ГЛУБОКИЙ АНАЛИЗ ДАННЫХ ПРИ РАБОТЕ С ПРИБЛИЖЕННЫМИ МНОЖЕСТВАМИ

Вишняков А.С.[1], Макаров А.Е.[2], Уткин А.В.[3], Зажогин С.Д.[4], Бобров А.В.[5]
Email: Vishniakov17139@scientifictext.ru

[1]Вишняков Александр Сергеевич – ведущий инженер,
системный интегратор «Крастком»;
[2]Макаров Анатолий Евгеньевич – архитектор решений,
Российская телекоммуникационная компания «Ростелеком»,
г. Москва;
[3]Уткин Александр Владимирович – старший инженер,
Международный системный интегратор «EPAM Systems», г. Минск, Республика Беларусь;
[4]Зажогин Станислав Дмитриевич - старший разработчик,
Международный IT интегратор «Hospitality & Retail Systems»;
[5]Бобров Андрей Владимирович – руководитель группы,
группа технической поддержки,
Компания SharxDC LLC,
г. Москва

Аннотация: *в статье проводится детальный анализ данных, которые исследуются с использованием теории приближенных множеств, в плане нахождения разнообразных свойств объектов на основе исследования связей между их атрибутами. Рассматриваются популярные научные работы в рамках тематики данной статьи.*

Приводится описание работы основных современных технологий Data Mining на основе использования концепции шаблонов, которые способны отыскать многофункциональные взаимоотношения в исследуемых данных. Аргументируется то, что использование основных методов Data Mining играет главную роль во время глубокого анализа данных при работе с приближенными множествами, поскольку позволяет решать задачи разной природы происхождения.

В статье подробно анализируются основные этапы протокола CRISP-DM, который позволяет разработать предсказательные модели, последние, в свою очередь, способны решить большой круг задач, которые возникают при построении бизнеса. Исследованы основные шесть этапов, где основным является третий этап, на котором подготавливаются данные, которые должны соответствовать формату поставленной задачи за различными качественными свойствами.

Осуществляется обзор подходов к обобщению и приводится сравнительный анализ по поводу их применения при работе с реальными массивами данных. Для визуализации уровней, извлекаемых из данных знаний, предложен рис. 1, а также на рис. 2. приведены основные дисциплины, которые входят в систему Data Minincg.

Ключевые слова: *машинное обучение, глубокий анализ, массив, данные, модель, бизнес-процесс, шаблон, математическая статистика, выборка, алгоритм, система анализа.*

DEEP DATA ANALYSIS WHEN WORKING WITH APPROXIMATE SETS

Vishniakov A.S.[1], Makarov A.E.[2], Utkin A.V.[3], Zazhogin S.D.[4], Bobrov A.V.[5]

[1]Vishniakov Alexandr Sergeevich – Lead System Engineer,
SYSTEM INTEGRATOR «KRASTCOM»;
[2]Makarov Anatoly Evgenevich – Solutions Architect,
ROSTELECOM INFORMATION TECHNOLOGY,
MOSCOW;
[3]Utkin Alexander Vladimirovich – Senior Enaineer.
INTERNATIONAL SYSTEM INTEGRATOR EPAM SYSTEMS, MINSK, REPUBLIC OF BELARUS;
[4]Zazhogin Stanislav Dmitrievich – Senior Software Engineer,
International IT Integrator Hospitality & Retail Systems;
[5]Bobrov Andrei Vladimirovich – Team leader,
TECHNICAL SUPPORT GROUP,
SHARXDC LLC,
MOSCOW

Abstract: *The article provides a detailed analysis of the data that are studied using the theory of approximate sets, in terms of finding the various properties of objects based on the study of the relationships between their attributes. Considered popular scientific work in the framework of this article.*

Describes the work of the main modern Data Mining technologies based on the use of the concept of templates that are able to find multifunctional relationships in the studied data. It is argued that the use of basic methods of Data Mining plays a major role during in-depth data analysis when working with approximate sets, because it allows solving problems of different origin.
The article analyzes in detail the main stages of the CRISP-DM protocol, which allows you to develop predictive models, the latter, in turn, are able to solve a wide range of tasks that arise when building a business. The basic six stages are investigated, where the main is the third stage on which the data are prepared, which should correspond to the format of the task for different qualitative properties.
A review of approaches to generalization is carried out and a comparative analysis is given about their use when working with real data arrays. To visualize the levels of knowledge extracted from these data, Figure 1 is proposed, as well as Figure 2. Shows the main disciplines that are part of the Data MinInc system.
Keywords: *machine learning, in-depth analysis, array, data, model, business process, template, mathematical statistics, sampling, algorithm, analysis system.*

УДК 331.225.3

Введение: Стремительное развитие информационных технологий влечет за собой новые перспективы в развитии машинного обучения и других направлений, связанных с ним, таких, как: нейронные сети, глубинное обучение и др. Отметим, то что, машинное обучение является приоритетным направлениям при создании искусственного интеллекта, а так же объединение основных рабочих аспектов машинного и глубинного обучения предоставляют возможность работать в многослойных сетях, то есть успешно принимать решения на основе неточной или же неполной информации, в качестве примера приведем систему глубокого обучения DeepStack, которой удалось обыграть одиннадцать опытных игроков в покер, поскольку после каждого раунда ставок данная система пересчитывала ранее использованные стратегии.

В современном мире постоянных открытий набирают популярность модели, которые предназначенные для предсказания бизнес-процессов [1, 3]. Предсказательные модели являются приоритетными по сравнению с моделями, которые предназначенные для установления связей между предикторами и некоторой переменной-откликом.

Популярность первых связана из тем, что возможность предсказывать итог необходимого бизнес-процесса является главным заданием бизнеса, а кроме этого позволяет активно конкурировать на рынке товаров или услуг, в отдельных случаях предсказания могут быть основной задачей бизнеса, в качестве примера несложно представить суть работы таких платформ как: Google, Amazon, Netflix, и т.д.

Инновационное развитие методов предназначенных для обработки данных способствовало возникновения таких синонимических терминов как: Data Mining («раскопка данных»), «выявления знаний в базах данных», «интеллектуальный анализ данных». Отметим, что сфера анализа данных не ограничена и присутствует везде где есть какие либо данные [4].

Анализ последних исследований и публикаций. Анализ литературных источников показывает, что в теории существуют многие правила для реализации проектов, но они не всегда подтверждаются на практике, поскольку успешное использование предлагаемых многими учёными правил состоит в их практическом применении.

Отметим, что наиболее известным и широко применяемым аппаратом для построения предсказательных моделей является – межиндустриальный стандартный протокол глубинного анализа данных – **Cross-Industry Standard Protocol for Data Mining** (CRISP-DM).

История зарождения протокола CRISP-DM начинается в далеких 1990-х годах, к этому причастны специалисты консорциумома из таких компании как: SPSS, Teradata, Daimler AG, NCR Corporation и OHRA, отметим что, несмотря на «возраст»данный протокол пользуется большой популярностью среди всех известных методологий, которые решают аналогичные задачи [1, 5, 6].

Детальным изучением нейронных сетей, которые обученные на случайных подмножествах данных посвященные работы Габора Мелиса. Если говорить о исследовании бустированных деревьев решений то следует отдать должное место Тиму Салимансу. Данный метод реализовывается с путем сканирования каждой переменной и предложением решения, которое разделяет пространство на определенный значения переменных, которое создает наибольшую разницу между двумя классами, в научных исследованиех это принято называть – бинарным разветвлением.

Пиер Куртиол занимался изучением интеллектуального анализа данных с использованием нейронных сетей, которые способны аппроксимировать любую непрерывную функцию и главной сложностью является количество скрытых уровней и количеством скрытых нейронов в уровне [6].

При использования глубокого анализа для задач прикладного характера, которые, в основному связаны из физикой, принято использовать такие методы как: глубокие нейронные сети, бустированные деревья решения, метод опорных векторов, байесовские нейронные сети [2, 4].

Изучение литературных источников показало, что тема глубокого исследования данных на основе теории приближенных множеств – это сравнительно новое направление в мире эпохи больших данных и требует дополнительных исследований с целью расширения методов глубинного анализа данных, который активно нашел свое применение у многих классификационных, регрессивных задачах теоретического и практического характера.

Формулирование целей статьи (постановка задачи). Провести исследование интеллектуального анализа данных при работе с приближенными множествами с использованием машинного обучения.

Изложение основного материала исследования. Для того чтобы разработать успешную предсказательную модель определенной бизнес-задачи необходимо детально провести исследование всех областей начиная с конкретного бизнес-домена и до баз данных, которые берут участие в рассматриваемом процессе, а также нужно провести экспертизу ИТ-инфраструктуры, вчасности изучить методы статистики и ключевые аспекты машинного обучения.

Исходя из описания построения предсказательной модели, становится ясно, что для выполнения всех исследований и запуску предсказательной модели необходима целая команда специалистов, поскольку построить успешный бизнес возможно в том случае, когда будут структурированные основные этапы работы и «обкатанные» на практике бизнес-процессы [1].

В качестве примера глубокого анализа данных рассмотрим межиндустриальный стандартный протокол, а в вчасности проанализируем основные этапы работы протокола CRISP-DM. На первом этапе необходимо четко сформулировать предстоящую для решения бизнес-задачу на языке, который будет понятен соответствующим специалистам, а уже потом описать её с помощью статистических методов, то есть определится с необходимыми данными и предсказательной переменной-откликом, для решения статистической задачи успешно используются методы машинного обучения [3].

Не зависимо от поставленной задачи, при построении квалификационной модели, для проведения оценки верности модели можно использовать долю правильно классифицированных объектов, в частности долю ложноположительных и ложноотрицательных случаев.

В зависимости от исходной задачи, если требуется определение количественной величины то можно использовать корень из среднеквадратической ошибки. Команда исследователей которые работают над решением задачи должны быть в курсе о том, каким образом будет проводится оценка «эксплутационных свойств» модели и какой уровень точности предсказания будет достаточно для реализации модели на практике на каком либо производстве [5].

Второй этап предвидит понимание имеющихся данных, работа на этом этапе начинается из предварительного извлечения небольшого набора данных для того чтобы аналитики смогли определить необходимые качества для предстоящих исследований, такие как: упущенные переменные, значения, ошибочные записи, выбросы и т.п.

Практика показывает, что в большинстве случаев присутствуют какие либо проблемы данных, какие необходимо обнаружить на начальных этапах работы с помощью разведочного анализа данных иначе они возникнут рано или поздно, когда уже будет приложено много материальных и других ресурсов. Когда аналитик находит проблемы с данными, то нужно вернутся к первому этапу с целью перепостановки бизнес-задачи которая будет соответствовать достоверным, проверенным данным.

На третьем этапе необходимо учитывая исходные данные подготовить конечный набор данных, на основе которого будет построенная модель. Для проведения данного этапа используют следующие процедуры: очистка данных от лишних наблюдений, отбор потенциально информативных переменных, построение производных переменных на основе существующих, третий этап является наиболее продолжительным из всех шести этапов.

Четвертый этап предусматривает создание самой модели, в зависимости от того какой метод машинного обучения (статистики) используется, происходит формулирование требований к входным данным, также может возникнуть необходимость возвращения к предыдущему этапу и обработать данные согласно требованиям конкретного метода [4].

На пятом этапе исследователь будет иметь в распоряжении одну или даже несколько построенных моделей, которые могут быть перспективными с позиции оценки статистики. При проведении оценки модели главным вопросом становится то – насколько точна модель в своих предсказаниях из учетом сформированных ранее критериев качества, а также нужно промониторить насколько лучше работает построенная Вами модель по сравнению с уже существующими моделями, исследование перечисленных вопросов необходимо для того чтобы подвести итог работы в плане практического применения новой модели предсказания, если возникают противоречия при поиске ответов, то следует обратится к первому этапу поскольку придется переформулировать начальную задачу.

Отметим, что в случае отрицательных ответов на поставленные вопросы лучшим решением будет перепостроение модели нежели запустить неадекватную модель в производство, поскольку тогда есть большая вероятность в материальных потерях.

Последним шестым этапом является запуск модели, при этом даже если Вы проделали несложную работу на основных этапах построения модели, то необходимо понимать что результаты Вашей работы должны быть представлены в понятной форме в той среде и тем людям которые будут с ней работать. Модель проверяют как она работа на производстве, когда она начинает принимать новые данные на вход и выдает предсказания, которые помогают для увеличения прибыли.

Для того чтобы запустить модель в уже существующую среду информационных технологий на производстве присутствуют соответственные специалисты, потому что может возникнуть ситуация когда модель была построенная на одном языке программирования, но в процессе адаптации на производстве ее необходимо перевести на другой язык программирования [5].

Следует уточнить, что при смене условий, в которых модель успешно работала ранее, она может работать менее эффективно и впоследствии будет снижение качества предсказаний (как пример, изменения на фондовых рынках). Для того чтобы избежать описанной проблемы необходимо регулярно исследовать насколько качественные предсказания каждой модели запущенной в производство, а также обучать ее самостоятельно, но существуют модели способны к самообучению, но их также нужно исследовать и проверять на качество работы.

Науке известны такие методы классификации глубокого анализа данных как: дискриминант Фишера; квадратический дискриминант; метод опорных векторов; деревья решений; нейронные сети; Байесовские нейронные сети; генетические алгоритмы; случайный лес. Одним из ключевых заданий является выбор переменных, для этого необходимо выбирать такие переменные, которые имеют индивидуальные свойства, и кроме того, по возможности нужно покрыть все степени свободы задачи, при этом не обязательно обращать внимание на количество переменных. Потому что, при осуществлении анализа данных при работе с приближенными множествами с использованием машинного обучения, а именно методов системы анализа Data Minincg лишние переменные будут убраны [1, 6].

Методы Data Minincg позволяют определить основные типы закономерностей, такие как: ассоциация, последовательность, классификация, кластеризация, прогнозирование. Использование машинного обучения основывается на работе основных методов математической статистики, но в отдельных случаях возникает проблема – концепция усреднения по выборке, что приводит к проведениям операций над неправдоподобными величинами.

На практике статистика успешно себя зарекомендовала при проверке заранее поставленных гипотез, а также для исследования глубокого анализа данных. Если говорить о современных технологиях глубокого анализа данных, то в их основу заложены концепции определенных шаблонов, которые способны отобразить основные аспекты многофункциональных взаимоотношений в данных. Такие шаблоны подаются в форме некоторых закономерностей, которые характеризуют подвыборку данных и их можно интерпретировать на понятный для пользователя язык.

Для отыскания шаблонов используют методы, где не присутствует ограничение предположений о структуре выборки и виде распределения значений исследуемых показателей.

Исследование показало, что главной чертой положения Data Minincg является, то что искомые шаблоны не будут тривиальными. Исходя из этого конечные результаты у виде шаблонов будут отражать скрытые знания в плане неочевидности, неожиданности регулярности в данных.

Исследователи пришли к выводу, что необработанные данные содержат глубокий пласт информационных знаний, и если осуществить грамотную «раскопку» то можно отыскать очень ценную информацию.

Для лучшего представления описанной технологии приведем рис. 1, на котором приведены уровни извлекаемых из данных знаний.

Рис. 1. Уровни извлекаемых из данных знаний

Глубокий интеллектуальный анализ данных в современном мире должен соответствовать следующим требованиям: данные могут иметь неограниченный объем; рассматриваемые данные могут быть разной природы происхождения и иметь разнородные качества (количественные, текстовые); интерпретация результатов должна быть четкой и понятной; инструментарий для обработки сходящих данных должен быть тривиален в использовании [3].

При построении алгоритма для проведения глубокого анализа данных при работе с приближенными множествами необходимо проработать следующие этапы:

1) нахождение классов эквивалентности отношения неразличимости;
2) поиск приближений (верхнего, нижнего);
3) поиск среза решающей системы;
4) конструирование определенного набора решающих правил.

Отметим также, что для того чтобы работать с атрибутами, которые имеют непрерывные области значений необходимо применить дискретизацию. Если входящая информация неполная или противоречивая, то алгоритм будет строить две системы решающих правил, одна система предоставит искомую классификацию, а вторая – возможную. В плане трудоемкости описанного алгоритма наиболее сложными являются шаги по поиску среза и выполнение дискретизации.

На практике, как правило, информационная система характеризируется более чем одним срезом, учитывая это, возникает вопрос выбора наилучшего среза, и в большинстве случаев ним является наиболее короткий срез, при этом задача по поиску такого среза считается *NP*-сложной. Для ее решения возможно использовать несколько подходов.

Одним из наиболее популярных подходов является, то, чтобы исследовать в качестве главных признаков (атрибутов) такие, которые находятся в пересечении всех существующих срезов информационной системы [4].

Вторым не менее известным подходом, является подход, основанный на динамических срезах, то есть множествами условных атрибутов, которые весьма часто встречаются среди срезов подвыборок первоначальной решающей таблицы. Те атрибуты, которые относятся к «большинству» динамических срезов считаются существенными. Определение границ для понятий «достаточно часто» и «большинство» необходимо осуществлять на конкретных данных.

Третий подход построен на введении определения значимости атрибутов, он предоставляет возможность, учитывая вещественные значения из замкнутого интервала [0,1] отразить, насколько главную роль отыгрывает тот или иной атрибут в решающей таблице.

Необходимость проведения этапа дискретизации возникает для большинства современных алгоритмов обобщения, это можно аргументировать тем, что постает задача по преобразовании непрерывной области значений атрибута в дискретную. Отметим, что во время осуществления выбора необходимых интервалов и построения непрерывных областей для значений атрибутов возникает математическая сложность провидения описанных действий, поскольку к числам к которым необходимо применить дискретизацию возникает возрастающая сложность в экспоненциальной зависимости от числа атрибутов.

Опишем процесс постановки задачи дискретизации. Припустим, что $\mathbb{S} = (U, A \cup \{d\})$ – непротиворечивая решающая система. Во время исследования предполагается, что область значений для

каждого атрибута $a \in A$ является вещественным интервалом, т.е. $V_a = [l_a, r_a) \subset R$. Произвольную пару вида $p^a = (a, c)$, где $a \in A$ и $c \in R$, будем называть *делением* на области V_a. Множество деления атрибута a будеть иметь следующий вид – $P_a = \{p_1^a, p_2^a, ..., p_{s_a}^a\}$, где p_i^a – деления атрибута a. Полное множество делений P определяется как $\bigcup_{a \in A} P_a$.

После того, как мы ввели основные обозначения, можно определить новую решающую систему $\mathbb{S}^P = (U, A^P \cup \{d\})$ на основе множества делений P и исходной решающей системы, где $A^P = \{a^P : a \in A\}$ и $a^P(x) = i \Leftrightarrow a(x) \in [c_i^a, c_{i+1}^a)$ для любого объекта $x \in U$ и $i \in \{0, ..., s_a\}$. В конечном результате решающую таблицу $\mathbb{S}^P$ будем называть *дискретизацией* таблицы $\mathbb{S}$ на основе множества делений P.

Главная цель осуществления процесса дискретизации состоит в том, чтобы построить множество делений P. Несложно понять, что алгоритм описанного процесса предвидит построение неограниченного количества множеств делений.

Поиск оптимального множества делений P для исходной решающей системы $\mathbb{S}$ является *NP*-сложной, исходя из этого следует, что построение эффективных эвристических алгоритмов касающихся поиска субоптимального множества делений является значимой задачей.

В целом, алгоритм дискретизации строится на том, что произвольное несократимое множество делений одной решающей таблицы $\mathbb{S}$, служит срезом другой решающей таблицы $\mathbb{S}^* = (U^*, A^* \cup \{d^*\})$, построенной на основе исходной таблицы $\mathbb{S}$, следующим образом: если припустить, что $\mathbb{S} = (U, A \cup \{d\})$ – исходная решающая таблица, а всякий атрибут $a \in A$ задает последовательность $v_1^a < v_2^a < ... < v_{n_a}^a$, где $\{v_1^a, v_2^a, ..., v_{n_a}^a\} = \{a(x) : x \in U\}$ и $n_a \le n$. Тогда, переменными новой решающей таблицы $\mathbb{S}^*$ являются все пары объектов из $\mathbb{S}$ с различными решениями, а множество объединения всех условных атрибутов будет задаваться некими делениями областей значений атрибутов исходной решающей таблицы, аналитически описанная процедура будет иметь следующий вид – $A^* = \bigcup_{a \in A} \left\{ p_i^a : p_i^a = (a, c_i^a), \text{ где } c_i^a = (v_i^a + v_{i+1}^a)/2,\ 1 \le i \le n_a - 1 \right\}$.

Отметим, что рассматриваемые атрибуты будут бинарными. Множество A^* принято называть начальным множеством делений. При исследовании приближенных множеств деление $p_i^a = (a, c_i^a)$ различает объекты x и y из разных классов решений, если $min(a(x), a(y)) < c_i^a < max(a(x), a(y))$. Значение нового атрибута, соответствующего делению p_i^a, для пары (x, y) равно 1, если с помощью этого деления различаются объекты x и y, и 0 в противоположном случае. Так же необходимо учитывать, что к объектам новой решающей таблицы добавляется еще один объект $\perp$, для которого все условия и решение d^* будут иметь значение 0. Для других объектов построенной решающей таблицы значение нового решения равно 1. Нетрудно понять, что срезы новой решающей таблицы $\mathbb{S}^*$ определяют все несократимые множества делений исходной решающей таблицы $\mathbb{S}$.

Исследование литературных источников показывает, что интеллектуальный (глубокий) анализ данных является комбинированной областью, которая зародилась на базе успешных результатов прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и других схожих дисциплин [3-5].

Мультидисциплинарность системы анализа Data Minincg объясняет наличие огромного количества алгоритмов, которые возможна реализовать на практике в различных действующих системах, а некоторые из них способны объединять в себе сразу несколько подходов, но вместе с этим каждая система содержит главный компонент на который делается решающая ставка.

На рис. 2 проиллюстрируем основные дисциплины, которые входят в систему Data Minincg:

Рис. 2. Основные дисциплины

Таким образом, учитывая приведенные результаты исследования, видим что использование современных технологий при проведении многофункционального исследования данных с использованием теории приближенных методов набирает стремительный темп развития, поэтому мы можем спрогнозировать повышенный интерес к использованию машинного обучения, а в вчасности аппарата математической статистики для анализа гипотез и т.п.

Выводы. В статьи показано, что главным инструментом глубокого анализа данных с привлечением теории приближенных множеств является математическая статистика, особенно успешно она используется для проверки заранее сформулированных гипотез. Отметим, что аппарат математической статистики, в отдельных случаях не весьма полезен, по причине – концепции усреднения по выборке и как результат приводит к операциям над фиктивными переменными. На рис. 2 проиллюстрированные основные дисциплины, которые входят в систему Data Mincg, следует отметить, что каждая система содержит главный компонент на который делается решающая ставка.

Проведен анализ основных этапов протокола CRISP-DM, указанный протокол предназначенный для построения предсказательных моделей, которые предоставляют возможность решать бизнес-задачи различной природы возникновения с привлечением машинного обучения.

Установлено, что общий подход для большей части алгоритмов дискретизации основан на том, что произвольное несократимое множество делений решающей таблицы является срезом другой решающей таблицы, построенной на основе исходной таблицы.

На основе изложенного материала, можно утверждать, что машинное обучение предоставляет возможность осуществлять глубокий анализ данных, поскольку его главной целью является создание таких систем, которые способны получать знания из данных, и кроме этого имеют способность с помощью обучения улучшать показатели своей работы. Исходя из этого, видим, что машинное обучение является одной из областей науки о данных. Для лучшего представления уровней извлекаемых из данных знаний предложен рис. 1.

Список литературы / References

1. Базы данных. Интеллектуальная обработка информации / В. Корнеев, А. Гареев, С.В. Васютин, В.В. Райх. М.: Нолидж, 2001. 653 с.
2. *Вагин В.Н.* Знание в интеллектуальных системах // Новости искусственного интеллекта, 2002. № 6 (54). [Электронный ресурс]. Режим доступа: http://www.raai.org/about/persons/vagin/pages/vagin_zn.doc/ (дата обращения: 10.06.2019); Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта, 2004. № 3.

3. *Куликов А.В., Фомина М.В.* Разработка алгоритма обобщения понятий с использованием подхода, основанного на теории приближенных множеств / Труды шестой международной конференции по технологии программирования на основе знаний // Под ред. В. Стефанюка и К. Каджири. IOS Press, 2004. С. 261–268 (на англ. яз.).
4. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Юнюков, Л.Д. Мешалкин. М.: Финансы и статистика, 2003. 483 с.
5. *Lan A.S. et al.* Mathematical languageprocessing: Automatic grading and feedback for open response mathematical questions // Proceedings of the Second (2015) ACM Conference on Learning@ Scale. ACM, 2015. С. 167–176.
6. *Parsaye K.A* Characterization of Data Mining Technologies and Processes // The Journal of Data Warehousing, 1998. № 1.