Проблемы извлечения скрытых знаний из распределенных баз данных с использованием алгоритмов машинного обучения: методы повышения согласованности и эффективности

Р. А. Турусов

Санкт-Петербургский политехнический университет Петра Великого

turusov.ra@edu.spbstu.ru

О. Ю. Сабинин

Санкт-Петербургский политехнический университет Петра Великого

sabinin_oyu@spbstu.ru

Аннотация. **В условиях стремительного роста количества хранимых гетерогенных данных особую актуальность приобретает задача извлечения скрытых знаний из распределенных баз данных. Существующие методы и подходы, включающие федеративное обучение и мультиагентные системы, демонстрируют ограничения при работе с мультимодальными данными и не обеспечивают необходимый уровень интерпретируемости результатов. В статье обосновывается понятие скрытого знания, определяются ключевые проблемы, препятствующие получению знаний из распределенных баз данных, дается теоретическое обоснование методологии, основанной на децентрализованной архитектуре интеллектуальных агентов, которая позволяет преодолеть ключевые ограничения современных решений в рамках поиска знаний в распределенных системах. Основу предлагаемой методологии составляет система специализированных агентов, осуществляющих глубокий анализ данных различных типов с помощью алгоритмов машинного обучения с последующей семантической интеграцией результатов через динамический граф знаний. Особенностью подхода является сохранение распределенного характера данных при одновременном обеспечении целостного представления выявленных закономерностей.**

Ключевые слова: распределенные базы данных, извлечение знаний, федеративное обучение, искусственный интеллект, машинное обучение

I. Введение

В настоящее время, благодаря росту количества используемых технологий, генерируется огромное количество разнородной информации. Это вызывает необходимость в хранении, которая решается использованием распределенных баз данных (РБД). РБД позволяют обрабатывать большие объемы информации за счет горизонтального и вертикального масштабирования, повышают отказоустойчивость благодаря репликации и децентрализованному хранению, а также соответствуют требованиям безопасности и законодательства. Они поддерживают различные форматы данных, что делает их удобными для практических применений в различных отраслях.

Данные в РБД представляют собой ценный ресурс, содержащий скрытые знания, которые могут быть использованы в процессе принятия решений в различных сферах жизни общества и бизнеса. Для их извлечения применяются статистические методы и машинное обучение. Однако извлечение знаний в РБД сталкивается с проблемами: гетерогенность данных, несогласованность из-за асинхронного обновления на узлах, потеря эффективности при передаче данных между узлами, ограничения безопасности, связанные с рисками утечки конфиденциальной информации.

Задача извлечения знаний из РБД требует компромисса между эффективностью анализа и ограничениями распределенной среды. В статье рассматриваются ключевые проблемы, описываются существующие подходы к извлечению знаний из РБД, решающие эти проблемы, предлагается новая методология для повышения эффективности извлечения скрытых знаний.

II. Получение скрытых знаний из данных

Скрытые знания представляют собой сложные, нетривиальные взаимосвязи, отражающие глубинные взаимодействия между данными. Они неочевидны при поверхностном классическом анализе и требуют применения специализированных методов для выявления. В конкретном смысле под знаниями стоит понимать правила, с высокой вероятностью выполняющиеся в данных, ценные признаки, влияющие на процессы классификации объектов и обучение моделей. Эти знания позволяют получать новые паттерны в данных, прогнозировать события и оптимизировать процессы в различных отраслях, таких как бизнес, медицина, финансы и промышленность [1].

Для целей извлечения скрытых знаний в классическом смысле в системах с централизованным хранением данных используются методы машинного обучения, которые включают в себя кластеризацию, классификацию, ассоциативные правила, графовый анализ, поиск аномалий, тематическое моделирование, анализ временных рядов, интерпретацию результатов глубокого обучения нейронных сетей, а также методы уменьшения размерности данных. Эти методы работают с различными типами данных, включая числовые, категориальные, текст и изображения.

Кластеризация, является методом обучения без учителя и группирует объекты на основе схожести признаков. Классификация напротив идентифицирует зависимости между признаками и целевой заданной переменной, что относит ее к обучению с учителем, но она может использоваться в рамках получения скрытых знаний в контексте выделения важности признаков. Ассоциативные правила обнаруживают частые комбинации элементов в наборах данных, выявляя скрытые взаимосвязи между объектами. Тематическое моделирование автоматически выделяет семантические темы в текстовых данных, позволяя находить нетривиальную информацию. Графовые методы анализируют структуру данных, выявляя ключевые узлы и аномальные связи. Методы же поиска аномалий направлены на обнаружение редких объектов или событий, которые существенно отклоняются от основного распределения данных. Анализ временных рядов позволяет выявлять тренды, сезонность и циклические закономерности в динамических данных, обеспечивая основу для прогнозирования и оптимизации процессов. Снижение размерности упрощает визуализацию и обработку многомерных данных, сохраняя их ключевые свойства и улучшая интерпретацию результатов [2].

В РБД, где данные хранятся на разных узлах, применение вышеописанных традиционных методов машинного обучения усложняется из-за физического распределения данных. Централизованная обработка затруднена, так как передача больших объёмов информации между узлами требует значительных ресурсов и времени, особенно при ограниченной пропускной способности сети. Дополнительные сложности возникают из-за проблем конфиденциальности и безопасности, так как данные могут принадлежать разным организациям или подчиняться строгим регуляторным требованиям, запрещающим их передачу. Гетерогенность данных является еще одной проблемой при извлечении скрытых знаний из РБД. Данные могут быть представлены в различных форматах и схемах, что требует сложной интеграции и предварительной обработки. Также важной проблемой является обеспечение согласованности данных, так как асинхронное обновление на разных узлах может приводить к противоречиям и устареванию информации. Кроме того, масштабируемость становится проблемой в системах с тысячами узлов, таких как IoT или мобильные приложения, где централизованная обработка непрактична из-за больших вычислительных затрат и необходимости постоянной синхронизации данных, особенно в реальном времени [3].

III. Современные подходы к получению скрытых знаний из распределенных баз данных

В настоящее время драйвером работы с распределенными базами данных стало федеративное обучение (ФО), которое представляет собой ключевой подход, позволяющий обучать модели на распределенных данных без их централизации. Метод предполагает локальное вычисление градиентов с последующей агрегацией обновлений на центральном сервере, что минимизирует передачу сырых данных и обеспечивает соответствие регуляторным требованиям [4]. Стоит отметить, что подходы ФО ориентированы на повышение предсказательной способности моделей, а не на прямое извлечение скрытых знаний. Интерпретация зависимостей, выявленных в распределенных системах, остается проблемой из-за физической изоляции данных и фрагментарности локальных градиентов [5]. Тем не менее, многие методы машинного обучения, ранее применявшиеся для централизованных данных, были адаптированы для распределенной среды с учетом федеративного подхода. Также мультиагентные системы активно используются в распределенных вычислениях, где каждый агент отвечает за обработку локального фрагмента данных и взаимодействие с другими агентами через онтологические модели, обеспечивающие семантическую совместимость данных [6, 7].

Что касается гетерогенности хранимых данных, то адаптированные под РБД методы справляются с вертикальной гетерогенностью, когда узлы содержат разные атрибуты одних объектов, используя ключи соединения без физического объединения данных. В случае горизонтальной гетерогенности, где данные на узлах имеют одинаковую структуру, но относятся к разным объектам, применяется локальный анализ с агрегацией результатов через метаобучение [2]. Однако мультимодальность (разный формат) данных остается вызовом, так как для совместного исследования информации, относящейся к одному и тому же объекту в разном формате, требуется перевод данных в единое представление [8]. В рамках федеративного обучения эта задача решается через преобразование мультимодальных данных (текст, изображения, временные ряды) в векторные эмбеддинги – компактные числовые представления, сохраняющие семантические и структурные особенности исходных данных. Для согласования эмбеддингов из разных модальностей применяется контрастивное обучение, которое максимизирует сходство векторов, относящихся к одному объекту, и минимизирует его для разнородных данных [9, 10]. Важно отметить, что данный подход используется исключительно для обучения моделей и не включается в методы поиска знаний.

Существующие методы, такие как федеративное обучение и мультиагентные системы, эффективно решают задачи работы с гетерогенными и распределенными данными, сохраняя конфиденциальность и снижая коммуникационные затраты. Однако их применение к мультимодальным сценариям ограничено из-за семантического разрыва между форматами, высокой вычислительной сложности и отсутствия стандартизированных подходов для интеграции разнородных данных. Для преодоления этих ограничений требуются инновационные решения, включающие гибридные архитектуры, способные объединять преимущества ФО и агентных систем.

IV. Предлагаемая методология

В рамках исследования были выявлены основные ограничения, возникающие при извлечении скрытых знаний из РБД. Основной не решенной проблемой на данном этапе развития науки является анализ мультимодальных, гетерогенных данных. Как показали исследования не существует универсального метода

способного работать с мультимодальными данными в рамках получения скрытых знаний в РБД.

Предлагаемая методология представляет собой целостный подход к анализу распределенных гетерогенных, мультимодальных данных, который принципиально отличается от традиционных решений. В основе подхода лежит децентрализованная архитектура, где каждый узел системы представлен интеллектуальным агентом, специализирующимся на обработке конкретного типа данных: табличных, текстовых, визуальных или временных последовательностей. Агенты не просто применяют заранее заданные алгоритмы, а адаптивно выбирают оптимальные методы обработки, учитывая специфику поступающей информации.

Работа метода начинается с глубокого анализа данных внутри каждого агента, в рамках которого применяются специализированные методы машинного обучения, соответствующие природе каждого типа данных. Агенты табличных данных выявляют комплексные статистические закономерности и ассоциативные зависимости, благодаря использованию алгоритмов ассоциативных правил и кластеризации, устанавливая количественные взаимосвязи между параметрами. Для визуальных данных предлагается выполнять выделение значимых визуальных признаков различного уровня абстракции с помощью легковесных сверточных нейронных сетей. Текстовые агенты планируется задействовать в рамках семантического анализа текстовых данных с помощью оптимизированных трансформерных моделей, позволяющих распознавать не только именованные сущности, но и их контекстуальные отношения между объектами. Агенты временных рядов идентифицируют устойчивые паттерны и тренды, учитывая фазовые переходы и нестационарные изменения, благодаря работе автоэнкодеров.

Результатом вышеописанной обработки становится формирование локальных онтологий – формальных семантических структур, где концепты представляют собой не просто изолированные сущности, а многомерные объекты с атрибутивными и реляционными характеристиками. Отношения между концептами аннотируются метаданными, включающими показатели достоверности, временные рамки и контекстные ограничения, что обеспечивает сохранение всей глубины выявленных закономерностей.

Дальнейшим важным шагом в процессе реализации предлагаемой методологии является механизм интеграции, полученных локальных онтологий. Данный пункт предлагается реализовывать, как многоступенчатый процесс семантического согласования. На первом этапе которого происходит автоматическое сопоставление концептов в разных онтологиях через анализ их векторных представлений в унифицированном пространстве признаков, где учитываются как лексические, так и статистические аспекты семантической близости. Затем предлагается установить кросс-модальные связи, используя алгоритмы реляционного вывода, способные выявлять нетривиальные зависимости, остающиеся скрытыми при изолированном анализе отдельных модальностей. Полученные данные служат строительными блоками для формирования глобального графа знаний - динамической семантической сети, где узлы представляют интегрированные концепты из различных модальностей, а ребра выявленные межмодальные зависимости. Реализованный граф знаний будет обладать свойством эволюционного развития, автоматически актуализируя свою структуру по мере поступления новых данных. Динамический механизм внимания, включенный в описываемую методологию, позволит проводить адаптивную регуляцию значимости различных типов данных, автоматически корректируя весовые коэффициенты связей в зависимости от текущего контекста анализа и решаемой задачи. На рис. 1 представлена схема предлагаемой методологии.

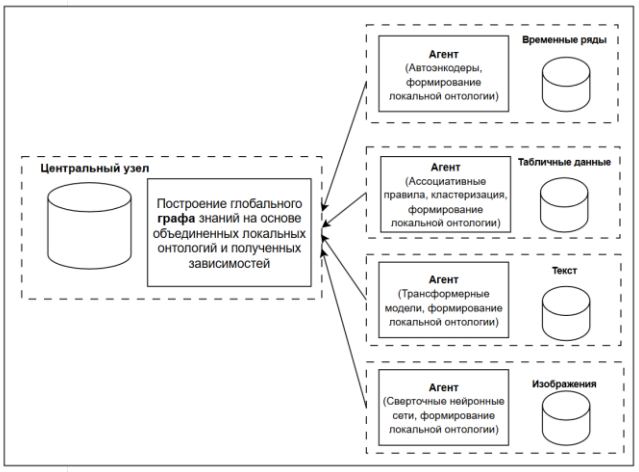

Рис. 1. Схема предлагаемой методологии

Федеративный принцип организации обмена информацией позволит сохранять данные распределенными, при этом граф знаний будет служить единой точкой доступа ко всем выявленным закономерностям, обеспечивая целостное представление без физической централизации данных. Это удовлетворяет не только требованиям конфиденциальности, но и эффективности обработки, так как каждый агент работает с данными локально, без необходимости их централизации. Предполагается, что в рамках методологии результирующий граф знаний способен поддерживать сложные семантические запросы, логический вывод новых знаний и генерацию понятных текстовых отчетов, что сделает систему удобной для конечных пользователей.

Главным преимуществом предлагаемой методологии получениях скрытых знаний из РБ перед традиционными подходами является способность выявлять комплексные кросс-модальные закономерности, остающиеся скрытыми при раздельном анализе данных. В отличие от специализированных существующих решений, работающих только с одним типом данных, предлагаемая методология обеспечивает целостное представление информации, сохраняя при этом распределенность исходных данных. А автоматизированный процесс построения онтологий и их интеграции значительно сократит временные затраты по сравнению с ручным моделированием предметных областей, а адаптивные механизмы обработки сделают систему гибкой и масштабируемой. Эти особенности откроют новые

возможности для междисциплинарных исследований и принятия обоснованных решений в различных предметных областях.

V. Заключение

В заключение проведенного исследования можно констатировать, что проблема извлечения скрытых знаний из распределенных баз данных остается одной из ключевых в условиях стремительного роста объемов получаемой гетерогенной информации и ужесточения регуляторных требований к безопасности данных. Анализ современных подходов показал, что, несмотря на значительный прогресс в области федеративного обучения и мультиагентных систем, сохраняются серьезные ограничения, связанные с обработкой мультимодальных данных, их семантической интеграцией и интерпретацией получаемых результатов. Традиционные методы машинного обучения, доказавшие свою эффективность в централизованных системах, оказываются малопригодными для распределенной среды из-за проблем с передачей данных, разнородностью форматов и необходимостью соблюдения конфиденциальности.

Предлагаемая методология, основанная на использовании интеллектуальных агентов для обработки различных типов данных и последующей интеграции результатов через семантические онтологии и графа знаний, позволяет преодолеть эти ограничения. Ее ключевое преимущество заключается в способности выявлять комплексные кросс-модальные закономерности без необходимости физической централизации данных, что обеспечивает как сохранение конфиденциальности, так и высокую интерпретируемость результатов. Адаптивный характер предложенного подхода, выражающийся в автоматическом подборе методов анализа в зависимости от типа данных и динамическом обновлении графа знаний, делает его особенно перспективным для применения в условиях быстро меняющихся распределенных систем.

Дальнейшее развитие данных исследований заключается в практической реализации предложенной методологии и опробовании ее на различных данных.

Список литературы

[1] Shu X., Yiwan Ye. Knowledge Discovery: Data Mining and Machine Learning Methods // Social Science Research. 2022. Vol. 110. P. 102817.

[2] Büchner A.G., Anand S.S., Bell D.A., Hughes J.G. Knowledge Discovery Methodology in Distributed Heterogeneous Databases // Proceedings of the Data Mining Colloquium. 1996. No. 198. Pp. 5-12.

[3] Waseem M., Abidin S. Issues and Challenges of the KDD Model for Distributed Data Mining Techniques and Architectures // Proceedings of the 10th International Conference on Computing for Sustainable Global Development. 2023. Pp. 1612-1617.

[4] Yang Q., Liu Y., Chen T., Tong Y. Federated Machine Learning: Concept and Applications // ACM Transactions on Intelligent Systems and Technology. 2019. Vol. 10, No. 12. Pp. 1-19.

[5] Li T., Sahu A.K., Talwalkar A., Smith V. Federated Learning: Challenges, Methods, and Future Directions // IEEE Signal Processing Magazine. 2020. Vol. 37, No. 3. Pp. 50-60.

[6] Pathak B., Sinha M. Analytical Study of Agent Based Distributed Data Mining and its Ontology // Proceedings of the International Conference on Computing for Sustainable Global Development. 2014. Pp. 400-404.

[7] Sánchez San Blas H., Sales Mendes A., García Encina F., Silva L.A., Villarubia González G. A Multi-Agent System for Data Fusion Techniques Applied to the Internet of Things Enabling Physical Rehabilitation Monitoring // Applied Sciences. 2021. Vol. 11, No. 1. P. 331.

[8] Kairouz P., McMahan H.B., Avent B., Bellet A., Bennis M., Bhagoji A.N. et al. Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning, vol. 14, no. 1-2. Boston, MA: Now Publishers, 2021. 210 pp.

[9] Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sutskever I. Learning Transferable Visual Models with Natural Language Supervision // Proceedings of the 38th International Conference on Machine Learning. 2021. Pp. 8748-8763.

[10] Besold T.R., Garcez A.A., Bader S., Bowman H., Domingos P., Hitzler P., Lamb L.C. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation // arXiv Preprint. 2017. No. 1711.03902.