УДК 681.5

Н.В. МАКСЮТА, НТУ “ХПИ” (г. Харьков),
А.И. ПОВОРОЗНЮК, канд. тех. наук, НТУ “ХПИ” (г. Харьков)

ТОЧНОСТЬ ПОСТАНОВКИ ДИАГНОЗА ПРИ ИСПОЛЬЗОВАНИИ СТРУКТУРНОЙ ИДЕНТИФИКАЦИИ ДИАГНОСТИЧЕСКИХ ПРИЗНАКОВ

Приведений короткий огляд методів синтезу діагностичного вирішального правила. Показані можливості застосування ієрархічної структури діагностичних ознак що до синтезу комп’ютерного діагнозу за показниками реологічних властивостей крові на прикладі деяких терапевтичних захворювань. Приведений порівняльний аналіз точності прогнозу по вихідному лінійному простору ознак і по інтегральним показникам, які побудовані за результатами структурної ідентифікації діагностичних ознак.

The brief review of methods of synthesis of a diagnostic deciding rule is resulted. Opportunities of application of hierarchical structure of diagnostic attributes for synthesis of the computer diagnosis on parameters rheology are shown blood by the example of some therapeutic diseases. The comparative analysis of accuracy of the forecast on initial linear space of attributes and on the integrated parameters constructed by results of structural identification of diagnostic attributes is resulted.

Постановка проблемы. Появившиеся сравнительно недавно интеллектуальные медицинские компьютерные системы, представляющие собой программно-аппаратные комплексы диагностики и терапии, получили достаточно широкое распространение благодаря простоте использования и высокой эффективности. Их проектирование включает в себя 3 основных этапа: сбор и первичная обработка исходных данных – этап, включающий построение таблиц экспериментальных данных, а также анализ их типов и связей; классификация исходных показателей и отбор полезной информации; построение диагностического решающего правила.

Отбор информативных признаков является необходимым специальным этапом классификации показателей, так как каждый признак несет в себе как положительный вклад в разделение, так и шумовую составляющую. Включение же малоинформативных переменных может заметно ухудшить качество диагностического правила [1 – 4]. Это особенно актуально при ограниченности исследуемой выборки. При этом полученная система информативных признаков должна обеспечить необходимое качество распознавания диагноза.

Авторами в [5] разработан оригинальный алгоритм построения иерархической структуры диагностических признаков с учетом характера их связей минимально необходимого объема, основанный на представлении множества исходных показателей в виде потоковой модели и иерархической кластеризации, для проверки работоспособности которого целесообразно провести сравнительный анализ его эффективности по сравнению с наиболее известными методами.

Анализ литературы. На сегодняшний день общеприменимыми являются следующие методы классификации и отбора информативных показателей: кластер-анализ, расщепление смесей распределений, факторный анализ, метод главных компонент, многомерное шкалирование, метод корреляционных плеяд, дискриминантный и регрессионный анализы [1, 2, 4, 6]. Их краткое описание, достоинства и недостатки приведены в [3]. При этом работа дискриминантного и регрессионного анализов основана на критерии внешней информативности, который нацелен на максимальное „выжимание” информации из исходного набора признаков относительно некоторых внешних показателей (например, классов заболеваний). Остальные же методы основаны на критерии автоинформативности, нацеленного на максимальное сохранение информации, содержащейся в исходном наборе признаков относительно их самих [1, 5, 6].

Из рассмотренных выше методов классификации по сравнению с разработанным алгоритмом структурной идентификации признаков по подобной схеме работает метод корреляционных плеяд, однако при его применении необходимо задаваться некоторым пороговым значением классификации и в некоторых случаях допускается построение только незамкнутых графов. Преимущество же предлагаемого подхода состоит в том, что он снимает ограничения на размерность задачи, позволяет наглядно представить структуру связей показателей, а также вскрывать неочевидные, но важные для решения диагностических задач связи и влияния и не требует эвристического задания порога кластеризации.

Качество компьютерного диагноза определяется как системой информативных признаков, так и методом синтеза решающего правила, наиболее распространенными из которых являются следующие [6, 7]:

1. Интенсиональные методы, основанные на операциях с признаками. К ним, прежде всего, относятся методы, основанные на оценках плотностей распределения значений признаков, которые базируются на байесовской схеме принятия решений: дискриминантный анализ, метод евклидова расстояния, метод потенциальных функций и др.; методы, основанные на предположениях о классе решающих функций, в которых считается известным общий вид решающей функции и задан функционал ее качества: алгоритм Ньютона, алгоритмы перцептронного типа, метод группового учета аргументов и др.; логические методы, в основе которых лежит аппарат алгебры логики: древообразные классификаторы, метод тупиковых тестов, алгоритм „Кора” и др.; лингвистические (структурные) методы, основанные на использовании специальных грамматик, с помощью которых может описываться совокупность свойств распознаваемых объектов: грамматики без ограничений, бесконтекстные, автоматные и другие типы грамматик.

2. Экстенсиональные методы, работающие на основе операций с объектами: метод сравнения с прототипом, который применяется в том случае, когда распознаваемые классы отображаются в пространстве признаков компактными геометрическими группировками; метод *k*-ближайших соседей, основанный на нахождении *k*-геометрически ближайших объектов в пространстве признаков; алгоритмы вычисления оценок (голосования), принцип которых состоит в вычислении приоритетов, характеризующих близость распознаваемого и эталонных объектов по определенной системе признаков.

3. Эвристический подход, который основывается на трудно формализуемых знаниях и интуиции исследователя, который определяет какую информацию и каким образом нужно использовать для достижения требуемого эффекта распознавания, например, Патохарактерологический диагностический опросник.

В [8] представлены результаты апробации разработанного алгоритма структурной идентификации диагностических признаков на примере реологических показателей и сравнительный анализ с результатами кластеризации этого же набора данных кластерным анализом, исходя из которого доказана работоспособность и эффективность структурной идентификации признаков с точки зрения классификации без обучения.

Цель статьи. Проверка адекватности применения структурной идентификации признаков при синтезе компьютерного диагноза.

Постановка диагноза с использованием иерархической структуры диагностических признаков. Адекватность применения структурной идентификации диагностических признаков для синтеза диагностического решающего правила можно проверить, проанализировав точность распознавания объектов по исходному множеству признаков и по отобранному информативному подмножеству. Исследование проводилось на выборке из 137 (N = 137) лиц мужского пола, здоровых и имеющих терапевтические заболевания (2 группы заболеваний), при этом анализировались 16 показателей реологических свойств крови [8, табл.1].

Отбор информативных признаков выполнялся методами факторного анализа [3, формула 1] в 5 группах реологических показателей, полученных в результате классификации исходных 16 переменных с помощью структурной идентификации диагностических показателей реологических свойств крови [8], представленной на рис.

Далее с помощью дискриминантного анализа производилась классификация объектов (отнесение к классу здоровых или имеющих заболевания) по исходным 16 показателям реологии крови и по 5 интегральным (ИП): ИП3, ИП5 – ИП8, и выполнялась проверка несовпадений исходных и полученных в результате классификации групп заболеваний. Выбор дискриминантного анализа, как метода распознавания образов, основывался на том, что данный метод можно применять в случае нелинейной зависимости внешнего критерия от исходных показателей и авторы имеют опыт работы с ним. Тем более, что в постановке данной задачи не стоял вопрос анализа достоинств и недостатков описанных выше методов распознавания образов с целью выбора наиболее оптимального из них.

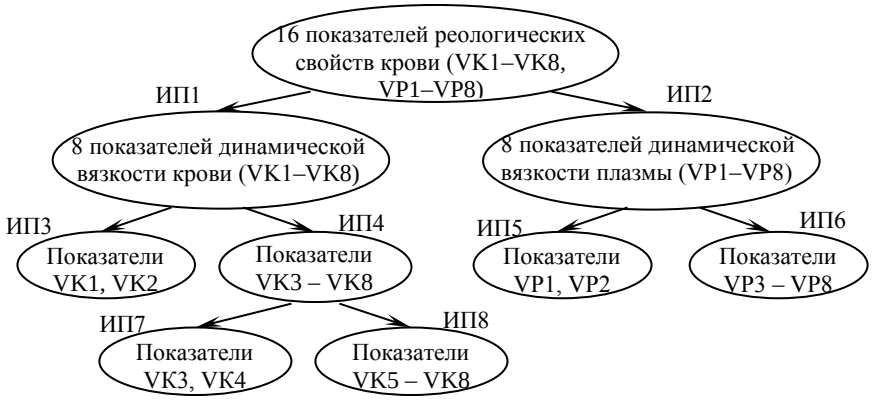

Рис. Иерархическая структура диагностических показателей реологических свойств крови

Сравнительный анализ качества распознавания групп заболеваний по исходному множеству реологических показателей и по их иерархической структуре. Прежде чем приступать к данному этапу, необходимо иметь представление о точности постановки диагноза с помощью дискриминантного анализа в отличие от исходных групп заболеваний. Проценты правильного распознавания групп терапевтических заболеваний по 16 показателям реологии крови в исходной выборке (N = 137) представлены в табл. 1. Анализируя табл. 1 можно сказать, что применение дискриминантного анализа с целью классификации объектов адекватно на 94,14 % для данной выборки, а это довольно высокий процент.

В табл. 2 представлены проценты правильного распознавания групп терапевтических заболеваний по 5 ИП реологии крови в исходной выборке (N = 137). В этом случае применение дискриминантного анализа с целью классификации объектов адекватно на 78,07 % для данной выборки.

Таким образом, в исходной выборке процент корректного распознавания здоровых и имеющих заболевания по 5 интегральным показателям реологии крови снизился с 94,14% до 78,07% по сравнению с распознаванием по 16 исходным показателям реологии крови, в то время как число показателей уменьшилось в 3,2 раза.

Для получения более достоверных результатов анализа точности постановки диагноза общепринятым является то, что исходную выборку разбивают на две – обучающую и контрольную. Контрольную выборку считают „неопределенной", диагностируют ее и производят проверку несовпадений исходных и полученных в результате классификации групп заболеваний. Последнее и будет характеризовать адекватность применения структурной идентификации диагностических признаков при синтезе компьютерного диагноза.

Таблица 1

Точность распознавания групп терапевтических заболеваний по 16 показателям реологии крови

	Процент распознавания	Количество объектов, отнесенных к группе „Здоровые"	Количество объектов, отнесенных к группе „Имеющие заболевания"
1	2	3	4
Группа „Здоровые"	97,1	67	2
Группа „Имеющие заболевания"	91,18	6	62
Общее	94,14	73	64

Таблица 2

Точность распознавания групп терапевтических заболеваний по 5 интегральным показателям реологии крови

	Процент распознавания	Количество объектов, отнесенных к группе „Здоровые"	Количество объектов, отнесенных к группе „Имеющие заболевания"
Группа „Здоровые"	82,61	57	12
Группа „Имеющие заболевания"	73,53	18	50
Общее	78,07	75	62

Исходя из этого, исходная выборка N была разбита на две: $N1$ – обучающая и $N2$ – контрольная. При этом обучающая выборка включала 69 лиц, а контрольная – 68. После произведенной классификации объектов, принадлежащих выборке $N2$, по исходным 16 показателям реологии крови выявилось 10 несовпадений по сравнению с исходными заданными группами заболеваний, что составляет 14,7%.

Однако изначально в выборке N присутствует 8 „некорректных" объектов с точки зрения классификации дискриминантным анализом, что составляет 5,8% (см. табл. 1), при этом 2,9% или 4 „некорректных" объекта приходится на выборку $N2$ (исходя из детального анализа выборки). С учетом этого можно сказать, что при классификации выборки $N2$ по 16 показателям реологии крови выявилось 6 несовпадений, что составляет 8,8%.

При классификации объектов, принадлежащих выборке $N2$, по 5 интегральным показателям реологических свойств крови, построенным с помощью их структурной идентификации и факторного анализа, выявилось 16 несовпадений по сравнению с исходными заданными группами заболеваний, что составляет 23,5%. С учетом же исходной „некорректности" объектов можно принять 12 несовпадений или 17,6%.

Выводы. Таким образом, точность постановки терапевтических диагнозов по 5 интегральным показателям реологических свойств крови, построенным с помощью их иерархической структуры (см. рис.) и факторного анализа, снизилась на 8,8% по сравнению с распознаванием тех же классов по исходному множеству реологических показателей (16 показателей), в то время как число показателей уменьшилось в 3,2 раза, что свидетельствует о высокой эффективности разработанного подхода с точки зрения классификации с обучением.

Перспективы дальнейших исследований состоят в том, чтобы провести анализ качества распознавания конкретных видов терапевтических заболеваний или результатов обследования другими специалистами, для чего необходимо наличие соответствующих медицинских баз данных.

Список литературы: 1. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с. **2.** *Корбинский Б.А.* Принципы математико-статистического анализа данных медико-биологических исследований // Российский вестник перинатологии и педиатрии, 1996. – Вып. 4. – С. 60 – 64. **3.** *Максюта Н.В., Поворознюк А.И.* Алгоритмы и методы снижения пространства диагностических признаков // Вісник Національного технічного університету „ХПІ ”. – Х.: НТУ ”ХПІ”. – 2005. – № 46. – С. 126 – 131. **4.** *Брандт З.* Анализ данных; статистические и вычислительные методы для научных работников и инженеров. – М.: Мир: АСТ, 2003. – 686 с. **5.** *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. **6.** *Дюк В.А*. Компьютерная психодиагностика. – С.-Петербург: Братство, 1994. – 364 с. **7.** *Поспелов Д.А.* Данные и знания. Представление знаний // Искусственный интеллект. Кн. 2: Модели и методы. – М.: Радио и связь, 1990. – С. 7 – 13. **8.** *Будянская Э.Н., Поворознюк А.И., Максюта Н.В.* Применение кластерного анализа для структурной идентификации диагностических признаков // Системи обробки інформації. – Х.: ХВУ, 2004. – Вип. 6. – С. 23 – 28.

Поступила в редакцию 25.10.2005