УДК 004.8 (075)

РОЛЬ KDD В РАЗВИТИИ ИНСТРУМЕНТОВ АНАЛИЗА И УПРАВЛЕНИЯ ДАННЫМИ

Ф.М. Алимова
Доцент кафедры Конвергенция цифровых технологий ТУИТ имении Мухаммада ал Хоразмий, доктор философии технических наук, PhD
tuit.alimova@gmail.com

Ф.М. Алимова
Окончила Ташкентский университет информационных технологий. Область научных интересов связана с исследованием проблем оптимизации бизнес процессов, построения информационных систем, организацией учебного и научно-исследовательского процессов в техническом университете.

Аннотация. В статье рассматривается роль процесса извлечения знаний из данных (KDD) в анализе и управлении образовательными данными. Выявлены особенности методов интеллектуального анализа данных (Data Mining), такие как кластеризация, классификация, прогнозирование и анализ ассоциаций. Освещаются анализ успеваемости учащихся, моделирование образовательных данных и интеграцию аналитических методов в учебные процессы. Рассматриваются преимущества и ограничения различных методов KDD, а также их практическое применение для оптимизации образовательных систем. В статье подчеркивается важность автоматизированных методологий анализа данных для выявления полезных знаний и принятия обоснованных решений в сфере образования.

Ключевые слова: Data Mining, big data, анализа данных, знания.

Введение. В сфере образования большие объемы данных собираются в системах управления базами данных и хранилищах данных из всех вовлеченных областей. Анализ данных стал важным инструментом для сбора знаний из учебных баз данных. Основные функции интеллектуального анализа данных, которые должны быть выполнены, включают в себя характеристику и описание, ассоциацию, классификацию, прогнозирование, кластеризацию и анализ эволюции. Было показано, что в последние годы наблюдается быстрый рост применения интеллектуального анализа данных в контексте образовательных процессов и учреждений. Этот обзор раскрывает прогрессивные применения и существующие пробелы, выявленные в контексте интеллектуального анализа данных в образовательном сфере.

В последнее время большое внимание уделяется важному аспекту инновации-большие данные. Учитывая важность сферы образования, в настоящее время наблюдается тенденция к изучению роли больших данных в этом секторе. На данный момент проводиться множество исследований, посвящённых применению больших данных в различных областях для разных целей. Однако в сфере образования до сих пор нет всестороннего обзора больших данных.

На сегодняшний день проводимые исследования охватывают четыре основные темы, связанные с большими данными в образовании: поведение и успеваемость учащихся,

моделирование и хранилища образовательных данных, совершенствование образовательной системы и интеграция больших данных в учебную программу. Большинство исследований в области образования, связанных с большими данными, были сосредоточены на поведении и успеваемости учащихся.

Большие объемы данных в образовательных базах данных, которые содержат большое количество записей со множеством атрибутов, которые необходимо одновременно исследовать для обнаружения полезной информации и знаний, делают ручной анализ непрактичным. Все эти факторы указывают на необходимость интеллектуальных и автоматизированных методологий анализа данных, которые могли бы обнаруживать полезные знания из данных. Поэтому обнаружение знаний в базах данных (KDD) и интеллектуальный анализ данных (DM) стали чрезвычайно важными инструментами в реализации цели интеллектуального и автоматизированного анализа данных. Интеллектуальный анализ данных является особым шагом в процессе KDD, включающим применение определенных алгоритмов для извлечения шаблонов (моделей) из данных. Дополнительные шаги в процессе KDD, такие как подготовка данных, очистка данных, выбор данных, включение соответствующих предварительных знаний и правильная интерпретация результатов интеллектуального анализа, гарантируют, что полезные знания будут извлечены из данных.

KDD включает в себя теории, алгоритмы и методы из пересечения нескольких областей исследований, включая технологию баз данных, машинное обучение, статистику, искусственный интеллект, системы, основанные на знаниях, и визуализацию данных.

KDD отличается от других областей, таких как машинное обучение или искусственный интеллект или смежных областей тем, что эти области предоставляют конкретные инструменты добычи данных, которые могут использоваться на различных этапах процесса KDD. В последнее время, с ростом технологии добычи данных, исследователи и практики в различных аспектах производства и логистики начали применять эту технологию для поиска скрытых связей или закономерностей, которые могли бы использоваться для оснащения их систем новыми знаниями. Ранние приложения добычи данных в основном применялись к финансовым приложениям, например, Чжан и Чжоу (2004) описали добычу данных в контексте финансовых приложений как с технической, так и с прикладной точки зрения. Главным преимуществом интеллектуального анализа данных перед другими экспериментальными методами является то, что необходимые данные для анализа могут быть собраны во время нормальной работы изучаемого производственного процесса. Поэтому, как правило, нет необходимости специально выделять машины или процессы для сбора данных.

Большинство исследований в области образования, связанных с большими данными, были сосредоточены на поведении и успеваемости учащихся. Это исследование служит руководством для будущих исследований и освещает новые идеи и направления для успешного использования больших данных в образовании.

Методы интеллектуального анализа даёт возможность выполнять систематизацию данных по критериям количества и качества. При обработке массивов информации наиболее часто используется следующие способы:

- кластерный анализ (иерархический и неиерархический);
- байесовские сети;
- поиск явных и неявных ассоциаций;
- линейная регрессия;
- представление сведений в визуальной форме;
- эволюционное программирование;
- генетические алгоритмы.

Каждый из перечисленных методов применяется для выборки конкретных данных, построения прогностической модели или очистки БД от ошибочных сведений.

Хан и Камбер (2001) отметили, что вид знаний, которые необходимо добыть, определяет функции добычи данных, которые необходимо выполнить. Возможные виды знаний включают описание концепций (характеристику и дискриминацию), ассоциацию, классификацию, кластеризацию и прогнозирование.

KDD– это процесс получения из данных знаний, в виде зависимости правил моделей позволяющих моделировать и прогнозировать различные процессы.

Технологии KDD описывают не конкретный алгоритм обработки данных или математический аппарат, а последовательность действий которую необходимо выполнить с данными для построения модели и получения знаний.

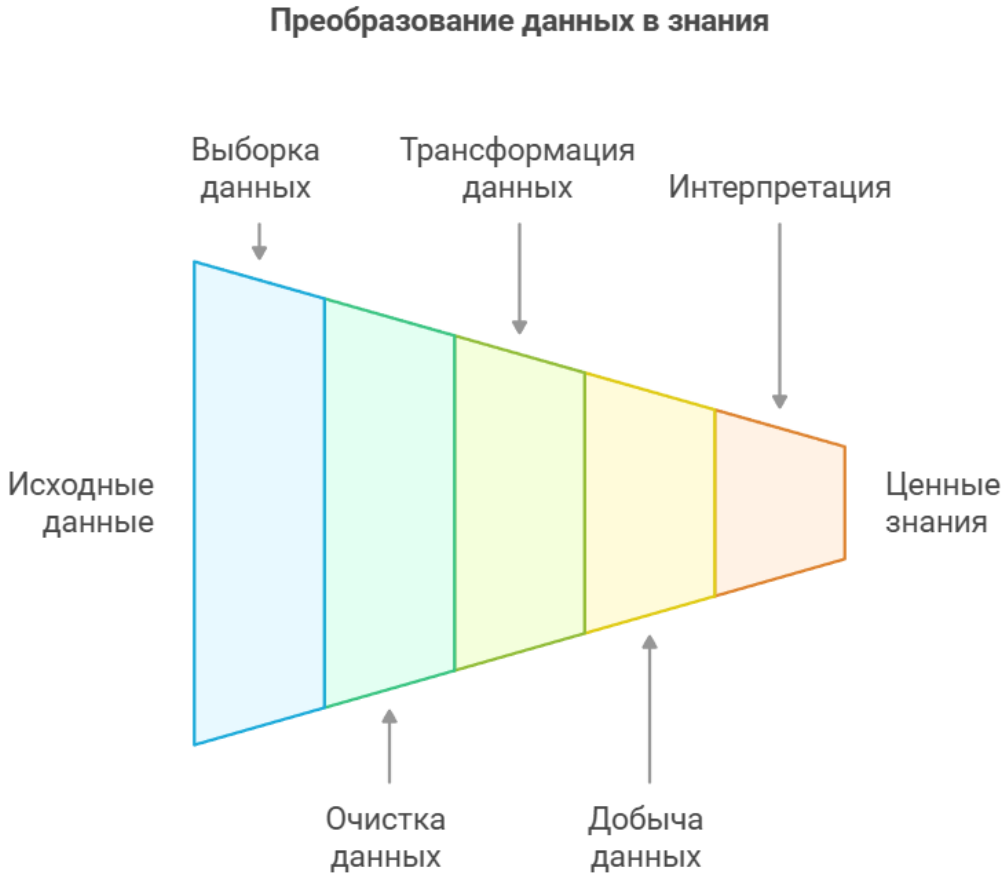

Рисунок 1. Процесс KDD в сфере образования

Процесс KDD является интерактивным и итеративным процессом которая предполагает динамическое взаимодействие и повторение действий для достижения наилучшего результата включающим более или менее следующие этапы.

1 Понимание области: Это включает в себя соответствующие предыдущие знания, связанные с применением образования и целевой целью.

2 Сбор целевых данных: Это включает в себя сбор необработанных данных, выбор набора данных и фокусирование на наборе переменных, влияющих на процесс образования.

3 Очистка, предварительная обработка и преобразование данных: Это включает предварительную обработку данных, такую как удаление шума, замена пропущенных значений и очистка данных. Данные консолидируются в формы, подходящие для добычи данных.

4 Интеграция данных: Это включает интеграцию нескольких образовательных и производственных гетерогенных источников данных.

5 Выбор функций добычи данных: На основе типа требуемых знаний необходимо выполнить различные функции добычи данных (кластеризация, классификация, прогнозирование, ассоциация, регрессия, суммирование и т. д.) для получения модели.

6 Выбор подходящего алгоритма интеллектуального анализа данных: это включает выбор методов для выполнения желаемой функции поиска закономерностей в данных.

7 Интеллектуальный анализ данных: это включает поиск закономерностей, представляющих интерес, в определенной репрезентативной форме или наборе таких представлений.

8 Интерпретация и визуализация: это включает интерпретацию и визуализацию закономерностей для получения новых знаний.

9 Внедрение обнаруженных знаний: обнаруженные знания включаются в систему производительности области образования. Получается обратная связь, и знания могут быть дополнительно изменены на основе обратной связи.

10 Хранение, повторное использование и интеграция знаний в образовательную систему: Это включает хранение обнаруженных знаний для будущего повторного использования и возможной интеграции в образовательной системе

Алгоритмы поиска бывают двух типов: поиск параметров по заданной модели и поиск моделей в пространстве моделей. Поиск наилучших параметров часто сводится к задаче оптимизации. Алгоритмы добычи данных, как правило, опираются на относительно простые методы оптимизации, хотя в принципе используются и более сложные методы оптимизации. Проблемы с локальными минимумами являются обычными и решаются обычным способом (например, множественные случайные перезапуски и поиск нескольких моделей). Поиск в пространстве моделей обычно выполняется жадным способом. Важным моментом является то, что каждый метод обычно подходит для некоторых задач лучше, чем для других. Например, классификаторы деревьев решений могут быть очень полезны для поиска структуры в многомерных пространствах, а также полезны в задачах со смешанными непрерывными и категориальными данными (поскольку методы деревьев не требуют метрик расстояний). Однако деревья классификации с одномерными пороговыми границами решений могут не подходить для задач, где истинные границы решений являются нелинейными многомерными функциями. Таким образом, не существует универсально наилучшего метода интеллектуального анализа данных. Выбор конкретного алгоритма для конкретного приложения — это своего рода искусство. На практике большая часть усилий приложения может быть направлена на правильную формулировку проблемы (постановку правильного вопроса), а не на оптимизацию алгоритмических деталей конкретного метода добычи данных.

Высокоуровневые цели добычи данных, как правило, являются предиктивными, описательными или комбинацией предиктивных и описательных. Чисто предиктивная цель фокусируется на точности предиктивной способности. Чисто описательная цель фокусируется на понимании основного процесса генерации данных — тонкое, но важное различие. При прогнозировании пользователю может быть все равно, отражает ли модель реальность, если она имеет предиктивную силу. Описательная модель, с другой стороны, интерпретируется как отражение реальности. На практике, большинство приложений KDD требуют некоторой степени как предиктивного, так и описательного моделирования.

Основные исследовательские и прикладные проблемы для KDD включают:

1 Массивные наборы данных и высокую размерность. Многогигабайтные базы данных с миллионами записей и большим количеством полей (атрибутов и переменных) являются обычным явлением. Эти наборы данных создают комбинаторно взрывные пространства поиска для индукции модели и увеличивают вероятность того, что алгоритм добычи данных найдет ложные шаблоны, которые в целом не являются допустимыми. Возможные решения включают очень эффективные алгоритмы, выборку, методы аппроксимации, массивно-параллельную обработку, методы снижения размерности и включение предыдущих знаний.

2 Взаимодействие с пользователем и предыдущие знания. Аналитик обычно не является экспертом KDD, а лицом, ответственным за осмысление данных с использованием

доступных методов KDD. Поскольку процесс KDD по определению интерактивен и итеративен, сложно обеспечить высокопроизводительную, быстродействующую среду, которая также помогает пользователям в правильном выборе и сопоставлении соответствующих инструментов и методов для достижения их целей. Необходимо больше внимания уделять взаимодействию человека и компьютера и меньше — полной автоматизации — с целью поддержки как опытных, так и начинающих пользователей. Многие текущие методы и инструменты KDD не являются по-настоящему интерактивными и не позволяют легко включать предыдущие знания о проблеме, за исключением простых способов. Использование знаний о предметной области важно на всех этапах процесса KDD. Например, байесовские подходы используют априорные вероятности по данным и распределениям как один из способов кодирования предыдущих знаний. Другие используют возможности дедуктивной базы данных для обнаружения знаний, которые затем используются для руководства поиском по добыче данных.

3 Интеграция. Отдельная система обнаружения может быть не очень полезной. Типичные проблемы интеграции включают интеграцию с СУБД (например, через интерфейс запросов), интеграцию с электронными таблицами и инструментами визуализации, и размещение показаний датчиков в реальном времени.

Высоко интерактивные среды человек-компьютер, как описано в процессе KDD, допускают как компьютерное обнаружение с помощью человека, так и компьютерное обнаружение с помощью человека. Разработка инструментов для визуализации, интерпретации и анализа обнаруженных шаблонов имеет первостепенное значение. Такие интерактивные среды могут обеспечить практические решения многих реальных проблем гораздо быстрее, чем люди или компьютеры, работающие независимо. Существует потенциальная возможность и проблема разработки методов интеграции инструментов OLAP сообщества баз данных и инструментов интеллектуального анализа данных машинного обучения и статистических сообществ.

4 Нестандартные, мультимедийные и объектно-ориентированные данные. Значительной тенденцией является то, что базы данных содержат не только числовые данные, но и большие объемы нестандартных и мультимедийных данных. Нестандартные типы данных включают нечисловые, нетекстовые, геометрические и графические данные, а также нестационарные, временные, пространственные и реляционные данные, а также смесь категориальных и числовых полей в данных. Мультимедийные данные включают многоязычный текст в свободной форме, а также оцифрованные изображения, видео, речевые и аудиоданные.

Эти типы данных в значительной степени выходят за рамки текущей технологии KDD.

Заключение. Несмотря на быстрый рост, область KDD все еще находится в зачаточном состоянии. Необходимо преодолеть множество проблем, но некоторые успехи были достигнуты. Поскольку потенциальная отдача от приложений KDD высока, на рынке наблюдается спешка с предложением продуктов и услуг. Большой проблемой, стоящей перед областью, является то, как избежать ложных ожиданий, преследующих другие зарождающиеся (и связанные) технологии. Исследователи и практики в этой области обязаны обеспечить, чтобы потенциальный вклад KDD не был преувеличен и чтобы пользователи понимали истинную природу вклада вместе с его ограничениями. Фундаментальные проблемы, лежащие в основе этой области, остаются нерешенными. Например, основные проблемы статистического вывода и открытия остаются такими же сложными и затруднительными, какими они были всегда. Овладение искусством анализа и способностью человеческого мозга синтезировать новые знания из данных по-прежнему непревзойденно ни одной машиной. Однако объемы данных, которые необходимо проанализировать, делают машины необходимостью. Эта ниша для использования машин в качестве вспомогательного средства для анализа и надежда на то, что огромные наборы

данных содержат крупицы ценных знаний, стимулируют интерес и исследования в этой области. Объединяя набор различных областей, KDD создает плодородную почву для роста новых инструментов для управления, анализа и, в конечном итоге, получения превосходства над потоком данных, с которым сталкивается современное общество. Тот факт, что область движима сильными социальными и экономическими потребностями, является стимулом для ее дальнейшего роста. Проверка реальности реальных приложений будет действовать как фильтр, отсеивающий хорошие теории и методы от менее полезных.

Список литературы

[1] A.Abdullaye, F.M. Alimova, U.Giyosov. Control of knowledge in the test. “XXI аср ва технология соҳасидаги устувор йўналишлар”, VII Халқаро илмий конференцияси бўйича мақолалар тўплами, 2021. – 477 с.

[2] F.M. Alimova. Data mining in customer relationship managemen. TATU xabarlari, 2014. – 11 с.

[3] Usmanov Jonibek Turdikulovich, Pulatova Ziyoda Mahmudjonovna, Naim Nodira Abdujalolovna, Makhbuba Kushmanova Abdunabiyevna, Alimova Fotima Muratovna. Developing a model for comprehensive and similarity analysis of systems. International Journal of Mechatronics and Applied Mechanics, 2022, Issue 11c.

THE ROLE OF KDD IN THE DEVELOPMENT OF DATA ANALYSIS AND MANAGEMENT TOOLS

F.M. Alimova
Associate Professor, Department of Convergence of digital technologies, PhD of Technical sciences, Associate Professor

Abstract. The article discusses the role of knowledge-driven data mining (KDD) in analyzing and managing educational data. The article highlights the features of data mining methods, such as clustering, classification, forecasting, and association analysis. It covers the analysis of student performance, educational data modeling, and the integration of analytical methods into educational processes. The advantages and limitations of various KDD methods are discussed, as well as their practical application for optimizing educational systems. The article emphasizes the importance of automated data mining methodologies for discovering useful knowledge and making informed decisions in the field of education.

Keywords: Data Mining, big data, data analysis, knowledge.