степень влияния на субъекта ПД потери конфиденциальности, целостности, доступности данных) в качественной шкале (низкий; средний; высокий; очень высокий);

– определяется контекстный риск с учетом максимальной вероятности атаки и уровня воздействия.

Комплексная оценка риска предполагает расчет произведения риска раскрытия информации и контекстного риска. Выбор показателя риска раскрытия информации при оценке комплексного риска зависит от типа публикации ОБД. В случае открытой публикации (наименьшая степень защиты данных) используют максимальные показатели для оценки риска раскрытия информации, а контекстный риск принимают за единицу. В случае закрытой публикации выбирают средние показатели риска раскрытия информации, но для обеспечения защиты редких записей среднее значение должно быть «строгим» средним, т.е. вероятность повторной идентификации отдельной записи не должна превышать заданного порогового значения. В международных рекомендациях предлагается пороговое значение 0.33, т.е. наименьший размер класса эквивалентности равен 3.14. На практике также может использоваться максимальная вероятность повторной идентификации 0.5, что в случае использования «строгого» среднего гарантирует отсутствие уникальных записей в наборе данных и приемлемый средний риск. Таким образом, показатель *k*-анонимности задается равным 2-3.

3. Разработана методика сравнения эффективности разных вариантов обезличивания данных. Сформулированы критерии сравнения вариантов обезличивания, которые объединены в 4 группы: соблюдение требований к данным; риски раскрытия информации; информационные потери; показатели времени (скорости). Разработан алгоритм сравнения эффективности разных вариантов обезличивания данных и выбора оптимального варианта ОД. Выбор оптималь-

ного варианта ОД сводится к решению оптимизационной задачи в одной из постановок:

– минимизация информационных потерь при выполнении ограничений на пороговые значения показателей риска раскрытия информации;

– минимизация риска раскрытия информации при выполнении ограничений на пороговые значения показателей информационных потерь;

– минимизация информационных потерь и риска раскрытия информации; введение вектора предпочтений;

– минимизация времени (скорости) обезличивания при соблюдении ограничений на пороговые значения риска раскрытия информации.

Должны быть установлены допустимые, пороговые значения показателей риска раскрытия информации после обезличивания БД, также важно найти компромисс между риском раскрытия ПД субъекта и информационными потерями. Уменьшение риска раскрытия информации сопровождается ростом информационных потерь, что, в свою очередь, может привести к непригодности ОБД для целей дальнейшей обработки.

Приведен пример сравнения разных вариантов обезличивания данных и выбора оптимального варианта на примере анализа медицинской БД по инфекционной заболеваемости.

Результаты исследования будут использованы в дальнейших работах Программы создания и развития Центра компетенций Национальной технологической инициативы «Технологии доверенного взаимодействия». Результаты будут положены в основу работ при выполнении последующих мероприятий проекта «Технология интеллектуального управления данными для платформы «Доверенная среда обмена информацией», включающая в себя автоматизированные системы обезличивания и обогащения данных» (Приложение 5 Программы развития ЦК НТИ ТДВ) на 2024 г: 2.5.2.10 «Разработка математических моделей на основании разработанных методик пп. 2.5.2.7-2.5.2.9», «Разработка модулей подсистемы интеллектуального управления, обогащения и обезличивания данных».

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Методические рекомендации по применению приказа Роскомнадзора от 5 сентября 2013 г. N 996 «Об утверждении требований и методов по обезличиванию персональных данных». – URL: https://base.garant.ru/70541864/ (дата обращения 15.03.2024).

2. Математический энциклопедический словарь / гл. ред. Ю.В. Прохоров. – М.: Сов. энциклопедия, 1988. – 847 с.

3. Исследование существующих методов оценки уровня доверия к субъекту информационного обмена в недоверенной среде [Текст]: отчет о НИР (промежуточ.) / ФГБОУ ВО НГТУ; рук. Иванов А.В. – Новосибирск, 2022 – 79 с. – Исполн.: Архипова А. Б., Рева И. Л., Селифанов В. В., Огнев И. А., Никрошкин И. В., Лысенко М. В., Прыткова О. В., Булатов А. Д., Якименко А.А., Альсова О.К., Малявко А. А., Медведев М. А.

4. Приказ Росстата от 19.04.2013, №165 «Методологические положения по формированию массивов деперсонифицированных микроданных годового структурного обследования по форме федерального статистического наблюдения №1 – предприятие «Основные сведения о деятельности организации» общего пользования для представления пользователям в аналитических целях». – URL: https://www.garant.ru/products/ipo/prime/doc/70270390/ (дата обращения 15.03.2024).

5. ISO/IEC 27018:2019 – Information technology – Security techniques. – URL: https://www.iso.org/standard/76559.html (дата обращения 15.03.2024).

6. ISO/IEC 20889:2018 – Privacy enhancing data de-identification terminology and classification of techniques. – URL: https://www.iso.org/standard/69373.html (дата обращения 15.03.2024).

7. Федеральный закон от 27 июля 2006 года, №152-ФЗ «О персональных данных». – URL: https://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения 15.03.2024).

8. Приказ Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций от 05.09.2013 № 996 «Об утверждении требований и методов по обезличиванию персональных данных». – URL: https://base.garant.ru/70451476/ (дата обращения 15.03.2024).

9. Методический документ. «Методика оценки угроз безопасности информации» (утв. ФСТЭК России 05.02.2021). – URL: https://dshi-ras-svet.rnd.muzkult.ru/media/2022/04/06/1295759064/Metodicheskij_dokument._Metodika_ocenki_ugroz_bezopasnosti_in.pdf (дата обращения 01.03.2023).

10. De-identification Guidelines for Structured Data / Information and Privacy Commissioner of Ontario. – Toronto, Ontario: 2016.

11. ENISA Data Pseudonymization: Advanced Techniques & Use Case / Athena Bourka (ENISA): ENISA, 2021. – 53 p.

12. Evaluating the level of risk for a personal data processing operation. – URL: https://www.enisa.europa.eu/risk-level-tool/risk (дата обращения 15.03.2024).

13. Борисов Р.С., Ефименко А.А. Паспорт наборов данных и результатов исследований для публикации в открытых источниках// Правовая информатика. – 2022. – №2 – С. 66-79.

14. Борисов Р.С., Ефименко А.А. Протокол анонимизации наборов данных для публикации в открытых источниках// Правовая информатика. – 2023. – №2 – С. 54-66.

15. Борисов Р.С., Ефименко А.А. Протокол обработки наборов данных для их публикации в открытых источниках// Правовая информатика. – 2021. – №2 – С. 59-70.

16. Flexible Data Anonymization Using ARX – Current Status and Challenges Ahead/J Software Pract Exper 50 – 2020. – Vol. 7, P. 1277-1304.

17. Benschop T., Welch M. Statistical Disclosure Control for Microdata: A Practice Guide. Retrieved. – URL: https://sdcpractice.readthedocs.io/en/latest/ (дата обращения 15.03.2024).

18. Hundepool A., Willenborg L. ARGUS: software packages for statistical disclosure control/ Physica, Heidelberg – 1996. – P. 341-345.

19. Osborne J.W. Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data/ Sage – 2012. — 275 p.

20. PostgreSQL : Документация: 12: 24.1. Регламентная очистка: Компания Postgres Professional. – URL: https:/postgrespro.ru/docs/postgresql/12/routine-vacuuming (дата обращения 22.03.2024).

21. Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data/ Wiley – 2002. — 408 p.

22. Olinsky A., Chen S., Harlow L. The comparative efficacy of imputations methods for missing data in structural equation modeling // European Journal of Operational Research — 2003. — Т. 151, № 1. — P. 53–79. — doi:10.1016/S0377-2217(02)00578-7.

23. Peugh J. L., Enders C. K. Missing data in educational research: A review of reporting practices and suggestions for improvement // Review of Educational Research — 2004. — Vol. 74. — P. 525—556.

24. Van Buuren S. Flexible Imputation of Missing Data/ Chapman and Hall – 2012. — 342 p. – doi:10.1201/b11826.

25. Menard S. Applied Logistic Regression Analysis/ SAGE – 2002. – 111 p.

26. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. "10. Boosting and Additive Trees"/New York: Springer – P. 337-384.

27. Мищенко Е.Ю., Соколов А.Н. Алгоритмы реализации методов обезличивания персональных данных в распределенных информационных системах// Доклады ТУСУР. – 2019. – Т.22. – С. 66-70.

28. Мищенко Е.Ю., Соколов А.Н. Количественный анализ процедуры обезличивания персональных данных. Метод введения идентификаторов // Вестник ЮУрГУ. Сер.: Компьютерные технологии, управление, радиоэлектроника. – Челябинск: Изд. центр ЮУрГУ, 2015. – Т. 15. – №3. – С. 18–25.

29. Мищенко Е.Ю., Соколов А.Н. Количественный анализ процедуры обезличивания персональных данных. Метод перемешивания. // Вестник УрФО.

Безопасность в информационной сфере. – Челябинск: Изд. центр ЮУрГУ, 2016. – №3(21). – С. 30–37.

30. Templ M. Statistical Disclosure Control for Microdata: Methods and Applications in R/ Cham, Switzerland: Springer – 2017.

31. Мищенко Е.Ю., Соколов А.Н. Количественный анализ процедуры обезличивания персональных данных. Метод изменения состава или семантики//Вестник УрФО. – 2016. – №1(19). – С. 30-38.

32. El Emam K. Guide to the De-Identification of Personal Health Information/ Boca Raton, FL: CRC Press – 2013. – doi: 10.1201/b14764.

33. El Emam K., Arbuckle L. Anonymizing Health Data Case Studies and Methods to Get You Started/ Sebastopol, CA: O'Reilly Media – 2013.

34. Benedetti R., Franconi L. Statistical and technological solutions for controlled data dissemination// Pre-proceedings of New Techniques and Technologies for Statistics. – 1998. – Vol.1, P. 225-232.

35. Franconi L., Polettini S. Individual risk estimation in mu-Argus: a review/ Privacy in Statistical Databases. Lecture Notes in Computer Science. – 2004. – P. 262–272.

36. Prasser F., Kohlmayer F., Kuhn K.A. The Importance of Context: Risk-based De-identification of Biomedical Data/Methods Inf Med. – 2016. – Vol. 55(4), P. 347-55. – doi: 10.3414/ME16-01-0012. Epub 2016 Jun 20. PMID: 27322502.

37. Kniola L. Calculating the Risk of Re-Identification of Patient-Level Data Using Quantitative Approach// PhUSE Annual Conference – 2016.

38. El Emam K., Dankar F. Protecting privacy using k-anonymity// Journal of the American Medical Informatics Association. – 2008. – Vol. 15, P. 627–637.

39. Dankar F., El Emam K. A method for evaluating marketer re-identification risk// Proceedings of the 3rd International Workshop on Privacy and Anonymity in the Information Society – 2010.

40. Shlomo N. Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility// Journal of Privacy and Confidentiality – 2010. – Vol. 2(1). – doi: https://doi.org/10.29012/jpc.v2i1.584.

41. Children’s Hospital of Eastern Ontario Research Institute Pan-Canadian DeIdentification Guidelines for Personal Health Information/ Canada: Office of the Privacy Commissioner of Canada, 2007. – 87 p.