ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПОСТРОЕНИЯ ДВУМЕРНЫХ ПРОЕКЦИЙ ПРИЗНАКОВОГО ПРОСТРАНСТВА ДЛЯ ВЫДЕЛЕННОГО КЛАССА

Г.Г. Коновалов, *студент*
Научный руководитель: *Н.О.К. Зарбалиева, доцент*
Волгоградский государственный университет
(Россия, г. Волгоград)

DOI:10.24412/2500-1000-2025-10-2-221-226

Аннотация. *В статье рассматриваются теоретические основы построения двумерных проекций признакового пространства с акцентом на анализ выделенного класса объектов. Обосновывается актуальность задачи снижения размерности при наличии высокой размерности данных и сложности их интерпретации. Проведён обзор основных методов построения проекций, выделены их преимущества и недостатки. Сформулированы критерии качества двумерных проекций, включающие сохранение внутриклассовой компактности, межклассовой разделимости, глобальной структуры данных и интерпретируемости визуализаций. Особое внимание уделено специфике построения проекций для выделенного класса и возможным стратегиям усиления его информативности.*

Ключевые слова: *снижение размерности, двумерные проекции, признаковое пространство, выделенный класс, интерпретируемость, t-SNE, UMAP, PCA, LDA, оптимизационная модель.*

Современные задачи анализа данных характеризуются высокой размерностью признакового пространства. В таких условиях исследователь сталкивается с проблемами интерпретации, вычислительной сложности и избыточности информации. Для решения этих задач активно применяются методы снижения размерности, позволяющие представить многомерные данные в более компактной форме. Особенно важную роль играют двумерные проекции, так как именно они обеспечивают наглядную визуализацию структуры данных и упрощают выявление закономерностей.

Отдельное значение имеет построение двумерных проекций для выделенного класса объектов. В ряде приложений – от биоинформатики и медицины до анализа текстов и изображений – исследователь интересуется не столько общей структурой данных, сколько особенностями одного конкретного класса. В таких случаях стандартные методы проекций оказываются недостаточно эффективными, так как они ориентированы на сохранение общей структуры пространства, а не на выделение интересующего класса.

Теоретические основы признаковых пространств

В основе большинства методов анализа и обработки данных лежит представление объектов в виде точек многомерного пространства признаков. Под признаковым пространством будем понимать декартово пространство R^n, каждая ось которого соответствует одному измеримому или вычисляемому признаку объекта. В таком представлении каждый объект описывается вектором:

$$x = (x_1, x_2, \dots, x_n),$$

где n – размерность пространства, а x_i – значение i-го признака.

С увеличением числа признаков возрастает сложность анализа данных. Высокая размерность приводит к ряду проблем, известных как «проклятие размерности» (curse of dimensionality). Среди них часто выделяют [4]:

- разреженность точек в пространстве, затрудняющую оценку плотностей распределений;
- снижение эффективности традиционных метрических функций расстояния (например, евклидовой);
- экспоненциальный рост вычислительных затрат при анализе.

В результате интерпретация данных в исходном пространстве становится затруднительной, что делает актуальными методы снижения размерности и визуализации.

Также в задачах классификации и распознавания данных важную роль играет понятие выделенного класса. Под выделенным классом будем понимать подмножество объектов в признаковом пространстве, обладающих общими свойствами и объединённых по заранее определённому критерию. Например, в медицинской диагностике таким классом могут быть пациенты с конкретным заболеванием, а в задаче анализа изображений – объекты определённого типа [7].

Исследование выделенного класса требует не только понимания его внутренней структуры, но и анализа его положения относительно остальных объектов пространства. Визуализация в двумерных проекциях позволяет наглядно оценить компактность, разделимость и взаимное расположение классов.

Так как человек способен воспринимать данные в основном в двумерной или трёхмерной форме, для анализа многомерных данных применяется переход к пространствам меньшей размерности. При этом важно, чтобы проекции сохраняли ключевые свойства исходных данных [5]:

- внутриклассовую структуру (компактность и взаимосвязи внутри выделенного класса);
- межклассовую разделимость (чёткие границы между различными классами);
- информативность признаков, определяющих характерные особенности данных.

Методы двумерных проекций

Построение двумерных проекций признакового пространства основано на применении методов снижения размерности, позволяющих преобразовать многомерные данные $X \in R^n$ в пространство низкой размерности $Y \in R^2$ при этом максимально сохранив существенные свойства исходных данных. В общем случае задача проекции формулируется как построение отображения:

$$f : R^n \to R^2,$$

которое минимизирует потерю информации.

Методы построения двумерных проекций принято разделять на две большие группы[3]:

- линейные методы – используют линейные преобразования исходных признаков и сохраняют глобальные структуры данных;
- нелинейные методы – позволяют выявлять сложные зависимости и локальные структуры, что особенно актуально для данных высокой размерности.

Наиболее известными линейными методами являются:

1) Метод главных компонент (PCA, Principal Component Analysis).

Основан на ортогональном преобразовании данных с целью выделения направлений наибольшей дисперсии. Двумерная проекция формируется за счёт выбора двух главных компонент, объясняющих наибольшую долю вариации данных.

2) Линейный дискриминантный анализ (LDA, Linear Discriminant Analysis).

Используется для задач классификации. Метод находит проекцию, которая максимизирует разделимость классов, т.е. отношение межклассовой дисперсии к внутриклассовой. Для выделенного класса LDA позволяет улучшить визуализацию за счёт более явного отделения целевого множества объектов.

Преимущества линейных методов заключаются в простоте вычислений и интерпретации. Однако они показывают крайне низкий результат в случаях, когда данные имеют сложную нелинейную структуру.

Для анализа данных с более сложной внутренней организацией применяются нелинейные методы:

1) t-SNE (t-distributed Stochastic Neighbor Embedding).

Основан на минимизации расхождения между вероятностными распределениями расстояний в исходном и целевом пространстве. Хорошо выявляет локальные кластеры, однако может искажать глобальную структуру.

2) UMAP (Uniform Manifold Approximation and Projection).

Более современный метод, опирающийся на топологическое приближение многомерного многообразия. Обеспечивает высокую ско-

рость работы и сохранение как локальных, так и глобальных свойств данных.

Нелинейные методы особенно полезны при визуализации выделенного класса, так как способны подчеркнуть его особенности даже в случае сложного распределения в пространстве.

Сравнительная характеристика методов:

- линейные методы лучше подходят для задач, где важна интерпретируемость признаков и глобальная структура данных;
- нелинейные методы более эффективны при выявлении скрытых закономерностей и локальных особенностей классов, но требуют тщательной настройки гиперпараметров.

Для выделенного класса часто оказывается целесообразным использовать комбинацию методов: сначала линейное преобразование для снижения шумов и избыточности, затем нелинейное – для уточнения структуры.

Критерии качества двумерных проекций

Оценка качества построенной двумерной проекции является ключевым этапом в анализе многомерных данных. Даже при использовании мощных алгоритмов снижения размерности результат может оказаться малоинформативным, если он не отражает значимых особенностей исходного пространства. Для выделенного класса это имеет особое значение: недостаточно просто перенести данные в двумерное пространство, необходимо при этом сохранить как внутреннюю структуру класса, так и его разделимость от остальных объектов.

Одним из важнейших критериев является сохранение компактности выделенного класса. В проекции объекты одного класса должны располагаться относительно близко друг к другу, что указывает на их сходство в исходном пространстве. Нарушение этого свойства ведёт к размытости границ и снижению интерпретируемости [5].

Формально компактность можно оценивать с помощью средних внутриклассовых расстояний или дисперсии проекций:

$$S_W = \frac{1}{|C|} \sum_{x_i, x_j \in C} d(y_i, y_j),$$

где C – множество объектов выделенного класса, y_i, y_j – их образы в двумерной проекции, $d(y_i,\ y_j)$ – функция расстояния.

Наряду с компактностью важна межклассовая разделимость. В проекции выделенный класс должен занимать область, отделённую от других классов, что облегчает задачу классификации и визуального анализа.

Для количественной оценки разделимости могут применяться метрики на основе расстояния между центрами классов, а также индексы качества кластеризации, например:

- индекс Дэвиса-Булдина (DB-index);
- коэффициент силуэта;
- коэффициент Фишера, отражающий отношение межклассовой и внутриклассовой дисперсий.

Хотя приоритет отдаётся выделенному классу, важно учитывать и глобальные свойства данных. Хорошая двумерная проекция должна не только отображать целевой класс, но и сохранять взаимное расположение остальных объектов. Это обеспечивает целостность восприятия и предотвращает искажения, которые могут привести к ложным выводам.

Помимо формальных критериев важную роль играет субъективная оценка исследователя. Визуально понятная проекция должна [2]:

- демонстрировать чёткие границы между классами;
- выявлять локальные кластеры и аномалии;
- быть устойчивой к изменению параметров метода.

Интерпретируемость часто связывают с простотой восприятия: двумерная проекция должна позволять быстро сформировать гипотезы о структуре данных, которые затем могут быть подтверждены количественными методами.

Качество проекций также определяется их эффективностью при:

- выбору случайных инициализаций (особенно актуально для t-SNE и UMAP);

- шуму и выбросах в данных;
- изменению числа объектов выборки.

Успешная эффективность гарантирует, что получаемые визуализации отражают реальные свойства данных, а не артефакты алгоритма.

Особенности построения проекций для выделенного класса

В отличие от общих задач снижения размерности, ориентированных на сохранение глобальной структуры данных, задача построения двумерных проекций для выделенного класса имеет ряд специфических особенностей. Главная цель заключается не только в упрощении визуализации, но и в создании представления, которое подчёркивает свойства целевого множества объектов относительно остальной выборки.

При построении проекций необходимо учитывать, что выделенный класс может занимать относительно небольшую область в признаковом пространстве. Если алгоритм ориентирован исключительно на глобальные характеристики (например, PCA), то локальные закономерности внутри класса могут быть размыты. И напротив, методы, акцентирующие внимание на локальных соседствах (t-SNE, UMAP), позволяют лучше визуализировать внутреннюю структуру выделенного класса, но при этом могут искажать глобальное расположение других объектов.

Таким образом, особенность задачи заключается в необходимости нахождения компромисса – проекция должна сохранять локальную целостность выделенного класса и одновременно обеспечивать корректное отображение его положения среди других классов [6].

Для повышения наглядности проекций могут использоваться специальные приёмы:

- взвешивание объектов выделенного класса в функции потерь алгоритма снижения размерности, что позволяет придать целевому множеству большее значение;
- модификация метрик расстояния, учитывающих принадлежность к классу;
- выбор признаков с высокой дискриминативной способностью относительно выделенного класса, что уменьшает влияние нерелевантных данных.

Применение этих подходов позволяет улучшить визуальную интерпретацию целевого множества и повысить его разделимость.

При акценте на выделенный класс возможны следующие сложности [1]:

- риск потери информации о других классах, что может привести к искажению общей картины;
- чрезмерная концентрация на локальных свойствах, что снижает интерпретируемость в глобальном контексте;
- необходимость подбора параметров алгоритма, чтобы избежать артефактов в визуализации.

Перспективы и направления дальнейших исследований

Исследование теоретических основ построения двумерных проекций признакового пространства для выделенного класса открывает широкие возможности для дальнейшего развития методов визуализации и анализа данных. Несмотря на значительный прогресс в данной области, остаётся ряд нерешённых задач и направлений, требующих углублённого изучения.

Одним из перспективных направлений является создание гибридных алгоритмов, объединяющих преимущества линейных и нелинейных подходов. Например, предварительное линейное преобразование (для устранения шумов и корреляций) может сочетаться с нелинейными методами, обеспечивающими сохранение локальной структуры выделенного класса.

Традиционные методы проекций универсальны и не учитывают специфику конкретного класса. Перспективным является развитие адаптивных моделей, в которых параметры алгоритма автоматически настраиваются под свойства выделенного класса. Это позволит формировать проекции, оптимизированные для конкретной исследовательской задачи.

Важным направлением является разработка методов, обеспечивающих объяснимость проекций. Исследователь должен не только видеть структуру выделенного класса на плоскости, но и понимать, какие признаки вносят наибольший вклад в полученную визуализацию. Это особенно актуально для медицины, биологии и социальных наук, где интерпретируемость играет ключевую роль.

С ростом объёмов данных возрастает потребность в методах, способных строить двумерные проекции в условиях больших выбо-

рок. Перспективным направлением является разработка алгоритмов, обеспечивающих линейную или квазилинейную зависимость времени вычислений от числа объектов при сохранении качества визуализации.

Заключение

В статье были рассмотрены теоретические основы построения двумерных проекций признакового пространства с акцентом на анализ выделенного класса объектов. Показано, что задача визуализации в условиях высокой размерности имеет не только прикладное, но и фундаментальное значение, так как напрямую связана с проблемами интерпретируемости, устойчивости и качества анализа данных.

В целом, построение двумерных проекций для выделенного класса представляет собой важный теоретический и практический инструмент, способный повысить эффективность анализа данных и расширить возможности их интерпретации.

Библиографический список

1. Алябьев С.А. Требования к организации измерений для формирования признакового пространства при построении моделей систем / С.А. Алябьев, В.В. Севрюкова // Программная инженерия: современные тенденции развития и применения. – Курск: ЗАО «Университетская книга», 2019. – С. 284-288.

2. Баянов Е.В. Двумерное пространство, как основа геометрических построений / Е.В. Баянов // Актуальные научные исследования в современном мире. – 2020. – № 8-1(64). – С. 122-124.

3. Блажко А.В. Теоретические основы построения математических моделей для классов сложных систем / А.В. Блажко, В.С. Смородин // Новые математические методы и компьютерные технологии в проектировании, производстве и научных исследованиях. – Гомель: ГГУ им. Ф. Скорины, 2015. – С. 94-95.

4. Мирзаев Н.М. Построение модели распознавания в условиях большой размерности признакового пространства / Н.М. Мирзаев, А.Ш. Мухамадиев // Нейрокомпьютеры и их применение. – Москва: МГППУ, 2018. – С. 378-379.

5. Мунасыпов Н.А. Построение оптимального признакового пространства в системе распознавания при наличии ресурсных ограничений / Н.А. Мунасыпов // Научно-технический вестник Поволжья. – 2019. – № 4. – С. 156-159.

6. Немирко А.П. Сокращение размерности признакового пространства на основе критерия разделимости классов / А.П. Немирко // Интеллектуализация обработки информации ИОИ-2016. – Москва-Барселона: ООО «ТОРУС ПРЕСС», 2016. – С. 36-37.

7. Сазонов А.Д. Исследование применимости методов снижения размерности признакового пространства для задачи медицинской диагностики / А.Д. Сазонов, И.Н. Яковина // Наука. Технологии. Инновации. – Новосибирск: НГТУ, 2019. – С. 104-107.

THEORETICAL BASIS FOR CONSTRUCTING TWO-DIMENSIONAL FEATURE SPACE PROJECTIONS FOR AN SELECTED CLASS

G.G. Konovalov, *Student*
Supervisor: *N.O.K. Zarbalieva, Associate Professor*
Volgograd State University
(Russia, Volgograd)

Abstract. *This article examines the theoretical foundations of constructing two-dimensional feature space projections, with an emphasis on the analysis of a selected class of objects. The relevance of dimensionality reduction in the presence of high-dimensional data and the complexity of its interpretation is substantiated. A review of the main projection construction methods is provided, highlighting their advantages and disadvantages. Quality criteria for two-dimensional projections are formulated, including preserving intra-class compactness, inter-class separability, global data structure, and interpretability of visualizations. Particular attention is paid to the specifics of constructing projections for a selected class and possible strategies for enhancing its information content.*

Keywords: *dimensionality reduction, two-dimensional projections, feature space, extracted class, interpretability, t-SNE, UMAP, PCA, LDA, optimization model.*