УДК 004.023

ВЫБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ ОЦЕНКИ ТЯЖЕСТИ ЗАБОЛЕВАНИЯ

Капустина С.В.[1], Кирякова О.В.[1], Капустина А.В.[1], Лапина Л.А.[1], Ступина А.А.[1]

[1]ФГАОУ ВПО «Сибирский федеральный университет», Красноярск, Россия (660041, Красноярск, пр. Свободный, 79), e-mail: sv_kapustina@mail.ru

Одним из подходов повышения качества распознавания образов и снижения вычислительных затрат является проведение предварительного анализа информации. Целью такого анализа является оценка информативных характеристик обучающей выборки, в частности оценка информативности признаков, оценка значений признаков, выделение наиболее представительных объектов. В задачах медицинской диагностики в роли объектов выступают пациенты. Признаки характеризуют результаты обследований, симптомы заболеваний и применявшиеся методы лечения. Накопив достаточное количество прецедентов, можно решить различные задачи: классифицировать вид заболевания (дифференциальная диагностика), определять наиболее целесообразный способ лечения, предсказывать длительность и исход заболевания, оценивать риск осложнений, находить синдромы — наиболее характерные для данного заболевания совокупности симптомов. Разработанная информационно-обучающая система базируется на современных технологиях и программных средствах.

Ключевые слова: медицинская диагностика, информативность признака, информационная система

CHOICE INFORMATIVE TO FEATURES TO ASSESS THE SEVERITY DISEASES

Kapustina S.V.[1], Kapustina A.V.[1],Kiryakova O.V.[1]., Lapina L.A.[1], Stupina A.A.[1]

[1]Siberian federal university, Krasnoyarsk, Russia (660041, Krasnoyarsk, Svobodnii av. 79), e-mail: purik28@yandex.ru

One approach to improve the quality of pattern recognition and to reduce the computational cost is to conduct a preliminary analysis of the information. The purpose of this analysis is to assess informative characteristics of training sample, in particular evaluation of informative, evaluation of characteristic values, selection of the most representative objects. The task of medical diagnostics in the role of objects appear, patients. Signs characterize the survey results, the symptoms of the diseases and the treatments. Having accumulated sufficient number of precedents can solve a variety of tasks: to classify the type of illness (differential diagnosis), to determine the most appropriate method of treatment, to predict the duration and outcome of the disease, to assess the risk of complications, finding syndrome - the most characteristic symptoms of the disease together. Developed information and training system based on modern technologies and software.

Keywords: medical diagnostics, informative features, information system

В современном обществе все большие объемы информации сохраняются в электронном виде в базах данных. Общим для всех этих данных является то, что эта база содержит большое количество скрытых закономерностей, являющихся весьма важными для принятия стратегических решений. Таким образом, существует необходимость в компьютерных системах, способных анализировать подобного рода данные и представлять новые знания в удобной для восприятия человеком форме. Основным отличием извлечения знаний из баз данных от традиционных методов машинного обучения является использование базы данных в качестве обучающего множества. Базы данных обычно очень велики как в смысле количества атрибутов, так и в смысле количества объектов, представленных в базе данных. С одной стороны, большое количество атрибутов дает больше шансов на то, что можно найти подходящие описания классов. С другой стороны, увеличение числа атрибутов приводит к увеличению размеров пространства поиска.

Очевидно, для любой реальной базы данных размер пространства поиска будет очень большим, так что ни один из методов полного перебора не может быть применен. Необходимо использовать знания о предметной области и эвристики для сокращения перебора.

Практическая реализация

В задачах медицинской диагностики в роли объектов выступают пациенты. Признаки характеризуют результаты обследований, симптомы заболеваний и применявшиеся методы лечения.

Специфика современных требований к обработке данных с целью обнаружения знаний следующая: данные имеют большой объем, являются разнородными (бинарными, порядковыми, количественными), результаты должны быть конкретны и понятны. Примерами бинарных признаков являются пол, наличие головной боли, слабости, тошноты и т.д. Порядковый признак – тяжесть состояния (легкое, средней тяжести, тяжелое, угрожающее жизни). Количественными признаками являются возраст, пульс, артериальное давление, содержание гемоглобина в крови, частота дыхательных движений, доза препарата и т.д. Признаковое описание пациента является, по сути, формализованной историей болезни. Накопив достаточное количество прецедентов, можно решить различные задачи: классифицировать вид заболевания (дифференциальная диагностика), определять наиболее целесообразный способ лечения, предсказывать длительность и исход заболевания, оценивать риск осложнений, находить синдромы — наиболее характерные для данного заболевания совокупности симптомов.

При изучении объектов, характеризуемых большим числом факторов, часто бывает важно определить, какие из этих факторов в большей степени влияют на интересующие нас свойства объектов. В частности, определение информативности факторов – это один из важных этапов анализа изучаемого объекта [2].

Установка оценки тяжести заболевания бронхиальной астмой влияет на дальнейшее лечение в амбулаторном или в клиническом отделении. В процессе лечения больному выбирается базисная терапия [3, 4]. Целью данной работы является выявление наиболее информативных признаков в оценке тяжести обострений бронхиальной астмы методами Кульбака, Шеннона и накопленных частот. В методах Кульбака и Шеннона использовались качественные, а в методе накопления частот — количественные признаки. Данные для анализа были извлечены из базы медицинской информационной системы qMS и собраны в текстовый файл, где каждая строка представляет собой вектор из 30 значений-признаков. Признаки не равнозначны, поэтому важной задачей является поиск и отбор признаков, достаточно информативных для распознавания.

Сущность метода накопленных частот состоит в том, что если имеются две выборки признака x, принадлежащие двум различным классам, то по обеим выборкам в одних координатных осях строят эмпирические распределения признака x и подсчитывают накопленные частоты (сумму частот от начального до текущего интервала распределения). Оценкой информативности служит модуль максимальной разности накопленных частот.

Алгоритм программной реализации нахождения информативности признака методом накопленных частот представлен на рисунке 1.

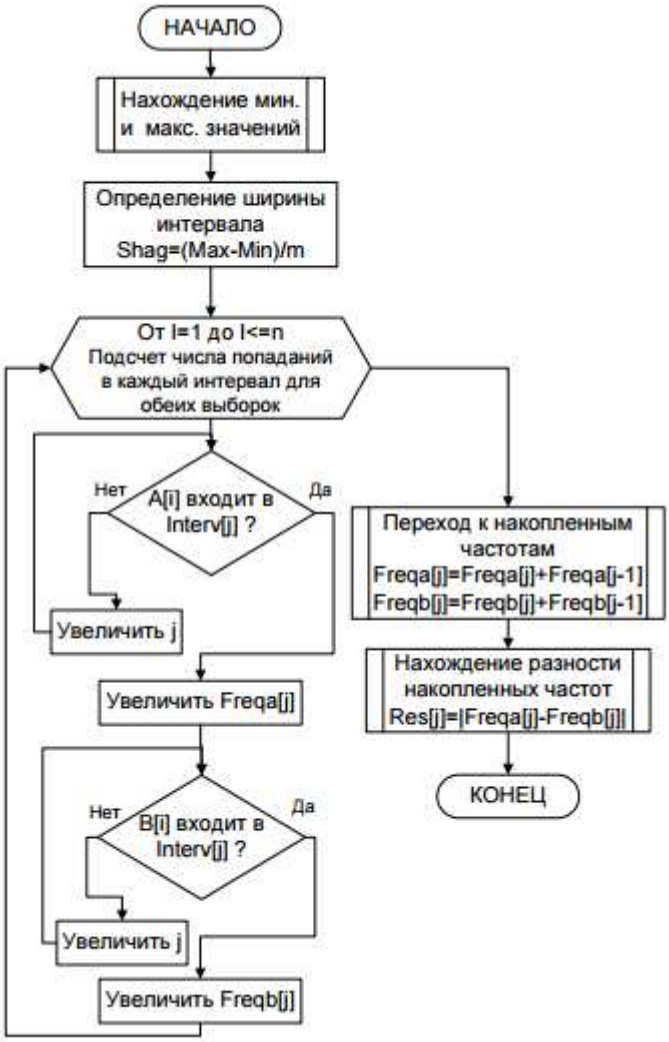

Рис. 1. Схема алгоритма

Метод Шеннона предлагает оценивать информативность как средневзвешенное количество информации, приходящееся на различные градации признака. Под информацией в теории информации понимают величину устраненной энтропии.

$$I(x_i) = 1 + \sum_{i=1}^{G}\left(P_i \cdot \sum_{k=1}^{K} P_{i,k} \cdot log_K P_{i,k}\right),$$

где G – количество градаций признака;

K – количество классов;

P_i – вероятность i-ой градации признака $P_i = \frac{\sum_{k=1}^{K} m_{i,k}}{N}$, где

$m_{i,k}$ — частота появления *i*-ой градации в *K*-ом классе, *N*- общее число наблюдений;

$P_{i,k}$ — вероятность появления *i*-ой градации признака в *K*-ом классе.

$$P_{i,k} = \frac{m_{i,k}}{\sum_{k=1}^{K} m_{i,k}}$$

Метод Кульбака предлагает в качестве оценки информативности меру расхождения между двумя классами, которая называется дивергенцией. Метод анализа признаков путем оценки информативности критерием Кульбака получил широкое применение в медицине при рассмотрении отдельных факторов, влияющих на постановку диагноза. Согласно этому методу информативность (или дивергенция Кульбака) вычисляется по формуле:

$$I(x_j) = \sum_{i=1}^{G} [P_{i1} - P_{i2}] \cdot log_2 \frac{P_{i,1}}{P_{i2}},$$

где *G* – число градаций признака;

P_{i1} – частота появления i-ой градации в первом классе;

$$P_{i,1} = \frac{m_{i,1}}{\sum_{k=1}^{K} m_{i,1}},$$

где $m_{i,1}$ — частота появления *i*-ой градации в 1-ом классе;

$P_{i,2}$ — вероятность появления *i*-ой градации признака в 2-ом классе.

$$P_{i,k} = \frac{m_{i,k}}{\sum_{k=1}^{K} m_{i,k}}$$

где $m_{i,2}$ — частота появления *i*-ой градации в 2-ом классе.

Информативность, определяемая всеми тремя методами, – величина положительная, однако в методе накопленных частот и методе Кульбака она не является нормированной, поэтому об информативности, определенной этими методами, можно говорить только в относительном плане – более высокая или более низкая по сравнению с информативностью другого признака.

Метод Шеннона дает оценку информативности как нормированной величины, которая изменяется от 0 до 1. Поэтому об информативности признака, определенной методом Шеннона, можно говорить в абсолютном плане: ближе к 1 – высокая; ближе к 0 – низкая.

После проведения расчетов получили признаки с наибольшей информативностью. В их число входят: пульс, ЧДД (частота дыхательных движений), ПСВ (пиковая скорость выхода), PaO2 (парциальное давление кислорода в артериальной крови), PaCO2 (артериальное напряжение двуокиси углерода), SaO2 (текущее содержание кислорода в крови, наличие/отсутствие свистящих хрипов, сознание (нет нарушения, возбуждение, спутанность), разговор (нет нарушения, фразы, слова)).

Результаты методов поиска логических закономерностей в данных выражаются в виде IF-THEN правил [5]. С помощью таких правил решим задачу интерпретации полученных результатов.

Если {пульс<110, ЧДД<20, ПСВ>80, PaCO2<45, SaO2>95, свистящие хрипы=да, сознание=нет нарушения, разговор=нет нарушения} Обострение=легкое;

Если {пульс=[110,120], ЧДД<25, ПСВ=[50,80],PaO2>60, PaCO2<45, SaO2[91,95], свистящие хрипы=да, сознание=возбуждение, разговор=фразы} Обострение=средней тяжести;

Если {пульс>120, ЧДД>25, ПСВ<50, PaO2<60 PaCO2>45, SaO2<90, свистящие хрипы=да, сознание=возбуждение, разговор=слова} Обострение=тяжелое;

Если {пульс>120, ЧДД>25, ПСВ<33, PaO2<60, PaCO2>45, SaO2<90, свистящие хрипы=да, сознание=спутанность, разговор=слова} Обострение=угрожающее жизни;

Заключение

Сравнивая методы определения информативности признака, следует отметить, что метод накопленных частот, в отличие от методов Шеннона и Кульбака, зависит от способа кодировки признака [1]. Метод Шеннона позволяет определить информативность признака, участвующего в распознавании произвольного числа классов. Все рассмотренные методы не зависят от числа градаций. В методах Кульбака и Шеннона объемы выборки наблюдений по двум распознаваемым классам могут быть различны.

Разработанная информационно-обучающая система на языке программирования C++Builder XE5, имеет практическую ценность и успешно используется в учебном процессе КрасГМУ на кафедре медицинской информатики и инновационных технологий с курсом ПО.

Список литературы

1. Голованова И.С. Выбор информативных признаков. Оценка информативности. — Томск: ТПУ, 2003. – 18 с.

2. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. — Новосибирск: Изд. Института математики, 1999. – 270 с.

3. Назаренко Г.И., Осипов Г.С. Основы теории медицинских технологических процессов. Ч. 1. — М.: Наука, Физматлит, 2005. — 144 с.

4. Назаренко Г.И., Осипов Г.С. Основы теории медицинских технологических процессов. Ч. 2. Исследование медицинских технологических процессов на основе интеллектуального анализа данных. — М.: Наука, Физматлит, 2006. — 144 с.

5. Назаренко Г.И., Осипов Г.С., Назаренко А.Г. Интеллектуальные системы в клинической медицине; синтез плана лечения на основе прецедентов // Информационные технологии и

вычислительные системы. — 2010. — № 1. – С. 24–35.

Рецензенты:

Пашков Г.Л., д.т.н., профессор, советник Научно-образовательного центра ИХХТ СО РАН, г. Красноярск;

Антамошкин А.Н., д.т.н., профессор кафедры системного анализа и исследования операций Института информатики и телекоммуникаций СибГАУ, г. Красноярск.