достигнутые договоренности и находить дополнительные возможности для сбыта.

Библиографический список

1. Указ Президента РФ от 28.02.2022 N 79 "О применении специальных экономических мер в связи с недружественными действиями Соединенных Штатов Америки и примкнувших к ним иностранных государств и международных организаций"
2. Инструкция Банка России от 30.03.2004 N 111-И "Об обязательной продаже части валютной выручки на внутреннем валютном рынке Российской Федерации"
3. Ефимова Л.А., Каменева А.М. Бухгалтерский финансовый учет (бухгалтерский расчет денежных средств и расчетов): Учебное пособие / Л.А. Ефимова, А.М. Каменева. М.: Изд-во РГАУ-МСХА, 2016. – 142 с.

УДК 311, 004.432

КЛАСТЕРНЫЙ АНАЛИЗ СРЕДСТВАМИ ЯЗЫКА PYTHON

***Быков Денис Витальевич**, ассистент кафедры статистики и кибернетики ФГБОУ ВО РГАУ-МСХА имени К.А. Тимирязева, bykovdv@rgau-msha.ru*
*Научный руководитель: **Уколова Анна Владимировна**, канд. экон. наук, доцент кафедры статистики и кибернетики ФГБОУ ВО РГАУ-МСХА имени К.А. Тимирязева, statmsha@rgau-msha.ru*

***Аннотация**: в статье описываются возможности реализации кластерного анализа (КА) данных, в том числе на основе искусственных нейронных сетей, с помощью специализированных библиотек языка программирования Python.*

***Ключевые слова:** кластерный анализ, нейронные сети, Python.*

Одним из распространенных методов анализа данных является кластерный анализ (КА), заключающийся в нахождении кластеров как групп близкородственных объектов, с последующей их классификацией по найденным кластерам. В отличии от обычной классификации, при решении задачи кластеризации описание классов заранее неизвестно [1].

КА позволяет упростить обработку данных в результате разбиения набора объектов на группы схожих объектов, сократить объем хранимых данных путем удаления из групп однотипных объектов значительной части типичных представителей, выделить нетипичные объекты, построить иерархию множества объектов [2].

Выделяют две основные группы методов (алгоритмов) КА: четкие, нечеткие. Четкие методы КА, в свою очередь, делятся на иерархические (алгоритм на основе построения иерархического дерева кластеров) и неиерархические (k-menas). Отдельно выделяются методы КА, основанные на нейронных сетях (самоорганизующиеся карты Кохонена) [6].

С точки зрения математической статистики, КА является методом многомерной классификации, то есть частью многомерного анализа [5, с. 165]. Все методы многомерного анализа базируются на алгоритмах приведения различных по содержанию и единицам измерения признаков в сопоставимый вид.

Можно выделить следующие этапы КА:

1) отбор признаков для анализа;
2) формирование матрицы исходных данных *X*, в которой число столбцов равно числу признаков (*m*), а число строк – числу объектов (единиц наблюдения) (*n*);
3) нормирование исходных данных по формуле (формирование матрицы нормированных отклонений *T*):

$$t_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_{x_j}}, \quad (1)$$

где t_{ij} – нормированное значение признака *j* для объекта *i*,

x_{ij} – исходное значение признака *j* для объекта *i*,

$\bar{x_j}$ – среднее значение признака *j*,

σ_{x_j} – стандартное отклонение признака *j*.

4) 4) выбор функции (меры) близости;
5) выбор метода объединения.

Этапы 1-3 являются общими для многомерного анализа, тогда как этапы 4 и 5 отражают специфику КА.

Можно выделить следующие функции близости нормированных отклонений: Евклидово расстояние, квадрат Евклидова расстояния, расстояние «городских кварталов («манхеттенское» расстояние), расстояние Чебышева и др.

Евклидово расстояние определяется по формуле:

$$\alpha_{kq} = \sqrt{\sum_{j=1}^{m} (t_{kj} - t_{qj})^2}, \quad (2)$$

где α_{kq} – расстояние между объектами *k* и *q*,

t_{kj} – нормированное отклонение признака *j* для объекта *k*,

t_{qj} – нормированное отклонение признака *j* для объекта *q*.

К методам объединения относят: метод одиночной связи (метод ближайшего соседа), метод полной связи (метод наиболее удаленных соседей), невзвешенное попарное среднее, взвешенное среднее расстояние, невзвешенный центроидный метод, взвешенный центроидный метод, метод медианной кластеризации, метод Варда.

Различают две схемы проведения КА: «без обучения», или иерархическая, и «с обучением», основанная на методе k-средних (итерационная).

Иерархический КА реализуется через n – 1 итераций, где n – число объектов (единиц) в исходной совокупности. На первой итерации строится матрица функций близости каждого объекта с каждым. Объекты с

минимальным значением функции близости объединяются в первый кластер. На последующих итерациях последовательно происходит объединение на основе выбранного метода объединения. Критерием объединения является минимальное значение функции близости на этой итерации.

Кластеризация с обучением (метод k-средних) предполагает, что число кластеров заранее известно. По этим кластерам устанавливаются центры тяжести, т.е. среднее значение нормированных отклонений по каждому из признаков. На основе выбранных функции близости и метода объединения производят распределение всех единиц совокупности по намеченным кластерам. Объект относится к тому кластеру, с которым он имеет минимальное значение функции близости. Основной проблемой данной схемы КА является выбор центров тяжести. Для решения этой задачи может быть использовано несколько подходов.

1. Первый подход к выбору центров тяжести. «Максимизация расстояния между кластерами»:

1) случайный отбор *k* объектов, где *k* – число будущих кластеров;
2) формирование изначальных центров тяжести как нормированных отклонений по признакам выбранных объектов;
3) расчет функции близости каждого невыбранного объекта с центрами тяжести;
4) корректировка центров тяжести – объект становится новым центром тяжести, если одновременно выполняется два условия:
    - расстояние между текущим объектом и текущим центром тяжести минимально по сравнению с расстоянием между текущим объектом и всеми другими центрами тяжести (объект расположен ближе к текущему центру тяжести, чем к остальным центрам тяжести);
    - расстояние между текущим объектом и текущим центром тяжести больше, чем наименьшее из расстояний между намеченными ранее центрами тяжести (объект позволит увеличить расстояние между наиболее близкими центрами тяжести).

Рассмотренный подход позволяет максимизировать расстояние между намечаемыми центрами тяжести, однако он может привести к образованию кластера, состоящего из одного наблюдения.

2. Второй подход к выбору центров тяжести (на основе «Матрицы ранжированных расстояний»).

1) расчет матрицы функции близости между объектами (матрицы расстояний);
2) ранжирование значений функции близости;
3) отбор ранжированных значений функции близости в качестве центров тяжести через определенный интервал, величина которого зависит от максимального и минимального значения функции близости, а также от числа кластеров *k*.

3. Третий подход предполагает наличие информации о предполагаемых значениях центров тяжести. Из всех объектов выбираются в качестве центров тяжести те, у которых нормированные отклонения наиболее близки к гипотетическим центрам тяжести.

Рассмотренные алгоритмы кластерного анализа свидетельствуют о том, что расчеты достаточно трудоемки, но они могут быть реализованы с использованием пакетов прикладных программ [5], в том числе с помощью языка программирования Python.

Кластеризация также связана с машинным обучением. Кластеризация – метод машинного обучения без учителя, когда метки данных неизвестны, либо они игнорируются [3]. Таким образом, в результате кластеризации формируются кластеры, после чего данные классифицируются по полученным кластерам, то есть исходным объектам присваиваются метки кластеров [4].

Для решения задачи кластерного анализа также применяются искусственные нейронные сети. Широкое применение нашли такие модели нейронных сетей, как, например, самоорганизующаяся карта Кохонена, сети адаптивного резонанса.

Самоорганизующаяся карта Кохонена (*англ.* KNC – Kohonen Clastering Network) используется для отображения нелинейных зависимостей на двумерные (чаще всего) сетки, представляющие метрические и топологические зависимости входных векторов, объединяемых в кластеры.

Нейронная сеть Кохонена имеет один слой нейронов. Количество входов каждого нейрона равно размерности входного вектора. Количество нейронов непосредственно определяет, сколько различных кластеров сеть может распознать.

Основная цель обучения в KNC состоит в выявлении структуры в n-мерных входных данных и предоставлении ее на карте в виде распределенных нейронных активностей. Каждый нейрон несет информацию о кластере, объединяющем в группу схожие по критерию близости входные вектора, формируя для данной группы собирательный образ.

Подобные вектора активизируют подобные нейроны, т. е. KNC способна к обобщению. Конкретному кластеру может соответствовать и несколько нейронов с близкими значениями векторов весов, поэтому выход из строя одного нейрона не так критичен к ошибке распознавания, как это имеет место в сети Хемминга.

В большинстве случаев каждый выходной нейрон связан со своими соседями. Эти внутрислойные связи играют важную роль в процессе обучения, так как корректировка весов происходит не для всех весов сети, а только в окрестности этого элемента.

Сеть Кохонена использует состязательный конкурентный алгоритм обучения. Выигрывает тот нейрон, чей вектор весов наиболее близок к текущему входному вектору. Близость определяется, например, евклидовой метрикой [7].

Сетями адаптивной резонансной теории (АРТ) (*англ.* ART – Adaptive Resonance Theory Network) называется семейство сетей на основе теории адаптивного резонанса, разработанное Гроссбергом, применительно к биологическим структурам и обладающее свойством «стабильности – пластичности».

Пластичность заключается в способности к восприятию новых образов, а стабильность – в способности к сохранению старых образов. Например, в многослойном персептроне, после предъявления нового входного вектора изменяются весовые коэффициенты, и нет гарантии, что старые образы не разрушатся. Аналогичная ситуация имеет место в сетях Кохонена, обучающихся на основе самоорганизации. Данные сети всегда выдают положительный результат при классификации и не способны отделить новые образы от искаженных образов [7].

Реализовать кластерный анализ можно с помощью языка программирования Python и специальных библиотек для данного языка.

Например, библиотеки scikit-learn, hdbscan позволят провести кластеризацию данных без использований нейронных сетей. Применить нейросетевой подход к КА можно при помощи библиотек PyTorch, Tensorflow, Keras, Theano.

Библиотека scikit-learn имеет специальный модуль для кластерного анализа (scikit-learn.cluster) и поддерживает такие методы, как: K-Means, Affinity propagation, Mean-shift, Spectral clustering, Ward hierarchical clustering, Agglomerative clustering, DBSCAN, OPTICS, Gaussian mixtures, BIRCH, Bisecting K-Means. Библиотека hdbscan представляет собой набор инструментов для машинного обучения без учителя, в том числе для поиска кластеров. Основным алгоритмом является HDBSCAN.

Кластеризация является эффективным методом анализа неразмеченных данных. Указанные в настоящей статье инструменты ее реализации позволяют разрабатывать компьютерные программы для существенного ускорения и облегчения обработки больших наборов данных, построения кластеров на их основе и последующей классификацией объектов по полученным кластерам. Результаты применения подобных программ помогут, например, выявить типичные группы объектов, и, как следствие, принимать наиболее оптимальные управленческие решения.

Библиографический список

1. Бураков, М.В. Нейронные сети и нейроконтроллеры: учеб. пособие / М. В. Бураков. – СПб.: ГУАП, 2013. – 284 с.

2. Воронцов К.В. Методы кластеризации. Машинное обучение (курс лекций), 2013. – 36 с. – Текст : электронный // MachineLearning.ru : информационно-аналитический ресурс. – URL: http://www.machinelearning.ru/wiki/images/archive/2/28/20150427184336%21Voron-ML-Clustering-slides.pdf.

3. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. – 2-е изд., перераб.и доп. – СПб.: БХВ-Петербург, 2021. - 416 с.

4. Демидова, Л. А. Кластерный анализ. Python : учебное пособие / Л. А. Демидова. – Москва : РТУ МИРЭА, 2022. – 103 с. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/240092. – Режим доступа: для авториз. пользователей.

5. Зинченко А.П. Математическая статистика : учебник / А. П. Зинченко, М. В. Кагирова, Ю. Н. Романцева [и др.]. – Москва : РГАУ-МСХА имени К. А. Тимирязева, 2018. – 199 с.

6. Пастухов А.А. Применение алгоритмов кластеризации к формированию представительской выборки для обучения многослойного персептрона / А.А. Пастухов, А.А. Прокофьев // Научно-технические ведомости СПБГПУ. Физико-математические науки. Т. 10. № 2. – 2017. – С. 58–68.

7. Ростовцев, В. С. Искусственные нейронные сети : учебник для вузов / В. С. Ростовцев. – 2-е изд., стер. – Санкт-Петербург : Лань, 2021. – 216 с. – ISBN 978-5-8114-7462-2. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/160142. – Режим доступа: для авториз. пользователей.

УДК 31:33

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ТИПИЗАЦИИ ФЕРМ В ЕВРОПЕ

Ульянкин Александр Евгеньевич, *ассистент кафедры статистики и кибернетики ФГБОУ ВО РГАУ-МСХА имени К.А. Тимирязева, aeulianckin@rgau-msha.ru*
Научный руководитель: ***Уколова Анна Владимировна****, канд. экон. наук, доцент кафедры статистики и кибернетики ФГБОУ ВО РГАУ-МСХА имени К.А. Тимирязева, statmsha@rgau-msha.ru*

Аннотация. *Целью данного исследования является изучение зарубежного опыта выделения типов ферм для разработки типологии сельскохозяйственных организаций в Российской Федерации*

Ключевые слова: *сельскохозяйственная перепись, типология, фермы.*

В соответствии с государственными программами и проектами Министерства сельского хозяйства Российской Федерации сельское хозяйство является одной из приоритетных отраслей российской экономики.

Одной из важнейших задач для экономики страны является активное развитие агропромышленного комплекса, способного конкурировать на мировом уровне, что позволит обеспечить промышленность сырьем, а жителей страны – продуктами высокого качества по доступным ценам.

Отсутствующая в Российской Федерации объективная типология сельскохозяйственных организаций, которая существует в Европейском союзе и США, затрудняет разработку государственной политики адресной поддержки сельхозтоваропроизводителей. Согласно закону «О развитии малого и среднего предпринимательства в Российской Федерации» [3], все предприятия делятся