.

УДК 004.021:004.75

АДАПТИВНАЯ ФИЛЬТРАЦИЯ МНОГОМЕРНЫХ ДАННЫХ ДЛЯ ОПТИМИЗАЦИИ БИЗНЕС-РЕШЕНИЙ НА ОСНОВЕ НЕЛИНЕЙНЫХ МЕТРИК ВЗАИМОЗАВИСИМОСТИ ПРИЗНАКОВ

В.Г. Евдокимов
Аспирант кафедры информационных технологий автоматизированных систем БГУИР
vigandvdk@gmail.com

А.А. Навроцкий
Заведующий кафедрой информационных технологий автоматизированных систем БГУИР, кандидат физико-математических наук, доцент
navrotsky@bsuir.by

А.А. Навроцкий
Окончил Минский радиотехнический институт. Область научных интересов включает программные системы, системы компьютерного зрения и СВЧ-устройства.

В.Г. Евдокимов
Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов включает оптимизацию алгоритмов искусственного интеллекта, применение ИИ в условиях неполноты данных и разработку адаптивных автономных интеллектуальных систем.

Аннотация. В статье предлагается концептуально новый подход к адаптивной фильтрации многомерных данных, основанный на переходе от представления данных в форме традиционного гиперкуба к сферической геометрии гиперсферы и обратно. Предложенный метод потенциально может позволить эффективно идентифицировать и исключать взаимозависимые параметры путем анализа их геометрических свойств в сферическом пространстве. Рассматриваются теоретические основы идеи, математический аппарат для оценки нелинейных зависимостей между признаками и потенциальные преимущества данного подхода в различных задачах бизнес-аналитики. Обсуждаются перспективы практической реализации и адаптации метода для конкретных предметных областей.

Ключевые слова: многомерные данные, гиперсфера, сферическая геометрия, взаимозависимость признаков, фильтрация данных, бизнес-аналитика, OLAP.

Введение. Современные системы бизнес-аналитики сталкиваются с постоянно возрастающими объемами и размерностью обрабатываемых данных. По оценкам IDC, глобальный объем данных к 2026 году превысит 180 зеттабайт, причем значительная часть этого объема будет приходиться на бизнес-данные [1]. Рост размерности данных приводит к увеличению вычислительной сложности аналитических моделей и затрудняет извлечение полезной информации для принятия решений.

Традиционно данные в бизнес-аналитике представляются в виде многомерных структур – гиперкубов, где каждое измерение соответствует определенному параметру или признаку. Такое представление является основой для OLAP-систем (Online Analytical Processing) и широко используется в системах поддержки принятия решений. Однако

стандартные методы работы с гиперкубами имеют существенные ограничения при анализе сложных нелинейных взаимосвязей между признаками.

Эти ограничения связаны с тем, что евклидова геометрия гиперкуба не позволяет в полной мере отразить нелинейный характер взаимодействия признаков в реальных бизнес-процессах. Как отмечают Ли и Верлейсен, «линейные структуры представления данных неадекватны для отражения сложных нелинейных взаимозависимостей, характерных для большинства реальных бизнес-систем» [2].

В данной статье предлагается принципиально новая идея – использовать переход от евклидовой геометрии гиперкуба к сферической геометрии гиперсферы для более эффективной фильтрации параметров и последующего возврата к форме гиперкуба, но уже с исключенными взаимозависимыми параметрами. Данный подход потенциально позволяет решить проблему избыточной размерности данных и повысить эффективность бизнес-аналитики.

Обзор существующих подходов к анализу многомерных данных

Традиционные методы работы с многомерными данными. В настоящее время наиболее распространенными методами работы с многомерными данными в бизнес-аналитике являются технологии OLAP, основанные на представлении данных в виде гиперкубов. Такое представление позволяет осуществлять многомерный анализ данных, включая различные операции агрегации, детализации, секционирования и поворота.

Однако OLAP-технологии имеют существенные ограничения при анализе сложных взаимозависимостей между параметрами. Как отмечает Ван дер Маатен, «традиционные OLAP-кубы не способны эффективно выявлять нелинейные паттерны во взаимосвязях параметров, что критически важно для многих бизнес-приложений» [3].

Методы снижения размерности. Для решения проблемы высокой размерности данных широко используются различные методы снижения размерности, такие как анализ главных компонент (PCA), многомерное шкалирование (MDS), t-SNE и другие. Эти методы позволяют преобразовать исходные данные в пространство меньшей размерности с сохранением ключевых характеристик данных.

Однако большинство существующих методов снижения размерности имеют существенные ограничения:

– линейность преобразований – нелинейные методы снижения размерности часто создают новые признаки, которые сложно интерпретировать в терминах исходных параметров;

– потеря интерпретируемости – MQTT-сервер, который принимает информацию от издателей и передает ее соответствующим подписчикам, в сложных системах может выполнять также различные операции, связанные с анализом и обработкой поступивших данных. Разные брокеры могут соединяться между собой, если они подписываются на сообщения друг друга;

– отсутствие адаптивности – большинство методов не обеспечивают адаптивной фильтрации признаков в зависимости от их взаимозависимости.

Как отмечает Амари в своей работе по информационной геометрии, «существующие методы снижения размерности не учитывают геометрические свойства пространства признаков, что приводит к неоптимальным результатам при анализе сложных данных» [5].

Постановка проблемы.

Анализ существующих подходов к обработке многомерных бизнес-данных позволяет сформулировать следующие ключевые проблемы:

Проблема идентификации взаимозависимых параметров. В реальных бизнес-данных многие параметры могут быть взаимозависимыми, причем эта зависимость часто имеет нелинейный характер. Традиционные методы, основанные на линейной корреляции, не способны эффективно выявлять такие зависимости. Это приводит к избыточности данных и снижению эффективности аналитических моделей.

Неэффективность евклидовой метрики для оценки нелинейных взаимосвязей. Евклидова метрика, используемая в традиционных методах анализа данных, не адекватна для оценки нелинейных взаимосвязей между параметрами. Как показали Баттерворт и соавторы, «евклидовы расстояния могут быть крайне обманчивыми при анализе сложных многомерных данных с нелинейными взаимодействиями» [4].

Отсутствие единого подхода к фильтрации параметров. В настоящее время отсутствует единый подход, который бы позволял:

– эффективно выявлять нелинейные взаимозависимости между параметрами;

– адаптивно фильтровать избыточные параметры;

– сохранять интерпретируемость результатов анализа;

– обеспечивать вычислительную эффективность при работе с большими данными.

Предлагаемая в данной статье идея трансформации данных из гиперкуба в гиперсферу с последующей фильтрацией и обратным преобразованием направлена на решение именно этих проблем.

Концептуальное описание предлагаемого подхода - идея перехода от гиперкуба к гиперсфере и обратно. Центральная идея предлагаемого подхода заключается в трансформации представления данных из традиционного гиперкуба в сферическое пространство (гиперсферу), где нелинейные взаимосвязи между параметрами могут быть более эффективно выявлены и проанализированы. После выявления и исключения взаимозависимых параметров данные преобразуются обратно в форму гиперкуба, но уже с оптимизированной размерностью.

Этот процесс можно представить в виде следующих этапов:

– формирование исходного гиперкуба данных – данные организуются в многомерную структуру, где каждое измерение соответствует определенному признаку или параметру;

– трансформация в сферическое пространство – каждый параметр представляется как точка на единичной гиперсфере $S^{(n-1)}$, где n – исходная размерность данных;

– анализ геометрических свойств в сферическом пространстве – выявление нелинейных взаимозависимостей между параметрами на основе геометрических характеристик (сферических избытков, углов, расстояний);

– фильтрация взаимозависимых параметров – исключение избыточных параметров на основе выявленных зависимостей;

– обратное преобразование к форме гиперкуба – построение оптимизированного гиперкуба с исключенными взаимозависимыми параметрами.

Сферическая геометрия как инструмент анализа нелинейных взаимосвязей. Сферическая геометрия предоставляет уникальные возможности для анализа нелинейных взаимосвязей между параметрами. В отличие от евклидовой геометрии, где прямые линии соответствуют кратчайшим расстояниям, в сферической геометрии кратчайшими путями являются дуги больших кругов. Это принципиально меняет характер метрических отношений между точками.

Ключевым элементом анализа в сферическом пространстве является сферический избыток треугольника, образованного тремя точками на сфере (формула 1):

$$E(p_i, p_j, p_k) = \alpha + \beta + \gamma - \pi, \quad (1)$$

где α, β, γ - углы сферического треугольника, образованного параметрами p_i, p_j, p_k.

Сферический избыток служит мерой нелинейности взаимодействия между параметрами. Чем больше сферический избыток, тем сильнее нелинейный характер взаимосвязи.

Формализация критериев взаимозависимости параметров. Для формализации критериев взаимозависимости параметров вводится понятие радиуса кривизны локального пространства параметров (формула 2):

$$r(\rho) = \sqrt{[A_{sph}/E]}, \quad (2)$$

где A_{sph} – площадь сферического многоугольника, образованного взаимодействующими параметрами, E – сферический избыток многоугольника.

Параметры считаются взаимозависимыми, если разница их радиусов кривизны меньше определенного порогового значения (формула 3):

$$\frac{|r(\rho_i) - r(\rho_j)|}{\max(r(\rho_i),\ r(\rho_j))} < \varepsilon, \quad (3)$$

где ε – пороговый параметр, определяющий степень чувствительности метода к выявлению взаимозависимостей.

Процедура фильтрации параметров гиперкуба. На основе выявленных взаимозависимостей производится фильтрация параметров гиперкуба. Для этого вводится понятие базовых и производных параметров:

– базовые параметры – минимальный набор параметров, необходимый для полного описания данных с учетом линейных зависимостей;

– производные параметры – параметры, которые могут быть выражены через базовые с учетом нелинейных зависимостей.

Процедура фильтрации заключается в выявлении и исключении производных параметров, которые могут быть выражены через базовые. Математически это можно представить следующим образом (формула 4):

$$N_{filtered} = N_{base} + \sum_{k=2}^{k_{crit}} F(k, r, \varepsilon), \quad (4)$$

где N_{base} – число базовых параметров, F(k, r, ε) – функция, определяющая число независимых параметров порядка k после фильтрации с радиусом кривизны r и порогом ε.

Алгоритм фильтрации взаимозависимых параметров. Алгоритм фильтрации взаимозависимых параметров можно представить в виде следующих шагов:

1 Определение базовых параметров
2 Инициализация множества фильтрованных параметров
3 Анализ комбинаций параметров
   a. для каждой комбинации вычислить сферический избыток E;
   b. для каждой комбинации вычислить радиус кривизны r.
4 Кластеризация параметров по радиусу кривизны
5 Фильтрация взаимозависимых параметров
6 Формирование оптимизированного гиперкуба

Этот алгоритм обеспечивает эффективную фильтрацию параметров с учетом их нелинейных взаимозависимостей и позволяет существенно снизить размерность данных без значимой потери информации.

Потенциальные приложения и преимущества подхода

Применение в бизнес-аналитике. Предлагаемый подход имеет широкий спектр потенциальных приложений в бизнес-аналитике:

– оптимизация OLAP-кубов – снижение размерности OLAP-кубов без значимой потери аналитической информации;

– улучшение прогностических моделей – исключение избыточных параметров повышает точность и стабильность прогнозов;

– оптимизация клиентской аналитики – выявление ключевых параметров, определяющих поведение клиентов;

– анализ цепочек поставок – идентификация критических параметров в сложных логистических системах;

– финансовый анализ – выявление ключевых факторов риска и доходности в финансовых инструментах.

Преимущества перед существующими методами. Предлагаемый подход обладает следующими преимуществами перед существующими методами:

– учет нелинейных взаимосвязей – в отличие от линейных методов, таких как PCA, предлагаемый подход эффективно выявляет нелинейные зависимости между параметрами;

– сохранение интерпретируемости – в отличие от многих нелинейных методов снижения размерности, результаты анализа сохраняют прямую интерпретацию в терминах исходных параметров;

– адаптивность – метод автоматически адаптируется к характеристикам конкретного набора данных;

– математическая обоснованность – подход имеет строгое математическое обоснование в рамках сферической геометрии и теории информации.

Ограничения и направления дальнейшей разработки. Несмотря на потенциальные преимущества, предлагаемый подход имеет определенные ограничения:

– вычислительная сложность – анализ в сферическом пространстве может требовать значительных вычислительных ресурсов, особенно для данных высокой размерности;

– эмпирический подбор параметров – эффективность метода зависит от правильного выбора параметров, таких как пороговое значение ε;

– необходимость экспериментальной проверки – теоретические преимущества требуют подтверждения на реальных бизнес-данных.

Направления дальнейшей разработки включают:

– разработку эффективных алгоритмов вычисления сферических метрик для больших данных;

– создание методов автоматического подбора оптимальных параметров метода;

– экспериментальную валидацию на различных типах бизнес-данных;

– интеграцию с существующими инструментами бизнес-аналитики.

Заключение. В данной статье предложена инновационная идея адаптивной фильтрации многомерных данных на основе перехода от представления данных в форме гиперкуба к сферической геометрии гиперсферы и обратно. Этот подход позволяет эффективно выявлять и исключать взаимозависимые параметры, учитывая их нелинейные взаимосвязи, что потенциально может значительно повысить эффективность бизнес-аналитики.

Ключевыми элементами предложенного подхода являются:

– трансформация представления данных из гиперкуба в гиперсферу;

– анализ нелинейных взаимосвязей на основе сферической геометрии;

– фильтрация взаимозависимых параметров с использованием критерия близости радиусов кривизны;

– обратное преобразование к гиперкубу с оптимизированной размерностью.

Данный подход имеет прочную математическую основу и потенциально может преодолеть ограничения существующих методов анализа многомерных данных. Однако его практическая эффективность требует дальнейшей экспериментальной проверки и разработки эффективных алгоритмов реализации.

Перспективы дальнейших исследований включают разработку конкретных алгоритмов и программных инструментов для реализации предложенного подхода, его интеграцию с существующими платформами бизнес-аналитики и валидацию на различных типах бизнес-данных.

Список литературы

[1] World Bank. Cloud Computing: Growth Drivers, Main Players, and Key Trends. Chapter 4 in Advancing Cloud and Data Infrastructure Markets. Washington, DC: World Bank, 2021. – 49 с.

[2] Holmström, Lasse. Review of Nonlinear Dimensionality Reduction by John A. Lee and Michel Verleysen. International Statistical Review, February 2008, 76(2):308–309.

[3] Van Der Maaten L., Hinton G. Visualizing Data using t-SNE // Journal of Machine Learning Research. 2008. Vol. 9. P. 2579-2605

[4] Butterworth R., Piatetsky-Shapiro G., Simovici D.A. On Feature Selection through Clustering // IEEE International Conference on Data Mining. 2015. P. 581-584.

[5] Amari S., Nagaoka H. Methods of Information Geometry. American Mathematical Society, 2007. – 206 p.

Авторский вклад

Навроцкий Анатолий Александрович – руководство исследованием и постановка задачи. Общее руководство научным проектом, формулировка научной проблемы.

Евдокимов Виталий Геннадьевич – разработка концепции, алгоритмического обеспечения и анализ практического применения.

ADAPTIVE FILTERING OF MULTIDIMENSIONAL DATA FOR BUSINESS DECISION OPTIMIZATION BASED ON NONLINEAR METRICS OF FEATURE INTERDEPENDENCE

V.G. Evdokimov
Postgraduate student of the Department of Information Technologies of Automated Systems, BSUIR

A.A. Navrotsky
Head of the Department of Information Technologies of Automated Systems, BSUIR, PhD in Physics and Mathematics, Associate Professor

Abstract. The article proposes a conceptually new approach to adaptive filtering of multidimensional data, based on the transition from data representation in the form of a traditional hypercube to the spherical geometry of a hypersphere and back. The method allows to effectively identify and exclude interdependent parameters by analyzing their geometric properties in spherical space. The theoretical foundations of the idea, mathematical apparatus for evaluating nonlinear dependencies between features, and potential advantages of this approach in various business analytics tasks are considered. The prospects for practical implementation and adaptation of the method for specific subject areas are discussed.

Keywords: multidimensional data, hypercube, hypersphere, spherical geometry, feature interdependence, data filtering, business analytics, OLAP.