ИНТЕЛЛЕКТУАЛЬНАЯ ВИЗУАЛЬНАЯ АНАЛИТИКА СЛОЖНЫХ ДАННЫХ

У. А. Кобзарь

Белорусский государственный университет, Минск, Беларусь
E-mail: kobzaruliana98@gmail.com

В работе исследуется возможность определения подходящего типа визуализации на основе признаков, извлекаемых из данных, с помощью алгоритмов машинного обучения. Для этого решение задачи разбивается на два этапа: определение типа визуализации и определение осей для отображения данных. Алгоритмы протестированы на наборах данных и соответствующих им визуализациях, полученных из Plotly Community Feed.

Ключевые слова: *визуальная аналитика; система рекомендации визуализаций; дерево принятия решений, оверсэмплинг, машинное обучение.*

Создание диаграмм для многомерного набора данных является распространенной задачей во многих областях: образование, исследования, инженерия, финансы. Достижение эффективной визуализации требует больших человеческих усилий и зависит от опыта в графическом дизайне, дизайне взаимодействия с пользователем и анализе данных. Однако «ручной» подход создания визуализаций является излишним для многих распространенных случаев, таких как предварительное исследование данных и создание базовых визуализаций.

В рамках этого исследования определим системы рекомендации визуализаций данных как инструменты, предлагающие эффективные визуализации. Под эффективной визуализацией понимается отображение данных в наиболее понятном виде, который выделяет их ключевые особенности. Системы рекомендации визуализаций делятся на две группы по основному подходу их построения:

- *Системы на основе правил* руководствуются принципами, которые опираются на экспериментальные данные и опыт экспертов.
- *Системы, использующие методы машинного обучения,* напрямую изучают взаимосвязь между данными и визуализациями.

Причиной создания систем, использующих методы машинного обучения [1, 3, 4], являются следующие выявленные ограничения систем на основе правил:

- определение правил системы вручную;
- нетривиальный и дорогостоящий процесс моделирования правил;
- низкая масштабируемость и адаптируемость системы.

Системы на основе методов машинного обучения повысили эффективность рекомендации визуализаций по сравнению с системами на основе правил. Однако в указанных исследованиях используются методы глубокого обучения, которые работают по принципу «черного ящика», что затрудняет интерпретацию рекомендуемых результатов.

Учитывая указанные недостатки систем на основе правил, предлагаемый подход опирается на классические алгоритмы машинного обучения. Эти алгоритмы позволяют получить удовлетворительные визуализации и выявить наиболее важные признаки, на основе которых было принято решение о типе визуализации (см. рис. 1).

Решение задачи построения визуализации по набору данных происходит в два этапа: определение типа визуализации, определение осей для отображения данных. Для решения обеих задачи использовались следующие методы машинного обучения: дерево принятия решений, случайный лес, метод градиентного бустинга, бэггинг. При оценивании результатов работы алгоритма кроме классических метрик, таких как accuracy, precision, recall, матрица ошибок, использовалась метрика top2-accuracy (доля правильных вариантов визуализации, оказавшихся в двух первых предполагаемых вариантах).

В ходе исследования использовался набор данных, который был собран с помощью Plotly API из общедоступных визуализаций Plotly Community Feed [2]. Данный набор содержит следующие типы диаграмм: точечная, столбиковая, линейная, гистограмма, диаграмма размаха, тепловая карта. Так как распределение диаграмм по типам визуализации является слабо сбалансированным, для поддержания баланса классов были применены техники оверсэмплинга и андерсэмплинга из модуля imblearn.

Согласно VizML [2], для каждого столбца из набора данных, предназначенного для визуализации, можно получить признаки, которые определяют тип столбца, рассчитывают статистические характеристики значений в столбце (распределение, выбросы и т. д.) и описывают название столбца. Извлеченные признаки используются для определения типа визуализации и осей для отображения данных.

Для проведения эксперимента было случайным образом выбрано 20.000 наборов данных для визуализации, которые суммарно содержат более 60.000 столбцов с данными. Перед решением задачи к тренировочным данным применялся алгоритм андерсэмплинга AllKNN для уменьшения количества экземпляров классов столбиковой, линейной и точечной диаграмм, а для увеличения количества экземпляров диаграммы размаха, гистограммы и тепловой карты – методы SMOTE, ADASYN, SVMSMOTE, SMOTETomek.

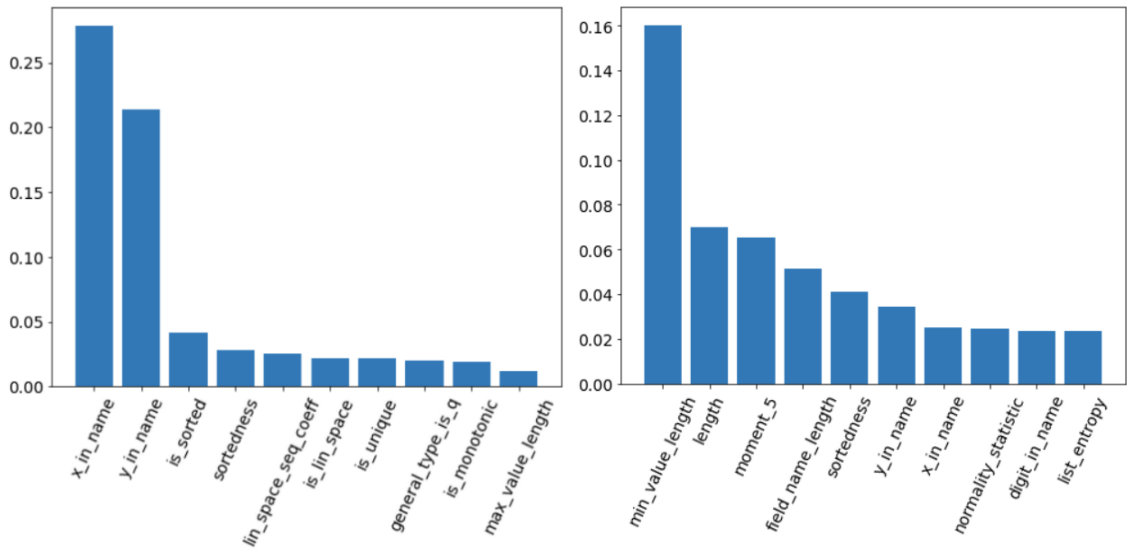

Рис. 1. Важность признаков для опеределения оси отобрежеия (слева) и определения типа визуализации (справа) в методах, показавших наилучший результат

Наилучший результат *top2-accuracy*=0.865 получен с помощью алгоритма градиентного бустинга с применением метода оверсэмплинга SMOTE. Для определения оси отображения данных использовались те же методы машинного обучения, что и для определения типа визуализации и также были применены методы оверсэмплинга для поддержания баланса классов, т. к. класс с меткой оси y преобладает над классом с меткой оси x. Наилучший результат *auc*=0.965 получен при применении алгоритма случайного леса с методом оверсэмплинга SMOTE.

Полученные модели имеют наибольшие неточности при определении классов точечной, линейной и столбиковых диаграмм. Эти неточности могут являться следствием отсутствия знаний контекста построения визуализации, поэтому для рекомендации визуализаций используются два наиболее вероятных типа визуализации, предложенных алгоритмом.

Для демонстрации результатов работы алгоритмов, показавших наилучший результат, было разработано веб-приложение, в котором для визуализации данных использовалась библиотека Highcharts JS (рис. 2). Приложение позволяет загрузить данные в формате JSON или csv.

После загрузки данные обрабатываются следующим образом:

- вычисляются необходимые метрики загруженных данных
- с помощью алгоритма градиентного бустинга определяется топ-2 типа визуализации
- с помощью алгоритма случайного леса определяются оси для каждого столбца данных
- на основе полученных результатов формируются визуализации и отображаются пользователю.

Рис. 2. Пример визуализаций данных, полученных с помощью демонстрационного приложения

Выводы. С помощью предложенных алгоритмов машинного обучения были отобраны наиболее важные признаки для определения типа визуализации. Несмотря на то, что эти алгоритмы не подбирают наиболее эффективную визуализацию, они предлагают достаточно эффективные представления, которые помогают проанализировать, выявить особенности и сделать выводы по предоставленным данным. Разработанное приложение можно использовать для подбора визуализаций к предоставляемым наборам данных, вопреки отсутствию знаний контекста.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Kaur P., Owonibi M. A Review on Visualization Recommendation Strategies // In Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017). 2017. V. 3. DOI: 10.5220/0006175002660273.
2. Hu K. Z., Bakker M. A., Stephen Li et al. VizML: A Machine Learning Approach to Visualization Recommendation // In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019. 1–12, DOI: 10.1145/3290605.33003581
3. Li H., Wang Y., Zhang S., et al. KG4Vis: A Knowledge Graph-Based Approach for Visualization Recommendation. // in IEEE Transactions on Visualization and Computer Graphics, V. 28, N. 1, P. 195-205. DOI: 10.1109/TVCG.2021.3114863.
4. Zhou M., Li Q., He et al. Table2Charts: Recommending Charts by Learning Shared Table Representations // KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery. 2021. P. 2389-2399. DOI: 10.1145/3447548.3467279.