УДК 004.65:004.8

124. ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИИ DATA MAINING В АНАЛИЗЕ ДАННЫХ

Кульбако А.Е., Рыбак К.А., студентки группы 377901
Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Нагулевич Р.С. – магистр экономических наук, старший преподаватель каф. ЭИ

Аннотация. В работе рассматривается технология Data Mining как инструмент извлечения скрытых знаний из больших массивов данных. Описаны основные типы выявляемых закономерностей: ассоциация, последовательность, классификация, кластеризация и прогнозирование временных рядов. Приведена классификация методов анализа – от статистических до нейросетевых алгоритмов и деревьев решений. Особое внимание уделено интеграции Data Mining с OLAP-технологиями для повышения эффективности поддержки принятия решений. Показано, что успешное применение технологии определяется не только выбором алгоритмов, но и культурой работы с данными.

Ключевые слова. Data Mining, ассоциация, последовательность, классификация, дискриминантный анализ, кластеризация, кластерный анализ, прогнозирование временных рядов, нейросетевые алгоритмы, деревья решений, OLAP Data Mining.

В современном мире объем накапливаемых данных растет экспоненциально. Предприятия, научные учреждения, государственные структуры сталкиваются с проблемой не столько сбора информации, сколько ее эффективного анализа и извлечения полезных знаний. Традиционные методы статистической обработки данных не всегда позволяют выявить скрытые, неочевидные закономерности, которые могут быть критически важны для принятия решений. В этих условиях технология Data Mining (интеллектуальный анализ данных) становится ключевым инструментом, позволяющим автоматически обнаруживать корреляции, тенденции и взаимосвязи в больших массивах информации.

Термин Data Mining [1] обозначает процесс поиска корреляций, тенденций и взаимосвязей посредством различных математических и статистических алгоритмов: кластеризации, регрессионного и корреляционного анализа и других, применяемых в системах поддержки принятия решений. При этом накопленные сведения автоматически обобщаются до информации, которая может быть охарактеризована как знания.

В основу современной технологии Data Mining положена концепция шаблонов, отражающих закономерности, свойственные подвыборкам данных и составляющие так называемые скрытые знания, представленных на рисунке 1. Ключевой особенностью Data Mining является то, что поиск шаблонов производится методами, не использующими никаких априорных предположений об этих подвыборках. Иными словами, средства Data Mining отличаются от инструментов статистической обработки данных и средств OLAP тем, что вместо проверки заранее предполагаемых пользователями взаимосвязей между данными, они на основании имеющихся данных способны самостоятельно находить такие взаимосвязи, а также строить гипотезы об их характере.

Рисунок 1 – Уровни знаний, извлекаемых из данных [2]

Процесс интеллектуального анализа данных в общем случае состоит из трех основных стадий, представленных на рисунке 2. Первая стадия – выявление закономерностей, или свободный поиск, на которой алгоритмы Data Mining исследуют массив данных с целью обнаружения статистически значимых паттернов. Вторая стадия – использование выявленных закономерностей для предсказания неизвестных значений, что называется прогностическим моделированием. На этой стадии построенные модели применяются для прогнозирования поведения объектов на основе известных характеристик. Третья стадия – анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях. Аномалии могут представлять особую ценность, поскольку часто указывают на нештатные ситуации, мошеннические действия или новые, ранее не известные явления.

Между стадиями нахождения и использования закономерностей иногда в явном виде выделяют промежуточную стадию – валидацию, то есть проверку достоверности найденных закономерностей. Валидация необходима для того, чтобы исключить случайные корреляции и убедиться в применимости полученных моделей к новым данным.

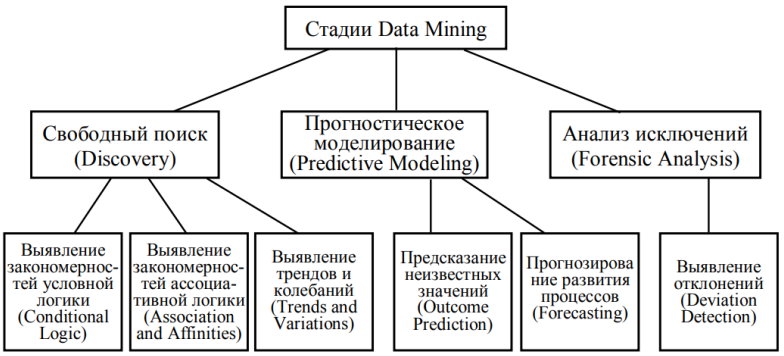

Рисунок 2 – Стадии процесса интеллектуального анализа данных [2]

Основные типы закономерностей

Методами Data Mining выделяют пять стандартных типов закономерностей, каждый из которых решает определенный класс задач.

Ассоциация позволяет выделить устойчивые группы объектов, между которыми существуют неявно заданные связи. Ассоциации записываются в виде правил A ⇒ B, где A – посылка, B – следствие. Для оценки значимости правил используются такие показатели, как распространенность (частота появления комбинации), доверительность (вероятность появления B при условии A) и мощность (сила влияния A на B).

Последовательность представляет собой метод выявления ассоциаций во времени. В данном случае определяются правила, описывающие последовательное появление определенных групп событий. Такие правила необходимы для построения сценариев и могут использоваться, например, для формирования типичного набора предшествующих продаж, которые влекут за собой последующие продажи конкретного товара.

Классификация является инструментом обобщения, позволяющим перейти от рассмотрения единичных объектов к обобщенным понятиям, характеризующим совокупности объектов (классы). Процесс классификации базируется на двух основных процедурах: обучении (построение классифицирующего правила на основе обучающей выборки) и проверке (оценка качества правила на новых данных). В рамках классификации особое место занимает дискриминантный анализ, который находит линейные или квадратичные комбинации признаков, наилучшим образом разделяющие заранее заданные классы. Дискриминантный анализ позволяет не только относить новые объекты к одному из классов, но и оценивать вклад каждого признака в разделение групп, что важно для интерпретации результатов.

Кластеризация – это распределение информации из базы данных по группам (кластерам) с одновременным определением этих групп. В отличие от классификации, здесь для проведения анализа не требуется предварительного задания классов. Кластерный анализ объединяет совокупность методов (k-means, иерархические методы, DBSCAN и др.), которые на основе мер сходства или расстояния между объектами формируют группы таким образом, чтобы объекты внутри кластера были максимально похожи, а между кластерами – максимально различны. Кластеризация широко применяется для сегментации клиентов, выделения групп риска, анализа схожести документов.

Прогнозирование временных рядов является инструментом для определения тенденций изменения атрибутов рассматриваемых объектов с течением времени. Анализ поведения временных рядов позволяет прогнозировать значения исследуемых характеристик на основе исторических данных. Данный тип закономерностей особенно востребован в финансовой сфере, при прогнозировании спроса, в экономическом и макроэкономическом анализе.

Методы и алгоритмы Data Mining

Data Mining развивалась на стыке таких дисциплин, как статистика, теория информации, машинное обучение и теория баз данных [3]. Соответственно, большинство алгоритмов и методов Data Mining были разработаны на основе методов этих дисциплин. Среди многообразия существующих методов исследования данных можно выделить несколько самых распространённых.

Статистические методы, включающие регрессионный, дисперсионный и корреляционный анализ, реализованы в большинстве современных статистических пакетов (SAS Institute, StatSoft и др.). Эти методы являются фундаментальными для количественной оценки взаимосвязей между переменными.

Нейросетевые алгоритмы представляют собой метод имитации процессов и явлений, позволяющий воспроизводить сложные нелинейные зависимости. Метод основан на использовании упрощенной модели биологического мозга: исходные параметры рассматриваются как сигналы, преобразующиеся в соответствии со связями между нейронами, а отклик всей сети является результатом анализа. Нейронные сети широко применяются для решения задач классификации, прогнозирования и распознавания образов.

Нечеткая логика применяется для обработки данных с размытыми значениями истинности, которые могут быть представлены лингвистическими переменными. Нечеткое представление знаний используется для решения задач классификации и прогнозирования в системах типа XpertRule Miner, AIS, NeuFuz.

Индуктивные выводы позволяют получать обобщения фактов, хранящихся в базах данных. В процессе индуктивного обучения может участвовать специалист, поставляющий гипотезы (обучение с учителем), либо поиск правил обобщения может осуществляться без учителя путем автоматической генерации гипотез. Современные программные средства, как правило, сочетают оба подхода.

Рассуждения на основе аналогичных случаев (Case-based reasoning, CBR) основаны на поиске в базе данных ситуаций, описания которых сходны с заданной ситуацией. Принцип аналогии позволяет предполагать, что результаты похожих ситуаций также будут близки. Недостаток метода заключается в том, что здесь не создается обобщающих моделей или правил, а надежность выводов зависит от полноты описания ситуаций.

Деревья решений представляют собой метод структурирования задачи в виде древовидного графа, вершины которого соответствуют продукционным правилам, позволяющим классифицировать данные или анализировать последствия решений. Метод дает наглядное представление о системе классифицирующих правил и эффективен для простых и средних по сложности задач.

Эволюционное программирование основано на поиске и генерации алгоритма, выражающего взаимозависимость данных, путем модификации изначально заданного алгоритма. Алгоритмы ограниченного перебора вычисляют частоты комбинаций простых логических событий в подгруппах данных.

Интеграция Data Mining и OLAP

Оперативная аналитическая обработка (OLAP) и интеллектуальный анализ данных (Data Mining) представляют собой две составные части процесса поддержки принятия решений. Однако большинство систем OLAP традиционно фокусируются на обеспечении доступа к многомерным данным, в то время как средства Data Mining работают с одномерными перспективами данных. Для повышения эффективности обработки данных эти два вида анализа должны быть объединены.

В настоящее время получает распространение составной термин «OLAP Data Mining» (многомерный интеллектуальный анализ), обозначающий такую интеграцию. Существует три основных способа формирования OLAP Data Mining.

Первый способ – «Cubing then mining» – предполагает выполнение интеллектуального анализа над любым результатом запроса к многомерному концептуальному представлению, то есть над любым фрагментом любой проекции гиперкуба показателей. Это позволяет пользователю сначала сформировать нужное многомерное сечение данных с помощью OLAP-инструментов, а затем применить методы Data Mining к полученному подмножеству.

Второй способ – «Mining then cubing» – заключается в том, что результаты интеллектуального анализа представляются в гиперкубической форме для последующего многомерного анализа. Таким образом, найденные закономерности, прогнозы или кластеры становятся новыми измерениями гиперкуба, доступными для дальнейшего исследования.

Третий способ – «Cubing while mining» – представляет собой гибкую интеграцию, позволяющую автоматически активизировать механизмы интеллектуальной обработки над результатом каждого шага многомерного анализа: при переходе между уровнями обобщения, извлечении нового фрагмента гиперкуба и т.д. Такой подход обеспечивает наиболее тесное взаимодействие между аналитическими инструментами.

Data Mining меняет сам подход к работе с данными. Раньше аналитик сначала придумывал гипотезу, а потом проверял её с помощью инструментов. Теперь же данные сами подсказывают, на что стоит обратить внимание, и помогают находить неожиданные связи. В итоге специалист может сосредоточиться на самом главном – на осмыслении того, что удалось обнаружить.

Успех Data Mining зависит не столько от сложности алгоритмов, сколько от того, насколько грамотно люди обращаются с данными: понимают ли их ограничения, видят ли возможные ошибки, умеют ли задавать правильные вопросы. В этом смысле технология становится не просто инструментом поиска ответов, а способом лучше понять то, что скрыто в данных.

Список использованных источников:

1. What is data mining? [Электронный ресурс] // IBM. – URL: https://www.trendmicro.com/ru_ru/what-is/machine-learning/data-mining.html (дата обращения: 29.03.2026).

2. Борисов, Д.Н. Корпоративные информационные системы : учебно-методическое пособие для вузов / Д.Н. Борисов // Государственное образовательное учреждение высшего профессионального образования «Воронежский государственный университет». – Воронеж : Издательско-полиграфический центр Воронежского государственного университета, 2007. – С. 37–42.

3. Технология Data Mining [Электронный ресурс] // Decosystems. – URL: https://www.decosystems.ru/tekhnologiya-datamining/ (дата обращения: 29.03.2026)

UDC 004.65:004.8

USING DATA MAINING TECHNOLOGY IN DATA ANALYSIS

Kulbako A.E., Rybak K.A.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Nagulevich R.S. – Senior Lecturer, Department of Economic Informatics

Annotation. The paper considers Data Mining technology as a tool for extracting hidden knowledge from large amounts of data. The main types of revealed patterns are described: association, sequence, classification, clustering, and time series forecasting. The classification of analysis methods is given, from statistical to neural network algorithms and decision trees. Special attention is paid to the integration of Data Mining with OLAP technologies to improve the efficiency of decision support. It is shown that the successful application of technology is determined not only by the choice of algorithms, but also by the culture of working with data.

Keywords. Data Mining, association, sequence, classification, discriminant analysis, clustering, cluster analysis, time series forecasting, neural network algorithms, decision trees, OLAP Data Mining.