РЕШЕНИЕ ЗАДАЧИ РАСПОЗНАВАНИЯ ОБРАЗОВ НА ПРИМЕРЕ ИНФОРМАЦИОННОЙ СИСТЕМЫ СКРИНИНГА ДЕВОЧЕК-ПОДРОСТКОВ

Коломойцева И.А.
Кафедра ПМИ ДонГТУ
kolomoit@r5.dgtu.donetsk.ua

Abstract

Kolomoitseva I.A. Cluster analysis by example information systems of girl-teenagers screening. *This article about the cluster analysis. It is describe the application of factor by distance and Z-factor by example of information system of premedical question by mass prophylactic examination of Donetsk region girls-teenagers.*

Введение

В настоящее время общая память всех компьютеров мира намного превысила тот объем информации, который способны запомнить все люди Земли. Все острее и острее поднимается вопрос обработки, обобщения и анализа накопленной информации. Большая часть этой информации носит статистический характер. Существует немало математических методов обработки таких данных. Несмотря на то, что эти методы легко алгоритмизируемы, их основным недостатком остается чрезмерная емкость (что выражается как в большом количестве требуемой памяти, так и в длительности процесса анализа). Современная ситуация требует разработки новых, не менее точных, но, в то же время, более быстрых и компактных методов обработки статистических данных.

Предметом изучения данной статьи является решение задачи распознавания образов в рамках статистического подхода. При решении данной задачи были использованы два подхода: стандартный и нестандартный. Стандартный подход предполагает применение линейного классификатора, а нестандартный - разработанного автором данной статьи Z-классификатора. Исследования проводились с использованием информационной системы (ИС) профилактических осмотров девочек-подростков Донецкого региона. Результатом проведенных исследований должны стать рекомендации по использованию стандартного или нестандартного подхода к решению задачи распознавания образов в предложенной предметной области.

1 Общие сведения о задаче распознавания образов

Под решением задачи распознавания образов понимается необходимость нахождения классификатора (дискриминантной функцию), который позволит произвести распознавание. В общем случае, классификатором может быть формула, алгоритм, эксперт и т. п.

Решение задачи распознавания образов осуществляется в два этапа: 1) попытка уменьшить размер задачи за счет выделения информативных признаков; 2) построение классификаторов [2, 3, 4].

Пути определения информативных признаков:

1) проверка с помощью критерия гипотез влияния каждого признака в отдельности на общую массу;

2) определение коррелированности признаков;

3) алгебраические преобразования такие, как, например, преобразование пространства.

Первый этап выходит за рамки изучения данной статьи, поэтому в применении к ИС профилактических осмотров девочек-подростков Донецкого региона он рассматриваться не будет.

Применительно ко второму этапу выделяют следующие стандартные виды классификаторов: классификатор по расстоянию; Байесовский классификатор; классификатор по максимальной апостериорной вероятности; минимаксный классификатор; классификатор максимального правдоподобия.

В данной статье приводится решение задачи распознавания образов на примере ИС профилактических осмотров девочек-подростков Донецкого региона с использованием классического классификатора по расстоянию и неклассического, разработанного автором данной статьи Z-классификатора.

2 Математическая постановка задачи

Все задачи распознавания образов можно представить одной математической моделью. Для этого используем случайный вектор X. Для базы данных ИС профилактических осмотров девочек-подростков Донецкого региона этот вектор состоит:

а) в случае применения классификатора по расстоянию - из 10 элементов: (X_1, X_2, ..., X_{10}), где X_1 - возраст, X_2 - рост, X_3 - вес, X_4 - балл полового развития (БПР), X_5 - возраст наступления менархе, X_6 - длина ноги, X_7 - окружность груди, X_8 - первый размер таза, X_9 - второй размер таза, X_{10} - третий размер таза.

б) в случае Z-классификатора, используемого в ИС профилактических осмотров девочек-подростков Донецкого региона - из 23 элементов: (X_1, X_2, ... X_{23}), где X_1 - возраст, X_2 - степень гирсутизма, X_3 - степень гипертрихоза, X_4 - степень ожирения, X_5 - признак низкой массы тела, X_6 - рост, X_7 - вес, X_8 - элемент половой формулы Ма, X_9 - элемент половой формулы Р, X_{10} - элемент половой формулы Ах, X_{11} - элемент половой формулы Ме, X_{12} - балл полового развития (БПР), X_{13} - возраст наступления менархе, X_{14} - состояние молочных желез, X_{15} - морфотип, X_{16} - трофические изменения кожи, X_{17} - длина ноги, X_{18} - окружность груди, X_{19} - первый размер таза, X_{20} - второй размер таза, X_{21} - третий размер таза, X_{22} - экстрагенитальная патология, X_{23} - перенесенные заболевания.

Размерность случайного вектора в дальнейшем будет обозначаться буквой p.

В качестве исходных данных в случае задачи распознавания образов используются:

1) обучающая выборка, состоящая из m случайных векторов;

2) контрольная выборка (экзаменуемая), состоящая из n случайных векторов (для ИС профилактических осмотров девочек-подростков Донецкого региона n=1).

Структура исходных данных для решения задачи распознавания образов приведена в таблице 1.

Таблица 1 - Исходные данные для решения задачи распознавания образов.

1	$X_1^{(1)}$	$X_1^{(2)}$	...	$X_1^{(m)}$	$X_1^{(m+1)}$	$X_1^{(m+2)}$	...	$X_1^{(m+n)}$
2	$X_2^{(1)}$	$X_2^{(2)}$	...	$X_2^{(m)}$	$X_2^{(m+1)}$	$X_2^{(m+2)}$	...	$X_2^{(m+n)}$
...	...	...	...	...	...	...	...	...
p	$X_p^{(1)}$	$X_p^{(2)}$	...	$X_p^{(m)}$	$X_p^{(m+1)}$	$X_p^{(m+2)}$	...	$X_p^{(m+n)}$

По обучающим выборкам строится дискриминантная функция в соответствии с выбранным классификатором и по этой функции проверяется контрольная выборка на принадлежность к какому-либо из классов.

Для работы с задачей распознавания образов необходимо знать количество классов r, к которым относят элементы из контрольной выборки. В случае ИС профилактических осмотров девочек-подростков Донецкого региона r=2: R_1 - «здоров», R_2 - «болен с некоторым диагнозом». Мы решим задачу распознавания образов для 9 диагнозов: первичная аменорея; ЗПР; вторичная аменорея; опсоменорея; ГСППС; альгодисменорея; ЮМК; воспалительные заболевания; мастопатия.

Объем выборки для класса R_1 (для здоровых) равен 25 человекам. С учетом этого, в случае диагноза «первичная аменорея» m=85, «ЗПР» - m=83, «вторичная аменорея» - m=48, «опсоменорея» - m=277, «ГСППС» - m=51, «альгодисменорея» - m=253, «ЮМК» - m=90, «воспалительные заболевания» - m=64, «мастопатия» - m=53.

Решим задачу распознавания образов на примере обследуемой со следующими параметрами: 1) возраст - 15 лет; 2) степень гирсутизма - "+"; 3) степень гипертрихоза - "-"; 4) степень ожирения - "-"; 5) признак наличия низкой массы тела - "+"; 6) рост - 168 см; 7) вес - 48 кг; 8) элемент половой формулы Ма - 2; 9) элемент половой формулы Р-2;10) элемент половой формулы Ах -2; 11) элемент половой формулы Ме - 3; 12) БПР - 9; 13) возраст наступления менархе - 13; 14) состояние молочных желез – "<N"; 15) морфотип – "женский"; 16) трофические изменения кожи - нет; 17) длина ноги - 85 см; 18) окружность груди - 78 см; 19) первый размер таза - 23,6 см; 20) второй размер таза - 26,8 см; 21) третий размер таза - 28,9 см; 22) экстрагенитальная патология – {хронический гастрит, хронический тонзлит, хронический пиелонефрит}; 23) перенесенные заболевания – {краснуха, пневмония, ветряная оспа, паротит}.

3 Решение задачи распознавания образов при помощи стандартного классификатора

Рассмотрим сначала решение задачи распознавания образов с помощью стандартного классификатора - классификатора по расстоянию.

Мы будем рассматривать пары классов: R_1 и R_2, R_1 и R_3, R_1 и R_4, R_1 и R_5, R_1 и R_6, R_1 и R_7, R_1 и R_8, R_1 и R_9, R_1 и R_{10}. Для получения достоверных результатов необходимо в случае, если в какой-то из пар R_1 и R_i (i=2..9) (кроме последней) обследуемая будет отнесена к классу R_1, то необходимо проверить ее на принадлежность классу R_{i+1} (i=3..10). Дискриминантная функция для этого классификатора имеет вид [3]:

$$f(x) = \sum_{i=1}^{p} a_i \cdot x_i \text{ . Если } f(x) \geq c \rightarrow x \in R_1,\ f(x) < c \rightarrow x \in R_2. \qquad (1)$$

В дискриминантной функции являются неизвестными коэффициенты a_i и с. Их находят из условия максимизации расстояния между R_1 и R_2 [3]:

$$\frac{((a,\mu_1)-(a,\mu_2))^2}{(a,a)} \to \max, \qquad (2)$$

где μ_1 и μ_2 - векторы размерности p математических ожиданий для классов R_1 и R_2.

Задача максимизации расстояния эквивалентна системе $K \cdot a = \mu_1 - \mu_2$, где K - корреляционная матрица, построенная по m векторам [3].

$$\text{Коэффициент } c = \frac{a,\mu_1+\mu_2}{2} = \frac{\sum_{i=1}^{p} a_k \cdot \mu_k^{(1)} + \sum_{i=1}^{p} a_k \cdot \mu_k^{(2)}}{2}. \qquad (3)$$

В случае информационной системы скрининга девочек подростков получаем систему из 10-ти уравнений.

$$\begin{cases} K11 \cdot a1 + K12 \cdot a2 + K13 \cdot a3 + K14 \cdot a4 + K15 \cdot a5 + \dots + K19 \cdot a9 + K1\,10 \cdot a10 = \mu_1^{(1)} - \mu_1^{(2)} \\ K21 \cdot a1 + K22 \cdot a2 + K23 \cdot a3 + K24 \cdot a4 + K25 \cdot a5 + \dots + K29 \cdot a9 + K2\,10 \cdot a10 = \mu_2^{(1)} - \mu_2^{(2)} \\ \dots \\ K101 \cdot a1 + K10\,2 \cdot a2 + K10\,3 \cdot a3 + K10\,4 \cdot a4 + K10\,5 \cdot a5 + \dots + K10\,9 \cdot a9 + K10\,10 \cdot a10 = \mu_{10}^{(1)} - \mu_{10}^{(2)} \end{cases} \qquad (4)$$

Для каждой пары R_1 и R_i (i=2..10) построили девять дискриминантных функций и нашли девять значений с. Коэффициенты a_i и с представлены в таблице 2.

Таблица 2 - Коэффициенты a_i и с для дискриминантных функций линейного классификатора, посчитанные при помощи ИС скрининга девочек-подростков

Номер функции	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	c
f_2	0,03	-0,02	0,01	-0,28	0,03	0,17	-0,09	0	0,09	0	12,7
f_3	0,08	0,001	-0,002	3,89	0	-0,16	-0,006	-0,31	0,003	-2,32	12,4
f_4	0,006	0,01	-0,002	0,02	0,09	0,004	0,01	-0,07	0,2	-0,06	4,15
f_5	-0,02	0,06	-0,02	0,09	0,001	-0,01	-0,02	-0,09	0,05	0,03	-3,34
f_6	-0,05	0,02	-0,02	0,09	-0,09	0,01	-0,02	0,03	0,01	0,05	-2,2
f_7	-0,02	-0,005	0,002	-0,044	0,03	-0,001	-0,003	0	-0,018	0	-1,2
f_8	-0,05	0	0	-0,11	0,1	0	0	0	0	0	-0,83
f_9	-2.1	0,32	-0,16	2,26	0	-0,913	8,562	0	-12,09	0	316,4
f_{10}	0,12	-0,03	0,013	-0,153	0	0,041	-0,128	0	0,183	0	-1,78

Определим к какому классу относится обследуемая. Для нее f_2=11.729443. $f_2<c_2$, следовательно, обследуемая принадлежит к классу R_2, то есть ей можно поставить диагноз "первичная аменорея".

4 Решение задачи распознавания образов при помощи нестандартного классификатора

Приведем решение задачи распознавания образов с использованием нестандартного, разработанного автором данной статьи Z-классификатора.

Решение задачи распознавания образов для ИС профилактических осмотров девочек-подростков Донецкого региона выполняется на основе модели "среднего больного", описанной в [1]. Затем определяется коэффициент риска Кр (его величина прямо пропорциональна количеству совпадений параметров проверяемой девочки с параметрами модели "среднего больного). Если Кр>=0.5, то считается, что девочка относится к группе риска, информация о ней заносится в БД и дальше используется при построении модели "среднего больного" по поставленному диагнозу. Если Кр<0.5, то считается, что девочка здорова. Алгоритм расчета Кр приведен в [1].

Параметры модели "среднего больного" для класса R_2: 1) средний возраст A=14.84; 2) средний рост H=159.35; 3) средний вес W=50.39; 4) средняя степень гирсутизма G="-"; 5) средняя степень гипертрихоза Gip="-"; 6) средняя степень ожирения O="-"; 7) средний признак наличия низкой массы тела L="-"; 8) совокупность трофических изменений кожи T="нет"; 9) среднее значение элемента половой формулы Ma="2"; 10) среднее значение элемента половой формулы P="3"; 11) среднее значение элемента половой формулы Ax="3"; 12) среднее значение элемента половой формулы Me="0"; 13) среднее значение БПР=6.64; 14) средний возраст наступления менархе AM=0; 15) среднее значение морфотипа Mt="женский"; 16) средняя окружность груди Og=80.45; 17) средняя длина ноги Lf=79.93; 18) среднее значение первого размера таза St1=22.81; 19) среднее значение второго размера таза St2=25.60; 20) среднее значение третьего размера таза St3=27.52; 21) совокупность экстрагенитальных патологий Ex="хронический тонзилит, хронический гайморит, хронический гастрит, диффузная струма первой-второй степени, сколиоз, астма, хронический пиелонефрит, цистит, эндокринопатия, диабет"; 22) совокупность перенесенных заболеваний I="паротит, краснуха, ветряная оспа, гепатит, ОРВИ, пневмония, скарлатина, корь, ангина, ОРЗ, эцефалит, бронхит, острый аппендицит". Коэффициент риска в данном случае равен 0.59. Диагноз "первичная аменорея", поставленный с применением классификатора по расстоянию, подтвердился

5 Программная реализация

ИС профилактических осмотров девочек-подростков Донецкого региона реализована с использованием двух языков программирования. Часть ИС, связанная с обработкой базы данных, реализована на Clipper'e. Вся необходимая информация при этом хранится в dbf-файлах. Обработка статистических данных выполнена на Си. Данные в этом случае берутся из текстовых файлов, формируемых путем конвертации из dbf-файлов.

Заключение

Для решения задачи распознавания образов были использованы два метода: основанный на применении стандартного классификатора по расстоянию и нестандартного, предложенного при создании ИС профилактических осмотров девочек-подростков Донецкого региона Z-классификатора [1]. Применение этих методов для решения задачи распознавания образов привело к одному и тому же результату. Но Z-классификатор использует большее количество параметров, поэтому можно говорить о более высокой точности полученного решения.

При решении задачи распознавания образов с использованием классического (линейного) классификатора требуется решение системы из p уравнений с p неизвестными. Кроме этого, в большинстве случаев перед решением требуется привести систему уравнений к диагональному преобладанию (около 3-4 операций умножения над матрицами коэффициентов). Таким образом, при использовании линейного классификатора мы получаем сложность решения порядка p^2. В случае применения неклассического Z-классификатора в решении системы уравнений нет необходимости. Нахождение средних показателей ограничивается выполнением простых арифметических действий (определение среднего арифметического и наиболее часто встречаемых значений, формирование множества элементов). Следовательно, использование Z-классификатора снижает сложность решения до порядка p.

Относительно затрат памяти, то при использовании линейного классификатора требуется выделение памяти для хранения корреляционной матрицы (размерность pxp), двух векторов математического ожидания (размерности p), коэффициентов дискриминантной функции (размерности p) и коэффициента c, т.е. в общей сложности - $p^2+3*p+1$. При применении нестандартного Z-классификатора затраты памяти минимальны и равны p +1 - количеству средних показателей плюс коэффициент риска.

Учитывая вышеизложенное, для ИС скрининга девочек-подростков Донецкого региона при решении задачи распознавания образов рекомендуется использование нестандартного Z-классификатора, так как он дает выигрыш в точности и быстродействии при меньших затратах памяти по сравнению со стандартным (линейным) классификатором.

Литература

1. Дацун Н.Н, Коломойцева И.А. Моделирование риска в репродуктивном здоровье подростков на основе массовых профилактических осмотров В кн.: Информатика, кибернетика и вычислительная техника (ИКВТ-97). Сборник научных трудов Донецкого государственного технического университета. Выпуск 1. Донецк: ДонГТУ, 1997.- с. 259 - 265.
2. Мелник М. Основы прикладной статистики.- М.: Энергия, 1983. - 414 с.
3. Фомин Я.А., Тарловский Г.Р. Статистическая теория распознавания образов. - М.: Радио и Связь, 1986. - 263 с.
4. Фор Ален. Восприятие и распознавание образов . - М., 1989. - 271 с.