Инженерный вестник Дона, №11 (2025)
ivdon.ru/ru/magazine/archive/n11y2025/10232

Экспериментальная оценка эффективности метода сопоставления схем данных на основе машинного обучения

К.Н. Кургузкин[1], А.Н. Марьенков[2]

[1]Астраханский государственный университет имени В. Н. Татищева, Россия, Астрахань
[2]Астраханский государственный технический университет, Россия, Астрахань

Аннотация: В статье изучается проблема сопоставления структурированных схем данных, агрегируются результаты, произведенные в рамках предыдущих этапов исследования. Систематизация результатов показала, что рассмотренные ранее подходы демонстрируют хорошие показатели, однако их результативность не всегда достаточно высока для применения в реальных условиях. Для дальнейших исследований был выбран один из наиболее эффективных методов. Произведен эксперимент по сопоставлению схем данных на основе пяти примеров. В результате эксперимента выявлены положительные и отрицательные аспекты рассматриваемого метода. Было дополнительно подтверждено, что выбранный метод демонстрирует недостаточную устойчивость и воспроизводимость результатов на разнородных реальных данных. Проведенная верификация метода подтвердила необходимость его оптимизации. В заключении статьи были определены направления для дальнейшего исследования в данной области.
Ключевые слова: управление данными, сопоставление схем, машинное обучение, классификация, кластеризация, машинное обучение, экспериментальный анализ, метрики данных

Введение

Процесс сопоставления информации является ключевым элементом в задачах интеграции данных. Сегодня количество разнообразных систем, схем и баз данных постоянно растёт, поэтому всё чаще возникает необходимость приводить входящие потоки данных к общей структуре.

Актуальность работы подтверждается статистическими данными, согласно которым рынок интеграции данных будет расти на 12–13% ежегодно, как минимум до 2032 года [1, 2]. При этом значительная часть проектов (около 40%) сталкивается с серьёзными трудностями при выполнении процедур интеграции.

В предыдущих исследованиях, выполненных в рамках общей работы по интеграции информационных систем [3], была предложена классификация методов сопоставления информации. Алгоритмические подходы были

разделены на два основных направления: основанные на графовых моделях и использующие методы машинного обучения.

Цель настоящей работы исследовать метод машинного обучения, который показывал наилучшие результаты согласно метрикам, указанным в исследованиях, проведенных в [4, 5]. Этот метод – метод самоорганизующейся карты (Self-Organizing Map – SOM). SOM – это процесс кластеризации без учителя, который использует конкурентное обучение.

Благодаря способности SOM группировать схожие атрибуты без необходимости наличия размеченных примеров, данный подход наиболее подходящим для анализа структурных и статистических признаков атрибутов из проанализированных в настоящем исследовании методов машинного обучения.

Метод SOM показал значение эффективности в 53 процента, то есть SOM указал чуть более половины пар атрибутов, которые действительно должны быть сопоставлены, что является наилучшим результатом в сравнении с другими методами машинного обучения. Таким образом, было принято решение выбрать данный метод в качестве основного для дальнейших экспериментов.

Описание исходных данных

Для анализа данных были выбраны следующие наборы: банк данных угроз безопасности информации [6]; показатели развития стран объединения, включающего Бразилию, Россию, Индию, Китай, Южно-Африканскую Республику и прочие страны (Brazil, Russia, India, China, South Africa – BRICS) [7]; экономические показатели согласно отчету Всемирного банка [8]; база данных рейсов российских авиакомпаний [9]; база данных статистики национальной баскетбольной ассоциации (The National Basketball Association – NBA) [10]; генеративная схема данных – систему управления взаимоотношениями с клиентами (Customer relationship management – CRM).

С подробным описанием схем, таблиц и атрибутов из первых пяти датасетов можно ознакомиться в ссылках-источниках [6 – 10]. В таблице №1 указаны примеры данных из указанных выше наборов.

Таблица № 1

Примеры данных

Схема	Атрибут	Значение
Банк данных угроз безопасности информации	thrlist.Идентификатор_УБИ	121
Банк данных угроз безопасности информации	thrlist.Наименование_УБИ	Угроза повреждения системного реестра
Показатели развития стран BRICS	EducationAndEnviron_Data.CountryCode	CHN
Показатели развития стран BRICS	EducationAndEnviron_Data.Year	1993.0
Экономические показатели согласно отчету Всемирного банка	Country.CountryCode	Panama
Экономические показатели согласно отчету Всемирного банка	Country.LongName	Republic of Panama
База данных статистики NBA	common_player_info.birthdate	1961-03-28 00:00:00
База данных статистики NBA	game.season_type	Regular Season

Генеративная схема данных составлена с помощью чат-ботов ChatGpt и Gemini. Атрибуты, сгенерированные в чат-ботах в общем наборе данных, указываются с префиксами «gpt» и «gemini». Атрибуты из остальных схем данных указываются с префиксами «small» и «big».

Схема «big» включает до 100 000 строк исходных данных для каждого атрибута, при этом, если общее количество записей по атрибуту меньше данного порога, используется весь доступный объем. Схема «small», в свою очередь, формируется выборкой из 100 случайных строк для каждого атрибута. Такой подход позволяет исследовать устойчивость и масштабируемость метода SOM при различной плотности данных, а также проанализировать влияние объема выборки на качество кластеризации признаков и точность последующего сопоставления атрибутов между схемами.

В схемах «gpt» и «gemini» содержатся таблицы: users, products, orders, employees, posts, accounts, events, students, vehicles, inventory. Набор таблиц описывает расширенную CRM-систему, которая объединяет не только управление клиентами и продажами, но и дополнительные бизнес-процессы: маркетинг, логистику, обучение и управление персоналом.

Итоговый набор содержит в себе 1188 атрибутов объемом приблизительно 200 Мб.

Описание критериев сопоставления

Все используемые критерии заимствованы из оригинального исследования «Schema Matching using Machine Learning» [11].

Первая группа критериев формируется на основе спецификаций схем данных. Данную группу критериев можно описать как набор параметров, структурно характеризующих схемы данных на основе описания структуры. Вторая группа критериев основывается на статистических показателях для числовых и строковых данных.

Для каждого атрибута формируется вектор, содержащий 20 критериев. В таблице №2 указан пример расчета критериев для атрибута «Наименование_УБИ» из схемы «Банк данных угроз безопасности информации». Часть критериев определена экспертным путем (Ключ, Уникальность, Не Null), так как изначально схема данных представлена в формате таблицы Excel и не содержит в себе полную информацию о структуре данных.

Таблица № 2

Пример полученного расчета критериев

Attribute	thrlist_big.Наименование_УБИ
Тип данных	2
Длина	100
Ключ	0
Уникальность	1
Не Null	1
Средняя используемая длина	67,92342342
Дисперсия длины	741,0166586
Коэффициент вариации длины	0,400769275
Среднее значение	0
Дисперсия	0
Коэффициент вариации	0
Минимум	0
Максимум	0
Количество пробелов	1
Количество специальных символов	0,162162162
Доля числовых символов	0
Доля буквенных символов	1
Количество обратных слэшей	0,04954955
Количество скобок	0,058558559
Количество дефисов	0,09009009

В таблице №3 указаны подобранные параметры модели, показавшие наилучшие результаты сопоставления атрибутов:

Таблица № 3

Экспериментальный набор параметров для модели SOM

Параметр	Значение
Размер карты	13
Начальная скорость обучения	0.5
Начальный радиус соседства	6.5
Количество итераций обучения	50

Анализ полученных результатов

В результате произведенного моделью расчета критериев на основе указанных выше параметров был получен 61 кластер атрибутов. Атрибуты, попавшие в один кластер SOM, должны обладать высокой степенью корреляции по множеству входных признаков. Атрибуты, находящиеся в близлежащих кластерах, в соответствии с порядковыми номерами, также могут указывать на возможную связь между рассматриваемыми атрибутами. Однако, в контексте задачи сопоставления, дополнительный анализ перекрестных пересечений между результатами различных кластеров только усложняет анализ, а не облегчает его.

В таблице №4 перечислены примеры полученных данных в различных кластерах. В каждом из кластеров вручную выделена корневая группа атрибутов, которые на самом деле могли бы быть сопоставимы в реальной ситуации при построении хранилищ данных.

Таблица № 4

Примеры кластеров с наборами атрибутов

Кластер	Количество групп	Корневая группа	Примеры атрибутов	Примеры данных	Результативность
0	16	Баскетбольные команды	CountryNotes_big.Country_code; draft_combine_stats_big.first_name; draft_history_big.team_city; employees_gpt.email; products_gpt.category; team_big.city	BEL; Thomas; Los Angeles; ivan.ivanov@example.com; Электроника; Houston	0,21
5	3	Общие	common_player_info_big.di	King; Frankie;	0,5

Кластер	Количество групп	Корневая группа	Примеры атрибутов	Примеры данных	Результативность
		описани	splay_last_comma_first; CountryNotes_big.Seriescode; events_gpt.event_description	SM.POP.REFG; A conference for tech enthusiasts.	
18	2	Экономические описания	Country_big.BalanceOfPaymentsManualInUse; SeriesNotes_small.Seriescode	IMF Balance of Payments Manual; 6th edition; SP.DYN.AMRT.FE	0,75
64	2	Веб-ссылки	Country_big.LatestAgriculturalCensus; url_big.gpt; game_summary_big.gamecode	2007; https://mail.ru; 19800214/NYKSAN	0,38
65	4	Даты	accounts_gpt.account_number; Footnotes_big.Year; users_gpt.email; vehicles_gpt.vin	AC12345678; YR2006; test.ivanov@example.com; 1HGBH41JXMN109186	0,4
99	2	Символьно-буквенные коды	flights_big.flight_no; tickets_big.passenger_id; boarding_passes_big.seat_no	PG0424; 8149 604011; 3A	0,67
120	6	Статистические показатели	common_player_info_big.team_id; common_player_info_small.person_id; Indicators_small.Year	1610612755; 2; 1994	0,26
168	5	Идентификаторы	Economy_Data_big.SeriesCode; flights_big.flight_id; tickets_small.ticket_no; students_gpt.id	NY.ADJ.DCO2.GN.ZS; 0005434264979; 1	0,56

На основе исходных данных задачи, как минимум 594 атрибута из схемы «big» и схемы «small» должны быть сопоставимы, так как составлены из одного и того же набора данных, но в разных масштабах. Точность сопоставления методом SOM для такого типа пар атрибутов из экспериментального датасета составила 88 процентов, что является очень хорошим результатом. Однако, необходимо также проанализировать какие атрибуты модель посчитала схожими и оценить результаты на наличие ложноположительных результатов, которые снижают качество расчетов.

Составленная карта кластеров получилась недостаточно точной. Это означает, что полученное разбиение на кластеры слабо соответствует предметной области и не позволяет сделать содержательные выводы о сопоставимости атрибутов.

В качестве примера можно рассмотреть email-адреса: они не только попали в разные кластеры, но и находятся на достаточно удаленном расстоянии между кластерами (кластер 0 и кластер 65).

Также можно рассмотреть группы атрибутов с экономическими показателями (CountryNotes.Seriescode, Indicators.Year, Country.LatestAgriculturalCensus, Country.NationalAccountsReferenceYear). Атрибуты с данными показателями разбросаны по разным кластерам (кластер 5 – кластер 18, кластер 64 – кластер 120).

Таким образом, можно отметить большое количество разделений данных по различным кластерам несмотря на то, что разнесенные значения логически входят в одну и ту же группу.

Средняя результативность по кластерам составила 61 процент. Результативность каждого кластера рассчитывалась по следующей формуле (1):

$$Result = \frac{TP}{TP + FP}, \quad (1)$$

где *TP* – действительно положительные сопоставления; *FP* – ложноположительные сопоставления.

В результате анализа полученных групп атрибутов, можно сделать вывод о большом количестве ложноположительных результатов, когда в один и тот же кластер попадают данные из категорий, которые логично было бы отнести к различным группам.

Заключение

Результаты проведенного контрольного исследования оказались неоднозначными. При высокой результативности сопоставления искусственно размеченных пар одинаковых атрибутов (с разным масштабом данных), обнаружилась средняя результативность при анализе остальных пар, сопоставленных методом SOM, а также довольно избыточное «дробление» сопоставляемых наборов данных по кластерам. Это свидетельствует о недостаточно эффективной работе рассматриваемого в исследовании метода.

Таким образом, при проведении исследований в данной области необходимо учитывать возможные недостатки механизмов машинного обучения.

Сложность интерпретации полученных данных – необходимо при помощи алгоритмического набора условий снизить степень неопределенности получаемых результатов.

Обучение моделей на большом наборе данных – возникает естественная необходимость снижать объем входных данных, особенно учитывая, что большое количество информации недоступно в открытом доступе, так как представляет коммерческую тайну или иные виды тайн. Уменьшение количества входных данных при этом не должно снижать качество получаемых моделью результатов.

Модели машинного обучения могут выдавать ложноположительные результаты – при большом количестве таких результатов трудно производить эффективное сопоставления данных, необходимо отсеивать действительно положительные результаты, что затрудняет процесс аналитики и увеличивает время обработки данных, а не уменьшает его, что является одной из основных задач процесса автоматизированного сопоставления схем данных.

Методы машинного обучения могут позволить автоматически анализировать контекст данных – при грамотном применении некоторых современных методов классификации и кластеризации теоретически возможно получать хорошие результаты сопоставления структурированных схем данных, что частично было доказано в эксперименте, проведенном в настоящем исследовании. При этом существующие методы сопоставления однозначно требуют доработки.

Литература

1. Big Data Analytics / Fortune Business Insights. – URL: fortunebusinessinsights.com/big-data-analytics-market-106179 (дата обращения: 25.12.2023).

2. Data integration in 2023 // Actioner. – URL: actioner.com/guides/data-integration-statistics (дата обращения: 20.12.2023).

3. Кургузкин К.Н., Марьенков А.Н. Сравнительный анализ методик сопоставления информации в контексте интеграции схем данных // Прикаспийский журнал: управление и высокие технологии. – 2024. – № 2 (66). – С. 16–31.

4. Sun Z., Huang J., Xu X., Chen Q., Ren W., Hu W. What Makes Entities Similar? A Similarity Flooding Perspective for Multi-sourced Knowledge Graph Embeddings // Proceedings of the 40th International Conference on Machine Learning (ICML’23). – 2023. – pp. 1–11.

5. Melnik S., Garcia-Molina H., Rahm E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching // Proceedings of the 18th International Conference on Data Engineering. – 2002. – pp. 1–12.

6. Банк данных угроз безопасности информации / Федеральная служба по техническому и экспортному контролю. – URL: bdu.fstec.ru/threat (дата обращения: 01.07.2024).

7. BRICS World Bank Indicators Dataset. – URL: kaggle.com/datasets/docstein/brics-world-bank-indicators (дата обращения: 10.08.2024).

8. World Development Indicators. – URL: kaggle.com/datasets/kaggle/world-development-indicators (дата обращения: 10.08.2024).

9. Airlines Dataset. – URL: kaggle.com/datasets/saadharoon27/airlines-dataset/data (дата обращения: 10.08.2024).

10. NBA Database. – URL: kaggle.com/datasets/wyattowalsh/basketball (дата обращения: 10.08.2024).

11. Sahay T., Mehta A., Jadon S. Schema Matching using Machine Learning. College of Information and Computer Sciences, University of Massachusetts, Amherst. – 2019. – pp. 1–7.

References

1. Kurguzkin K.N., Marienkov A.N. Prikaspiyskiy Zhurnal: Upravlenie i Vysokie Tekhnologii, 2024, No. 2 (66), pp. 16–31.

2. Fortune Business Insights. Big Data Analytics [online]. Available at: fortunebusinessinsights.com/big-data-analytics-market-106179 (Accessed 25 December 2023).

3. Actioner. Data integration in 2023 [online]. Available at: actioner.com/guides/data-integration-statistics (Accessed 20 December 2023).

4. Sun Z., Huang J., Xu X., Chen Q., Ren W., Hu W. Proceedings of the 40th International Conference on Machine Learning (ICML'23), 2023, pp. 1–11.

5. Melnik S., Garcia-Molina H., Rahm E. Proceedings of the 18th International Conference on Data Engineering, 2002, pp. 1–12.

6. Federal Service for Technical and Export Control. Information Security Threat Data Bank [online]. Available at: bdu.fstec.ru/threat (Accessed 1 July 2024).

7. Kaggle. BRICS World Bank Indicators Dataset [online]. Available at: kaggle.com/datasets/docstein/brics-world-bank-indicators (Accessed 10 August 2024).

8. Kaggle. World Development Indicators [online]. Available at: kaggle.com/datasets/kaggle/world-development-indicators (Accessed 10 August 2024).

9. Kaggle. Airlines Dataset [online]. Available at: kaggle.com/datasets/saadharoon27/airlines-dataset/data (Accessed 10 August 2024).

10. Kaggle. NBA Database [online]. Available at: kaggle.com/datasets/wyattowalsh/basketball (Accessed 10 August 2024).

11. Sahay T., Mehta A., Jadon S. College of Information and Computer Sciences, University of Massachusetts, Amherst, 2019, pp. 1–7.

Дата поступления: 17.09.2025

Дата публикации: 28.10.2025