МЕТОДЫ И ИНСТРУМЕНТЫ ЭКОНОМИЧЕСКОЙ АНАЛИТИКИ НА ОСНОВЕ БОЛЬШИХ ДАННЫХ

Ахметов Т.К., магистрант

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Беляцкая Т.Н. – докт. экон. наук, доцент

Аннотация. В данной статье рассматривается применение кластерного анализа (метод К-средних) для анализа вклада в мировую цифровую экономику 5 постсоветских стран (РФ, Эстония, Казахстан, Беларусь, Узбекистан) в различные временные промежутки.

Ключевые слова. Кластерный анализ, методы и инструменты экономической аналитики на основе больших данных

Кластерный анализ (англ. cluster analysis) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы[1]. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Кластерный анализ выполняет следующие основные задачи:

Разработка типологии или классификации.

Исследование полезных концептуальных схем группирования объектов.

Порождение гипотез на основе исследования данных.

Проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:

Отбор выборки для кластеризации. Подразумевается, что имеет смысл кластеризовать только количественные данные.

Определение множества переменных, по которым будут оцениваться объекты в выборке, то есть признакового пространства.

Вычисление значений той или иной меры сходства (или различия) между объектами.

Применение метода кластерного анализа для создания групп сходных объектов.

Проверка достоверности результатов кластерного решения.

Методы кластеризации:

1) Вероятностный подход. Предполагается, что каждый рассматриваемый объект относится к одному из k классов. Некоторые авторы (например, А. И. Орлов) считают, что данная группа вовсе не относится к кластеризации и противопоставляют её под названием «дискриминация», то есть выбор отнесения объектов к одной из известных групп (обучающих выборок).

2) K-средних

3) K-медиан

4) EM-алгоритм

5) Алгоритмы семейства FOREL

6) Дискриминантный анализ

Метод k-средних (англ. K-means) — наиболее популярный метод кластеризации. Был изобретён в 1950-х годах математиком Гуго Штейнгаузом [2] и почти одновременно Стюартом Ллойдом [3]. Особую популярность приобрёл после работы Маккуина [4].

Цели кластеризации:

1) Понимание данных путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»).

2) Сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера.

3) Обнаружение новизны (англ. novelty detection). Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

В первом случае число кластеров стараются сделать поменьше. Во втором случае важнее обеспечить высокую степень сходства объектов внутри каждого кластера, а кластеров может быть сколько угодно. В третьем случае наибольший интерес представляют отдельные объекты, не вписывающиеся ни в один из кластеров.

4) Во всех этих случаях может применяться иерархическая кластеризация, когда крупные кластеры дробятся на более мелкие, те в свою очередь дробятся ещё мельче, и т. д. Такие задачи называются задачами таксономии. Результатом таксономии является древообразная иерархическая структура. При этом каждый объект характеризуется перечислением всех кластеров, которым он принадлежит, обычно от крупного к мелкому.

В построенной математической модели были рассмотрены 5 стран постсоветского пространства (Эстония, Российская Федерация, Беларусь, Узбекистан и Казахстан). Данные были взяты с сайта Всемирного банка. Был рассмотрен вклад этих стран в процессы цифровой экономики, процент импорта и экспорта хайтек товаров этих стран. Были рассмотрены три временных промежутка 1995-2005, 2005-2015 и 2015-2019. Также при кластеризации этих стран по показателям участия цифровой экономики были выделены как показатели три года-1998, 2008, 2018.

При построении математической модели в программе Python использовались два метода для анализа вклада этих пяти стран в цифровую экономику:

1) анализ временных рядов.
2) кластерный анализ

Список использованных источников:

1. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности.//М.: Финансы и статистика, 1989-С.607

2. Беляцкая, Т. Н. Электронная экономика: теория, методология, системный анализ / Т. Н. Беляцкая. – Минск : Право и экономика, 2017. – 284 с.

3. Беляцкая, Т. Н. Методики сравнительного анализа систем электронной экономики / Т. Н. Беляцкая // Междунар. науч.-исслед. журн. – 2017. – № 10-2. – С. 74–83.

4. Беляцкая, Т. Н. Методологические основы управления электронной экономической системой / Т. Н. Беляцкая // Азимут науч. исслед.: экономика и упр. – 2018. – № 2. – С. 52–55.

Steinhaus HSur la division des corps materiels en parties.// Bull. Acad. Polon. Sc. 1956. i.,- C1. III vol IV: 801-804.

5. Lloyd SLeast square quantization in PCM's. //Bell Telephone Laboratories Paper. 1957.

6. MacQueen J. Some methods for classification and analysis of multivariate observations. //In Proc. 5th Berkeley Symp. on Math. Statistics and Probability, 1967 –P. 281—297.