УДК 681.5

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

АЛГОРИТМЫ И МЕТОДЫ СНИЖЕНИЯ ПРОСТРАНСТВА ДИАГНОСТИЧЕСКИХ ПРИЗНАКОВ

У статті обгрунтовано необхідність використання попередньої обробки даних в комп'ютерних діагностичних системах. Проведений огляд та порівняльний аналіз методів зниження простору діагностичних ознак.

In article necessity of application of preliminary data processing for computer diagnostic systems is proved. The comparative analysis of methods of decrease space of diagnostic attributes is carried out.

Постановка проблемы. В лечебной практике в последние годы компьютерные диагностические системы получили широкое внедрение. Как показано в [1, 2], для их использования в целях диагностики и прогнозирования необходимо решить ряд задач, среди которых немаловажное место занимает предварительная обработка данных и корректное снижение размерности пространства диагностических признаков с целью обнаружения информативной совокупности показателей без существенной потери значимой информации. Необходимость выполнения данной задачи обуславливается тем, что при исследовании состояния здоровья пациента во многих случаях анализируется большое число разнотипных показателей [1 – 3], несущих избыточную информацию и ухудшающих качество решающего правила [4].

Анализ литературы. Снижение размерности пространства диагностических признаков осуществляется с помощью методов классификации многомерных наблюдений. К ним, прежде всего, относятся следующие методы: главные компоненты, экстремальная группировка параметров, корреляционные плеяды, многомерное шкалирование; кластерный и факторный анализы [1, 4 – 9]. Работа данных методов основана на анализе некоторых мер связи между показателями или объектами, в качестве которых может выступать коэффициент корреляции или расстояние в пространстве признаков. При этом, как показано в [3, 9], для различных типов исходных данных (дихотомические, ранговые, численные) применяются соответствующие им меры связи, что в свою очередь порождает проблемы корректного применения того или иного метода снижения пространства признаков и увеличения возможностей диагностической системы.

Целью статьи является краткий обзор и сравнительный анализ существующих математико-статистических алгоритмов и методов снижения размерности пространства диагностических признаков.

Кластерный анализ основан на вычислении расстояний между показателями в пространстве признаков и заключается в формировании из исходного множества показателей m-подмножеств (кластеров), внутри которых показатели имеют максимальное сходство (минимальное расстояние), в то время как между кластерами наблюдается разнородность (максимальное расстояние) [4, 8]. При этом для различных наборов исходных показателей применяются соответствующие им меры сходства или расстояния [3, 9]. Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности (целевой функции). Известны следующие методы кластерного анализа: одиночных и полных связей, попарное среднее, центроидный, Варда, k-среднего [7]. Отличие названных методов заключается в способе вычисления расстояния между кластерами (объединением двух кластеров в один), причем в некоторых из них за точки кластеров принимаются либо центры тяжестей, либо наиболее или наименее удаленные точки, а также – в виде целевой функции.

Метод главных компонент основан на выделении линейных комбинаций случайных величин, имеющих максимально возможную дисперсию, с помощью матрицы парных корреляций [7, 9]. Линейные комбинации выбираются таким образом, что среди всех возможных линейных нормированных комбинаций исходных признаков первая главная компонента обладает максимальной дисперсией. Вторая – максимальной дисперсией среди оставшихся линейных преобразований, некоррелированных с первой компонентой и т.д. Таким образом, компоненты ранжированы по степени их вклада в суммарную дисперсию исходных факторов. В связи с этим появляется возможность выразить информацию, содержащуюся в большом наборе исходных факторов, с помощью меньшего числа независимых главных компонент. В основе метода лежит ортогональное преобразование исходных факторов к обобщенным (главным компонентам).

Цель факторного анализа – объяснить ковариационную матрицу (имеющиеся между признаками корреляции) минимальным числом факторов. Эти факторы выражаются переменными, которые не могут быть идентифицированы точно, поскольку известно лишь их число и интенсивность, с которой они действуют на первичные показатели [1, 9].

Основная модель факторного анализа записывается следующей системой равенств [9]:

$$x_i = \sum_{j=1}^{m} l_{ij} f_{ij} + \varepsilon_i, \qquad (1)$$

где i=1, ..., p; p – число признаков; $m<p$ – число факторов; l_{ij} – нагрузка i-й переменной на j-й фактор. Т.е. полагается, что значения каждого признака x_i могут быть выражены взвешенной суммой латентных переменных (простых

факторов) f_j, число которых меньше числа исходных признаков, и остаточным членом ε_i (специфическим фактором) с дисперсией σ^2, действующей только на x_i.

В самой простой модели факторного анализа считается, что факторы f_j взаимно независимы и их дисперсии равны единице, и случайные величины ε_i тоже независимы друг от друга и от какого-либо фактора. Максимально возможное число факторов m при заданном числе признаков p определяется неравенством $(p+m)<(p-m)^2$, которое должно выполняться, чтобы задача не вырождалась в тривиальную. При этом [9]:

$$\text{дисперсия признака} = \text{общность} \left(\sum_{j=1}^{m} l_{ij}^2\right) + \text{специфично сть} \left(\varepsilon_i^2\right), \qquad (2)$$

где общность является частью дисперсии признака, которую объясняют факторы, а специфичность показывает, какая часть дисперсии исходного признака остается необъясненной при используемом наборе факторов.

Основное соотношение факторного анализа показывает, что коэффициент корреляции любых двух признаков можно выразить суммой произведений нагрузок некоррелированных факторов [9]:

$$r_{ij} = l_{i1} l_{j1} + l_{i2} l_{j2} + \ldots + l_{im} l_{jm}. \qquad (3)$$

Задачу факторного анализа нельзя решить однозначно. Равенства (1) не поддаются непосредственной проверке, так как p исходных признаков задается через $(p+m)$ других переменных – простых и специфичных факторов. Поэтому представление корреляционной матрицы факторами (факторизацию) можно произвести бесконечно большим числом способов.

Вычисления нагрузок начинаются с $m = 1$, затем проверяется насколько корреляционная матрица, восстановленная по однофакторной модели в соответствии с (3), отличается от исходной. Если однофакторная модель признается неудовлетворительной, то испытывается модель с $m = 2$ и т.д., пока не будет достигнута адекватность или число факторов в модели не превысит максимально допустимое. В завершении всей процедуры факторного анализа с помощью математических преобразований выражают факторы f_j через исходные признаки, т.е. получают в явном виде параметры линейной диагностической модели.

Метод экстремальной группировки параметров используется когда изменение какого-либо общего фактора неодинаково на измеряемых частных признаках. В частности, возможна ситуация, когда исходная совокупность из p признаков обнаруживает расщепление на сравнительно небольшое число групп, при котором изменение признаков, относящихся к какой-либо одной группе, обуславливается в основном каким-то одним общим фактором, своим для каждой такой группы [7]. В этом случае разбиение на группы естественно

строить так, чтобы параметры, принадлежащие к одной группе, были коррелированны сравнительно сильно, а параметры, принадлежащие к разным группам, – слабо. После такого разбиения для каждой группы признаков строится случайная величина, которая в некотором смысле наиболее сильно коррелированна с параметрами данной группы. Эта случайная величина интерпретируется как искомый фактор, от которого существенно зависят все параметры данной группы. Отличие от моделей факторного анализа заключается в том, что выделение общего фактора делается на основе оптимизации (максимизации) некоторых эвристически введенных функционалов.

Метод корреляционных плеяд также как и метод экстремальной группировки, и кластерный анализ, предназначен для нахождения таких групп признаков – 'плеяд', когда корреляционная связь, которая определяется как сумма модулей коэффициентов корреляции между параметрами одной группы (внутриплеядная связь), достаточно велика, а связь между параметрами из разных групп (межплеядная) – мала [1, 5]. На основе корреляционной матрицы строится граф связей, который разбивается по тем или иным признакам на подграфы. Вершины подграфа и образуют плеяду. Задаваясь некоторыми пороговыми значениями коэффициента корреляции, исключают из исходного графа все ребра с коэффициентом корреляции по модулю меньше порогового. Поэтапно увеличивая пороговое значение, повторяют эту процедуру до разбиения графа на несколько подграфов. Для полученных таким образом плеяд внутриплеядные коэффициенты корреляции будут больше последнего порогового значения коэффициента корреляции, а межплеядные – меньше.

Многомерное шкалирование – совокупность методов, позволяющих по заданной информации о мерах различия (близости) между объектами приписывать каждому из этих объектов вектор характеризующих его количественных показателей; при этом размерность искомого координатного пространства задается заранее, а «погружение» в него анализируемых объектов производится таким образом, чтобы структура взаимных различий между ними, измеренных с помощью приписываемых им вспомогательных координат, в среднем наименее отличалась бы от заданной в смысле того или иного функционала качества [1, 6 – 7]. Определение координат объектов в пространстве и размерности пространства основано на преобразовании матрицы расстояний в матрицу скалярных произведений центрированных векторов. Таким образом, на выходе алгоритма получаются числовые значения координат, которые приписываются каждому объекту в некоторой новой системе координат (во "вспомогательных шкалах", связанных с латентными переменными), причем размерность нового пространства признаков существенно меньше размерности исходного.

Сравнительный анализ методов снижения размерности пространства диагностических признаков. Рассмотренные методы классификации имеют определенные достоинства и ограничения.

Кластерный анализ не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет наглядно интерпретировать результаты разбиения, но состав и количество кластеров зависит от выбираемых критериев разбиения. При этом считается, что выбранные характеристики допускают желательное разбиение на кластеры и единицы измерения (масштаб) выбраны правильно. Также при проведении классификации объектов часто игнорируется возможность отсутствия в рассматриваемой совокупности каких-либо значений кластеров.

Выбор масштаба играет большую роль для большинства рассматриваемых методов. Неоднородность единиц измерения и вытекающая отсюда невозможность обоснованного выражения значений различных показателей в одном масштабе приводит к тому, что величины расстояний между точками, отражающими положение объектов в пространстве их свойств, оказываются зависящими от произвольно избираемого масштаба. Чтобы устранить неоднородность измерения исходных данных, все их значения предварительно нормируются, т.е. выражаются через отношение этих значений к некоторой величине, отражающей определенные свойства данного показателя. Как правило, данные нормализуют вычитанием среднего и делением на стандартное отклонение.

Компонентный анализ в общем случае позволяет получить наивысший коэффициент сжатия, являясь оптимальным в смысле минимума энтропии и среднего квадрата ошибки описания. Однако его использование не позволяет объективно учесть ошибки измерения в наблюдаемых переменных.

Применение же факторного анализа наряду со сжатием данных позволяет осуществить выделение ошибок в специфические факторы и исключить их из дальнейшего описания. В отличие от метода главных компонент факторный анализ требует, чтобы исследуемые данные подчинялись многомерному нормальному распределению, а зависимости были линейными. Он также не может адекватно работать с сильно коррелированными между собой показателями, т.к. корреляционная матрица для такого набора переменных не может быть обращена. Это, в свою очередь, приводит к необходимости применения процедуры понижения всех корреляций в матрице путем добавления малой константы к диагональным элементам, что ведет к менее точным оценкам.

Многомерное шкалирование применимо в ряде случаев, когда непригодно большинство методов факторизации, т.к. метод менее чувствителен к типам расстояний или сходств, однако адекватно работает в пространстве небольшой размерности.

Метод экстремальной группировки параметров основан на наличии в каждой коррелированной группе общего фактора, характеризующего разбивку исходного множества показателей, что не всегда является обоснованным.

Применение метода корреляционных плеяд включает в себя выбор порогового значения коэффициента корреляции, выступающего критерием разбиения, что может привести к различным результатам классификации, как и в кластерном анализе.

Авторами в [3, 4] разработан оригинальный алгоритм классификации диагностических признаков, основанный на представлении множества исходных показателей в виде потоковой модели, и иерархической кластеризации, который в отличие от метода корреляционных плеяд не требует эвристического задания порога кластеризации и снимает ограничения на размерность задачи.

Выводы. Таким образом, при применении методов снижения пространства признаков обязательным является предварительный анализ исходных данных, причем процедура подготовки данных к математической обработке является индивидуальной для каждого рассмотренного метода, что вносит основную трудность при построении компьютерных диагностических систем, при этом каждый метод имеет свои достоинства и недостатки. А для различных структур диагностических данных оптимальное решение может дать любой из рассмотренных выше методов.

Список литературы: 1. *Корбинский Б.А.* Принципы математико статистического анализа данных медико-биологических исследований // Российский вестник перинатологии и педиатрии. – 1996. – Вып. 4. – С. 60–64. **2.** *Поворознюк А.И., Поворознюк Н.И.* Формализация диагностических признаков в компьютерных системах медицинской диагностики // Системи обробки інформації. – Х.: НАНУ, ПАНИ, ХВУ, 2002. – Вип. 6 (22). – С. 13 – 17. **3.** *Будянская Э.Н., Поворознюк А.И., Максюта Н.В.* Структурная идентификация диагностических признаков на основе алгоритма «дефекта» // Системи обробки інформації. – Х.: НАНУ, ПАНИ, ХВУ, 2003. – Вип. 3. – С. 159 – 164. **4.** *Будянская Э.Н., Поворознюк А.И., Максюта Н.В.* Применение кластерного анализа для структурной идентификации диагностических признаков // Системи обробки інформації. – Х.: НАНУ, ПАНИ, ХВУ, 2004. – Вип. 6. – С. 23 – 28. **5.** *Жилинская М.В., Овсенева Т.Л.* Метод корреляционных плеяд в изучении структуры связей показателей эритроцитарной системы при пиелонефрите // Медицинская и биологическая кибернетика. – М.: МОЛГМИ им. Н.И. Пирогова, 1997. – Вып. 2. – С. 145 – 150. **6.** *Дэйвисон М.* Многомерное шкалирование: Методы наглядного представления данных. – М.: Финансы и статистика, 1988. – 254 с. **7.** *Айвазян С.А., Бухштабер В.М., Енюков И.С.* Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с. **8.** *Жамбю М.* Иерархический кластер-анализ и соответствия. – М.: Финансы и статистика, 1988. – 342 с. **9.** *Дюк В.А.* Компьютерная психодиагностика. – СПб.: Братство, 1994. – 364 с.

Поступила в редакцию 11.04.2005