УДК 004.67

МЕТОДЫ И СРЕДСТВА КЛАСТЕРНОГО АНАЛИЗА ПРИ РАБОТЕ С БАЗАМИ ДАННЫХ

А. А. Мельник, Т. А. Гробер
Технологический институт (филиал ДГТУ) в г. Азове (г. Азов, Российская Федерация)

Преподаватели, желающие повысить свой уровень квалификации, могут воспользоваться любой образовательной платформой, расположенной в сети Интернет. Использование платформы дистанционного обучения позволяет полностью автоматизировать процесс обучения. Целью работы являлось изучение клиентской базы данных образовательной платформы компании «InformTeacher», выявление наиболее перспективных и востребованных курсов дистанционного обучения. Для этого применялись различные методы кластерного анализа данных: иерархические и неиерархические. Применены различные методы кластерного анализа в сочетании с информационными технологиями. Определены конкретные курсы образовательной платформы, которые следует активно продвигать на рынке для получения наибольшей прибыли.

Ключевые слова: образовательная платформа, кластеризация, кластер, неиерархический метод, иерархический метод.

METHODS AND TOOLS OF CLUSTER ANALYSIS WHEN WORKING WITH DATABASES

A. A. Melnik, T. A. Grober
Technological Institute (branch of DSTU) in the city of Azov (Azov, Russian Federation)

Teachers who want to improve their level of qualification can use any educational platform located on the Internet. Using the distance learning platform allows you to fully automate the learning process. The purpose of the work is to study the client database of the educational platform of the company "InformTeacher", to identify the most promising and popular distance learning courses. Various methods of cluster analysis in combination with information technologies are applied. Specific courses of the educational platform have been identified, which should be actively promoted on the market in order to obtain the greatest profit.

Keywords: educational platform, clustering, cluster, non-hierarchical method, hierarchical method.

Введение. Согласно требованиям министерства науки и образования, преподаватели различной квалификации обязаны регулярно повышать уровень профессионального мастерства. Использование платформы дистанционного обучения позволяет автоматизировать процесс обучения, начиная с оформления документов, заканчивая итоговым тестированием и получением диплома.

На данный момент существует определённая конкуренция среди образовательных платформ. Так как в разработку интерактивных курсов компания вкладывает средства, то необходимо среди множества направлений обучения выбрать наиболее востребованные. Изучение данной проблемы проводилось на основе имеющейся базы данных компании «InformTeacher».

Целью работы являлось изучение клиентской базы данных компании «InformTeacher», выявление наиболее перспективных и востребованных курсов. Для реализации цели необходимо было решить следующие задачи:

– применить иерархические методы кластерного анализа для небольшой выборки из базы данных;

– применить неиерархические методы кластерного анализа для выборки большого объёма данных из базы данных.

Научная новизна работы заключается в применении различных методов кластерного анализа в сочетании с информационными технологиями для исследования использования платформы дистанционного обучения.

Актуальность работы связана с применением современных методов и средств анализа данных, таких как кластерный анализ, информационные технологии, в том числе, программирование в среде R.

Основная часть. Для анализа информации баз данных разработаны определённые методы и средства. Одним из них является кластерный анализ.

Основной целью кластерного анализа является поиск групп сходных объектов в выборке данных. Актуальность кластерного анализа обусловлена тем, что этот метод полезен и эффективен в случаях, когда необходимо классифицировать и представить в форме, пригодной для дальнейшего исследования большого объема информации. По сути, кластеризация эффективна во всех сферах человеческой деятельности [1].

Классификация существующих методов кластеризации приведена на рис. 1. Выбор метода во многом определяется ожидаемым характером классификации данных и конечным результатом.

Рис. 1. Методы кластерного анализа

Иерархические методы кластеризации

Иерархические методы применяются в таких средствах, как MS Excel — компьютерной программе для статистической обработки данных SPSS [2].

Иерархические методы подразделяются на агломеративный и дивизионный (дивизимный). Иерархический агломеративный метод (Agglomerative Nesting, AGNES) характеризуется постепенной группировкой данных с соответствующим сокращением количества комбинаций. Алгоритм начинает свою работу с того, что объекты разбиты на кластеры со схожими характеристиками. На первом этапе наиболее схожие объекты будут объединяться в класс. Слияние классов продолжается до тех пор, пока объекты не образуют единую группу.

Иерархический дивизимный (делимый) метод (Divisive Analysis, DIANA) работает по принципу полной противоположности агломеративному методу.

Пример реализации иерархического метода кластерного анализа в MS Excel

В данном примере был проведён кластерный анализ 6-ти курсов по психологии по их востребованности среди двух возрастных категорий (первая группа — это люди от 23 до 40 лет

(столбец x), вторая — от 41 до 65 лет (столбец y на рис. 2) с помощью метода «ближайшего соседа» агломеративной иерархической кластеризации.

№ п/п	x	y
1	2	9
2	5	11
3	6	7
4	12	6
5	13	6
6	14	5

Рис. 2. Исходные данные для анализа

Каждый объект (курс) имеет два параметра x и y. Для вычисления расстояния между объектами в матрице применяют евклидову метрику, вычисляемую по формуле (1).

$$D_{ab} = \sqrt{(x_а - x_в)^2 + (y_а - y_в)^2} \quad (1)$$

По этой формуле были вычислены данные и занесены в таблицу, как показано на рис. 3.

№ п/п	1	2	3	4	5	6
1	0	3,61	4,47	10,44	11,40	12,65
2	3,61	0	4,12	8,60	9,43	10,82
3	4,47	4,12	0	6,08	7,07	8,25
4	10,44	8,60	6,08	0	1,00	2,24
5	11,40	9,43	7,07	1,00	0	1,41
6	12,65	10,82	8,25	2,24	1,41	0

Рис. 3. Полученные данные

На рис. 3 видно, что наиболее близко по значениям расположены друг к другу объекты 4 и 5. Соединяем их, оставляя наименьшие значения в качестве основных, как показано на рис. 4, и переходим к следующему шагу.

ШАГ 1					
№ п/п	1	2	3	4,5	6
1	0	3,61	4,47	10,44	12,65
2	3,61	0	4,12	8,60	10,82
3	4,47	4,12	0	6,08	8,25
4,5	10,44	8,60	6,08	0	2,24
6	12,65	10,82	8,25	2,24	0

Рис. 4. Шаг первый

На втором шаге объединяем уже имеющуюся группу объектов 4, 5 с объектом 6, оставляя значения кластера 4, 5. Пример приведен на рис. 5.

ШАГ 2				
№ п/п	1	2	3	4,5,6
1	0	3,61	4,47	10,44
2	3,61	0	4,12	8,60
3	4,47	4,12	0	6,08
4,5,6	10,44	8,60	6,08	0

Рис. 5. Шаг второй

На третьем шаге объекты 1 и 2 группируются в один кластер, так как их значения минимальны по отношению к другим объектам (рис. 6).

ШАГ 3			
№ п/п	1,2	3	4,5,6
1,2	0	4,12	8,60
3	4,12	0	6,08
4,5,6	8,60	6,08	0

Рис. 6. Шаг третий

Далее, как наглядно представлено на рис. 6, можно объединить кластер 1, 2 с объектом 3. Результат представлен на рис. 7.

ШАГ 4		
№ п/п	1,2,3	4,5,6
1,2,3	0	6,08
4,5,6	6,08	0

Рис. 7. Шаг четвертый

В итоге были получены два кластера. Кластерный анализ показал, что обучающиеся первой возрастной категории преимущественно отдают предпочтение курсам под номерами 1, 2, 3, тогда, как второй категории — курсам под номерами 4, 5, 6.

Неиерархические методы кластеризации

Если количество наблюдений очень велико, то иерархические методы будет неэффективны. Более подходящими являются неиерархические методы, основанные на разделении набора данных на n отдельных групп.

Неиерархические методы применяются в таких средах, как R и Python, а также в программном пакете для статистического анализа Statistica [3].

В основе таких методов находится выбор центра группы и объединение объектов, находящихся в пороговом значении. Далее выбирается новый центр группы, объединяем не сгруппированные объекты в кластеры. Алгоритм продолжает свою работу до тех пор, пока все объекты не будут разбиты на группы [4, 5].

Пример реализации неиерархических методов кластерного анализа в R

Для определения наиболее перспективных курсов использовался кластерный анализ данных средствами программной среды R. В начале создаётся программный код для определения оптимального количества кластеров по критерию «Каменистой осыпи», а затем продолжили кластерный анализ «Методом k-средних». Выборка содержит 228 значений.

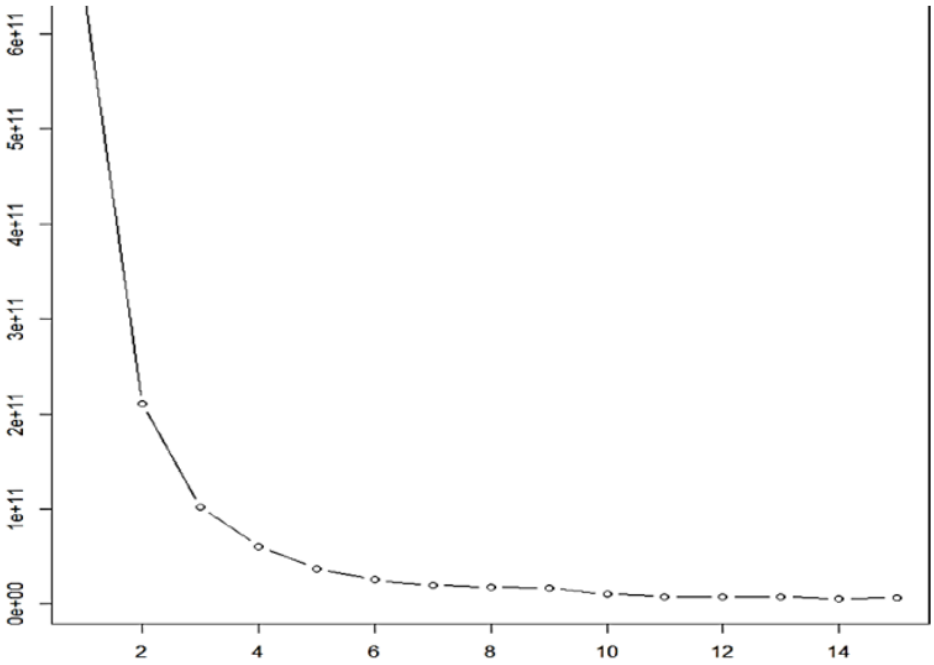

Рис. 8. Применение к выборке критерия «Каменистой осыпи»

По горизонтали откладываются номера собственных значений, по вертикали — собственные значения.

На рис. 8 видно, что точка графика, в которой убывание собственных значений слева направо максимально замедляется, соответствует значению аргумента 3, следовательно, лучше использовать для дальнейшего исследования 3 кластер.

Самым распространённым методом неиерархической кластеризации является «Метод k-средних», целью которого является разделение объектов на кластеры, при этом каждое наблюдение относится к тому классу, к центру которого минимальное расстояние. С помощью программного кода были преобразованы данные из исходной базы данных компании, которые представлены в таблице 1.

Таблица 1

Средние значения переменных в каждом кластере

Кластеры	Возраст	Сумма оплаченных курсов
1	36,42308	1411,282
2	43,56140	1855,789
3	30,89796	2129,592

Наиболее частым графическим представлением является паутинообразная диаграмма, как представлено на рис. 9. Она демонстрирует разброс значений каждого кластера относительно среднего значения.

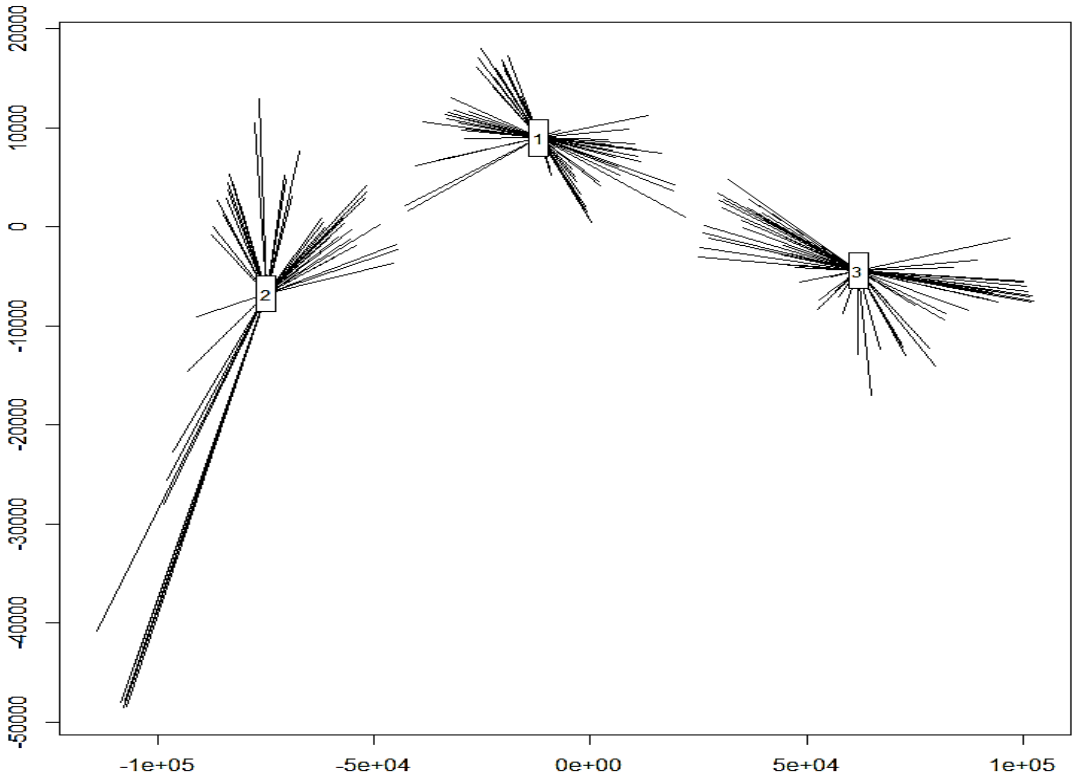

Рис. 9. Паутинообразная диаграмма

На рис. 9 по горизонтали отложены значения, участвующих в классификации переменных, по вертикали — средние значения переменных для каждого кластера.

Для кластеров 1 и 3 характерна большая близость к центру, что говорит о большей однородности этих групп.

Первый кластер представлен респондентами среднего возраста, кластер наиболее многочисленный. В него входят люди с разными интересами и потребностями, согласно возрасту.

Второй кластер представлен респондентами старше 40 лет, которые хотят обновить свои знания, осваивают инновационные методы обучения, также они изучают английский язык, который может требоваться как в связи с обновлением знаний по специальности, так и для освоения англоязычной литературы для работы.

Третий кластер представлен молодыми людьми, которые готовы больше других платить за курсы, им интересны основы первой помощи, педагогика дошкольного образования и инновационные методы обучения, представляющие собой базовые курсы.

Следует обратить внимание, что процент объясненной дисперсии изначального набора данных составляет 74 %, что свидетельствует о хорошем качестве классификации.

Кластерный анализ данных показал, что наиболее востребованными являются вариации следующих курсов: «Основы первой доврачебной помощи», «Внедрение бережливых технологий в деятельность образовательных учреждений в соответствии с ФГОС», «Инновационные методы и технологии обучения дисциплине в условиях реализации ФГОС», а также курсы переподготовки: «Педагогика и методика дошкольного образования» и «Педагогическое образование: английский язык».

Выводы

Кластеризация имеет большую значимость в анализе довольно больших баз данных благодаря превращению большого объема информации в структурированный и сжатый вид.

В результате работы были определены конкретные курсы образовательной платформы, которые следует активно продвигать на рынке для получения наибольшей прибыли.

В проведенном исследовании клиентской базы данных было получено пересечение между первым и вторым, первым и третьим кластерами за счет такой дисциплины как «Инновационные методы и технологии обучения в условиях реализации ФГОС». Поскольку на данный момент особое значение приобрели инновационные методы обучения, то все категории респондентов были заинтересованы данным курсом. Соответственно, можно предположить, что на страницу данного курса следует ожидать наибольший поток потребителей.

Применение современных методов кластерного анализа в сочетании с информационными технологиями позволяет улучшить конкретность и визуализацию результатов наблюдений для дальнейшего прогнозирования развития бизнеса.

Библиографический список

1. Дюран, Б. Кластерный анализ / Б. Дюран. — Москва : Книга по Требованию, 2012. — 128 с.

2. Наследов, А. Д. SPSS 15: профессиональный статистический анализ данных / А. Д. Наследов. — Санкт-Петербург : Питер, 2008. — 416 с.

3. Вуколов, Э. А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL / Э. А. Вуколов. — Москва : Форум, 2011. — 464 с.

4. Петрунин, Ю. Ю. Информационные технологии анализа данных: Dataanalysis: учебное пособие / Ю. Ю. Петрунин. — Москва : КДУ, 2008. — 292 с.

5. Просветов, Г. И. Управленческие решения: задачи и решения: учебно-практическое пособие / Г. И. Просветов. — Москва : Альфа-Пресс, 2009. — 319 с.

Об авторах:

Гробер Татьяна Александровна, и. о. зав. кафедрой «Вычислительная техника и программирование» Технологического института (филиал ДГТУ) в г. Азове (346780, РФ, г. Азов, ул. Промышленная, 1), кандидат физико-математических наук, groberta2020@mail.ru

Мельник Алина Александровна, бакалавр кафедры «Вычислительная техника и программирование» Технологического института (филиал ДГТУ) в г. Азове (346780, РФ, г. Азов, ул. Промышленная, 1) m.alina15@yandex.ru

About the Authors:

Grober, Tatyana A., Acting head, Department of Computer Engineering and Programming, Technological Institute (branch of DSTU) in Azov (1, Promyshlennaya str., Azov, 346780, RF), Cand.Sci. (Phys.-Math.), groberta2020@mail.ru

Melnik, Alina A., Bachelor’s degree student, Department of Computer Engineering and Programming, Technological Institute (branch of DSTU) in Azov (1, Promyshlennaya str., Azov, 346780, RF), m.alina15@yandex.ru