УДК 658.012

СТРУКТУРНАЯ ИДЕНТИФИКАЦИЯ ДИАГНОСТИЧЕСКИХ ПРИЗНАКОВ НА ОСНОВЕ АЛГОРИТМА ”ДЕФЕКТА”

к.м.н. Э.Н. Будянская, к.т.н., проф. А.И. Поворознюк, Н.В. Максюта

Проанализированы способы построения компьютерных медицинских диагностических систем. Показана необходимость применения структурной идентификации диагностических показателей. Предлагается адаптация алгоритма «дефекта», который основан на теории графов, и теоремы о максимальном потоке и минимальном разрезе для задачи кластеризации показателей.

Постановка проблемы. Компьютерная диагностика сегодня широко внедряется во многие области медицины. При этом в медицинской диагностике используется объемная и весьма разнотипная информация, которая, как показано в [1 – 4], отражает сложную структуру организма. Поэтому при построении компьютерной медицинской диагностической системы качество диагнозов в большей степени определяется диагностической ценностью обрабатываемых данных. В связи с этим часто возникает вопрос о способе компоновки (преобразования) исходных диагностических признаков в диагностические показатели.

Для определения параметров диагностической модели используются две стратегии эмпирико-статистического анализа данных [2].

Первая стратегия основывается на критерии автоинформативности обрабатываемых данных, который подразумевает, что диагностическую модель можно непосредственно определить путем аппроксимации геометрической структуры множества объектов в пространстве исходных признаков, не прибегая к сведениям об эмпирических (внешних) отношениях исследуемых объектов и признаков. Параметры диагностической модели можно определить с помощью таких известных методов, как метод главных компонент, факторный анализ, метод контрастных групп [2].

Вторая стратегия определения параметров диагностической модели основана на использовании дополнительной информации о диагностируемом свойстве исследуемых объектов. Критерии, по которым формируется дополнительная информация, называют критериями внешней информативности или внешними критериями. Главными представителями методов, использующих внешний критерий, являются методы регресси-

онного и дискриминантного анализа [2].

Рассматриваемые формальные методы определения параметров диагностической модели не учитывают особенности исследуемых объектов, их внутреннюю структуру, поэтому в данной работе предлагается перспективный подход, основанный на выявлении внутренних связей, который, как показано в [1], может быть реализован на теории графов [5].

Целью данной работы является структурная идентификация диагностических признаков на основе алгоритма "дефекта", которая базируется на критерии автоинформативности.

Разработка алгоритма структурной идентификации. В качестве первичного материала для алгоритма структурной идентификации диагностических признаков используется реальная база клинических и клинико-лабораторных данных пользователей видео-дисплейных терминалов (ВДТ), разработанная в лаборатории гигиены компьютерных и прецизионных технологий ГП "Харьковский научно-исследовательский институт гигиены труда и профессиональных заболеваний".

База клинических и клинико-лабораторных данных состоит из трех таблиц: профмаршрут, данные клинических и клинико-лабораторных показателей и данные гигиенической оценки условий труда (уровни электромагнитного поля и статического электричества ВДТ, за которым работает пользователь), связанных между собой по порядковому номеру. База включает в себя следующие подсистемы [3]: профмаршрут (ФИО, пол, дата рождения, предприятие, профессия, дата прихода в профессию, вредные факторы условий труда, дата обследования, количество часов работы за ВДТ в месяц, возраст, стаж); осмотры врачей-профпатологов (терапевт, ЛОР-врач, окулист, невропатолог, гинеколог, дерматолог, эндокринолог); общий клинический анализ крови (12 показателей); гормоны (6 показателей); биохимия (14 показателей); иммунология (38 показателей); реология (24 показателя); микроэлементы (6 показателей); гигиена (20 показателей). Исходя из структуры входных данных, целесообразно выполнять структурную идентификацию отдельно по каждой системе. В данной работе анализ проводится на примере реологических показателей организма человека.

Анализ исходных данных показал, что существует ряд диагнозов с некоторыми повторяющимися признаками, что свидетельствует о пересечении этих заболеваний в пространстве признаков и означает то, что некоторые данные могут быть избыточными. Поэтому возникает необходимость применения структурной идентификации диагностических признаков, для которой предлагается использование графа для получения структуры, максимальным образом отражающей структуру объекта [1]. Исходные показатели представляются вершинами полносвязного графа, дугам кото-

рого приписываются некоторые численные данные (веса), в качестве которых могут быть использованы статистические меры связи или расстояния в пространстве признаков [4]. В качестве статистической меры связи может использоваться коэффициент парной корреляции для численных признаков или его аппроксимация (коэффициент сопряженности для дихотомических и коэффициент ранговой корреляции для ранговых признаков) [1], а в качестве расстояния – евклидово расстояние (для количественных, качественных и дихотомических признаков), расстояние Минковского (для ординальных признаков) [2] и другие меры расстояний.

При наличии априорной информации о структуре объекта некоторые вершины графа могут быть объединены, а дуги исключены. Тогда задача структурной идентификации сводится к задаче группировки вершин графа в некоторые классы (кластеризация вершин) или к эквивалентной ей задаче разреза графа на подграфы. При такой формулировке задача сводится к потоковой, для решения которой предлагается адаптация алгоритма ”дефекта” [1], который является одним из перспективных алгоритмов решения потоковых задач с ограничением.

Так как сходство или различие объектов определяется мерами близости/удаления (расстояния), а признаков – мерами связи и реологические показатели являются количественными признаками, то в качестве весов дуг графа принимается значение коэффициента парной корреляции.

При определении весов графа используется корреляционная матрица, при этом берутся только значимые, в соответствии с критерием Стьюдента, значения коэффициента парной корреляции, что существенно снижает число дуг графа (граф перестает быть полносвязным).

Суть алгоритма ”дефекта” заключается в поиске циркуляции, минимизирующей суммарную стоимость потоков по всем дугам. Циркуляцией называется поток по дугам, для которого в каждом узле (вершине) выполняется условие сохранения, т.е. суммарный поток, входящий в узел, равен суммарному потоку, выходящему из узла [5]. При работе алгоритма используется 4 параметра, характеризующие дуги графа: f_{ij} – поток по дуге (i , j); L_{ij} – нижняя пропускная способность дуги (i, j); U_{ij} – верхняя пропускная способность дуги (i, j); c_{ij} – стоимость прохождения единицы потока из узла i в узел j.

Для рассматриваемой задачи структурной идентификации диагностических признаков на основе алгоритма ”дефекта” в качестве L_{ij} принимается значение, равное 0, в качестве U_{ij} – значение коэффициента парной корреляции. Значения потоков определяются после завершения работы алгоритма ”дефекта”, изначально они могут быть равны 0.

Задача нахождения циркуляции минимальной стоимости представляется в виде специальной задачи линейного программирования, кото-

рую называют прямой задачей [5]:

минимизировать $\sum_{(i,j\in S)} c_{ij} f_{ij}$, где S – множество дуг, при условии, что:

$\sum_{j\in N} f_{ij} - \sum_{j\in N} f_{ji} = 0,$ для всех $i \in N$ (условие сохранения потока);

$f_{ij} \le U_{ij},\ (i, j) \in S$ (ограничения на потоки сверху);

$f_{ij} \ge L_{ij},\ (i, j) \in S$ (ограничения на потоки снизу);

$f_{ij} \ge 0$ (условие неотрицательности потока).

Согласно известному в линейном программировании результату, для любой прямой задачи всегда существует соответствующая ей двойственная задача. В рассматриваемом случае двойственная задача формулируется следующим образом [5]: минимизировать $\sum_{(i,j\in S)} U_{ij} \cdot \alpha_{ij} - L_{ij} \cdot \delta_{ij}$, где S – множество дуг, при условии, что: $\pi_i - \pi_j + \alpha_{ij} - \delta_{ij} \ge -c_{ij}$ для всех (i, j) $\in$ S: π_i не имеют ограничений по знаку для всех $i \in N$; $\alpha_{ij} \ge 0$ для всех (i, j) $\in$ S; $\delta_{ij} \ge 0,$ для всех (i, j) $\in$ S.

В данной формулировке переменные π соответствуют ограничениям, описывающим условие сохранения потока для прямой задачи, и могут принимать произвольные значения. Переменные α в двойственной задаче соответствуют ограничениям сверху на потоки по дугам в прямой задаче, а переменные δ – ограничениям снизу. Каждой переменной f_{ij} в прямой задаче соответствует некоторое ограничение в двойственной задаче.

При работе алгоритма "дефекта" определяются значения π_i и f_{ij}, для которых выполнены условия оптимальности: k1: если $\overline{c_{ij}} < 0$, то $f_{ij} = U_{i,j}$; k3: если $\overline{c_{ij}} = 0$, то $U_{ij} \ge f_{ij} \ge L_{ij}$; k2: если $\overline{c_{ij}} > 0$, то $f_{ij} = L_{ij}$; k4: условие сохранения потока. Здесь $\overline{c_{ij}} = c_{ij} + \pi_i - \pi_j$.

Если одно из условий k_1, k_2, k_3 нарушено, то дуга называется "дефектной". Решение является оптимальным, если устранены дефекты всех дуг и выполнено условие сохранения потока (условие k_4). Если же таких потоков по дугам не существует, то допустимого решения задачи не существует.

Работу алгоритма можно начать, приписывая дугам произвольные потоки, удовлетворяющие условию сохранения, а узлам – произвольные величины π_i. Проверяя состояния дуг, можно изменять потоки по ним до тех пор, пока не будут выполнены условия оптимальности. По значению

$\overline{c_{ij}}$ можно однозначно определить, является ли дуга дефектной или нет, а также выявить, что нужно делать – увеличивать или уменьшать поток по дуге, для того, чтобы она перестала быть дефектной. Однако не все так просто, как кажется, поскольку условия оптимальности включают в себя также условие сохранения потока. Поэтому если поток по некоторой дуге (i, j) увеличивается (или уменьшается), то в инцидентных ей узлах нарушается условие сохранения потока. А для того чтобы оно не нарушалось, необходимо найти *другой* путь из узла j в узел i (или из i в j, если поток уменьшается). Дуга (i, j) и этот путь из j в i вместе образуют цикл. Изменение потока по циклу не влияет на сохранение потока ни для какого узла. Путь из j в i следует выбрать так, чтобы, во-первых, ни одна бездефектная дуга не стала бы дефектной и, во-вторых, ни одна из дефектных дуг не стала бы более дефектной [4]. Процедура, позволяющая определить каким образом следует изменить потоки по всем рассматриваемым дугам и какой путь следует выбрать, называется *процедурой расстановки пометок*.

После завершения работы алгоритма "дефекта", при наличии оптимального решения каждой дуге соответствует оптимальное (максимальное) значение потока. Далее выполняется разрез графа на подграфы (кластеризация признаков) с помощью теоремы о максимальном потоке и минимальном разрезе, которая формулируется в [5] следующим образом: для любой сети с ограничениями с одним источником (s) и одним стоком (t) величина максимального потока от источника к стоку равна величине минимального разреза, т.е., если для некоторого потока величиной F и некоторого разреза $V_{st} = (N_c, \overline{N_c})$, где $s \in N_c, t \in \overline{N_c}$ выполнено равенство $F = V_{st}$, то данный поток является максимальным [5]. При этом выполняется следующее равенство:

$$\sum_{i \in N_c} \sum_{j \in \overline{N_c}} f_{ij} = \sum_{i \in N_c} \sum_{j \in \overline{N_c}} U_{ij} .$$

Таким образом, разрез выполняется по минимальной сумме пропускных способностей (значений коэффициента парной корреляции) дуг, принадлежащих разрезу, т.е. граф разбивается на подграфы при минимальной связи вершин между ними и максимальной связи вершин внутри подграфов. Так выполняется кластеризация вершин. После разбивки исходного графа на 2 подграфа, к каждому из подграфов применяется кластеризация для получения подграфов второго уровня и т.д., в результате получается иерархическая структура подграфов, в каждом из которых собраны коррелированные признаки $\{x^i_l\}_G$. Окончательным этапом кластеризации является выделение в группе $\{x^i_l\}_G$ коррелированных признаков в подграфах последнего уровня иерархии наиболее информативного признака $\{x^i_l\}_G \rightarrow x^i_l$ (снижение

размерности пространства признаков) или замена группы интегральным диагностическим признаком $\{x^i_l\} \rightarrow y^i_l$ (переход к новому пространству признаков). Полученные в результате кластеризации диагностические признаки используются для построения диагностических решающих правил [1].

Выбор источника и стока выполняется при минимальной связи между вершинами с учетом экспертной оценки в соответствии со значением приведенного коэффициента

$$K_{\text{прij}} = K_{ij} \cdot r_{ij},$$

где K_{ij} – коэффициент парной корреляции; r_{ij} – экспертная оценка в баллах.

При $K_{\text{прij}} \rightarrow \min$, вершина i является источником, а вершина j – стоком.

Итак, разработан подход структурной идентификации диагностических признаков на основе алгоритма ”дефекта”, позволяющий выполнять кластеризацию. Разработана программная реализация алгоритма кластеризации на языке высокого уровня C++. В настоящее время проводится тестовая проверка работы алгоритма применительно к реологическим показателям организма и набор статистики. Предполагается использовать алгоритм применительно к показателям иммунной системы организма.

Выводы. Показана адаптация алгоритма ”дефекта” и теоремы о максимальном потоке и минимальном разрезе к задаче структурной идентификации диагностических признаков.

ЛИТЕРАТУРА

1. *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.*
2. *Дюк В.А. Компьютерная психодиагностика. – С.-Пб.: Братство, 1994. – 364 с.*
3. *Баран Н.В. Архитектура базы клинических и клинико-лабораторных данных пользователей ВДТ // Інформаційні технології: наука, техніка, технологія, освіта, здоров'я. – Х.: НТУ ”ХПІ”. – 2002. – С. 340 – 341.*
4. *Поворознюк А.И. Синтез моделей биологических объектов на основании декомпозиции структур // Вестник НТУ «ХПИ». – Х.: НТУ «ХПИ». – 2001. – Вып. 4. – С. 213 – 215.*
5. *Филлипс Д., Гарсиа-Диас А. Методы анализа сетей: Пер. с англ. – М: Мир, 1984. – 648 с.*

Поступила 7.05.2003

***БУДЯНСКАЯ Элеонора Николаевна**, канд. мед. наук, ст. научн. сотр., зав. лаб. гигиены труда и профессиональных заболеваний. Область научных интересов – разработка методов и алгоритмов построения компьютерных систем медицинской диагностики.*

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

***МАКСЮТА Наталья Валерьевна**, магистр НТУ "ХПИ".*