

А.С. ЗАМКОВОЙ, канд. техн. наук, ХНУРЭ (г. Харьков),
Т.Н. МУСТЕЦОВ, ХНУРЭ (г. Харьков)

УЧЕТ ЛИНГВИСТИЧЕСКИХ ПЕРЕМЕННЫХ ПРИ СОЗДАНИИ СИСТЕМ ОБЪЕКТИВНОЙ МЕДИЦИНСКОЙ ДИАГНОСТИКИ

Розглянуто питання оптимізації кількості і розподілу по числовій шкалі значень лінгвістичних змінних при створенні систем медичної діагностики. Запропоновано вираження для обчислення імовірності діагнозу з урахуванням кореляції між показниками.

Questions of optimization of quantity and distribution on a numerical scale of linguistic variables are considered at creation of systems of medical diagnostics. Expression for calculation of probability of the diagnosis in view of correlation between parameters is offered.

Постановка проблемы. Медицинская диагностика – распознавание заболеваний представляет собой процесс познания двух видов. Во-первых, врач старается узнать (вспомнить) то, что он когда-то видел. В этом случае совершается процесс простого узнавания, чем познание в полном смысле слова. Второй вид процесса диагностики заключается в познании врачом нового, ещё не известного, т. е. вскрытие новых связей, причинных отношений между симптомами и установление в конкретном случае условий происхождения заболевания. При этом врач генерирует новую информацию. Новая информация – производная продуктивного мышления. Продуктивное мышление есть творчество. И.А. Кассирский [1], выражая свои сомнения в возможностях компьютерной диагностики, писал: "как запрограммировать машине умение дать диагностическую оценку выражения лица, глаз больного, его походке, манере вести себя и разговаривать, его скрытым переживаниям, смутно ощущаемым врачом". Чем менее уместны алгоритмы (репродуктивное мышление), тем более уместно творчество (продуктивное мышление). Средствами искусственного интеллекта обеспечиваются функции доступного формализации репродуктивного мышления, но не обеспечиваются функции продуктивного, творческого, "которые не могут быть формализованы".

В настоящее время многие функции человека переданы технике, и человек используется для большей её отдачи. Что же касается интеллектуальной функции генерирования новой информации, творчества, то самым совершенным средством её осуществления является мозг человека, и здесь главной задачей проектирования человеко-машинных комплексов, направленных на осуществление этих функций, является создание техники для большей отдачи от мышления человека.

В частности, в отношении медицинской диагностики, важно отметить, что классификация болезней, описания симптомов, синдромов, нозологических единиц проведены человеком, следовательно, человеком постижимы и, следовательно, проблема диагностики есть проблема понимания между людьми, между авторами классификаций и врачами с их индивидуальным опытом. Здесь проявляется роль личности, индивидуальности в медицине – классификация может быть разработана самим практикующим врачом, в большей или меньшей мере расходящаяся с существующей, формируя основы новой клинической школы. Одним из путей выхода из ситуации, является создание средств, способствующих появлению, поддержанию и стимулированию продуктивного мышления врача.

Анализ литературы. На практике для решения медицинских задач применяют, в основном, два типа интеллектуальных систем: информационно-поисковые и экспертные. Причем некоторые из них достаточно эффективны. Например, в системе «Компас» используется метод морфологического анализа и производится компьютерное моделирование творческих процедур. Авторы [2] утверждают, что с помощью этой системы было получено более 60 патентов на новые механизмы принятия решения.

Система Sounding Board – часть образовательного программного комплекса, пытающегося обучить пользователя выработке творческих решений задач через обучение его задавать себе наводящие вопросы. Система Sounding Board не является генератором новых решений, но объединение Sounding Board и пользователя может быть более творческим, чем пользователь сам по себе. Sounding Board может помочь пользователю объединить факты и правила для творческого использования при диагностике [3]. Система Sounding Board основана на концепции тренера, использует метод контрольных вопросов. Система Sounding Board воплощает теорию повышения творческой активности, изложенную в книге "The Creative Attitude" ("Творческие способности"). Основное утверждение, стоящее за теорией, равно как и за программой, основано на том, что решение творческих задач зависит от способности иметь мышление, "систематически блуждающее" в разных направлениях (аспектах) задачи, и что задавание себе стандартного набора вопросов может улучшить способность делать это.

Система Case-Adaptation Coach объединяет идеи системы ABE [3] в плане использования базы знаний конкретных ситуаций и механизмов их адаптации к конкретному примеру, и системы Sounding Board в плане использования концепции тренера и метода вопросов. На практике в большинстве существующих медицинских интеллектуальных системах используются только формализованные данные [4]. Результат работы таких систем служит только для получения вспомогательной "осведомительной" информации. Повышение эффективности медицинских систем принятия решений происходит за счет формализации и учета качественных знаний.

Цель статьи. Формализация качественных знаний (данных) на основе теории информации и определение метода их учета при создании медицинских систем принятия решения.

Формализация качественных знаний. Распределение лингвистических переменных по цифровой шкале. Особенностью лингвистических переменных используемых в медицине является полное отсутствие их числового значения. Кроме того, набор значений переменной часто зависит от субъективного восприятия врача. Использование теории нечетких множеств Л. Заде для решения задач медицинской диагностики возможно в случае четкого определения числовых соответствий таким понятиям естественного языка как "всегда – часто – иногда – редко – никогда". Однако наличие психологического и лингвистического аспектов при формировании симптомокомплекса оставляет слабо разработанными следующие вопросы:

- необходимое и достаточное количество значений лингвистических переменных для описания непрерывных свойств объекта от полного отсутствия свойства до максимального присутствия;
- распределение значений лингвистических переменных по цифровой шкале, измеряющей данное свойство.

Рассмотрим принципы учета значений лингвистических переменных по шкале вероятностей в системах медицинской диагностики. Области, в которых человек производит, в основном несознательно, градацию (шкалирование) разнообразны и трудно описываемы математически. Доказательство тому – существование в психологии двух противоречащих друг другу законов: Вебера-Фехнера и Стивенса (рис. 1).

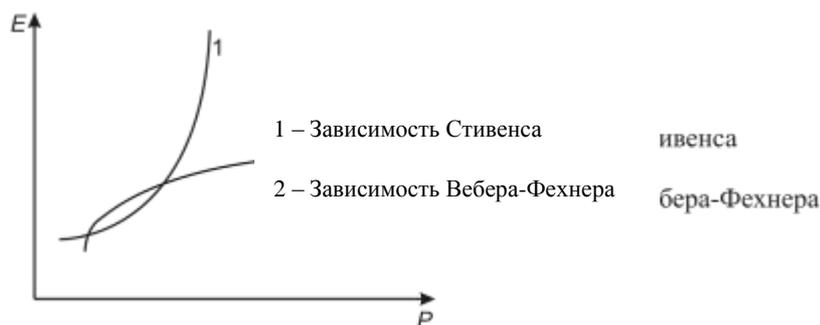


Рис. 1.

Первый утверждает, что зависимость силы ощущения E от физической интенсивности раздражителя P – логарифмическая, т.е. отношение $\Delta E/\Delta P$ с ростом P уменьшается. Второй говорит, что зависимость степенная, а значит отношение $\Delta E/\Delta P$ с ростом P увеличивается. За силой ощущения стоит лингвистическая переменная. Для шкал ощущений характерно наличие верхнего и нижнего порогов ощущений. При этом они могут быть болевыми: оба порога (температура); один порог (сила звука); ни одного, оба порога связаны с прекращением ощущения. С приближением к порогам, в так называемой "пороговой зоне", отношение $\Delta E/\Delta P$ увеличивается, в связи с чем можно предположить, что закон Стивенса справедлив в пороговых зонах, скорее даже Вебера-Фехнера лишь для среднего участка диапазона воспринимаемых значений раздражителя [5]. Например, при лингвистических значениях $L1 =$ боль обморожения, $L2 =$ очень холодно, $L3 =$ холодно, $L4 =$ нормально, $L5 =$ тепло, $L6 =$ жарко, $L7 =$ боль ожога, предполагаемая зависимость может быть представлена в соответствии с рис. 2.

Сложность описания этой области связана также с действием механизмов адаптации, которые являются отражением индивидуальных врождённых качеств и профессиональной деятельности. Исследователи в области искусственного интеллекта, использующие в основном равномерное распределение, также отмечают, что "люди часто искажают оценки, например, сдвигают их в направлении концов оценочной шкалы" [6].

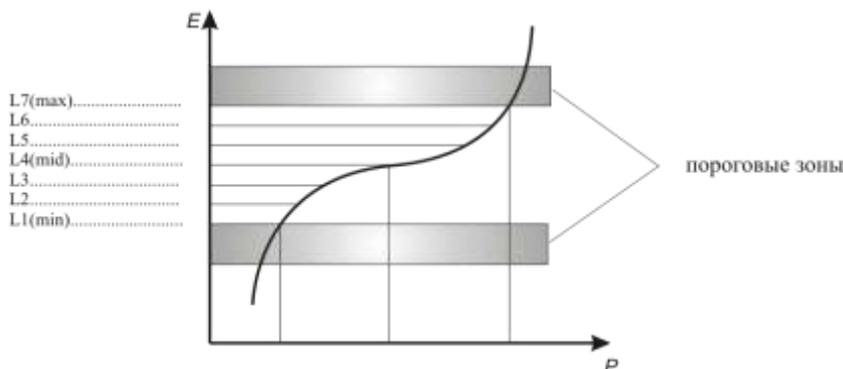


Рис. 2.

Объяснение этого явления возможно на основе вышеизложенных зависимостей. Несмотря на неравномерное распределение значений лингвистических переменных по количественной шкале, они равномерно распределены по шкале информативности: с увеличением информативности ощущений в пороговых зонах возрастает количество описывающих их значений лингвистических переменных. Если описываемое свойство вероятность P , то может быть прямо применён закон К. Шеннона [7]:

$$I = P \log_2 \frac{1}{P} + (1 - P) \log_2 \frac{1}{1 - P},$$

где I – информативность, прямо связанная с лингвистической переменной. По графику формулы (рис. 3) можно определять численные значения переменных, в данном случае вероятности.

Для сравнения, в шкале Харрингтона [8], одной из распространённых методик оцифровки качественной информации, используется следующая градация 0 – 0.2 – 0.37 – 0.64 – 0.8 – 1. Информационно значимые пороги могут располагаться как на краях шкалы, так и в её середине и шкала Харрингтона может рассматриваться как обобщённый вариант, хотя, несомненно, более точным решением будет использование в каждом из двух случаев своей шкалы, одна из которых предлагается.

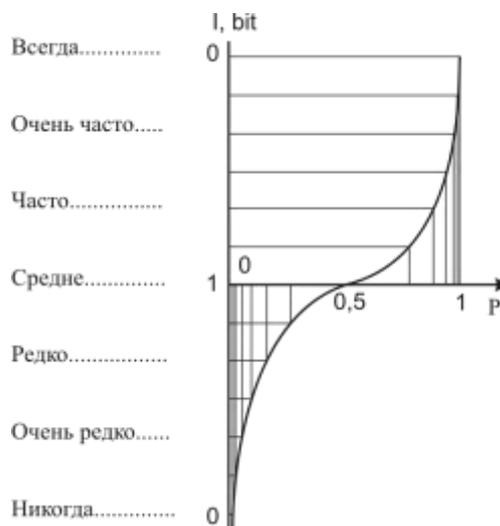


Рис. 3.

Объединение мнений, выраженных в вероятностях P , должно проводиться с учетом неравномерного распределения значений лингвистических переменных $L(P)$:

$$L(P_1, P_2, \dots, P_i, \dots) = L\left(P\left(\frac{\sum I(P_i)w_i}{\sum w_i}\right)\right),$$

где w_i – вес мнения i -го эксперта.

Количество значений лингвистических переменных. Как было указано выше, верхний предел количества составляет 11 – 13. С другой стороны "при определении степени принадлежности множество исследуемых объектов должно содержать, по крайней мере, два объекта, численные представления которых на интервале $[0; 1]$ – 0 и 1 соответственно".

Третья обязательная точка шкалы, обладающая значительно большим разбросом, чем две предыдущие – 0.5. Количество точек между 0 и 0.5 (и симметричном интервале 0.5 ... 1) может варьироваться от 0 до 4 (при 11 делениях). На практике приходится сталкиваться с четвёртой обязательной точкой, лежащей вне шкалы, а точнее, охватывающей всю шкалу – это значение лингвистической переменной "неизвестно", "нет информации". Её легко спутать с переменной "средне" (3-я точка), но у последней область значительно уже.

В приводимых в литературе примерах количество значений лингвистических переменных варьируется в пределах от 3 до 11 [9] с преобладанием в промежуточных значениях 5–7–9. Наиболее часто встречается 7-интервальная шкала. Для большинства случаев в решении проблемы количества приемлемы выводы психолингвиста Джорджа Миллера, данные в книге "Магическое число семь, плюс или минус два: некоторые ограничения на наши возможности обработки информации". Предлагаем количество значений лингвистических переменных ограничить числом семь.

Использование лингвистических переменных при создании медицинских систем. При постановке диагноза врач редко бывает в нём полностью уверен. В связи с этим более целесообразно говорить о диагнозе с точки зрения его вероятности. Всё ещё очень часто эта вероятность выражается не в форме

процентов, а с помощью таких выражений, как "почти всегда", "обычно", "иногда", "редко". Поскольку разные люди вкладывают различную степень вероятности в одни и те же термины, это ведёт к возникновению недопонимания между врачами или врачом и пациентом. Врачам следует как можно более точно и, по возможности, используя цифры, давать свои заключения и, если это осуществимо, использовать для выражения вероятности количественные методы. Например, вместо того чтобы говорить, что вероятность обнаружения карциномы толстой кишки рентгенологическими методами мала, лучше, если это возможно, представить точные данные о диагностике новообразования по данным рентгенографии. Вероятность обнаружения карциномы в 10–15% случаев может быть интерпретирована как "небольшая", однако с клинических позиций в этом случае требуется дальнейшее уточнение диагноза, поскольку последствия гиподиагностики потенциально опережающей опухоли могут быть самыми серьёзными. Хотя наличие таких количественных показателей было бы очень желательно, они обычно отсутствуют в клинической практике. Имеется тенденция к гипердиагностике относительно редких заболеваний.

Выражение "особенно трудно бывает количественно оценить вероятность, которая может быть очень высокой или очень низкой" следует интерпретировать как "особенно трудно разбить значения лингвистических переменных на краях шкалы на ещё более мелкие (это трудно сделать не только на краях шкалы), но особенно важно, поскольку именно на краях шкалы происходит скачок от неопределённости к определённости, т.е. к "0" (нет) или "1" (да), и незначительное изменение вероятности вблизи этих точек приводит к значительному изменению информации". Что касается связи вероятности и информации в середине шкалы, то "например, уменьшение вероятности диагноза коронарной болезни с 75 до 50% может иметь значение для статистика, но для врача это не будет поводом для изменения терапевтической тактики".

Предлагаем использовать вычислительный метод обработки корреляции между показателями. База знаний экспертной системы с вероятностной логикой содержит условные вероятности определённого симптома при определённой болезни. Симптомы считаются независимыми друг от друга там, где зависимость не указана явно. Перед проведением корреляционного анализа "на основании эмпирических соображений (врачебного опыта) составляется сознательно завышенный перечень признаков" [4]. Необходимо отметить, что "завышенный перечень признаков" часто составляется несознательно и поэтому всегда есть потребность в анализе корреляции.

Предлагаемый паллиативный метод позволяет проводить такую обработку автоматически. Теоретической основой метода являются статистические методы обработки распределений, теория вероятностей. Коэффициент корреляции между симптомами S_k, S_l вычисляется по формуле:

$$r_{k,l} = \frac{\text{cov}(x_k, x_l)}{\delta_k \delta_l}, \quad \text{cov}(x_k, x_l) = (1/n) \sum_{i=1}^n (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l); \quad \delta_k = \left(\frac{1}{n} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2 \right)^{1/2},$$

где $x_{ik} = P(S_k/D_i)$ – условная вероятность k -го симптома при i -й болезни; $\bar{x}_k = \bar{P}(S_k | D)$ – условная средняя вероятность k -го симптома при обрабатываемом множестве болезней; n – количество болезней в множестве; δ_k – среднее квадратическое отклонение степени для вероятности k -го симптома; $\text{cov}(x_k, x_l)$ – взаимная ковариация вероятностей симптомов; x_k – вариация k -го симптома.

Следующий шаг – введение коэффициента корреляции в формулу Байеса [4] для коррекции вероятности заболевания – уменьшение изменения этой вероятности в случае высокой корреляции (как положительной, так и отрицательной). Предлагается следующее эмпирическое соотношение:

$$P(D | S_1, \dots, S_k) = \frac{P(D) \prod_{i=1}^k P(S_i | D)}{\prod_{i=1}^k (P(S_i | D) + (\prod_{j=1}^{i-1} (1 - r_{ij}^2)) P(\bar{D} | S_1, \dots, S_{k-1}) (P(S_i | \bar{D}) - P(S_i | D)))}, \text{ где } P(D/S_1, \dots, S_k) -$$

условная вероятность диагноза D при условии выбора симптомов S_1, \dots, S_k ; $P(D)$ – априорная вероятность диагноза D ; $P(\bar{D} | S_1, \dots, S_k)$ – условная вероятность отсутствия диагноза D при условии выбора симптомов S_1, \dots, S_k ; $P(S_i | \bar{D})$ – условная вероятность i -го симптома S при отсутствии диагноза D .

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

В случае, когда $r < 0$, действие формулы аналогично вышеизложенному, 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.

Выводы. В работе рассмотрены вопросы учета значений лингвистических переменных при создании медицинских интеллектуальных систем принятия решения. Рассмотрены проблемы учета значений лингвистических переменных при формировании, как входных данных так и при принятии решений. Предлагается использовать неравномерное распределение значений лингвистических переменных по числовой шкале, с учетом их информативности.

Необходимое и достаточное число значений лингвистических переменных равно семи. Предлагается соотношения для вычисления вероятности диагноза, в котором учитывается корреляция между симптомами, что позволит учитывать лингвистические переменные.

Список литературы: 1. Тарасов К.Е., Великов В.К., Фролова А.И. Логика и семиотика диагноза (методологические проблемы). – М.: Медицина, 1989. – 272. 2. Андрейчиков А.Н., Андрейчикова О.Н. Компьютерное моделирование творческих процедур синтеза принципиально новых механизмов (система "Компас") // Программные продукты и системы. – 1995. – № 2. – С. 35 – 37. 3. Kaas Alex. Question asking, artificial intelligence, and human creativity // In: Question and information systems. Edited by Th. W. Lauer. Hillsdale, New Jersey. – 1992. – P. 374. 4. Минцер О.П., Кнышов Г.В., Цыганый А.А. Кибернетика в сердечной хирургии. – К.: Вища школа. – 1984. – 140 с. 5. Психология. Словарь / Под ред А.В. Петровского, М.Г. Ярошевского. – М.: Политиздат, 1990. – 494 с. 6. Искусственный интеллект. – В 3-х кн. Кн. 2. Модели и методы: Справочник / Под ред. Д.А. Поспелова. – М.: Радио и связь, 1990. – 304 с. 7. Кузьмин И.В., Кедрус В.А. Основы теории информации и кодирования. – К.: Вища школа, 1986. – 234 с. 8. Литвак Б.Г. Экспертные оценки и принятие решений. – М.: Патент, 1996. – 271 с. 9. Алиев Р.А. Интеллектуальные работы с нечеткими базами знаний. – М.: Радио и связь, 1994. – 176 с.

Поступила в редакцию 31.03. 2006