

МЕТОДОЛОГИЯ ОЦЕНКИ СЛОЖНОСТИ ПАТТЕРНА

Мячин А.Л.

Национальный исследовательский университете «Высшая школа экономики»
Институт проблем управления им. В.А. Трапезникова РАН

Введение

Методы анализа паттернов, популярные в задачах анализа данных, имеют несколько ключевых направлений развития. Одним из них является минимизация воздействия шума на конечные результаты. Ошибки измерений, выбросы и другие аномалии искажают анализ, поэтому важным является создание робастных методов, устойчивых к таким факторам. Другим направлением является снижение размерности данных, что позволяет уменьшить вычислительную сложность и ускорить обработку больших массивов информации. Третье направление – обеспечение интерпретируемости и оценка качества результатов. Использование метрических показателей и визуализации помогает оценивать значимость и достоверность выводов, делая процесс анализа прозрачным. Масштабируемость методов также является приоритетной задачей. Разработка распределённых алгоритмов и использование облачных вычислений позволяют эффективно обрабатывать большие данные и сокращать время анализа. Оптимизация количества паттернов играет важную роль в интерпретируемости результатов. Слишком мало паттернов может исказить структуру данных, а избыточное их количество усложняет интерпретацию. Наконец, интеграция с традиционными методами анализа расширяет возможности междисциплинарных исследований, повышает адаптивность моделей и их практическую значимость (к примеру, в таких областях, как медицина или финансовый анализ).

Настоящая работа посвящена частичному развитию третьего направления, а именно: повышению интерпретируемости конечных результатов. В связи с этим, рассматриваются такие понятия, как «пустой паттерн» и «сложность паттерна».

Предлагаемые методы

Общая постановка задачи анализа паттернов приведена во множестве работы, к примеру [Aleskerov 2014; Myachin 2019]. В качестве исходных данных исследуется n объектов $x_i = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{im})$, которые необходимо разбить на непересекающиеся подмножества (паттерны p) со схожей структурой по выбранной системе показателей.

Поскольку в работе предлагается к рассмотрению два новых термина, приведем обоснования необходимости их использования. Методы анализа паттернов успешно применялись при работе с прогностическими моделями (к примеру, при анализе уровня заболеваемости короновирусной инфекцией [Aleskerov et al 2022], а также ценообразования в российских ВУЗах [Дмитриенко, Мячин 2023]). Одной из ключевых задач анализа паттернов является выделение динамических групп объектов, обладающих сходной структурой развития. При этом каждый паттерн может представлять собой группу объектов, для которых требуется внесение определённых улучшений в соответствии с заранее заданными критериями. Для таких задач возможно возникновение трёх сценариев: повышение или снижение значений выбранных показателей приводит к сохранению принадлежности объекта к исходному паттерну; изменение значений показателей объекта вызывает переход к другому, уже существующему паттерну; изменение значений показателей создаёт новую структуру, формируя новый паттерн, ранее не существовавший.

В последнем сценарии возникает необходимость введения концепции «пустого паттерна» для прогнозирования смены траектории развития исследуемого объекта. Такой подход позволяет учитывать возможность появления нового паттерна, который изначально не содержал ни одного объекта. При этом использование вероятностных оценок позволяет корректно оценивать вероятность перехода объекта в новое состояние с учетом изменения траектории его развития. Далее, добавим понятие «пустого паттерна» и «сложности паттерна».

Определение 1. Пустой паттерн представляет собой теоретически возможное пустое подмножество, формирование которого обусловлено результатами парных сравнений объектов, выполненных в рамках заданного метода анализа паттернов.

Определение 2. Сложность паттерна определяется как величина, равная единице плюс минимальное число паттернов, необходимых для преобразования в другой паттерн, принимая во внимание существование пустых паттернов. Данная методология особенно актуальна для методов анализа паттернов, основанных на парном сравнении показателей [Мячин 2016а; Мячин 2016б].

Таким образом, введение понятия «пустого паттерна» и анализ его роли в прогнозировании, а также определение сложности паттернов, способствует более точному описанию и моделированию процессов, происходящих в сложных динамических системах. Эти концепции обеспечивают методологическую основу для проведения глубокого анализа структуры и динамики развития объектов, относящихся к различным паттернам.

Заключение

Исследование демонстрирует важность введения и использования понятий «пустой паттерн» и «сложность паттерна» в задачах анализа данных. Данные концепции позволяют более точно моделировать сложные динамические процессы, учитывать появление новых структур и прогнозировать их развитие. Особое значение имеют предложенные подходы к оценке вероятностных характеристик перехода объектов между паттернами, что способствует улучшению интерпретируемости и обоснованности результатов анализа.

Благодарность. Исследование выполнено за счет гранта Российского научного фонда № 24-61-00030, <https://rscf.ru/project/24-61-00030/>

Литература

[Aleskerov 2014] Aleskerov F. et al. A Method of Static and Dynamic Pattern Analysis of Innovative Development of Russian Regions in the Long Run //Models, Algorithms and Technologies for Network Analysis: From the Third International Conference on Network Analysis. – Springer International Publishing, 2014. – С. 1-8.

[Aleskerov et al 2022] Aleskerov F. et al. Short-Term Covid-19 Incidence Prediction in Countries Using Clustering and Regression Analysis //International Conference on Computers Communications and Control. – Cham : Springer International Publishing, 2022. – С. 333-342.

[Myachin 2019] Myachin A. L. Pattern analysis in parallel coordinates based on pairwise comparison of parameters //Automation and Remote Control. – 2019. – Т. 80. – С. 112-123.

[Дмитриенко, Мячин 2023] Дмитриенко А. С., Мячин А. Л. Ценообразование в российских вузах: как влияют на стоимость обучения вузы-конкуренты? //Университетское управление: практика и анализ. – 2023. – Т. 27. – №. 2. – С. 75-88.

[Мячин 2016a] Мячин А. Л. Анализ паттернов: порядково-инвариантная паттерн-кластеризация // Управление большими системами: сборник трудов. – 2016. – №. 61. – С. 41-59.

[Мячин 2016b] Мячин А. Л. Анализ паттернов: диффузионно-инвариантная паттерн-кластеризация // Проблемы управления. – 2016. – №. 4. – С. 2-9.