УДК 004.82

АВТОРАЗМЕТКА ВРЕМЕННЫХ РЯДОВ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ ПАТТЕРНОВ

Мещеряков И.А. (ИТМО)
Научный руководитель – кандидат технических наук, ст. преподаватель Ходненко И.В. (ИТМО)

Введение. Кластеризация данных до сих пор остается важной частью процесса применения методов машинного обучения и глубокого обучения [1]. Благодаря качественно размеченным данным исследователи могут совершенствовать инструменты анализа, обучать модели и делать глубокие, обоснованные выводы. Однако разрыв в соотношении размеченных и неразмеченных данных остается огромным, и создание репрезентативного размеченного датасета требует крупных финансовых вложений и времени [2]. Более того, кластеризация временных рядов связана со следующими сложностями: высокая размерность, вычислительная сложность, выбор из большого числа методов, несопоставимость результатов исследований в рамках одной области, зависимость наблюдений во времени [3]. Целью данной работы является разработка метода автоматической разметки временных рядов с возможностью ручной настройки гиперпараметров, который улучшит качество результатов алгоритмов при решении задачи классификации.

Основная часть. В ходе предварительного анализа исследований в нескольких областях и доступных реализаций методов кластеризации был сформулирован подход и параметры, которые позволяют проводить кластеризацию автоматически. Одним из ключевых этапов является подготовка данных и разбиение на фрагменты. Для предварительного анализа были использованы синтетические данные 7 различных распределений. Далее к ним были применены статистические преобразования, анализ пиков, частотные методы (DFT, DWT) и методы, основанные на глубоком обучении. После анализа полученных представлений они были оценены на способность выделения характеристик любого временного ряда: тренд, сезонность, цикличность, шум. Разбиение на фрагменты было реализовано методом сопоставление последовательностей (subsequence matching) [4] с помощью тестирования нескольких конфигураций. После этапа подготовки и разбиения данных был проведен количественный анализ с использованием алгоритмов кластеризации. В тестировании были выделены 3 группы методов кластеризации: partitioning, hierarchical, density-based методы. Для оценки эффективности применялись классические метрики: Индекс Дэвида-Болдуина (Davies-Bouldin), Коэффициент Силуэта (Silhouette Coefficient), Индекс Дунна (Dunn Index). Для оптимизации параметров алгоритмов кластеризации использовался сетка параметров. Для удобства тестирования были составлены короткий и длинные списки параметров. Короткий список параметров использовался в любом случае и состоял их параметров, которые оптимизировали внутрикластерное расстояние. Длинный список использовался для достижения максимальных значений метрик. По итогу проведенного анализа для различных сценариев алгоритм предлагает пользователю наилучший вариант кластеризации и методов предобработки, а по необходимости ручную настройку параметров. Настроенный алгоритм затем был протестирован на данных цен криптовалют, сенсорных данных нескольких типов.

Выводы. В рамках работы был разработан и реализован настраиваемый алгоритм и частично автоматический алгоритм для разметки данных, основанный на применении алгоритмов кластеризации. Основной целью исследования было создание инструмента, способного автоматизировать процесс разметки неразмеченных данных временных рядов и улучшить качество моделей классификации. Для реальных данных качество классификации при использовании автоматической разметки выросло в среднем на 5%. Предложенный метод может быть использован в различных областях, таких как анализ финансовых данных, анализ сенсоров и других временных рядов.

Дальнейшие исследования могут быть направлены на оптимизацию алгоритма при работе с многомерными временными рядами, а также на дополнения методами глубокого обучения в процессе предобработки.

Список использованных источников:

1. Ezugwu A. E. et al. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects //Engineering Applications of Artificial Intelligence. – 2022. – Т. 110. – С. 104743.
2. Aghabozorgi S., Shirkhorshidi A. S., Wah T. Y. Time-series clustering–a decade review //Information systems. – 2015. – Т. 53. – С. 16-38.
3. Ezugwu A. E. et al. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature //Neural Computing and Applications. – 2021. – Т. 33. – С. 6247-6306.
4. Fu T. A review on time series data mining //Engineering Applications of Artificial Intelligence. – 2011. – Т. 24. – №. 1. – С. 164-181.

Автор _____ Мещеряков И.А.

Научный руководитель _____ Ходненко И.В.