УДК 519.237.8, 004.42

О.К. Альсова

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ РАЗНОТИПНЫХ ДАННЫХ НА ПРИМЕРЕ РЕШЕНИЯ МЕДИЦИНСКОЙ ЗАДАЧИ

Альсова О.К. **Алгоритмы кластеризации разнотипных данных на примере решения медицинской задачи.**
Аннотация. Описан оригинальный алгоритм кластеризации разнотипных данных, основанный на комплексном применении набора мер расстояний и методов кластеризации и многоэтапной кластеризации. В рамках алгоритма реализовано ранжирование признаков объекта по их значимости для группировки и выбор оптимального набора признаков, ансамблевый подход для получения более устойчивого итогового кластерного решения. Алгоритм реализован в программной системе MixDC (Mixed Data Clustering). Приведены методика и результаты решения реальной задачи кластеризации медицинских данных средствами программной системы.
Ключевые слова: кластеризация, разнотипные данные, мера расстояния, алгоритм кластеризации, ансамблевый подход.

Alsova O.K. **Algorithms for Clustering of a Heterogeneous Data on the Example of Solution of the Medical Task.**
Abstract. The paper describes the original algorithm of a heterogeneous data clustering is based on complex application of a set of measures of distances and clustering methods and multi-stage clustering. In the algorithm we use ranging of attributes the object on their importance for group and a choice of an optimum attributes set, ensemble approach to get the final clustering solution. The algorithm is realized in MixDC (Mixed Data Clustering) software system. The technique and results of the solution of a real problem of a medical data clustering in software system are described.
Keywords: clustering, heterogeneous data, measure of distance, algorithm of clustering, ensemble approach.

1. Введение. Одной из основных задач интеллектуального анализа данных на настоящий момент является кластеризация – процесс получения «кластеров» или групп похожих объектов [1-4]. Вычислительные процедуры кластерного анализа данных применяются в самых разных предметных областях (медицина, социология, экономика) и позволяют обнаружить в данных неизвестные ранее закономерности.

Разработано множество методов и алгоритмов кластерного анализа данных. Однако, применение большей части методов требует, чтобы описывающие объект признаки (на основе которых выполняется кластеризация) были количественного типа. Значительно меньше разработано методов, которые позволяют работать с качественными признаками в процессе кластеризации. Наибольшую сложность при кластеризации объектов представляет обработка совокупности разнотипных признаков (количественных и качественных: порядковых или номинальных). Тогда как в реальных задачах, как правило, объект описывается совокупностью разнотипных признаков, причем при класте-

ризации необходимо учесть все признаки. Например, такая ситуация характерна для обработки медицинских данных.

Еще одна ключевая проблема кластеризации заключается в выборе и обосновании набора признаков, которые используются при группировке, оценка их статистической значимости и ранжирование.

И наконец, проблема устойчивости группировочных решений [2]. Результаты группировки могут сильно меняться в зависимости от выбора алгоритма и его параметров, меры сходства и т.п. Один из способов повышения устойчивости – разработка и применение ансамблевых алгоритмов. При этом используются результаты группировки, полученные различными алгоритмами или одним алгоритмом, но с разными параметрами настройки, по различным подсистемам признаков и т.д. После построения ансамбля проводится нахождение итогового коллективного кластерного решения.

В статье предложен оригинальный алгоритм кластеризации разнотипных данных, основанный на комплексном применении набора мер расстояний, методов кластеризации и многоэтапной кластеризации, направленный на решение вышеуказанных проблем. Разработанный алгоритм реализован в программной системе MixDC [5] и применен для анализа реальных медицинских данных. Средствами системы анализировались данные о хирургическом лечении больных с патологией восходящего отдела и дуги аорты (данные предоставлены ФГБУ «ННИИПК им. акад. Е.Н. Мешалкина» Минздрава России).

В работе [6] приведено описание первого варианта программной системы, без реализации многоэтапной кластеризации и ансамблевого подхода.

2. Общий алгоритм кластеризации. Алгоритм кластеризации разнотипных данных, реализованный в программной системе MixDC, обобщенно можно представить в виде последовательности выполнения следующих этапов.

Этап 1. «Постановочный». На этапе выполняется выбор исходных данных (объектов кластеризации и описывающих их признаков), возможен учет априорного веса признака, который задается пользователем. Также на первом этапе выбираются меры сходства, методы (алгоритмы) кластеризации и выполняется настройка параметров алгоритмов.

Этап 2. «Кластеризация». Проводится кластеризация объектов, на основе выбранных на предыдущем этапе мер сходства и методов кластеризации. В результате формируется кластерное решение (разбиение объектов на группы) для каждого из примененных методов.

Этап 3. «Определение набора значимых признаков». Выбираются признаки из исходной совокупности, определенной на этапе 1, которые вносят статистически значимый вклад в кластерное решение. В результате

формируется апостериорный набор признаков для каждого метода, которые следует учитывать при повторной кластеризации. Далее происходит ранжирование признаков по значимости. Итоговый набор признаков формируется при участии пользователя, который может задать при каком минимальном ранге признак будет включен в итоговый набор.

Этап 4. «Повторная кластеризация». После того как определен набор значимых признаков происходит повторная кластеризация, при которой учитываются лишь отобранные на этапе 3 признаки.

Этап 5. «Применение ансамбля алгоритмов». На последнем этапе выполняется нахождение итогового решения с применением ансамблевого подхода. Разбиения, полученные при помощи каждого из выбранных пользователем методов (этап 4), интегрируются в обобщенную матрицу сходства. Затем сформированная матрица используется в качестве исходных данных для иерархического метода, при помощи которого находится итоговое решение.

Далее подробно описываются методы (алгоритмы) и подходы, используемые на каждом из этапов.

3. Меры сходства и методы кластеризации. Кластеризация выполняется с использованием двух мер сходства: модифицированные мера согласования разнотипных шкал [4] и мера Гауэра [3]. Меры предназначены для вычисления сходства между объектами на основе разнотипных признаков.

Модифицированная мера, согласующая разнотипные шкалы, по сути, является мерой расстояния. По сравнению с классическим вариантом, в меру введен учет априорного веса признака. Вес признака вводится пользователем системы, если у него есть соответствующая информация, при отсутствии таковой, все признаки учитываются с равным весом. Мера вычисляется следующим образом для *i*–ого и *j*–ого объектов.

Для количественных признаков, описывающих объект, расстояние вычисляется по формуле:

$$d_{num}(i,\ j) = \frac{|x_i - x_j|}{x_{max} - x_{min}}. \quad (1)$$

Для номинальных признаков:

$$d_{nom}(i,j) = \frac{1}{m}\sum_{k=1}^{m} d(i,j)_s, \quad (2)$$

где m – число объектов в выборке; $d(i,\ j)_s$ – различия в отношениях *i*–го и *j*–го объектов к некоторому *s*–ому объекту.

Для порядковых признаков:

$$d_{ord}(i,j) = \frac{1}{m-1}\sum_{k=1}^{m} d(i,j)_s. \quad (3)$$

Суммарное расстояние между *i*–ым и *j*–ым объектом вычисляется по формуле:

$$d_{res}(i,j) = \sqrt{(v_1 d_{num}(i,j))^2 + \left(v_2 d_{nom}(i,j)\right)^2 + (v_3 d_{ord}(i,j))^2}, \qquad (4)$$

где v_1, v_2, v_3 – веса признаков.

Формула (4) приведена для случая, когда каждый объект описывается одним количественным, одним порядковым и одним номинальным признаком и может быть расширена для случая *p* признаков разных типов.

В мере Гауэра сходство между *i*–ым и *j*–ым объектом вычисляется по формуле:

$$s_{ij} = \sum_{k=1}^{p} s_{ijk} \Big/ \sum_{k=1}^{p} v_{ijk}, \qquad (5)$$

где s_{ijk} – сходство объектов по k–му признаку; v_{ijk} – весовая переменная; p – число учитываемых признаков. Вес задается так, чтобы $\sum_{k=1}^{p} v_{ijk} = 1$. Вклад для качественных (номинальных и порядковых) признаков вычисляется следующим образом:

$$s_{ijk} = 1, \text{ если } X_{ik} = X_{jk}, \text{ иначе } s_{ijk} = 0. \qquad (6)$$

Для количественных признаков:

$$s_{ijk} = 1 - \frac{|X_{ik} - X_{jk}|}{R_k}, \qquad (7)$$

где R_k – размах значений переменной.

Кластеризация выполняется при помощи либо порогового метода [1], либо оригинального алгоритма кластеризации. Последний алгоритм итеративный, в качестве априорной информации для работы алгоритма указывается число кластерных центров. В качестве центров выбираются объекты, находящиеся на максимальном расстоянии друг от друга.

Затем выполняется распределение объектов по кластерам. Расстояние до кластера вычисляется как среднее расстояние от включаемого объекта до каждого из объектов в кластере.

Пусть на первом шаге выбрано k объектов $O_1, \ldots, O_k$, которые объявлены соответственно кластерными центрами: $E_1, \ldots, E_k$.

На втором шаге выполняется отнесение объекта O_{k+1} к одному из кластеров: вычисляется расстояние между O_{k+1} и центрами кластеров $E_1, \ldots, E_k$ отдельно по каждому признаку на основе формул (1-3).

Объект O_{k+1} относится к кластеру, расстояние до которого (4) минимально.

На m-ом шаге вычисляется среднее расстояние для каждого i-го кластера между объектом O_{k+m} и объектами кластера отдельно по каждому признаку. Для количественных признаков:

$$\bar{d}_{num}(i) = \frac{1}{n} \sum_{j=1}^{n} d_{num}(O_{k+m}, j), \quad (8)$$

где n – число объектов в кластере.

Для номинальных признаков:

$$\bar{d}_{nom}(i) = \frac{1}{n} \sum_{j=1}^{n} d_{nom}(O_{k+m}, j). \quad (9)$$

Для порядковых признаков:

$$\bar{d}_{ord}(i) = \frac{1}{n} \sum_{j=1}^{n} d_{ord}(O_{k+m}, j). \quad (10)$$

Суммарное среднее расстояние между объектом O_{k+m} и i-ым кластером вычисляется по формуле:

$$\bar{d}_{res}(i) = \sqrt{\left(v_1 \bar{d}_{num}(i)\right)^2 + \left(v_2 \bar{d}_{nom}(i)\right)^2 + \left(v_3 \bar{d}_{ord}(i)\right)^2}. \quad (11)$$

Объект O_{k+m} относится к тому кластеру, до которого суммарное среднее расстояние (11) минимально.

Процесс продолжается, пока не будет выполнена кластеризация всех объектов.

4. Определение набора значимых признаков. Для каждого из полученных на втором этапе кластерных решений выполняется оценка статистической значимости признаков, используемых при кластеризации. Для признаков различных типов используются разные критерии оценки статистической значимости. Для количественных признаков в случае их нормальной распределенности – одномерный дисперсионный анализ [1], для порядковых и для количественных в случае отклонения гипотезы о нормальном распределении значений признака – критерий Крускалла-Уоллиса [7], для номинальных – критерий однородности χ^2 Пирсона [7].

После того, как статистическая значимость признаков оценена, вычисляется общая оценка значимости (по всем выбранным методам):

$$Sim_{res} = \frac{\sum_{i=1}^{L} Sim_i}{L}, \quad (12)$$

где L – число методов, $Sim_i = 1$, если показатель признан статистически значимым для кластерного решения метода i, и $Sim_i = 0$ в противном случае.

Далее признаки ранжируются по значениям Sim_{res}, признаки с наиболее высокими значениями Sim_{res}, превышающими заданный пользователем уровень, учитываются при повторной кластеризации.

5. Ансамблевый подход. Для получения более устойчивых кластерных решений после этапа повторной кластеризации (этап 4) применяется ансамблевый подход [2], основанный на комплексном использовании набора алгоритмов. Для получения итогового решения находится согласованная матрица подобия объектов.

Пусть используются L методов кластеризации, формирующих некоторые группировки. Далее для каждой *i*-й группировки рассчитывается бинарная матрица подобия $S_i = \{S_i(j, m)\}$ размерностью $N \times N$, где j,m=1,2,…,N, а i=1,2,…,L, N – число объектов в исходной выборке, L – число применяемых методов в ансамбле. Причем, $S_i(j, m) = 1$, если объекты O_j и O_m принадлежат одному кластеру, иначе $S_i(j, m) = 0$.

Затем формируется согласованная матрица подобия $S = \{S(j, m)\}$:

$$S(j, m) = \frac{1}{L}\sum_{i=1}^{L} S_i(j, m). \quad (13)$$

Величина $S(j, m)$ будет равна частоте классификации объектов в одну и ту же группу в наборе группировок G. Близкое к единице значение величины означает, что данные объекты имеют большой шанс попадания в одну и ту же группу. Близкое к нулю значение этой величины говорит о том, что шанс оказаться в одной группе у этих объектов незначителен.

Полученная матрица используется в качестве исходных данных для иерархического алгоритма, результат применения которого является итоговым согласованным решением. В программной системе MixDC реализовано использование следующих стандартных иерархических методов: метод средней связи (Average Linkage); метод полной связи (Complete Linkage); метод одиночной связи (Single Linkage); метод Варда (Ward Linkage); метод взвешенной средней связи (Weighted Average Linkage).

6. Кластеризация медицинских данных на основе разработанных алгоритмов и программных средств. Система MixDC применялась для анализа реальных медицинских данных о хирургическом лечении больных с патологией восходящего отдела и дуги аорты. База включает 18 признаков (факторов), оцененных у 124 пациентов, оперированных на восходящем отделе и дуге аорты с использованием разных хирургических технологий. В БД представлено 11 номинальных (см. таблицу 1, номера: 1-11), 1 порядковый (номер 12) и 6 количественных признаков (номера: 13-18).

Таблица 1. Оценка статистической значимости признаков

№	Признак (фактор)	Р-значение				Значимость
		B1	B2	B3	B4	
1	Пол	0,01	0,39	0,66	0,49	1
2	Этиология	0,06	0,00	0,73	0,51	1
3	Предыдущие операции	0,38	0,00	0,21	0,01	2
4	ИБС – ишемическая болезнь сердца	0,33	0,04	0,26	0,34	1
5	АГ – артериальная гипретензия	0,01	0,00	0,23	0,00	3
6	ХОБЛ – хроническая обструктивная болезнь легких	0,19	0,08	0,41	0,61	0
7	Патология почек	0,15	0,08	0,24	0,02	1
8	Тип реконструкции дуги	0,00	0,00	0,01	0,00	4
9	Операция Борста	0,00	0,00	0,07	0,00	3
10	Перфузия ГМ – головного мозга	0,00	0,00	0,01	0,00	4
11	Вмешательство на корне аорты	0,16	0,00	0,00	0,01	3
12	НМК – нарушения мозгового кровообращения	0,01	0,03	0,00	0,08	3
13	Возраст	0,34	0,78	0,21	0,04	1
14	рост	0,22	0,78	0,29	0,36	0
15	вес	0,47	0,03	0,04	0,25	2
16	ИК – искусственное кровообращение	0,01	0,02	0,03	0,12	3
17	ОА – окклюзия аорты	0,00	0,04	0,03	0,04	4
18	ЦА – циркуляторный арест	0,00	0,00	0,01	0,00	4

Одним из тяжелых осложнений после операций на проксимальной аорте являются нарушения мозгового кровообращения (НМК) различной степени выраженности. Для практической медицины представляется важным выявить факторы риска возникновения НМК в раннем послеоперационном периоде и связать степень выраженности НМК со значениями выделенных факторов риска. Анализ взаимосвязи между отдельно взятыми факторами и НМК не позволил выявить статистически значимых связей.

В ходе дальнейшего исследования была выдвинута гипотеза, согласно которой на НМК влияет одновременно комплекс (взаимосвязанная совокупность) факторов.

Применение кластерного анализа позволяет разбить случаи заболеваний на группы, выявить наиболее значимые признаки для группировки и описать образы каждой группы по степени тяжести НМК и применяемых хирургических технологий. Тем самым, проанализировав какие случаи заболеваний, с какими значениями признаков, попадают в один кластер, можно сделать выводы о связи степени выраженности НМК с комплексом выделенных факторов риска.

Проведенное исследование включало следующие этапы:

1. Кластеризация с учетом всех имеющихся в базе признаков: с использованием порогового алгоритма и меры Гауэра; с использованием оригинального алгоритма и меры согласования разнотипных шкал; с использованием порогового алгоритма и меры согласования разнотипных шкал, с использованием оригинального алгоритма и меры Гауэра.
2. Формирование набора значимых признаков на основе оценки их статистической значимости.
3. Повторная кластеризация с учетом выбранных признаков.
4. Получение итогового решения с использованием ансамблевого подхода.

На первом этапе для кластеризации использовались все признаки (см. таблицу 1) с заданием их априорного веса. Наименьший вес имели признаки пол, вес, рост, возраст (по 0.5), остальные признаки учитывались с равным весом (1).

В таблице 1 приведены результаты оценки статистической значимости признаков (p-значение) для каждой комбинации метода и меры сходства: В1 – вариант 1 (пороговый алгоритм, мера Гауэра); В2 – вариант 2 (оригинальный алгоритм, мера согласования разнотипных шкал); В3 – вариант 3 (пороговый алгоритм, мера согласования разнотипных шкал); В4 – вариант 4 (оригинальный алгоритм, мера Гауэра). Был принят уровень значимости 0,05: при $p<0,05$ признак статистически значимо влияет на кластерное решение. Итоговая значимость для каждого из признаков представлена в таблице 1, в столбце «значимость»: отмечено количество методов, при использовании которых признак был оценен как статистически значимый. При повторной кластеризации (этап 4) использовались признаки со значимостью 3 и 4 и таким образом было выделено 9 статистически значимых из 18 исходных признаков.

При повторной кластеризации также, как и на втором этапе, использовались 4 комбинации методов кластеризации и мер сходства.

На заключительном этапе в рамках ансамблевого подхода для получения итогового решения использовался иерархический метод Варда.

На рисунке 1 представлена дендрограмма метода Варда. На ней видно 4-х кластерное разбиение. Ниже, на рисунках 2-6, представлены круговые и лепестковые диаграммы значений признаков, учитываемых при кластеризации, для итогового 4-кластерного разбиения. Признаки на рис. 2-6 закодированы в порядке возрастания сложности операционного вмешательства и тяжести пациента. Например, НМК 0 – нет нарушений; НМК 1 – гипоксическая энцефалопатия; НМК 2 – транзиторные ишемические атаки; НМК 3 – инсульт.

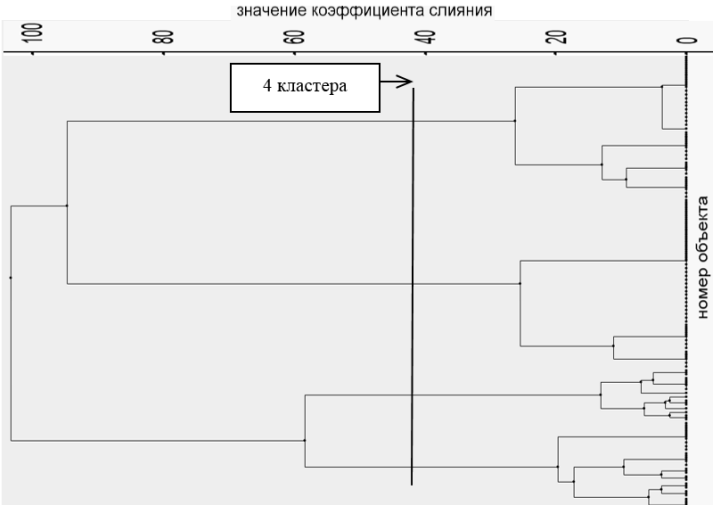

Рис. 1. Дендрограмма метода Варда для итогового решения

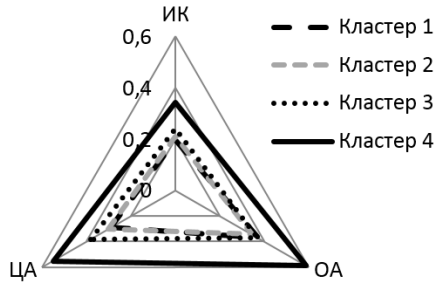

Рис. 2. Лепестковая диаграмма средних значений количественных признаков (ИК, ОА, ЦА) для итогового кластерного решения

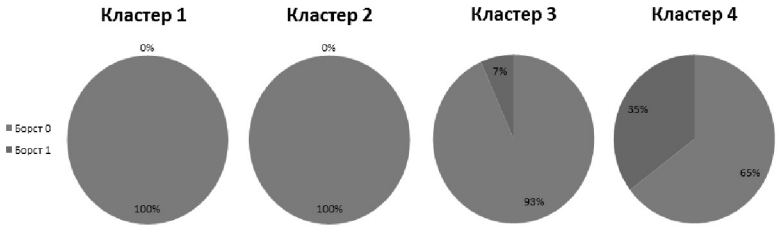

Рис. 3. Круговая диаграмма значений признака «Операция Борста» для итогового решения

Рис. 4. Круговая диаграмма значений признака «Перфузия ГМ» для итогового решения

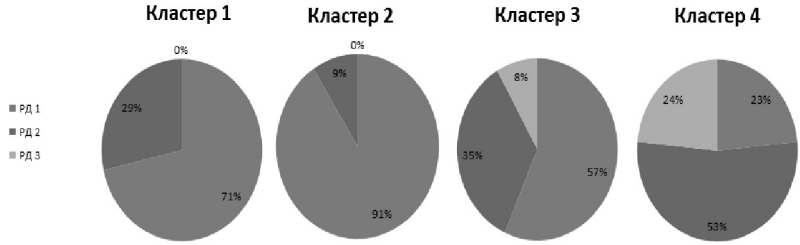

Рис. 5. Круговая диаграмма значений признака «Тип реконструкции дуги» для итогового решения

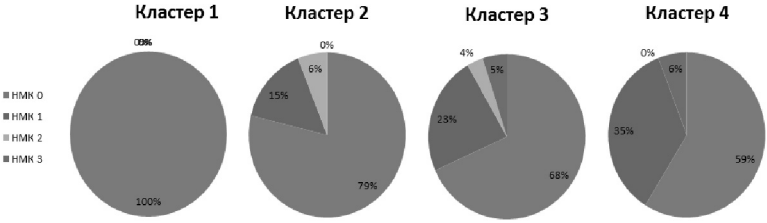

Рис. 6. Круговая диаграмма значений признака «НМК» для итогового решения

Полученные результаты имеют логическое объяснение и могут быть содержательно проинтерпретированы, согласуются с мнением врачей [8]. Так, тяжесть пациентов, наличие осложнений, увеличивается от первого к четвертому кластеру. Например, в первом кластере вообще не наблюдалось нарушений мозгового кровообращения (НМК 0). При анализе выявлено, что в первом кластере выполнялись наиболее простые типы реконструкции дуги и корня аорты, с наименьшими временными затратами. Тогда как в четвертом кластере, с наибольшим количеством инсультов (НМК 4), применялись наиболее агрессивные с точки зрения сложности и длительности выполнения хирургические технологии. Наличие сопутствующих заболеваний и предыдущих опе-

раций (АГ, операция Борста) усложняет операцию на аорте, а, следовательно, может увеличивать риск НМК. Тип реконструкции дуги и тип перфузии – это техники, применяемые при операционном вмешательстве, а значит, данные факторы могут оказывать влияние на НМК.

7. Заключение. Предложен и описан общий алгоритм кластеризации разнотипных данных, который состоит из нескольких взаимосвязанных этапов и основан на многократной кластеризации с использованием набора мер сходства и методов кластеризации. В рамках алгоритма применен автоматизированный выбор признаков объекта, учитываемых при кластеризации, основанный на анализе статистической значимости признаков и их ранжировании и использован ансамблевый подход для получения более устойчивого итогового решения.

Предложенный алгоритм реализован в программной системе MixDC. Средствами программной системы исследовались реальные медицинские данные. Результаты исследования показали перспективность и эффективность применения реализованных в программной системе методов и подходов для решения задач кластеризации разнотипных данных.

Разработанная система MixDC расширяема, планируется включить в ее состав дополнительные меры сходства и новые методы кластеризации. Другим перспективным направлением развития системы может быть ее интеллектуализация – оценка качества кластеризации и выдача рекомендаций пользователю на естественном языке относительно выбора алгоритма, меры сходства и т.п.

Литература

1. *Айвазян С.А.*, *Мхитарян В.С.* Прикладная статистика и основы эконометрики // М.: Юнити. 2001. 656 с.
2. *Бериков В.Б.* Современные тенденции в кластерном анализе // URL: http://www.ict.edu.ru/ft/005638/62315e1-st02.pdf/ (дата обращения: 14.05.2014).
3. *Ким Дж.-О.* Факторный, дискриминантный и кластерный анализ / пер. с англ. А.М. Хотинского, С.Б. Королева // М.: Финансы и статистика. 1989. 215 с.
4. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний // Новосибирск: Издательство института математики. 1999. 270 с.
5. *Ускова К.С., Альсова О.К.* Программная система кластерного анализа данных смешанного типа (MixDC-Mixed Data Clustering) // Свидетельство о государственной регистрации программы для ЭВМ № 2014618830. 2014.
6. *Альсова О.К., Ускова К.С.* Программная система кластерного анализа данных смешанного типа // Автоматика и программная инженерия. 2013. №1(3). С. 75–81.
7. *Кобзарь А.И.* Прикладная математическая статистика // М.: Физматлит. 2006. 816 с.
8. *Чернявский А.М., Альсов С.А., Ляшенко М.М. и др.* Анализ неврологических осложнений после хирургической реконструкции дуги аорты у пациентов с проксимальным расслоением // Патология кровообращения и кардиохирургия. 2013. №2. С. 35–39.

References

1. Ajvazjan S.A., Mhitarjan V.S. *Prikladnaja statistika i osnovy jeconometriki* [Applied statistics and econometrics]. M.: Unity. 2001. 656 p. (In Russ).
2. Berikov V.B. Sovremennye tendencii v klasternom analize [Current trends in cluster analysis]. 2008. Available at: www.ict.edu.ru (accessed 05.14.2014). (In Russ).
3. Kim J.-O. *Faktornyi, diskriminantnyi i klasternyi analiz* [Factor, discriminant and cluster analysis]. M.: Finansy i statistika. 1989. 215 p. (In Russ.)
4. Zagorujko N.G. *Prikladnye metody analiza dannyh i znanij* [Applied methods of data analysis and knowledge]. Novosibirsk: Izdatel'stvo instituta matematiki. 1999. 270 p. (In Russ).
5. Uskova K.S., Alsova O.K. Programmnaja sistema klasternogo analiza dannyh smeshannogo tipa (MixDC-Mixed Data Clustering) [Software system of cluster mixed type data analysis (MixDC-Mixed Data Clustering)]. Patent RF. №2014618830. 2014. (In Russ).
6. Alsova O.K., Uskova K.S. [Software system of cluster mixed type data analysis]. *Avtomatika i programmnaja inzhenerija – Automation and software engineering*. 2013. vol. 1(3). pp. 75–81. (In Russ).
7. Kobzar' A.I. *Prikladnaya matematicheskjy statistica* [Applied mathematical statistics]. M.: Fizmatlit. 2006. 816 p.
8. Chernyvsky A.M., Alsov S.A., Lyashenko M.M. et al. [The analysis of neurological complications after aortic arch reconstruction in patients with proximal aortic dissection]. *Patologiy krovoobrasheniy i kardiohirurgiya – Pathology of circulation and cardiac surgery*. 2013. vol. 2. pp. 35–39. (In Russ).

Альсова Ольга Константиновна — к-т техн. наук, доцент, ФГБОУ ВПО Новосибирский государственный технический университет (НГТУ). Область научных интересов: исследование и разработка методов и средств прогнозирования временных рядов, компьютерное моделирование систем, интеллектуальный анализ данных. Число научных публикаций — 30. alsowa@mail.ru; 630073, г. Новосибирск, пр. К. Маркса, 20; р.т. +7(383)346-04-92.

Alsova Olga Konstantinovna — Ph.D., associate professor, Computer Sciences Department, Novosibirsk State Technical University (NSTU). Scientific interests: research and development of methods and means of time series forecasting, computer modeling of systems, intellectual data analysis. The number of publications — 30. alsowa@mail.ru; 20, Prospekt K. Marksa, Novosibirsk, 630073, Russia; office phone +7(383)346-04-92.

РЕФЕРАТ

Альсова О.К. **Алгоритмы кластеризации разнотипных данных на примере решения медицинской задачи.**

Одной из основных задач интеллектуального анализа данных на сегодняшний день является кластеризация (процесс получения «кластеров» или групп очень похожих объектов). Вычислительные процедуры кластерного анализа данных применяются в самых разных предметных областях. Наибольшую сложность при кластеризации объектов представляет обработка совокупности разнотипных признаков (количественных и качественных). Еще одна ключевая проблема кластеризации заключается в выборе и обосновании набора признаков, которые используются при группировке, оценка их статистической значимости и ранжирование. И наконец, проблема устойчивости группировочных решений. Результаты группировки могут сильно меняться в зависимости от выбора алгоритма и его параметров, меры сходства и т.п. Один из способов повышения устойчивости – разработка и применение ансамблевых алгоритмов.

В статье предложен оригинальный алгоритм кластеризации разнотипных данных, основанный на комплексном применении набора мер расстояний, методов кластеризации и многоэтапной кластеризации, направленный на решение вышеуказанных проблем. В рамках алгоритма реализовано ранжирование признаков объекта по их значимости для группировки и выбор оптимального набора признаков, ансамблевый подход для получения более устойчивого итогового кластерного решения. Разработанный алгоритм реализован в программной системе MixDC.

Средствами системы анализировались данные о хирургическом лечении больных с патологией восходящего отдела и дуги аорты (данные предоставлены ФГБУ «ННИИПК им. акад. Е.Н. Мешалкина» Минздрава России). База включает 18 признаков (факторов), оцененных у 124 пациентов, оперированных на восходящем отделе и дуге аорты с использованием разных хирургических технологий. В БД представлено 6 количественных, 1 порядковый и 11 номинальных признаков.

В результате было получено четырех-кластерное решение. Полученные результаты имеют логическое объяснение и могут быть содержательно проинтерпретированы.

В целом, исследования показали перспективность и эффективность применения разработанных алгоритмов и программной системы для решения задач кластеризации разнотипных данных.

SUMMARY

Alsova O.K. **Algorithms for Clustering of a Heterogeneous Data on the Example of Solution of the Medical Task.**

Clustering (the process of obtaining of «clusters» - groups of resembling objects) is one of the main problems of data mining today. Cluster analysis computational procedures are used in different areas of interest. Clustering of heterogeneous data (when numeric, ordinal and nominal attributes are combined in one dataset) is the most difficult thing in this sphere. Another clustering problem includes the choice and validation of attributes set which is used in the grouping process, their statistical significance evaluation and ranging.

And finally, there is a problem of grouping solutions stability. Clustering results can vary depending of the algorithm and its parameters, similarity measure and so on. Development and usage of ensemble algorithms is one of the ways to raise stability.

The paper describes the original algorithm of a heterogeneous data clustering based on complex application of a set of measures of distances and clustering methods and multi-stage clustering. The algorithm aims at solving the above problems. In the algorithm we use ranging of attributes the object on their importance for group and a choice of an optimum attributes set, ensemble approach to get the final clustering solution. The algorithm is realized in MixDC (Mixed Data Clustering) software system.

MixDC was used for analysis of real medical data about surgical treatment of patients of pathology of ascending aorta and aortic root (the dataset was furnished by Academishian E.N. Meshalkin Novosibirsk Research Institute of Circulation Pathology). Database includes 18 attributes (factors) estimated for 124 patients operated on ascending aorta and aortic arch with different surgical techniques. Database contains 6 numeric, 1 ordinal and 11 nominal attributes.

As a result, the four-cluster solution was obtained. The achieved results can be explained logically and meaningfully interpreted. In general, the research has shown promising and effective application of the developed algorithms and software systems for clustering of heterogeneous data.