УДК 681.518.54

Д.А. НИЦЫН

МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ПРИЗНАКОВ В БАЙЕСОВСКОМ КЛАССИФИКАТОРЕ МЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

Предлагается модель представления диагностических признаков в виде графа, который связывает значения признаков, описывающих состояние здоровья, с числами их возможных сочетаний. Предложенная модель упрощает определение условных вероятностей данных сочетаний диагностических признаков при применении формулы Байеса для классификации медицинских изображений.

Ключевые слова: диагностические признаки, формула Байеса, медицинские изображения.

Постановка проблемы. К решению задач медицинской диагностики, как правило, привлекаются данные, которые являются неполными или неточными. Поэтому наилучшим подходом к классификации медицинских изображений является применение вероятностных методов [1 – 3]. Например, в основу классификации, которая позволяет распознавать состояние здоровья пациентов по рентгеновским изображениям их внутренних органов, можно положить формулу Байеса [1]:

$$P(H_1 / S_k) = \frac{P(H_1)P(S_k / H_1)}{P(H_1)P(S_k / H_1) + P(H_2)P(S_k / H_2)}. \quad (1)$$

Результатом вычислений по формуле Байеса является оценка апостериорной вероятности $P(H_1 / S_k)$ гипотезы о том, что данное медицинское изображение не имеет симптомов заболевания, при условии, что диагностические признаки имеют данное сочетание S_k. При этом $P(H_1)$ – априорная вероятность гипотезы о том, что рентгенограмма не содержит патогенных зон; $P(H_2)$ – априорная вероятность гипотезы о том, что рентгенограмма содержит патогенные зоны; $P(S_k / H_1)$ – условная вероятность данного сочетания диагностических признаков при условии, что рентгеновское изображение не содержит симптомов болезни; $P(S_k / H_2)$ – условная вероятность данного сочетания диагностических признаков при условии, что рентгеновское изображение содержит симптомы болезни.

Одна из проблем применения формулы Байеса состоит в том, что вычисление апостериорной вероятности $P(H_1 / S_k)$ требует подсчета числа сочетаний диагностических признаков. Эта задача имеет достаточно простое решение, если размерность *E* пространства диагностических признаков равна двум. Действительно, пусть классификация состояний здоровья выполняется по двум независимым признакам S^1 и S^2, которые могут принимать следующие значения $S^1 = S_1^1, \ \ldots, \ S_i^1, \ \ldots, \ S_n^1, \ S^2 = S_1^2, \cdots, S_j^2, \cdots, S_m^2$. Тогда результаты вычислений условных вероятностей $P(S_k / H_1)$, $P(S_k / H_2)$ можно

представить в виде таблиц, в ячейки которых заносятся объекты диагноза, имеющие данное сочетание диагностических признаков $S_k = S_i^1 \,\&\, S_j^2$. Поэтому в таблице, представленной на рис. 1, содержимое каждой ячейки соответствует количеству объектов наблюдения, которые имеют данное сочетание диагностических признаков $S_k = S_i^1 \,\&\, S_j^2$. Однако в случае, если евклидова размерность пространства диагностических признаков $E \geq 3$, представление результатов статистических вычислений в виде таблицы становится невозможным.

	$S^1{}_1$	⋯	$S^1{}_i$	⋯	$S^1{}_n$
$S^2{}_1$	N_{11}	⋯	N_{i1}	⋯	N_{n1}
⋯	⋯	⋯	⋯	⋯	⋯
$S^2{}_j$	N_{1j}	⋯	N_{ij}	⋯	N_{nj}
⋯	⋯	⋯	⋯	⋯	⋯
$S^2{}_m$	N_{1m}	⋯	N_{im}	⋯	N_{nm}

Рис.1

Анализ литературы. Приложению формулы Байеса к решению задач медицинской диагностики посвящено достаточное количество публикаций. Например, в работе [1] приведена модификация формулы Байеса, которая представляет собой попытку найти решение проблемы подсчета данных сочетаний диагностических признаков. Вычисление вероятности диагноза $P(H_1 / S_k)$ при условии, что состояние пациента определяется набором признаков S_k, основывается на предположении, что признаки S_i^1, S_j^2 могут принимать дискретные значения в интервале $S_i^1 = [0,1]$, $i = 1, \ldots, n$, и $S_j^2 = [0,1]$, $j = 1, \ldots, m$. При этом условные вероятности $P(S_k / H_1)$, $P(S_k / H_2)$, входящие в выражение (1), определяются по следующим формулам:

$$P(S_k / H_1) = \prod_{i=1}^{n} \left(S^1{}_i P\left(S^1{}_i / H_1\right) + \left(1 - S^1{}_i\right)\left(1 - P\left(S^1{}_i / H_1\right)\right)\right) \cdot \tag{2}$$

$$\cdot \prod_{j=1}^{m} \left(S^2{}_j P\left(S^2{}_j / H_1\right) + \left(1 - S^2{}_j\right)\left(1 - P\left(S^2{}_j / H_1\right)\right)\right);$$

$$P(S_k / H_2) = \prod_{i=1}^{n} \left(S^1{}_i P\left(S^1{}_i / H_2\right) + \left(1 - S^1{}_i\right)\left(1 - P\left(S^1{}_i / H_2\right)\right)\right) \cdot \tag{3}$$

$$\cdot \prod_{j=1}^{m} \left(S^2{}_j P\left(S^2{}_j / H_2\right) + \left(1 - S^2{}_j\right)\left(1 - P\left(S^2{}_j / H_2\right)\right)\right).$$

Формулы, по которым подсчитываются условные вероятности $P(S_k / H_1)$ и $P(S_k / H_2)$, выведены в предположении, что значения диагностических признаков равняются $S_i^1 = 1$ и $S_j^2 = 1$, если данный признак у наблюдаемого пациента присутствует, и значения диагностических признаков равняются $S_i^1 = 0$ и $S_j^2 = 0$, если указанный признак у диагностируемого пациента отсутствует.

Однако к данному способу вычисления условных вероятностей $P(S_k / H_1)$ и $P(S_k / H_2)$ можно предъявить существенное замечание. Это замечание состоит в том, что вероятности $P(S_k / H_1)$ и $P(S_k / H_2)$ данного сочетания диагностических признаков $S_k = S_i^1 \,\&\, S_j^2$ не могут равняться произведениям условных вероятностей того, что признаки S_i^1, S_j^2 принимают данные значения:

$$P(S_k / H_1) = P\left(S^1{}_i \,\&\, S^2{}_j / H_1\right) \neq P\left(S^1{}_i / H_1\right) P\left(S^2{}_j / H_1\right);$$
$$P(S_k / H_2) = P\left(S^1{}_i \,\&\, S^2{}_j / H_2\right) \neq P\left(S^1{}_i / H_2\right) P\left(S^2{}_j / H_2\right).$$

Это обусловлено тем, что данное сочетание значений диагностических признаков является событием, а не совокупностью независимых событий, состоящих в присвоении диагностическим признакам данных значений. Поэтому приведенный выше способ вычисления условных вероятностей не решает проблему, связанную с подсчетом числа данных сочетаний диагностических признаков [4 – 7].

Целью статьи является разработка модели представления диагностических признаков, которая позволяет придать процедуре подсчета числа данных сочетаний диагностических признаков наглядный и удобный для вычислений вид.

Метод определения условных вероятностей данных сочетаний диагностических признаков. Пусть задана статистическая выборка, которая

состоит из N рентгенограмм, не содержащих признаков заболевания. Кроме того, пусть классификация состояний здоровья также выполняется по двум признакам S^1 и S^2, как и классификация, представленная на рис. 1. Выполним процедуру подсчета числа данных сочетаний диагностических признаков в следующей последовательности:

– распределим число N объектов наблюдения по значениям $S^1 = S_1^1, \cdots, S_i^1, \cdots, S_n^1$ первого признака. Получим числа $N_1^1, \ldots, N_i^1, \ldots, N_n^1$ объектов наблюдения, которые имеют данные значения первого признака;

– распределим каждое число $N_1^1, \ldots, N_i^1, \ldots, N_n^1$ объектов наблюдения, которые имеют данные значения первого признака, по значениям $S^2 = S_1^2, \ldots, S_j^2, \ldots, S_m^2$ второго признака. Получим числа $N_{11}, \ldots, N_{ij}, \ldots, N_{nm}$ сочетаний данных значений первого S^1 и второго S^2 признаков.

Представим результаты подсчета числа данных сочетаний диагностических признаков в виде графа [8], показанного на рис. 2. Этот граф образован совокупностью значений двух признаков S^1 и S^2, причем связи между его узлами описываются числом их возможных сочетаний. Заметим, что представление результатов расчета в виде графа позволяет установить следующие зависимости между числом объектов наблюдения, имеющих данное сочетание диагностических признаков, и числом объектов наблюдения, которые имеют данное значение диагностического признака:

$$N_1^2 = N_{11} + \cdots + N_{i1} + \cdots + N_{n1}; \quad (4)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$N_j^2 = N_{1j} + \cdots + N_{ij} + \cdots + N_{nj}; \quad (5)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$N_m^2 = N_{1m} + \cdots + N_{im} + \cdots + N_{nm}; \quad (6)$$

$$N_1^1 = N_{11} + \cdots + N_{1j} + \cdots + N_{1m};$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$N_i^1 = N_{i1} + \cdots + N_{ij} + \cdots + N_{im};$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$N_n^1 = N_{n1} + \cdots + N_{nj} + \cdots + N_{nm};$$

$$N = N_1^1 + \cdots + N_i^1 + \cdots + N_n^1.$$

Покажем, что определение числа сочетаний диагностических признаков не зависят от порядка, в котором перечисляются диагностические признаки. Пусть процедура вычисления начинается с того, что объекты наблюдения распределяются по значениям второго признака S^2, после чего объекты

наблюдения, сгруппированные по значениям $S^2 = S_1^2, \ldots, S_j^2, \ldots, S_m^2$, распределяются по значениям первого признака S^1.

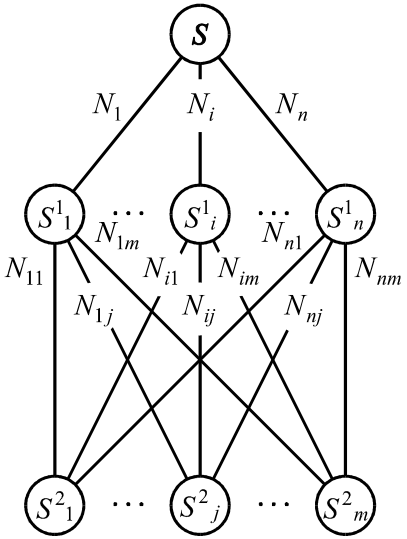

Рис.2

Представим результаты распределения объектов наблюдения в виде графа, приведенного на рис. 3. При этом связи между узлами графа, характеризующими значения диагностических признаков, позволяют установить следующие зависимости:

$$N_1^2 = N_{11} + \cdots + N_{1i} + \cdots + N_{1n}; \quad (7)$$

……..

$$N_j^2 = N_{j1} + \cdots + N_{ji} + \cdots + N_{jn}; \quad (8)$$

……

$$N_m^2 = N_{m1} + \cdots + N_{mi} + \cdots + N_{mn}. \quad (9)$$

Сравним соотношения (4) – (6) с соотношнниями (7) – (9). Поскольку число объектов наблюдения, имеющих данное значение второго признака S^2, одинаково как при составлении соотношений (4) – (6), так и при составлении соотношений (7) – (9), будут справедливы следующие равенства:

$$N_{ij} = N_{ji},\ i = 1,\cdots,n,\ j = 1,\cdots,m.$$

Следовательно, результаты определения числа сочетаний диагностических признаков действительно не зависят от порядка, в котором перечисляются диагностические признаки.

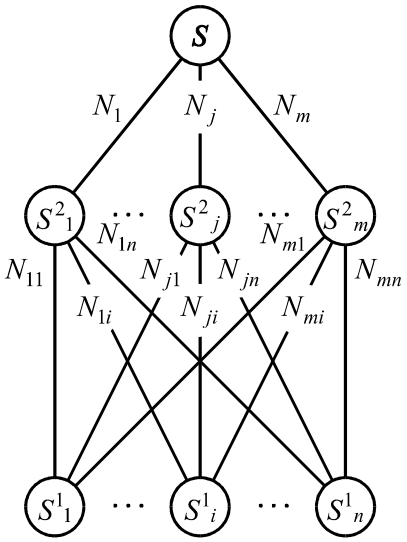

Рис.3

Введем третий диагностический признак, принимающий значения $S^3 = S_1^3, \ldots, S_k^3, \ldots, S_l^3$. При этом граф, описывающий процедуру определения числа данных сочетаний диагностических признаков S^1 и S^2, дополняется строкой, моделирующей процедуру распределения объектов наблюдения, сгруппированных по значениям второго признака $S^2 = S_1^2, \ldots, S_j^2, \ldots, S_m^2$, по значениям третьего признака S^3. Представим процедуру определения числа данных сочетаний диагностических признаков S^1, S^2, S^3 в виде графа, показанного на рис. 4. При этом, если узлы графа соответствуют отдельным значениям диагностических признаков, а связи между ними – числу возможных сочетаний значений диагностических признаков, то данный граф можно описать следующей системой линейных уравнений:

$$N_k^3 = \sum_{i=1}^{n} \sum_{j=1}^{m} N_{ijk} \ , \ k = 1, \ldots, l \ ; \tag{10}$$

$$N_j^2 = \sum_{i=1}^{n} \sum_{k=1}^{l} N_{ijk} \ , \ j = 1, \ldots, m \ ; \tag{11}$$

$$N_i^1 = \sum_{j=1}^{m} \sum_{k=1}^{l} N_{ijk} \ , \ i = 1, \ldots, n \ ; \tag{12}$$

$$N = \sum_{i=1}^{n} N_i^1 \ . \tag{13}$$

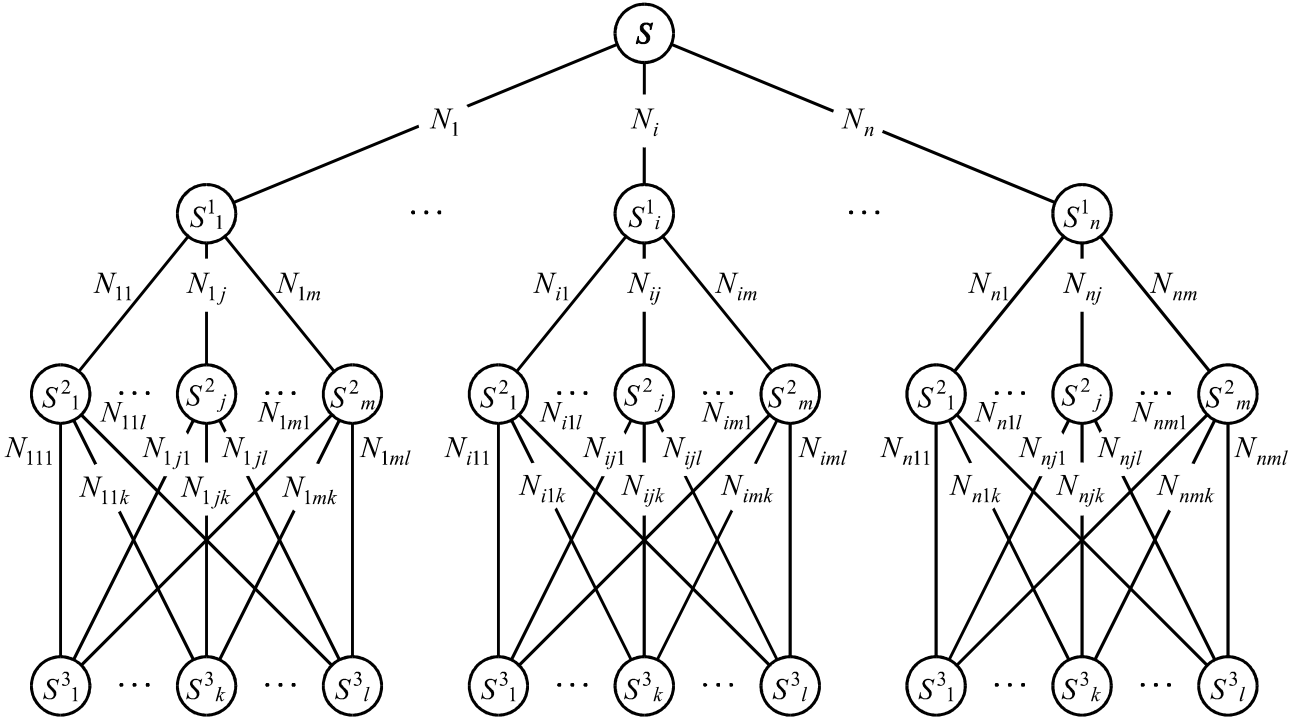
S
N1
Ni
Nn
S1 1
S1 i
S1 n
N11
N1j
N1m
Ni1
Nij
Nim
Nn1
Nnj
Nnm
S2 1
S2 j
S2 m
N111
N11l
N1m1
N1ml
N1j1
N1jl
N11k
N1jk
N1mk
Ni11
Ni1l
Nim1
Niml
Nij1
Nijl
Ni1k
Nijk
Nimk
Nn11
Nn1l
Nnm1
Nnml
Nnj1
Nnjl
Nn1k
Nnjk
Nnmk
S3 1
S3 k
S3 l

Рис.4

Заметим, что данную графическую модель представления диагностических признаков можно распространить на случай, когда евклидова размерность пространства диагностических признаков $E > 3$. Действительно, введем дополнительный признак S^4. При этом граф, соответствующий размерности $E = 4$, наследует структуру графа, построенного для размерности $E = 3$, и дополняет его строкой, состоящей из узлов, отображающих значения диагностического признака S^4. Кроме того, связи между узлами описываются системой линейных алгебраических уравнений, количество которых равно сумме чисел значений всех диагностических признаков, увеличенной на единицу $q = n + m + l + \cdots + 1$, а количество неизвестных равно произведению чисел значений всех диагностических признаков $p = n \times m \times l \times \cdots \times 1$. Если ввести $p - q$ дополнительных условий, то процедуру подсчета числа возможных сочетаний диагностических признаков можно заменить решением системы линейных алгебраических уравнений (10) – (13).

Преимуществом данного подхода является то, что достаточно сложная задача на определение числа данных сочетаний значений диагностических признаков сводится к решению более простой задачи на определение числа данных значений диагностических признаков. При этом определение числа данных сочетаний значений диагностических признаков можно представить как решение системы уравнений (10) – (13).

Выводы. Таким образом, впервые разработана графическая модель представления диагностических признаков в виде графа, узлы которого являются значениями диагностических признаков, а связи между ними выражают условные вероятности того, что объект наблюдения имеет данное сочетание значений диагностических признаков. Эта модель позволяет придать процедуре подсчета числа данных сочетаний диагностических признаков наглядный и удобный для вычислений вид и распространяется на случай, когда евклидова размерность пространства диагностических признаков больше или равна трем. Кроме того, впервые выведены соотношения (10) – (12), которые связывают условные вероятности того, что объект наблюдения имеет данные сочетания значений диагностических признаков, с условными вероятностями того, что объект наблюдения, имеет данные значения диагностических признаков. Эти соотношения доказывают несостоятельность применения формул (2) – (3) для выбора гипотезы с помощью формулы Байеса. Направление дальнейших исследований связано с поиском дополнительных условий, необходимых для решения системы линейных уравнений, описывающей граф чисел данных сочетаний диагностических признаков.

Список литературы: 1. *Постнова Т.Б.* Информационно-диагностические системы в медицине. – М.: Наука, 1972. – 376 с. **2.** *Максимов Г.К., Синицын А.Н.* Статистическое моделирование многомерных систем в медицине. – Л.: Медицина, 1983. – 144 с. **3.** *Завалишин Н.В., Мучник И.Б.*

Модели зрительного восприятия и алгоритмы анализа изображений. – М.: Наука, 1976. – 402 с. **4.** *Форсайт Д, Понс Ж.* Компьютерное зрение. Современный подход. – М.: Издательский дом "Вильямс", 2004. – 928 с. **5.** *Хайкин С.* Нейронные сети: полный курс, 2-е издание. – М.: Издательский дом "Вильямс", 2006. – 1104 с. **6.** *Потапов А.С.* Распознавание образов и машинное зрение. – СПб.: Политехника, 2007. – 548 с. **7.** *Вентцель Е.С.* Теория вероятностей. – М.: Наука, 1969. – 576 с. **8.** *Кристофидес Н.* Теория графов. Алгоритмический подход. – М.: Мир, 1978. – 432 с.

УДК 681.518.54

Модель подання ознак у байєсовскому класифікаторі медичних зображень / Ніцин Д.О. // Вісник НТУ "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ", 2008. – № 49. – С. 105 – 113.

Пропонується модель подання діагностичних ознак у вигляді графа, який зв'язує значення параметрів, що описують стан здоров'я, із числами їх можливих сполучень. Запропонована модель спрощує визначення умовних ймовірностей даних сполучень діагностичних ознак при застосуванні формули Байєса для класифікації медичних зображень. Іл.: 4. Бібліогр.: 8 назв.

Ключові слова: діагностичні ознаки, формула Байєса, медичні зображення.

UDC 681.518.54

Model of representation of attributes in the Bayesian qualifier of the medical images / Nitsyn D.A. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modeling. – NTU "KhPI". – 2008. – №. 49. – P. 105 – 113.

The model of representation of diagnostic attributes as the column is offered which connects meanings of parameters describing a condition of health, to numbers of their probable combinations. The offered model simplifies definition of conditional probabilities of the given combinations of diagnostic attributes at application of the Bayesian formula for classification of the medical images. Figs: 4. Refs: 8 titles.

Key words: diagnostic attributes, Bayesian formula, medical images.

Поступила в редакцию 16.10.2008