Проблемы представления данных в задаче управления розничными продажами

Н. В. Размочаева[1], Д. М. Клионский[2]
Санкт-Петербургский государственный электротехнический
университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]nvrazmochaeva@etu.ru, [2]dmklionsky@etu.ru

***Аннотация*. Особое место в автоматизации процесса управления розничными продажами занимают вопросы представления данных большой размерности. Проблема заключается в том, что для описания данных имеет место большое число характеристик, определяющее размерность пространства признаков. Отмечается, что существует множество методов сокращения размерности пространства признаков, например, корреляционный анализ, метод главных компонент или метод опорных векторов. В настоящей работе были исследованы некоторые методы. Так, например, корреляционный анализ позволил исключить параметры с сильными линейными зависимостями. Результаты применения методов снижения размерности подтверждены специалистами экспертной группы. Проведенный кластерный анализ показал состоятельность полученных результатов.**

Ключевые слова: представление данных; анализ данных; интерпретация данных; корреляционный анализ; автоматизация процессов; розничные продажи; управление; машинное обучение

I. Введение

Информационные потоки, которые ежедневно окружают человека, обеспечивают его профессиональную деятельность и быт, с каждым днем увеличиваются как снежный ком. Причиной этому послужило скоротечное развитие информационных технологий, высокий современный технический и научный уровень. Однако даже в сегодняшних условиях насыщенного рынка инструментов для информационного обеспечения имеет место проблема, связанная с большим количеством информации. Проблема заключается в трудоемкости проведения быстрого и качественного анализа информации, что существенно замедляет процесс принятия решения.

Для различных областей деятельности эта проблема имеет различные последствия. Особо стоит отметить важность для области медицины и экономики [1], [2], когда принятие решения необходимо сделать в кратчайшие сроки. С другой стороны, например, такие области как строительство и геология имеют гораздо меньше ситуаций, где приятие решение требуется осуществить как можно быстрее.

Решения проблемы анализа больших данных различны. В некоторых задачах можно использовать усовершенствованные технические средства, например, суперкомпьютеры, облачные технологии, распределенные вычислительные системы. Но такие решения могут оказаться дорогостоящими. Другой подход – качественная предобработка данных.

Предварительная обработка данных – широкая научная область, включающая как методы статистической обработки [3], так и методы машинного обучения [4]. С одной стороны, препроцессинг может выполняться для извлечения полезной информации, знаний, неявных зависимостей. С другой стороны, он может послужить инструментом для выявления избыточной информации. Зачастую предварительную обработку выполняют для сокращения размерности данных, то есть для выявления достаточного количества признаков (характеристик), с помощью которых можно описать объекты предметной области (объекты исследования).

В данной работе будет проведен обзор различных методов, относящихся к предварительной обработке данных и выполняющих сокращение размерности пространства признаков.

II. Обзор предметной области

Объект исследования настоящей работы – данные о розничных продажах, выполненных через автоматизированные торговые системы – торговые автоматы. Такая торговля называется вендинговой [5]. Предмет исследования – сокращение размерности данных. Различные подходы к формализации задачи исследования данных о продажах, описание структуры данных и история обсуждаемой проблемы изложены в [6].

В работе [7] проведен корреляционный анализ, который показал, что без потери качества анализа пространство признаков возможно сократить в 2 раза. В [8] исследованы методы t-SNE и PCA (метод главных компонент использует линейное преобразование), результаты которых визуализированы с помощью кластеризации K-means, Dendrogram и Affinity Propagation. Кластеризация обработанных данных позволила разделить товары на группы согласно маркетинговым свойствам: напитки, снеки, еда и т.д. Точность результатов

удовлетворительная, об этом говорят значения метрик качества. Как в случае с корреляционным анализом, так и в случаях использования алгоритмов t-SNE и PCA, все получаемые результаты передавались экспертной группе для оценки их состоятельности. Экспертная группа состоит из квалифицированных маркетологов с большим опытом работы в сфере вендинга.

Экспертная группа в обоих случаях подтвердила адекватность результатов. На основе проведенных исследований, перечисленных выше, было разработано программное приложение [9], в ходе эксплуатации которого выяснилось, что еще не все методы по предобработке данных исследованы и, может быть, не вся полезная информация извлечена. В связи с этим работа в этом направлении продолжается и в настоящей статье.

Анализ способов решения проблемы представления данных и итоговый выбор методов снижения пространства признаков являются сложными задачами. Ситуация затруднена тем, что не существует единой классификации и подробного перечня всех существующих методов.

III. Обзор подходов к решению

Проблему представления данных в рамках задачи сокращения размерности можно рассматривать с двух точек зрения: экстракция признаков (выявление из данных скрытых, но полезных признаков, методы PCA, ICA, *Auto-Encoders with bottleneck* (метод для поиска аномалий в данных)) и отбор признаков (выделение наиболее полезных признаков из числа уже имеющихся, методы корреляционного анализа, алгоритмы основанные на переборе, embedded methods: *Sparse Regression*, *Decision Trees with pruning*, *Regularized Random Forest* (*RTT*), *Regularized gradient boosting*, *Regularized Neural Nets*) [10].

В группе методов машинного обучения без учителя выделяют ряд способов снижения размерности пространства, среди которых уже известный нам PCA, а также методы случайных проекций и метода агломерации признаков (включающих иерархические способы кластеризации, такие как дендрограммы).

Однако, в большинстве случаев исследователи используют PCA. Особенную популярность данный метод набрал в задачах анализа изображений. Почти такой же уровень применимости имеет и метод SVM (метод опорных векторов). Оба эти метода нашли широкое применение при проектировании нейронных сетей, когда ставится задача подбора наилучших параметров для обучения сети (задачи кросс-валидации).

Особенно стоит отметить тот факт, что методы машинного обучения без учителя наиболее подходят для исследуемых в настоящей работе данных. Причина заключается в том, что исследуемые данные не позволяют сформировать выборки для обучения и тестирования, как это принято в машинном обучении с учителем. Так же важно отметить, что методы обучения без учителя применяются и для задач снижения размерности пространства. Некоторые источники все статистические методы обработки данных, которые не требуют обучающих выборок, относят к методам обучения без учителя, используя эти два понятия как синонимы.

Другие методы снижения размерности статистических данных и методы прикладной статистики для анализа данных рассмотрены в [11] и [12].

Выше перечисленные методы, конечно же, не единственные. В следующем разделе будет на практике рассмотрено применение трех простых способов снижения размерности, основанных на неструктурированных случайных матрицах. Понятие неструктурированных случайных матриц рассмотрено в [13].

IV. Снижение размерности пространства признаков

В качестве инструмента для проведения анализа на практике был выбран язык программирования python 3.7 [14] и библиотека scikit-learn [15] (*sklearn*) в силу того, что там все необходимые алгоритмы и методы уже реализованы и готовы к применению.

Далее будут применены методы модуля «Random projections» библиотеки *scikit-learn*. Данный модуль реализует простые и эффективные в вычислительном плане способы уменьшения размерности ценой точности, величину которой можно изменять. Эти простые действия позволяют ускорить время обработки и уменьшить размеры представления данных, что решает поставленную задачу.

В составе рассматриваемого модуля реализованы гауссова случайная матрица [16], разреженная случайная матрица [17] и метод, основанный на преобразовании Джонсона-Линденштраусса (*Johnson-Lindenstrauss Transform, JLT* [18]).

Принцип, по которому работают вышеуказанные методы заключается в следующем: размеры и распределение матриц случайных проекций преобразуются так, чтобы сохранить попарные расстояния между любыми двумя выборками набора данных [19]. Таким образом, случайная проекция является подходящей техникой снижения размерности пространства для метода, основанного на расстоянии [20].

A. *Метод Джонсона-Линденштраусса*

Метод основан на лемме Джонсона-Линденштраусса. Метод заключается в отображении точек с низким искажением из многомерного пространства в евклидово пространство с меньшей размерностью. Лемма утверждает, что небольшой набор точек в многомерном пространстве может быть умещен в пространство гораздо меньшего размера таким образом, что расстояния между точками сохранятся почти точно.

B. *Случайная гауссовская проекция (Gaussian Random Projection)*

Данный метод позволяет уменьшить размерность пространства признаков, проецируя исходное пространство на случайно сгенерированную матрицу, где компоненты извлекаются из следующего распределения:

$N\left(0, 1/n\right)$, где n – это размер результирующего пространства.

C. *Разреженная случайная проекция (Sparse Random Projection)*

Данный метод позволяет уменьшить размерность, проецируя исходное пространство с использованием разреженной случайной матрицы.

Разреженные случайные матрицы являются альтернативой гауссовой матрице случайных проекций и гарантируют аналогичное качество отображения точек в пространство с меньшей размерностью, но при этом намного более эффективно используют память и позволяют быстрее вычислять проецируемые данные.

D. *Визуализация результатов путем кластеризации*

Метод кластеризации – *DBSCAN* (Density-Based Spatial Clustering of Applications with Noise) – это популярный алгоритм кластеризации [21], используемый в качестве альтернативы методу K-means.

Будем использовать реализацию из библиотеки *scikit-learn*. Уточним следующие устанавливаемые параметры метода: *eps* – это максимальное расстояние между двумя точками в наборе (по умолчанию 0,5), *min_samples* – это минимальное количество точек данных в окрестности, которое можно считать кластером (по умолчанию 5).

Предварительно подготовим данные для анализа – выборку из 205 товаров, которые описаны 12 параметрами. Выполним стандартизацию выборки известным способом с помощью модуля *preprocessing* библиотеки *sklearn* [22].

Имея в распоряжении 12 характеристик товаров итеративно по размерности целевого пространства (от 3 до 12) применим методы случайной гауссовской проекции (*GRP*) и метод разреженной случайной проекции (*SRP*).

Графическая интерпретация результатов представлена на рис. 1.

В качестве метрики качества будем использовать так называемый «коэффициент силуэта» – *Silhouette Coefficient* [23]. Коэффициент рассчитывается с использованием среднего внутрикластерного расстояния (a) и среднего расстояния до ближайшего кластера (b) следующим образом: $\frac{b-a}{\max(a,b)}$. Данный критерий подходит для рассматриваемой задачи, так как не требует наличия вектора образцовых значений.

Как видно из рис. 1 даже после снижения размерности пространства данные трудно интерпретируемы. Это обусловлено таким фактором как слишком высокая густота расположения образцов. Эти результаты также подтверждаются полученной оценкой качества – коэффициент силуэта изменялся по модулю от 0.26 (при количестве параметров в целевом пространстве 7 и методе снижения *SRP*) до 0.029 (при количестве параметров в целевом пространстве 11 и методе снижения *SRP*).

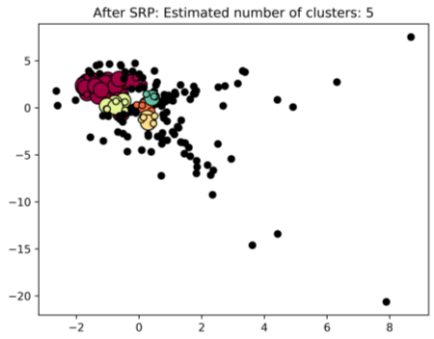

Рис. 1. Результаты кластеризации *DBSCAN* при заданной размерности целевого пространства признаков 3 и методе снижения пространства *SRP*

Несмотря на небольшой разброс значений метрики качества (таблица), результаты кластеризации нельзя однозначно интерпретировать.

ТАБЛИЦА I ОЦЕНКА МЕТРИКИ КАЧЕСТВА

Method	**Statistical characteristics of quality metrics**			
	Min	***Max***	***Mean***	***Var***
SPR	-0.173	0.442	-0.033	0.035
GPR	-0.273	0.191	-0.076	0.015

Полученные после снижения размерности пространства признаков результаты не дают однозначного ответа на вопрос о целесообразности применения данных методов в реальных условиях. Таким образом, можно сформулировать задачу для дальнейшего исследования – компоновка результатов, полученных в настоящей работе, и результатов других исследований.

V. ЗАКЛЮЧЕНИЕ

Проблема представления данных, для которых не применимы методы машинного обучения с учителем, остается актуальной по результатам настоящей работы. Отсутствие обучающей выборки и вектора образцовых значений сильно ограничивает свободу исследования. К сожалению, генерация образцовых значений для исследуемой задачи – вопрос отдельных исследований, т.к. исследуемую зависимость нельзя аналитически вывести.

В ходе проведения обзора была выявлена еще одна проблема, свойственная высокому уровню развития информационных технологий, – отсутствие однозначной классификации всех методов с учетом особенностей их применения. Используемый инструмент – язык python – частично решает данную проблему.

Полученные в настоящей работе результаты (результаты кластеризации) послужат основой для критического пересмотра постановки задачи снижения размерности и проблемы представления данных.

В продолжение исследования предварительной обработки данных и снижения пространства признаков планируется рассмотрение таких методов как неотрицательное матричное разложение, ядерный метод главных компонент (в том числе и на графах), линейный и

обобщенный дискриминантный анализ, канонический корреляционный анализ и др.

Список литературы

[1] Razmochaeva N.V., Semenov V.P., Bezrukov A.A. Role of Process Automation in Quality Management of Enterprises in Perfumery and Cosmetic Industry. Материалы Конференции молодых исследователей России по электротехнике и электронике IEEE (2019 ElConRus) (2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering). 28-31 Января 2019. СС: 1449-1452. [Электронный ресурс] URL: https://ieeexplore.ieee.org/document/8657085/ (дата обращения: 27.03.2019)

[2] Mikhailov Y.I., Razmochaeva N.V. The Problems of Quality Management Automation in Retail Sales. Материалы Международной научно–практическая конференция «Менеджмент качества, транспортная и информационная безопасность, информационные технологии» IT&MQ&IS – 2018 (The International Conference "Quality Management, Transport and Information Security, Information Technologies"). 24-30 Сентября 2018. СС. 372-375. [Электронный ресурс] URL: https://ieeexplore.ieee.org/document/8525056 (дата обращения: 27.03.2019)

[3] Семенов В.П., Чернокульский В.В., Размочаева Н.В. Исследование искусственного интеллекта в задачах управления розничной торговлей // II Международная научная конференция по проблемам управления в технических системах (CTS'2017). Материалы конференции. Санкт-Петербург. 25–27 октября 2017 г. СПб.: СПбГЭТУ «ЛЭТИ». С. 346 - 349.

[4] Klionskiy D.M., Chernokulsky V.V., Razmochaeva N.V. The Investigation of Machine Learning Methods in the Problem of Automation of the Sales Management Business-process. Материалы Международной научно–практическая конференция «Менеджмент качества, транспортная и информационная безопасность, информационные технологии» IT&MQ&IS – 2018 (The International Conference "Quality Management, Transport and Information Security, Information Technologies"). 24-30 Сентября 2018. СС. 376-381. [Электронный ресурс] URL: https://ieeexplore.ieee.org/document/8525008 (дата обращения: 27.03.2019)

[5] Размочаева Н.В., Клионский Д.М., Чернокульский В.В. Автоматизация бизнес-процессов в торговле с использованием методов интеллектуального анализа данных. Журнал «Качество. Инновации. Образование». 2018. №5. С. 77-85.

[6] Чернокульский В.В., Размочаева Н.В. Разработка подхода к решению задачи формирования ассортимента товаров точки розничной торговли // Известия СПбГЭТУ ЛЭТИ. 2018. №. 2. С. 5-10.

[7] Razmochaeva N.V., Klionskiy D.M. Data Presentation and Application of Machine Learning Methods for Automating Retail Sales Management Processes. Материалы Конференции молодых исследователей России по электротехнике и электронике IEEE (2019 ElConRus) (2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering). 28-31 Января 2019. СС: 1444 - 1448. [Электронный ресурс] URL: https://ieeexplore.ieee.org/document/8657077/ (дата обращения: 27.03.2019)

[8] Semenov V.P., Chernokulsky V.V, Razmochaeva N.V. Research of artificial intelligence in the retail management problems. Материалы II Международная научная конференция по проблемам управления в технических системах (ПУТС-2017) (2017 IEEE II International Conference on Control in Technical Systems (CTS)). 25 - 27 октября 2017. СС: 333 -336. [Электронный ресурс] URL: http://ieeexplore.ieee.org/document/8109560/ (Дата обращения: 27.03.2019)

[9] Семенов В.П., Чернокульский В.В., Размочаева Н.В. Программное приложение для оптимизации розничных продаж // XXI Международная конференция по мягким вычислениям и измерениям (SCM'2018). Материалы конференции. Санкт-Петербург. 23–25 мая 2018 г. СПб.: СПбГЭТУ «ЛЭТИ». С. 468-471

[10] Гулинин В. Лекция №8 "Методы снижения размерности пространства". 12 ноября 2014. [Электронный ресурс] URL: https://www.slideshare.net/Technosphere1/lecture-8-47107559 (дата обращения: 31.03.19)

[11] Орлов А.И., Луценко Е.В. Методы снижения размерности пространства статистических данных // Научный журнал КубГАУ. 2016. №119. [Электронный ресурс] URL: https://cyberleninka.ru/article/n/metody-snizheniya-razmernosti-prostranstva-statisticheskih-dannyh (дата обращения: 31.03.2019)

[12] Елисеева И. И., Юзбашев М. М. Общая теория статистики: Учебник / Под ред. И. И. Елисеевой. 4-е издание, переработанное и дополненное. – Москва: Финансы и Статистика, 2002. 480 с.

[13] Choromanski K., Rowland M., Weller A.. The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. [Электронный ресурс] URL: https://arxiv.org/pdf/1703.00864.pdf (дата обращения: 27.03.2019)

[14] scikit-learn: Machine Learning in Python. [Электронный ресурс] URL: http://scikit-learn.org/stable/ (дата обращения: 27.03.2019)

[15] scikit-learn: Random Projection. [Электронный ресурс] URL: https://scikit-learn.org/stable/modules/random_projection.html#random-projection (дата обращения: 27.03.2019)

[16] Ipsen J.R.. Products of Independent Gaussian Random Matrices. Doctoral dissertation. Department of Physics Bielefeld University. С. 145 [Электронный ресурс] URL: https://www.researchgate.net/publication/283118015_Products_of_Independent_Gaussian_Random_Matrices (дата обращения: 27.03.2019).

[17] Semerjian G., Cugliandolo L.F. Sparse random matrices: the eigenvalue spectrum revisited. Journal of Physics A General Physics 35(23). February 2002. [Электронный ресурс] URL: https://www.researchgate.net/publication/1833185_Sparse_random_matrices_The_eigenvalue_spectrum_revisited (дата обращения: 27.03.2019)

[18] Chen L.. Johnson-lindenstrauss transformation and random projection. . [Электронный ресурс] URL: https://www.math.uci.edu/~chenlong/MathPKU/JL.pdf (дата обращения: 27.03.2019)

[19] Dasgupta S.. Experiments with random projection. In Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence (UAI‘00), Craig Boutilier and Moisés Goldszmidt (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000. С. 143-151.

[20] Bingham E., Mannila H.. Random projection in dimensionality reduction: applications to image and text data. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ‘01). ACM, New York, NY, USA, 2001. С. 245-250.

[21] sklearn.cluster.DBSCAN. [Электронный ресурс] URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html(дата обращения: 27.03.2019)

[22] sklearn. Preprocessing data. [Электронный ресурс] URL: https://scikit-learn.org/stable/modules/preprocessing.html (дата обращения: 27.03.2019)

[23] sklearn.metrics.silhouette_score. [Электронный ресурс] URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html (дата обращения: 27.03.2019)