УДК 681.5

ОЦЕНКА ИНФОРМАТИВНОСТИ ДИАГНОСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ В КОМПЬЮТЕРНЫХ СИСТЕМАХ МЕДИЦИНСКОЙ ДИАГНОСТИКИ

к.т.н. А.И. Поворознюк, Т.В. Гуторова
(представил д.т.н., проф. В.Д. Дмитриенко)

Проведен анализ диагностической ценности симптомов в компьютерных системах медицинской диагностики. Рассмотрены вопросы оценки информативности признаков. Предлагается теоретико-информационный подход к определению диагностической значимости медицинских показателей и формированию системы информативных диагностических признаков на основании количественного выражения их значимости. Рассматривается процедура реализации предлагаемого подхода.

Постановка проблемы. Параллельно с познаванием функций и особенностей взаимодействия живого организма с окружающей средой, разработкой и внедрением в клиническую практику все новых и новых диагностических методов, осуществляется поиск ранее неизвестных симптомов заболеваний и происходит изучение их патогенеза. На сегодняшний день при диагностике с использованием вычислительной техники довольно часто приходится сталкиваться с обработкой очень большого количества показателей состояния больного. Поэтому использование алгоритмов компьютерной медицинской диагностики во многих случаях требует предварительного отбора информативных признаков.

Математическая обработка исходных медицинских данных позволяет определить диагностическую ценность показателей или их комплексов, что в дальнейшем помогает построить оптимальный план обследования, и существенно снизить количество необходимых для диагностики исследований и повысить в конечном итоге качество компьютерного диагноза.

Анализ литературы. Предварительная обработка диагностической информации в компьютерных диагностических системах основывается на формализации исходных признаков и выделении пространства диагностически ценных признаков. В [1 – 7] рассмотрены математические методы оценки информативности диагностических признаков. При этом исполь-

зуются традиционные методы, основанные на дисперсионном, регрессионном, корреляционном анализе [1], теоретико-информационный подход, основанный на вычислении условных вероятностей и количества информации [2 – 4, 6, 7], многомерный статистический анализ, который, как показано в [3], эффективен только при комплексном применении разных методов и относительно большом числе рассматриваемых параметров, методы синтеза иерархической структуры диагностических признаков [5].

В работе рассматривается теоретико-информационный подход, который является наиболее строгим и формализованным и может служить основой для построения более сложных методов. В классическом варианте метод применим для оценки простых (дихотомических) признаков, которые могут принимать только два значения (наличие или отсутствие признака).

Однако, при постановке диагноза, используется разнородная диагностическая информация, которая включает дихотомические, ранговые (принимают несколько значений) и численные (результаты инструментальных измерений) признаки.

Целью работы является разработка процедуры оценки информативности разнородных диагностических показателей на основе теоретико-информационного подхода.

Оценка информативности простых признаков. Для определения ценности полученных результатов обследования пациента на основе анализа простых независимых диагностических признаков целесообразно использовать основные положения теории информации [8], одним из которых является оценка количества информации.

Допустим, что существует некоторая система диагнозов D, которая состоит из n заболеваний. На основании статистической информации и исходя из медицинских данных еще до проведения обследования пациента можно вычислить априорные вероятности появления того или иного заболевания $P(D_i)$. Фактически такая вероятность будет отражать частоту встречаемости каждого диагноза в обрабатываемой выборке.

Неопределенность системы возможных диагнозов оценивается с помощью энтропии или количества информации [8]:

$$H(D) = -\sum_{i=1}^{n} P(D_i) \cdot \log_2 P(D_i), \qquad (1)$$

где H(D) – мера неопределенности (энтропия) системы диагнозов; $P(D_i)$ – априорная вероятность диагноза D_i.

Величина H также называется содержательностью и всегда является положительной величиной. При этом для n возможных равновероят-

ных составляющих ее значение будет максимальным, а формула (1) примет следующий вид:

$$H(D) = -\sum_{i=1}^{n} P(D_i) \cdot \log_2 P(D_i) = -\sum_{i=1}^{n} \frac{1}{n} \cdot \log_2 \frac{1}{n} = \log_2 n. \qquad (2)$$

Так как энтропия отражает меру неопределенности системы, ее величина будет изменяться при поступлении в систему новой информации. Такой информацией для диагнозов являются данные, полученные в результате обследования пациента. Уменьшение энтропии происходит на величину, равную количеству внесенной информации. Крайнее значение, которое может принимать энтропия, равняется нулю и имеет место для достоверного события. В этом случае ноль показывает отсутствие неопределенности в системе.

Соответственно количество поступившей в систему информации определяется как разница между величиной энтропии до и после обследования

$$Z_D(k_j) = H(D) - H(D/k_j), \qquad (3)$$

где $Z_D(k_j)$ – количество информации, внесенной в систему после проведения обследования пациента на признак k_j; $H(D)$ – начальная (первичная) энтропия системы диагнозов; $H(D / k_j)$ – энтропия системы после проведения обследования с учетом признака k_j.

Таким образом, величина $Z_D(k_j)$ характеризует диагностическую ценность симптома k_j по отношению к системе диагнозов D и основывается на количестве поступившей информации. Единицей измерения диагностической ценности признака или комплекса признаков является информация, внесение которой устраняет неопределенность при N равновероятных диагнозах. Диагностическая ценность простого признака, принимающего одно из двух возможных значений, определяется по формуле

$$Z_{D_i}(k_j) = \log_2 \frac{P(k_j / D_i)}{P(k_j)}, \qquad (4)$$

где $Z_{Di}(k_j)$ – диагностический вес признака k_j для заболевания D_i; $P(k_j / D_i)$ – частота встречаемости или априорная вероятность наличия признака при заболевании D_i; $P(k_j)$ – частота встречаемости или априорная вероятность наличия признака во всей системе возможных диагнозов D.

Величина $P(k_j / D_i)$ рассчитывается как отношение количества пациентов, у которых присутствует признак k_j при заболевании D_i, к общему числу пациентов с рассматриваемым заболеванием

$$P(k_j / D_i) = \frac{\sum_{k_j=1} k_j(D_i)}{\sum_{D_i=1} D_i}. \quad (5)$$

На основании (4) можно сделать вывод, что при одинаковом значении вероятностей наличия признака для конкретного заболевания и для всей системы диагнозов диагностический вес признака равен нулю и признак не несет никакой информативности.

Диагностический вес отсутствия простого признака определяется с помощью выражения, которое получается из формулы (4) путем внесения обратных величин вероятностей

$$Z_{D_i}(\overline{k_j}) = \log_2 \frac{1 - P(k_j / D_i)}{1 - P(k_j)}. \quad (6)$$

Следует учитывать, что диагностический вес признака может быть как положительной, так и отрицательной величиной, то есть как уменьшать, так и увеличивать вероятность того или иного диагноза.

Полный диагностический вес простого признака для заболевания D_i учитывает как наличие, так и отсутствие признака, и может быть рассчитан по выражению

$$Z_{D_i}(k_j) = P(k_j / D_i) \cdot \log_2 \frac{P(k_j / D_i)}{P(k_j)} + [1 - P(\overline{k}_j / D_i)] \cdot \log_2 \frac{1 - P(k_j / D_i)}{1 - P(k_j)}. \quad (7)$$

Диагностическая ценность простого признака для системы заболеваний

$$Z_D(k_j) = \sum_{i=1}^{n} P(D_i) \cdot Z_{D_i}(k_j). \quad (8)$$

Оценка информативности сложных признаков. К сложным признакам относятся ранговые, значение которых можно выразить конечным числом интервалов и числовые, которые тоже можно выразить конечным числом интервалов, так как любые измерения выполняются с конечной точностью. Если рассматривать каждый интервал (диагностический разряд) сложного признака в качестве простого признака, то диагностическая ценность s-го интервала сложного признака по (4) запишется в виде

$$Z_{D_i}(k_{js}) = \log_2 \frac{P(k_{js} / D_i)}{P(k_{js})}, \quad (9)$$

где $P(k_{js} / D_i)$ – частота встречаемости (априорная вероятность) s-го диагностического интервала сложного признака для диагноза D_i.

Величина

$$Z_{D_i}(k_j) = \sum_{s=1}^{m} P(k_{js}/D_i) \cdot \log_2 \frac{P(k_{js}/D_i)}{P(k_{js})} \qquad (10)$$

определяет диагностическую ценность сложного признака для диагноза D_i.

Для определении полной диагностической ценности сложного признака применяется формула

$$Z_D(k_j) = \sum_{i=1}^{n}\sum_{s=1}^{m} P(D_i) \cdot P(k_{js}/D_i) \cdot \log_2 \frac{P(k_{js}/D_i)}{P(k_{js})}. \qquad (11)$$

При определении информативности по (9 – 11), особенно при большом m, кроме увеличения сложности вычислений, предъявляются повышенные требования к объему и репрезентабельности обучающей выборки, что ограничивает практическое применение указанного подхода. При практическом применении теоретико-информационного подхода к реальным медицинским базам данных, авторы обнаружили следующие особенности:

1) для некоторых простых признаков диагностические веса по (7) принимают предельные значения, что свидетельствует о наличии детерминистической связи;

2) в случае сложных числовых признаков многие интервалы оказались пустыми, т.е. в (9) $P(k_{js}) = 0$.

С учетом отмеченных особенностей предлагается описанная ниже процедура оценки информативности.

Процедура оценки информативности разнородных диагностических признаков включает следующие этапы:

1) Признаки разбиваются на три группы – простые (дихотомические), ранговые (интервальные) и числовые.

2) Для числовых признаков определяется динамический диапазон изменения: $Y_{max} - Y_{min}$.

3) Динамический диапазон равномерно разбивается на L интервалов.

4) В каждом из интервалов подсчитывается априорная вероятность $P(k_{js})$, которая сравнивается с пороговым значением $P_{пред}$.

В зависимости от выполнения условия

$$P(K_{JS}) \geq P_{ПРЕД}\ , \qquad (12)$$

интервал помечается действительным – (12) выполняется, или недействительным – (12) не выполняется.

5) Все недействительные интервалы объединяются с рядом расположенными действительными (идущие подряд недействительные интервалы присоединяются к разным действительным интервалам с разных

сторон). Процедура объединения интервалов продолжается до тех пор, пока все интервалы станут действительными, т.е. для всех интервалов выполняется условие (12). Указанная процедура объединения интервалов обеспечивает разбивку динамического диапазона на неравномерные, статистически значимые интервалы и подсчет их априорных вероятностей для всей системы диагнозов. Так как вычислительная сложность подсчета априорных вероятностей для (12) невелика (накопление суммы), то начальное значение количества интервалов L можно взять "с запасом", что не влияет на конечный результат.

6) Расчет и проверка по (12) выполняется для интервалов ранговых признаков, которые, при необходимости, тоже могут объединяться.

7) Аналогичные расчеты и проверка по (12) выполняются и для простых признаков, но в случае невыполнения условия (12), признак исключается из системы признаков с выдачей сообщения "По признаку k в базе данных недостаточно информации".

8) Рассчитываются априорные вероятности диагнозов и определяется энтропия системы диагнозов по (1).

9) Рассчитываются условные вероятности наличия признаков (интервалов) для каждого из диагнозов D_i и проверяется условие

$$P(K_{JS}/D_I)\cdot P(D_I) \geq R_1\cdot P(K_{JS}), \qquad (13)$$

где r_1 – весовой коэффициент, который выбирается из диапазона (0,85 – 0,95).

При выполнении условия (13), фиксируется наличие детерминистической связи (диагноз D_i полностью определяется признаком (интервалом) k_{js}), и отмеченные данные (диагноз и признак) исключаются из базы данных с целью уменьшения объема дальнейших вычислений.

10) Рассчитывается диагностическая ценность признаков для системы диагнозов по (8) или (11) и признаки ранжируются по убыванию.

11) Рассчитывается суммарное количество поступившей в систему информации по

$$H(Z_\Sigma) = \sum_i Z_D(k_i), \qquad (14)$$

где $Z_D(k_i)$ вычисляются по выражению (8) или (11).

12) Проверяется условие

$$H(Z_\Sigma) \geq r_2 \cdot H(D), \qquad (15)$$

где r_2 – весовой коэффициент, который выбирается из диапазона (0,6 – 0,8).

При невыполнении условия (15) считается, что существующая система диагностических признаков на данной обучающей выборке вносит недостаточно информации для постановки достоверного диагноза, происходит отказ от дальнейшей обработки с выдачей соответствующего

заключения. Если условие (15) выполняется для всей группы признаков, то группа начинает уменьшаться путем отбрасывания малоинформативных признаков в упорядоченном ряду до тех пор, пока выполняется условие (15).

Выводы и рекомендации. Изложенная методика может быть использована как при задачах отбора диагностически ценных симптомов для целых клинических направлений в медицине (групп заболеваний), так и при диагностике пациентов. При применении методики в конкретной предметной области необходима адаптивная настройка пороговых и весовых коэффициентов.

ЛИТЕРАТУРА

1. *Дюк В.А. Компьютерная психодиагностика. – С-Пб.: Братство, 1994. – 364 с.*
2. *Весненко А.И., Попов А.А., Проненко М.И. Топо-типология структуры развернутого клинического диагноза в современных медицинских информационных системах и технологиях // Кибернетика и системный анализ. – 2002. – № 6. – С. 143 – 154.*
3. *Ахутин В.М., Шаповалов В.В., Иоффе М.О. Оценка качества формализованных медицинских документов // Медицинская техника. – 2002. – Вып. 2. – С. 27 – 31.*
4. *Величко О.Н., Мустецов Н.П. Формализация качественных знаний в медицинских экспертных системах // Вестник НТУ "ХПИ". – Х.: НТУ "ХПИ". – 2003. – № 19. – С. 26 – 33.*
5. *Поворознюк А.И. Синтез иерархической структуры диагностических признаков в компьютерных системах медицинской диагностики // Вестник НТУ "ХПИ".– Х.: НТУ "ХПИ". – 2003. – № 7, Т. 2. – С. 39 – 44.*
6. *Максимов Г.К., Синицын А.Н. Статистическое моделирование многомерных систем в медицине. – Л.: Медицина, 1983. – 144 с.*
7. *Постнова Т.Б. Информационно-диагностические системы в медицине. – М.: Наука, 1972. – 376 с.*
8. *Кузьмин И.В., Кедрус В.А. Основы теории информации и кодирования. – К.: Вища шк., 1986. – 268 с.*

Поступила 22.04.2004

***ПОВОРОЗНЮК Анатолий Иванович**, канд. техн. наук, доцент, проф. кафедры вычислительной техники и программирования НТУ "ХПИ". В 1977 году окончил Харьковский политехнический институт. Область научных интересов – разработка методов и алгоритмов построения компьютерных систем медицинской диагностики.*

***ГУТОРОВА Татьяна Викторовна**. В 2004 году окончила НТУ "ХПИ". Область научных интересов – разработка алгоритмов и программ реализации компьютерных систем медицинской диагностики.*