ПРИМЕНЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ДЛЯ АНАЛИЗА МЕДИКО-СТАТИСТИЧЕСКИХ ДАННЫХ

Тевелев А.Д., группа АСУ-00а

Руководитель: доц. Мокрый Г. В

За время функционирования Донецкого Центра медико-статистической информации накоплен достаточно обширный банк данных, автоматизация анализа которых способна дать гораздо больше информации для принятия решений в области здравоохранения. Для анализа данных возможно применение большого количества методов, направленных на решение следующих основных задач: выявление скрытых связей между данными – задача ассоциации; выявление некоторых признаков, характеризующих группу, к которой принадлежит тот или иной объект, другими словами, совокупность данных – задача классификации; задача построения математической модели по выбранным из хранилища данным, например построение модели развития заболевания – задача прогнозирования.

В данной статье рассматривается вопрос применения методов классификации для анализа медико-статистических данных. Методы классификации позволяют выявлять признаки, характерные для однотипных групп объектов — классов.

По известным значениям этих характеристик можно отнести новый объект к тому или иному классу. Ключевым моментом выполнения классификации является анализ множества классифицированных объектов. С помощью классификации можно определить, например, районы с опасным уровнем туберкулеза, гепатита, и других распространенных на сегодняшний день заболеваний, или районы, где заболеваемость имеет приемлемый уровень. В качестве методов решения задачи классификации могут использоваться алгоритмы k-ближайшего соседа (k-Nearest Neighbor), байесовские сети

(Bayesian Networks), индукция деревьев решений, индукция символьных правил, нейронные сети [1].

Рассмотрим применение методов классификации для решения задачи анализа данных о заболеваемости активным туберкулезом в Донецкой области.

Ниже приведен вид основной таблицы этой формы.

Таблица 1. Отчет о заболеваемости активным туберкулезом. (Форма F08).

Вид туберкулеза	Число больных с установленным диагнозом активного туберкулеза по возрастным группам.																			
	Всего	<1	1 -4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+

Как видим, полем данных в этой таблице является число больных. Данные представлены по возрастным категориям, разбитым на четырехгодичные интервалы. Таким образом, каждая ячейка данных лежит на пересечении следующих измерений: 1-е измерение – возрастная категория; 2-е измерение – вид заболевания.

Учитывая, что медицинская статистика собирает данные на определенных территориях и за отчетный год, добавим еще третье и четвертое измерения: 3-е измерение – территория (город, район, ЛПУ); 4-е измерение – год.

Применение методов классификации в данном случае позволит решать следующие задачи: классификация ситуации с заболеваемостью определенным видом туберкулеза на определенной территории; классификация определенных возрастных группы по подверженности заболеванию; возможность решения двух вышеописанных и других задач «без учителя», то есть без заданных заранее классов (кластеризация).

Рассмотрим достоинства и недостатки основных методов классификации.

Нейронные сети. Искусственная нейронная сеть (ИНС, нейросеть) - это набор нейронов, соединенных между собой. Как правило, передаточные

функции всех нейронов в сети фиксированы, а веса являются параметрами сети и могут изменяться. Некоторые входы нейронов помечены как внешние входы сети, а некоторые выходы - как внешние выходы сети. Подавая любые числа на входы сети, мы получаем какой-то набор чисел на выходах сети. Таким образом, работа нейросети состоит в преобразовании входного вектора в выходной вектор, причем это преобразование задается весами сети. В решаемой задаче входным вектором $P(p_1,p_2,..,p_n)$ для нейронной сети может являться набор показателей заболеваемости для каждой возрастной категории. На выходе при этом требуется получить число или вектор, характеризующие класс, к которому принадлежит набор входных параметров. Для того, чтобы сеть можно было применять в дальнейшем, ее прежде надо "натренировать" на полученных ранее данных, для которых известны и значения входных параметров, и правильные ответы на них. Эта тренировка состоит в подборе весов межнейронных связей, обеспечивающих наибольшую близость ответов сети к известным правильным ответам. С классификацией при заранее известном наборе классов хорошо справляется персептрон Розенблатта [3], с задачами кластеризации — сети Кохонена [2-4].

Среди преимуществ использования нейронных сетей в задачах классификации медико-статистических данных — большой объем обучающей информации, накопленной за время функционирования Центра Медико-статистической информации Донецкого УЗО, а также их высокая эффективность в задачах кластеризации. Недостаток нейронных сетей — их высокая требовательность к вычислительным ресурсам.

Деревья решений. Деревья решения создают иерархическую структуру классифицирующих правил типа «ЕСЛИ..., ТО...», имеющую вид дерева. Для того чтобы решить, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Вопросы имеют вид «значение параметра А больше х». Если ответ положительный, осуществляется переход к правому узлу следующего уровня,

если отрицательный – то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом.

Популярность подхода связана с наглядностью и понятностью. Но очень остро для деревьев решений стоит проблема значимости, так как отдельным узлам на каждом новом построенном уровне дерева соответствует все меньшее и меньшее число записей данных – данные разбиваются на большое количество частных случаев. Чем больше этих частных случаев, чем меньше обучающих примеров попадает в каждый такой частный случай, тем менее уверенной становится их классификация. Если построенное дерево слишком «кустистое» – состоит из неоправданно большого числа мелких веточек – оно не будет давать статистически обоснованных ответов. Как показывает практика, в большинстве систем, использующих деревья решений, эта проблема не находит удовлетворительного решения. Кроме того, общеизвестно, и это легко показать, что деревья решений дают полезные результаты только в случае независимых признаков. В противном случае, они лишь создают иллюзию логического вывода [2].

Метод K-ближайших соседей. Данный метод является одним из наиболее простых и в то же время эффективных способов классификации объектов. Отнесение объекта к тому или иному классу в данном методе зависит от заранее заданных классов объектов, находящихся наиболее близко к анализируемому. Достоинство этого метода – невысокая требовательность к вычислительным ресурсам. Недостатком является зависимость результата от параметра K – числа ближайших соседей, а, следовательно, необходимость его правильного выбора.

Рассмотрим данный метод на примере. Возьмем ситуацию с туберкулезом легких в ряде городов Донецкой области за 2002 год. Данные представлены в таблице:

Таблица 2. Данные о заболеваемости туберкулезом легких.

ыВозраст Город	*1 –4*	*5-9*	*10-14*	*15-19*	*20-24*	*25-29*	*30-34*	*35-39*	*40-44*	*45-49*	*50-54*	*55-59*	*60-64*	*65-69*	*70-74*	*75-79*	*80-84*
Макеевка (-)		1	4	6	8	18	26	20	20	22	22	24	20	16	2	1	1
Константи-новка (+)		2	3	6	4	22	24	24	25	25	26	25	24	21	9	7	0
Мариуполь								22	24				26				

На графике (рис. 1), по оси X отложим возрастную категорию граждан, по оси Y – количество заболевших для данной категории. В данном случае классификация проводится с помощью двух заранее известных классов – “Обычный уровень заболеваемости”, и “Высокий уровень заболеваемости”, которые были установлены специалистами управления здравоохранения области в 2002 году. Такие данные могут вводиться в систему, создавая базу данных классов. Опасный уровень заболевания зарегистрирован был в Константиновке, и для этого города количество заболевших, соответствующее каждой возрастной группе отмечено знаком плюс (+). В Макеевке уровень заболеваемости оставался в норме, и для него данные отмечены знаком минус (-). Кружками (•) отмечены данные по городу Мариуполь, подлежащие классификации. Для Мариуполя взяты только 3 возрастные категории – 35-39, 40-44, 60-64, так как для осуществления классификации достаточно считывать информацию, не во всех, а в некоторых ключевых точках. Классифицируемые точки обозначим A, B и C соответственно.

Принадлежность точки к тому или иному классу будем определять по преобладанию их среди ее ближайших соседей. Проще говоря, если среди ближайших соседей точки преобладают точки со знаком «+», классифицируемая точка получит знак «+» и наоборот. Ближайшим соседом будем считать всякую точку, расстояние от которой до исследуемой точки минимально среди всего множества точек, еще не выбранных в качестве ближайших соседей.

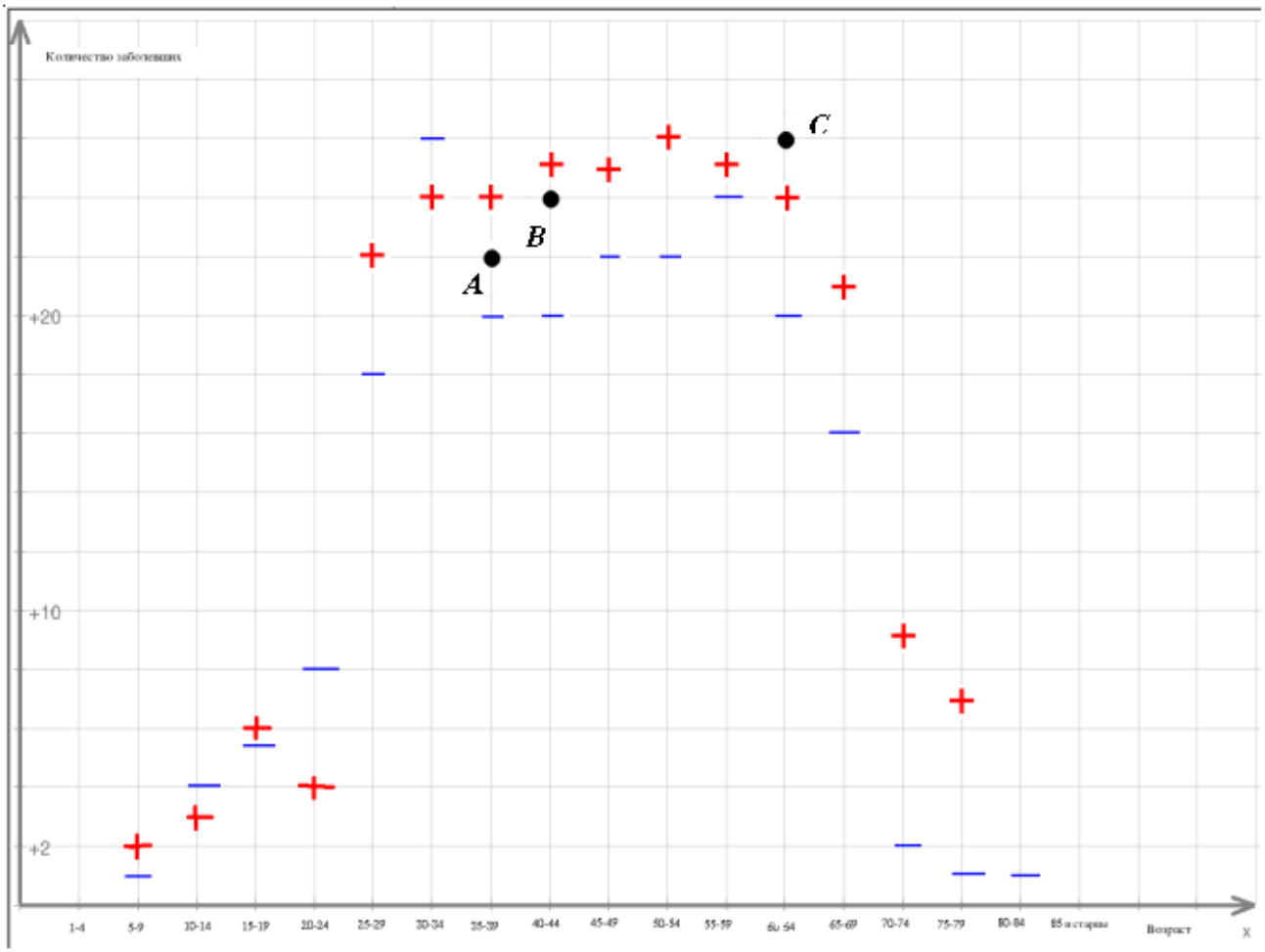

Рисунок 1. Результаты применения метода К-Ближайших соседей, для классификации данных о заболеваемости туберкулезом.

Изначально установим количество ближайших соседей K=1. Для точки A при этом возникает неопределенность, заключающаяся в том, какую точку выбрать в качестве ближайшего соседа – верхнюю или нижнюю, так как они обе отстоят от нее на одинаковом расстоянии, являющимся минимальным относительно всех других точек. Для точек B и C такой проблемы не возникает, каждая из них имеет по одному ближайшему соседу со знаками «+», следовательно и эти точки получают на данном этапе положительный знак. Так как для одной из точек результат не определен, увеличиваем количество ближайших соседей. При 2-х, 3-х, 4-х ближайших соседях знаки распределятся соответственно “неопределенность”, +, +. При пяти ближайших соседях получаем все три плюса. Таким образом, первые результаты были определены

при количестве ближайших соседей K=5. По полученным результатам можно сделать вывод, что в 2002 году в городе Мариуполь была опасная ситуация по туберкулезу. Для более детального анализа можно взять дополнительные точки.

Вышеописанный метод является наиболее подходящим для решения задачи классификации статистических данных о заболеваемости туберкулезом. Основное его достоинство – невысокая требовательность к вычислительным ресурсам и к обучающей выборке, что немаловажно, так как данные о заболеваемости зачастую разрознены и собраны не по всем объектам. Однако при необходимости можно применять и другие методы в совокупности, с целью более детального анализа результатов.

Для классификации объектов наблюдения по определенным признакам, а также для подтверждения результатов вышеописанных методов может быть использован аппарат математической статистики. Рассмотрим применение одного из ее методов - метода дискриминантного анализа - для предыдущего примера. Объектами наблюдения будут служить города Мариуполь, Макеевка, Константиновка, а также добавим еще три города – Дружковка, Краматорск, Артемовск. Возьмем 3 возрастные категории – 15-19 лет, 30-34 года, 45-49 лет.

Имеем следующие данные, подлежащие классификации.

Таблица 3. Данные о заболеваемости туберкулезом легких для дискриминантного анализа.

Город *Возраст*	Макеевка	Константиновка	Мариуполь	Дружковка	Краматорск	Артемовск
15-19	6	6	4	3	4	9
30-34	26	24	37	30	18	11
45-49	22	25	23	21	17	10

Имеется 2 группы: 1-я группа – «опасный уровень заболеваемости» и 2-я группа – «обычный уровень заболеваемости». Выдвинем гипотезу о наличие

4-х объектов в первой группе, и 2-х – во второй. Далее составим функции классификации, которые позволяют для каждой группы вычислить показатели классификации по формуле[5]:

$S_i = c_i + w_{i1}*x_1 + w_{i2}*x_2 + ... + w_{im}*x_m$,

где i – номер группы (в данном случае i=1,2), x_j – параметры каждой группы (в данном примере в качестве параметра выступает количество заболевших по определенной возрастной категории), c_i – константа для i-ой группы, w_{ij} – веса для j-ой переменной при вычислении показателя классификации для i-ой группы.

Воспользуемся одним из стандартных пакетов статистических вычислений для проведения классификации вышеприведенных данных.

Для каждой из групп получаем соответственно следующие дискриминантные функции:

$S_1=-131.09+10.37*x_1+3.26*x_2+5.17*x_3$

$S_2=-65.95+7.85*x_1+2.26*x_2+3.56*x_3$

В общем случае наблюдение считается принадлежащим той совокупности, для которой получен наивысший показатель классификации.

Полученные значения показателей классификации для каждого из объектов представлены в таблице:

Таблица 4. Результаты дискриминантного анализа

Объект	S1	S2
1 (Макеевка)	129.6	118.31
2 (Константиновка)	138.61	124.48
3 (Мариуполь)	149.83	131.04
4 (Артемовск)	106.32	100.24
5 (Дружковка)	56.94	66.71
6 (Краматорск)	49.82	65.19

Для объектов 1-4 значение функции S_1 больше чем значение функции S_2, для объектов 5-6 – наоборот. Учитывая результаты применения предыдущего

метода, можем выдвинуть предположение, что функция S_1 классифицирует группу с опасным уровнем заболеваемости, отсюда делаем вывод, что города Дружковка и Крамоторск имеют нормальный уровень заболеваемости, а остальные города – опасный.

Применение метода дискриминантного анализа целесообразно в совокупности с методом K-ближайших соседей и другими методами классификации с целью проверки гипотезы об отнесении объектов к тем или иным классам, и установления количественной меры различия между этими классами.

Классификация медико-статистических данных играет очень важную роль в медицине в первую очередь в условиях нехватки информации, когда правильно выдвинутая гипотеза позволит существенно снизить затраты, которые были бы необходимы для сбора дополнительных данных, а также затраты на выполнение профилактических мер.

Литература

1. Щавелев Л. В Способы аналитической обработки данных для поддержки принятия решений. СУБД. - 1998. - № 4-5.
2. Киселев М., Саламатин Е. Средства добычи знаний в бизнесе и финансах. - М.: Открытые системы - 1997. - №4.
3. Розенблатт Ф. Принципы нейродинамики. Перцептрон и теория механизмов мозга. - М.: Мир, 1965. 480 с.
4. Кохонен Т. Ассоциативные запоминающие устройства.- М.: Мир, 1982. 384 с.
5. Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р. Факторный, дискриминантный и кластерный анализ. – М., Финансы и статистика, 1989. 215 с.