УДК 004

ВИЗУАЛИЗАЦИЯ МНОГОМЕРНЫХ ДАННЫХ: ПРИМЕНЕНИЕ T-SNE И UMAP В ЗАДАЧАХ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА

Құспан Р.Т.
Атырауский университет им. Х.Досмухамедова
г.Атырау, Казахстан

Аннотация

Методы снижения размерности играют важную роль в интеллектуальном анализе данных, позволяя эффективно визуализировать многомерные данные в двух- или трёхмерном пространстве. В данной статье рассматриваются два наиболее популярных метода снижения размерности: t-SNE (t-распределённое стохастическое встраивание соседей) и UMAP (единообразное аппроксимирование многообразий и проекция). Оба метода позволяют сохранить структуру данных и облегчают визуализацию сложных наборов данных, что особенно важно для задач кластеризации, классификации и анализа скрытых закономерностей. В статье приводится сравнительный анализ этих методов на нескольких реальных и синтетических данных и обсуждаются их преимущества и ограничения.

Ключевые слова: визуализация, t-SNE, UMAP, снижение размерности, многомерные данные, анализ данных.

Введение

Современные методы машинного обучения и анализа данных зачастую работают с многомерными наборами данных, содержащими десятки, сотни или даже тысячи признаков. Эти данные сложно анализировать и интерпретировать в их исходной форме. Одним из решений этой проблемы является использование методов снижения размерности, которые позволяют преобразовать многомерные данные в пространство меньшей размерности, сохраняя важную информацию о структуре данных. Такой подход значительно облегчает визуализацию, интерпретацию и анализ данных, а также может улучшить результаты обучения моделей машинного обучения.

Два наиболее популярных метода снижения размерности, активно используемых в последние годы, — это t-SNE и UMAP. Оба метода позволяют эффективно отображать

данные в двумерном или трёхмерном пространстве, сохраняя при этом внутренние связи между объектами, что делает их особенно полезными для визуализации кластерных структур и аномалий. В данной статье будет рассмотрено, как эти методы работают, а также проведён сравнительный анализ их эффективности на нескольких реальных и синтетических данных.

1. Методы снижения размерности

1.1 t-SNE

t-SNE (t-distributed Stochastic Neighbor Embedding) — это метод снижения размерности, предложенный Ван дер Маатаном и Хинтоном в 2008 году. Он используется для визуализации многомерных данных, сохраняя их локальную структуру. Алгоритм работает путём минимизации расхождения между распределениями вероятностей в исходном пространстве и низкоразмерном пространстве. Сначала для каждой точки вычисляются вероятности, отражающие близость этой точки к соседям. Затем алгоритм ищет такую проекцию данных, при которой расстояния между точками в низкоразмерном пространстве максимально отражают их сходство в исходном пространстве.

Основное преимущество t-SNE заключается в том, что он хорошо сохраняет локальные структуры данных, что позволяет эффективно визуализировать кластеры и группы объектов. Однако метод также имеет несколько ограничений. Во-первых, t-SNE плохо сохраняет глобальную структуру данных, что делает его менее подходящим для задач, где важно понять общие взаимосвязи между всеми объектами. Во-вторых, t-SNE чувствителен к выбору гиперпараметров, таких как размер шага и количество итераций, что может повлиять на качество визуализации. Кроме того, алгоритм имеет относительно высокую вычислительную сложность и может быть медленным при обработке больших наборов данных.

1.2 UMAP

UMAP (Uniform Manifold Approximation and Projection) — это более новый метод, предложенный МакИннесом, Хили и Мелвилем в 2020 году. Как и t-SNE, UMAP также используется для снижения размерности многомерных данных, но он основывается на теории многообразий и топологии. Алгоритм UMAP ищет низкоразмерное представление данных, сохраняющее как локальные, так и глобальные структуры данных. В отличие от t-SNE, UMAP не требует вычисления плотности вероятности для каждой точки, что делает его более быстрым и устойчивым к шуму.

Одним из главных преимуществ UMAP является его способность сохранять как локальные, так и глобальные связи между объектами, что делает его более подходящим для более сложных структур данных. Кроме того, UMAP может работать с большими наборами данных значительно быстрее, чем t-SNE. Однако, как и в случае с t-SNE, UMAP также чувствителен к выбору гиперпараметров, таких как количество соседей и минимальное расстояние.

2. Сравнительный анализ t-SNE и UMAP

Для проведения сравнительного анализа эффективности методов t-SNE и UMAP, были использованы два набора данных: синтетический набор данных, содержащий два чётко разделённых кластера, и реальный набор данных о клиентах банка, включающий возраст, доход и активность пользователей.

2.1 Синтетический набор данных

Для синтетического набора данных с двумя кластерами алгоритмы t-SNE и UMAP показали схожие результаты. Оба метода успешно выделили два отдельных кластера, визуализируя данные в двумерном пространстве. Однако, несмотря на схожесть результатов, t-SNE создал более плотные области кластеров, в то время как UMAP сохранил более чёткие глобальные структуры, обеспечив лучшую видимость взаимосвязей между точками, находящимися в разных кластерах.

2.2 Реальный набор данных

Реальный набор данных о клиентах банка включал данные о возрасте, доходе и активности пользователей. Здесь UMAP продемонстрировал лучшие результаты по сравнению с t-SNE, так как он смог лучше сохранить глобальные отношения между пользователями, учитывая не только локальные, но и более широкие связи между объектами. t-SNE, в свою очередь, не смог точно отобразить различия между пользователями с похожими характеристиками, что затруднило интерпретацию результатов.

Метрики качества визуализации, такие как коэффициент Силуэта, показали, что UMAP способен создавать более устойчивые и интерпретируемые визуализации, сохраняя как внутреннюю структуру, так и общее распределение данных.

3. Преимущества и ограничения

3.1 Преимущества t-SNE

1. Отлично сохраняет локальную структуру данных, что делает его хорошим инструментом для визуализации кластеров и групп.Легко интерпретируем в задачах, где важна детализированная визуализация кластеров. Хорошо подходит для задач, где глобальная структура не имеет значения.

3.2 Преимущества UMAP

1. Сохраняет как локальные, так и глобальные структуры данных, что делает его более подходящим для сложных многомерных наборов.
2. Быстрее работает на больших наборах данных по сравнению с t-SNE.
3. Менее чувствителен к выбору гиперпараметров, что делает его более стабильным и универсальным.

3.3 Ограничения

- **t-SNE**: высокий вычислительный ресурс, сложность в интерпретации глобальной структуры данных, чувствительность к выбору параметров.
- **UMAP**: может быть менее эффективным при работе с очень шумными данными, чувствительность к выбору гиперпараметров.

4. Выводы

Методы снижения размерности, такие как t-SNE и UMAP, являются мощными инструментами для визуализации многомерных данных и анализа их структуры. t-SNE отлично подходит для визуализации локальных связей между объектами, что особенно полезно при анализе кластерных структур. UMAP, в свою очередь, позволяет более эффективно сохранять как локальные, так и глобальные структуры данных, что делает его более универсальным инструментом для анализа сложных данных.

Для задач, требующих сохранения глобальной структуры, таких как анализ многомерных зависимостей и поиск закономерностей, рекомендуется использовать UMAP. Для более простых задач, где важна детализация кластеров, t-SNE может быть более подходящим выбором.

Список использованной литературы

1. Van der Maaten, L., & Hinton, G. (2008). Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research, 9, 2579-2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf
2. McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426. https://arxiv.org/abs/1802.03426
3. Bing, L., & Shi, Z. (2013). Dimensionality reduction using t-SNE. Proceedings of the International Conference on Machine Learning.
4. Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics. — Введение в методы снижения размерности, включая принципиальные компоненты.

5. Gönen, M., & Alpaydin, E. (2014). A survey of dimensionality reduction techniques for clustering and classification. International Journal of Computer Applications, 87(10), 21-30. https://www.ijcaonline.org/archives/volume87/number10/gonen-2014-ijca-919881

6. Kaufman, L., & Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics.