УДК 681.5

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА ДЛЯ СТРУКТУРНОЙ ИДЕНТИФИКАЦИИ ДИАГНОСТИЧЕСКИХ ПРИЗНАКОВ

к.м.н. Э.Н. Будянская, к.т.н., проф. А.И. Поворознюк, Н.В. Максюта
(представил д.т.н., проф. В.М. Михайлов)

В статье рассмотрены методы снижения размерности пространства диагностических признаков, наиболее детально – методы кластерного анализа. Обосновывается необходимость их использования при структурной идентификации диагностических признаков. Показана апробация алгоритма структурной идентификации диагностических признаков на примере показателей реологических свойств крови.

Постановка проблемы. При построении компьютерных медицинских диагностических систем важным является вопрос формирования информативно полного пространства диагностических признаков, т.к. в медицине, как показано в [1], для постановки диагноза используется весьма разнородная информация. При этом исследуемые подсистемы организма, по сути, являются сложными иерархическими системами, поэтому адекватное описание таких систем возможно путем построения иерархической системы диагностических признаков посредством структурной идентификации диагностических признаков [2]. При структурной идентификации диагностических признаков необходимо также выполнять отбор информативных диагностических признаков из предлагаемого множества, поскольку включение в компьютерную диагностическую модель неинформативных показателей ухудшает качество компьютерного диагноза. Исходя из этого, на сегодняшний день актуальными проблемами являются разработка методов снижения пространства диагностических признаков и построение иерархической структуры диагностических признаков.

Анализ литературы. Для решения поставленной проблемы в медицине используются широко известные математико-статистические методы снижения размерности пространства диагностических признаков, а именно: кластерный, дискриминантный, факторный, дисперсионный, регрессионный анализы, многомерное шкалирование, метод главных компонент, метод контрастных групп. С их помощью устанавливают

также характер и структуру взаимосвязи исследуемых диагностических показателей [3 – 5]. Однако, как показано в [4], чаще всего для снижения пространства диагностических признаков используются методы разбиения общего числа исследуемых показателей на однородные группы, так называемые методы классификации, наиболее приемлемыми из которых являются кластерные методы.

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве X, разбить множество объектов G на m (m – целое) кластеров (подмножеств) Q_1, Q_2, … , Q_m, так, чтобы каждый объект G_j принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время, как объекты, принадлежащие разным кластерам, были разнородными. При этом существует достаточно много методов кластерного анализа: метод полных связей, метод максимального локального расстояния, метод Ворда, центроидный метод (k-среднего) [6]. Однако при их использовании необходимым условием является однотипность показателей, которая практически никогда не наблюдается в медицинской диагностике [7]. В [1] приведена разработка собственного алгоритма структурной идентификации диагностических признаков на основе алгоритма «дефекта», который выполняет кластеризацию диагностических признаков с учетом их разнотипности и внутренних связей.

Цель статьи. Апробация алгоритма структурной идентификации диагностических признаков на основе алгоритма «дефекта» на примере показателей реологических свойств крови лиц мужского пола, работающих и не работающих с видеодисплейными терминалами **(**ВДТ), с целью их кластеризации, а также построение иерархической структуры диагностических показателей реологических свойств крови.

Основной раздел. В [1] описана база клинических и клинико-лабораторных данных пользователей ВДТ, которая разработана специалистами лаборатории гигиены компьютерных и прецизионных технологий ГП «ХНИИ гигиены труда и профзаболеваний». База данных содержит около 150 показателей на каждого обследованного, в число которых входят и показатели реологических свойств крови. Для апробации алгоритма структурной идентификации диагностических признаков на основе алгоритма «дефекта» были взяты 16 показателей реологических свойств крови у 130 обследованных: 8 показателей динамической вязкости плазмы и 8 показателей динамической вязкости цельной крови. Далее для обращения к этим показателям будет использована аббревиатура соответственно названию поля в базе клинических и клинико-лабораторных данных (табл. 1).

Таблица 1

Аббревиатура показателей реологических свойств крови

Индекс	Аббревиатура	Показатель
1	VK1	динамическая вязкость цельной крови в точке 1
2	VK2	динамическая вязкость цельной крови в точке 2
3	VK3	динамическая вязкость цельной крови в точке 3
4	VK4	динамическая вязкость цельной крови в точке 4
5	VK5	динамическая вязкость цельной крови в точке 5
6	VK6	динамическая вязкость цельной крови в точке 6
7	VK7	динамическая вязкость цельной крови в точке 7
8	VK8	динамическая вязкость цельной крови в точке 8
9	VP1	динамическая вязкость плазмы в точке 1
10	VP2	динамическая вязкость плазмы в точке 2
11	VP3	динамическая вязкость плазмы в точке 3
12	VP4	динамическая вязкость плазмы в точке 4
13	VP5	динамическая вязкость плазмы в точке 5
14	VP6	динамическая вязкость плазмы в точке 6
15	VP7	динамическая вязкость плазмы в точке 7
16	VP8	динамическая вязкость плазмы в точке 8

В соответствии с алгоритмом структурной идентификации диагностических признаков на основе алгоритма «дефекта» [1] по значениям показателей реологических свойств крови была построена корреляционная матрица (табл. 2), значимые в соответствии с критерием t-Стьюдента значения которой использовались в графе показателей в качестве веса дуги между вершинами (показателями).

С помощью разработанной нами программной реализации алгоритма структурной идентификации диагностических показателей на основе алгоритма «дефекта» 16 показателей реологических свойств крови были распределены (кластеризованы) в различные группы. Сначала определились две группы: восемь показателей динамической вязкости цельной крови (VK1 – VK8) и восемь показателей динамической вязкости плазмы (VP1 – VP8). После выполнения процедуры кластеризации отдельно в этих двух группах определилось пять групп показателей реологических свойств крови. Первая группа состоит из двух показателей динамической вязкости цельной крови: VK1 и VK2; вторая группа – два показателя динамической вязкости цельной крови: VK3 и VK4; третья группа – четыре показателя динамической вязкости цельной крови: VK5, VK6, VK7 и VK8; четвертая группа –

два показателя динамической вязкости плазмы: VP1 и VP2; пятая группа – шесть показателей динамической вязкости плазмы: VP3 – VP8, т.е. явным образом просматривается иерархическая структура диагностических показателей реологических свойств крови (рис. 1).

Таблица 2

Корреляционная матрица показателей реологических свойств крови

Индексы показателей																
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	1,00	**0.65**	**0.70**	**0.64**	**0.65**	**0.60**	**0.60**	**0.53**	**0.34**	**0.24**	**0.25**	**0.29**	**0.20**	**0.18**	0.08	0.13
2	**0.65**	1,00	**0.84**	**0.81**	**0.79**	**0.74**	**0.75**	**0.65**	0.16	**0.23**	**0.23**	**0.27**	**0.29**	**0.19**	**0.19**	**0.23**
3	**0.70**	**0.84**	1,00	**0.86**	**0.92**	**0.86**	**0.88**	**0.78**	**0.25**	**0.28**	**0.29**	**0.33**	**0.33**	**0.18**	**0.21**	**0.26**
4	**0.64**	**0.81**	**0.86**	1,00	**0.93**	**0.94**	**0.91**	**0.79**	0.14	**0.26**	**0.25**	**0.33**	**0.33**	**0.20**	**0.21**	**0.22**
5	**0.65**	**0.79**	**0.92**	**0.93**	1,00	**0.95**	**0.96**	**0.85**	0.17	**0.23**	**0.27**	**0.31**	**0.33**	0.17	**0.19**	**0.22**
6	**0.60**	**0.74**	**0.86**	**0.94**	**0.95**	1,00	**0.96**	**0.86**	0.09	**0.22**	**0.22**	**0.31**	**0.32**	0.17	**0.20**	**0.21**
7	**0.60**	**0.75**	**0.88**	**0.91**	**0.96**	**0.96**	1,00	**0.88**	0.11	**0.23**	**0.26**	**0.31**	**0.33**	0.16	**0.20**	**0.20**
8	**0.53**	**0.65**	**0.78**	**0.79**	**0.85**	**0.86**	**0.88**	1,00	0.08	**0.18**	**0.20**	**0.24**	**0.26**	0.11	0.14	0.16
9	**0.34**	0.16	**0.25**	0.14	0.17	0.09	0.11	0.08	1,00	**0.40**	**0.58**	**0.36**	**0.31**	**0.35**	**0.27**	0.16
10	**0.24**	**0.23**	**0.28**	**0.26**	**0.23**	**0.22**	**0.23**	**0.18**	**0.40**	1,00	**0.68**	**0.69**	**0.65**	**0.58**	**0.60**	**0.41**
11	**0.25**	**0.23**	**0.29**	**0.25**	**0.27**	**0.22**	**0.26**	**0.20**	**0.58**	**0.68**	1,00	**0.69**	**0.71**	**0.64**	**0.57**	**0.39**
12	**0.29**	**0.27**	**0.33**	**0.33**	**0.31**	**0.31**	**0.31**	**0.24**	**0.36**	**0.69**	**0.69**	1,00	**0.77**	**0.76**	**0.67**	**0.52**
13	**0.20**	**0.29**	**0.33**	**0.33**	**0.33**	**0.32**	**0.33**	**0.26**	**0.31**	**0.65**	**0.71**	**0.77**	1,00	**0.84**	**0.77**	**0.57**
14	**0.18**	**0.19**	**0.18**	**0.20**	0.17	0.17	0.16	0.11	**0.35**	**0.58**	**0.64**	**0.76**	**0.84**	1,00	**0.80**	**0.57**
15	0.08	**0.19**	**0.21**	**0.21**	**0.19**	**0.20**	**0.20**	0.14	**0.27**	**0.60**	**0.57**	**0.67**	**0.77**	**0.80**	1,00	**0.72**
16	0.13	**0.23**	**0.26**	**0.22**	**0.22**	**0.21**	**0.20**	0.16	0.16	**0.41**	**0.39**	**0.52**	**0.57**	**0.57**	**0.72**	1,00

Примечание. Жирным шрифтом выделены значимые различия между показателям в соответствии с критерием t-Стьюдента

Для сравнения результатов апробации алгоритма структурной идентификации диагностических признаков на основе алгоритма «дефекта» была выполнена процедура кластеризации 16 показателей реологических свойств крови с помощью стандартного математического метода кластерного анализа – метода k-среднего (рис. 2).

Из дендограммы видно, что показатели динамической вязкости плазмы и динамической вязкости цельной крови сильно отличаются: это две разные группы показателей (1-й этап). В группе показателей динамической вязкости цельной крови есть различимые показатели, а показатели динамической вязкости плазмы являются сильно зависимыми. На рис. 2 пунктирными линиями показаны этапы кластеризации, принятые экспертно, исходя из анализа расстояний между показателями.

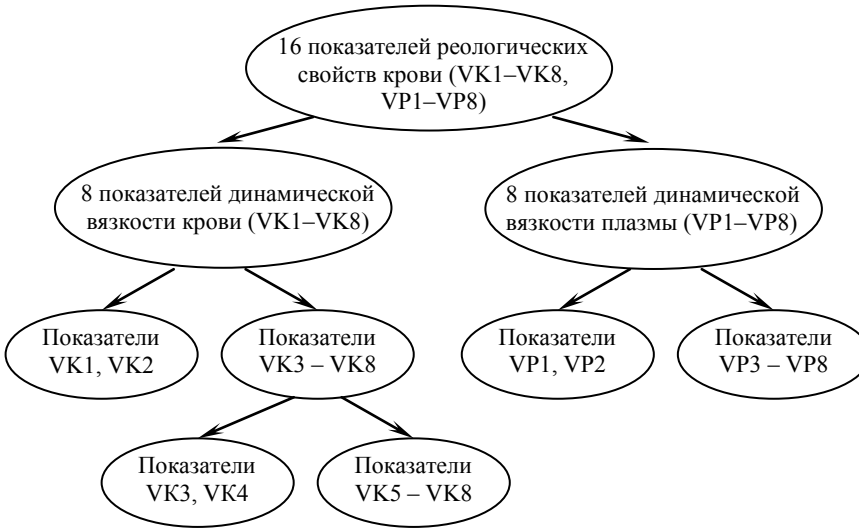

Рис. 1. Иерархическая структура диагностических показателей реологических свойств крови

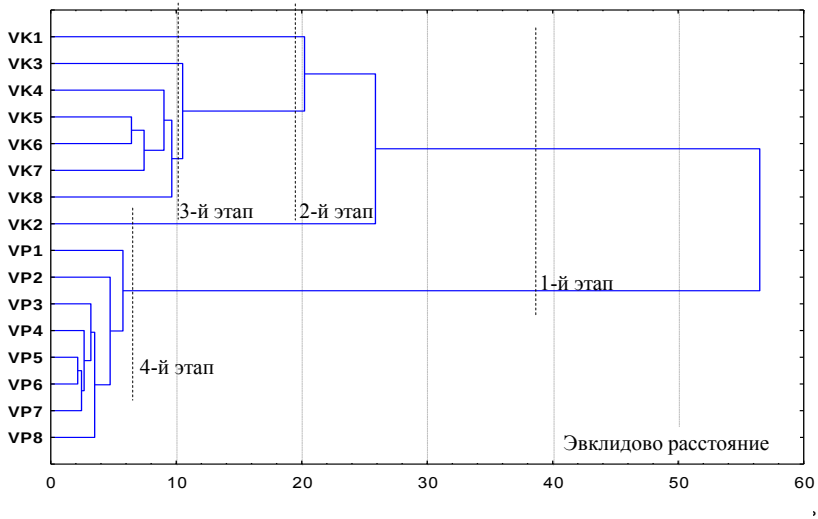

Рис. 2. Дендограмма показателей реологических свойств крови, построенная по методу k-среднего

Выводы. Таким образом, апробация разработанного в [1] алгоритма структурной идентификации диагностических признаков на основе алгоритма «дефекта» на примере показателей реологических свойств крови показала эффективность и работоспособность алгоритма «дефекта» при выполнении процедуры кластеризации и при построении иерархической структуры диагностических показателей, так как результаты апробации практически не отличаются от результатов метода k-среднего.

Перспектива дальнейших исследований в данном направлении заключается в нахождении критерия, позволяющего определить количество этапов (групп) кластеризации диагностических показателей.

ЛИТЕРАТУРА

1. *Будянская Э.Н., Поворознюк А.И., Максюта Н.В. Структурная идентификация диагностических признаков на основе алгоритма «дефекта» // Системи обробки інформації. – Х: ХВУ. – 2003. – Вип. 3. – С. 159 – 164.*
2. *Поворознюк А.И., Поворознюк Н.И. Формализация диагностических признаков в компьютерных системах медицинской диагностики // Системи обробки інформації. – Х: НАНУ, ПАНМ, ХВУ. – 2002. – Вип. 6 (22). – С. 13 – 17.*
3. *Максимов Г.К., Синицын А.Н. Статистическое моделирование многомерных систем в медицине. – Л.: Медицина, 1983. – 144 с.*
4. *Айвазян С.А., Бухштабер В.М., Енюков И.С. Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.*
5. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. Справ. изд. – М.: Финансы и статистика, 1985. – 487 с.*
6. *Жамбю М. Иерархический кластер-анализ и соответствия. – М.: Финансы и статистика, 1988. – 342 с.*
7. *Ахутин В.М., Шаповалов В.В., Иоффе М.О. Оценка качества формализованных медицинских документов // Медицинская техника. – М: Медицина, 2002. – Вып. 2. – С. 27 – 31.*

Поступила 15.04.2004

***БУДЯНСКАЯ Элеонора Николаевна**, канд. мед. наук, ст. научн. сотр., зав. лаб. гигиены компьютерных и прецизионных технологий ГП «ХНИИ гигиены труда и профзаболеваний». В 1963 году окончила ХМИ. Область научных интересов – изучение влияния комплекса факторов физической природы малой интенсивности на организм человека с целью разработки руководящих нормативных документов в Украине и России.*

***ПОВОРОЗНЮК Анатолий Иванович**, канд. тех. наук., проф., докторант НТУ «ХПИ». В 1977 году окончил ХПИ. Область научных интересов – разработка методов и алгоритмов построения компьютерных систем медицинской диагностики.*

***МАКСЮТА Наталья Валерьевна**, аспирант НТУ «ХПИ». В 2003 году окончила НТУ «ХПИ». Область научных интересов – разработка методов и алгоритмов построения компьютерных систем медицинской диагностики.*