Проблемы бионики
1998. Вып. 48. С. 60—63

УДК 519.7

В.А. МУСИЙЧЕНКО, Н.П. МУСТЕЦОВ

ВЫЧИСЛИТЕЛЬНЫЙ МЕТОД ОБРАБОТКИ КОРРЕЛЯЦИИ МЕЖДУ СИМПТОМАМИ В МЕДИЦИНСКИХ БАЗАХ ЗНАНИЙ

База знаний экспертной системы с вероятностной логикой содержит условные вероятности определенного симптома при определенной болезни. Симптомы считаются независимыми друг от друга там, где зависимость не указана явно. Неявная зависимость обнаруживается корреляционными методами. Если ее не учитывать, то, в предельном случае, задавая один и тот же вопрос пациенту (коэффициент корреляции в этом случае равен единице), в конечном счете, получим равную единице вероятность болезни, для которой этот симптом специфичен, но не определяет её однозначно, хотя после получения первого ответа никакой информации в систему не поступило.

Разные авторы по-разному подходят к решению этой проблемы. Наиболее распространена такая методика: "При коэффициенте корреляции или корреляционном отношении (в случае нелинейной зависимости рассматриваемых показателей) $r \geq 0.7$ два параметра заменяются обобщённым симптомом или же выбирается один из них. Если $0.3 \leq r \leq 0.7$, то для уменьшения погрешности внимание обращается лишь на экстремальные значения каждого из показателей, сопоставленных с возможной величиной второго. Наконец, при $r < 0.3$ считаем симптомы невзаимосвязанными и подлежащими дальнейшему изучению" [1, с.54].

Перед проведением корреляционого анализа "на основании эмпирических соображений (врачебного опыта) составляется сознательно завышенный перечень признаков, которые могут иметь значение в оценке тяжести состояния больного и риска намечаемого лечения" [1, с.54].

Необходимо отметить, что "завышенный перечень признаков" часто составляется несознательно и поэтому всегда есть потребность в обработке корреляции.

Предлагаемый метод позволяет проводить такую обработку автоматически. Теоретической основой метода являются статистические методы обработки распределений [2], теория вероятностей.

Коэффициент корреляции между симптомами S_k, S_l вычисляется по формуле таким образом:

$$r_{k,l} = \frac{\mathrm{cov}(x_k, x_l)}{\delta_k \delta_l}; \qquad (1)$$

$$\mathrm{cov}(x_k, x_l) = \frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l); \quad (2)$$

$$d_k = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2}, \quad (3)$$

где $x_{ik} = P(S_k / D_i)$ — условная вероятность k-го симптома при i-й болезни;

$\bar{x}_k = \bar{P}(S_k|D)$ — средняя условная вероятность k-го симптома при обрабатываемом множестве болезней;

n — количество болезней в множестве;

δ_k — среднеквадратическое отклонение степени для вероятности k-го симптома;

$\mathrm{cov}(x_k, x_l)$ — взаимная ковариация вероятностей симптомов;

x_k — вариация k-го симптома.

Следующий шаг — введение коэффициента корреляции в формулу Байеса [1] для коррекции вероятности заболевания — уменьшение изменения этой вероятности в случае высокой корреляции (как положительной, так и отрицательной). Предлагается следующее эмпирическое решение:

$$P(D/S_1 \ldots S_k) = \frac{P(D) * \prod_{i=1}^{k} P(S_1/D)}{\prod_{i=1}^{k} \left(P(S_1/D) + \left(\prod_{j=1}^{i=1} (1 - r_{ji}^2) \right) * P(\bar{D}/S_1 \ldots S_{k-1}) * (P(S_i/\bar{D}) - P(S_i/D)) \right)}, \quad (4)$$

где $P(D|S_1 \ldots S_k)$ — условная вероятность диагноза D при условии выбора симптомов $S_1 \ldots S_k$;

$P(D)$ — априорная вероятность диагноза D;

$P(\bar{D}\,|S_1..S_k)$ — условная вероятность отсутствия диагноза D при условии выбора симптомов $S_1..S_k$,

$P(S_i|\bar{D})$ — условная вероятность i-го симптома S при отсутствии диагноза D.

Формула (4) отличается от формулы Байеса наличием множителя $\prod_{j=1}^{i-1}(1-r_{ji}{}^2)$ во 2-м слагаемом знаменателя. Этот множитель позволяет учесть корреляцию между симптомами. Если коэффициент корреляции будет равен нулю, то формула (4) превращается в формулу Байеса, где события S_i считаются независимыми. Если r=1 хотя бы с одним из предыдущих симптомов, множитель превращается в 0 и значение $P(D|S_1...S_k)$ не изменяется по отношению к предыдущему своему значению $P(D|S_1...S_{k-1})$.

В случае когда r<0, действие формулы аналогично изложенному выше, т.е. осуществляется уменьшение изменения вероятности заболевания, но отрицательный коэффициент корреляции говорит о противоречии в базе знаний, о чем в некоторых случаях целесообразно извещать пользователя.

Поскольку учёт корреляции в формуле (4) на интервале $r \in]0;1[$ не является строгим, множитель, учитывающий корреляцию, может быть представлен в другом виде: $\prod_{j=1}^{i-1}(1-|r_{ij}|)$ или $\prod_{j=1}^{i-1}(1-r_{ij}^2)^{1/2}$, однако избрана вышеприведенная форма ввиду её согласованности с формулой (5) зависимости коэффициента множественной корреляции от коэффициентов частной(парной) корреляции [4]:

$$1-R_k^2=\prod_{j=1}^{k-1}(1-r_{kj}^2) \qquad (5)$$

Коэффициенты парной корреляции могут быть вычислены на этапе создания базы знаний и храниться в ней. Они могут служить критерием отбора вопросов.

Рассмотрим применение формулы (1) на фрагменте реальной медицинской базы знаний [3].

База знаний предназначена для диагностики заболеваний желудка. Допустим, пользователь выбрал первый симптом из группы взаимоисключающих симптомов:

- ригидность складок слизистой есть;
- ригидность складок слизистой отсутствует;

и первый симптом из группы:

- пониженный тонус желудка;
- нормальный тонус желудка;
- повышенный тонус желудка.

Условные вероятности симптомов (%)

Симптом	Номер диагноза																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Ригидность складок	13	05	13	10	16	06	03	02	02	03	06	60	04	20	13	03	05
Пониженный тонус	16	08	32	32	25	50	06	05	06	08	68	80	10	16	13	03	08

Избранные симптомы имеют следующие условные вероятности при каждой из 17 диагностируемых болезней (см. таблицу):

Коэффициент парной корреляции между симптомами, вычисленный по формулам (1)—(3), r=0.53.

Значение $P(D|S)$ для 1-го симптома вычисляем по теореме Байеса, т.е. по формуле(4) при r=0. Учитываем при этом формулы (5)—(7).

$$P(S|\bar{D}) = \frac{P(S) - P(S|D) * P(D)}{1 - P(D)}, \tag{5}$$

где $P(S)$ — априорная полная вероятность симптома S;

$$P(S) = \sum_{i=1}^{k} P(D_i) P(S|D_i); \tag{6}$$

$$\sum_{i=1}^{k} D_i = 1. \tag{7}$$

Например, для болезни №11 — "стенозирующая язва" — при априорной вероятности $P(D)$=0.1 условная вероятность диагноза $P(D|S_1)$=0.293. Значение $P(D|S_1S_2)$ вычисляем уже с учётом корреляции второго симптома с первым: $P(D|S_1S_2)$=0.352. Вычисление по теореме Байеса, т.е. без учёта корреляции, даёт результат, больше отличающийся от $P(D|S_1)$, $P(D|S_1S_2)$=0.382.

При наличии большого количества вопросов в базе знаний накопление незначительных погрешностей может приводить к существенным изменениям конечного результата, что доказывает необходимость учёта корреляции.

Предлагаемый вычислительный метод может быть полезен при обработке больших массивов информации, где коррекция данных экспертом требует больших затрат.

Список литературы: 1. *Кибернетика* в сердечной хирургии / Минцер О.П., Кнышов Г.В., Цыганий А.А. К.: Вища шк. Головное изд-во, 1984. 140 с. 2. *Павловский З.* Математическая статистика. М.: Статистика, 1969. 3. *Вишневский А.А., Артоболевский И.И., Быховский М.Л.* Машинная диагностика и информационный поиск в медицине. М.: Наука 1969. 242 с. 4. *Терехов Л.Л.* Экономико-математические методы. М.: Статистика. 1968. 300 с.

Поступила в редколлегию 30.10.97