Ю.А. БУЛАНОВА

Муромский институт (филиал) ФГБОУ ВПО «Владимирский государственный университет имени А.Г. и Н.Г. Столетовых», г. Муром

Исследование алгоритмов распознавания новообразований по маммограммам

УДК 004.932

Заболевания молочной железы представляют опасность для всего женского населения. По данным Минздрава РФ [1] показатель заболеваемости раком молочной железы ежегодно возрастает, что можно увидеть при сравнении показателей заболеваемости нескольких онкологических заболеваний, представленных на рис. 1.

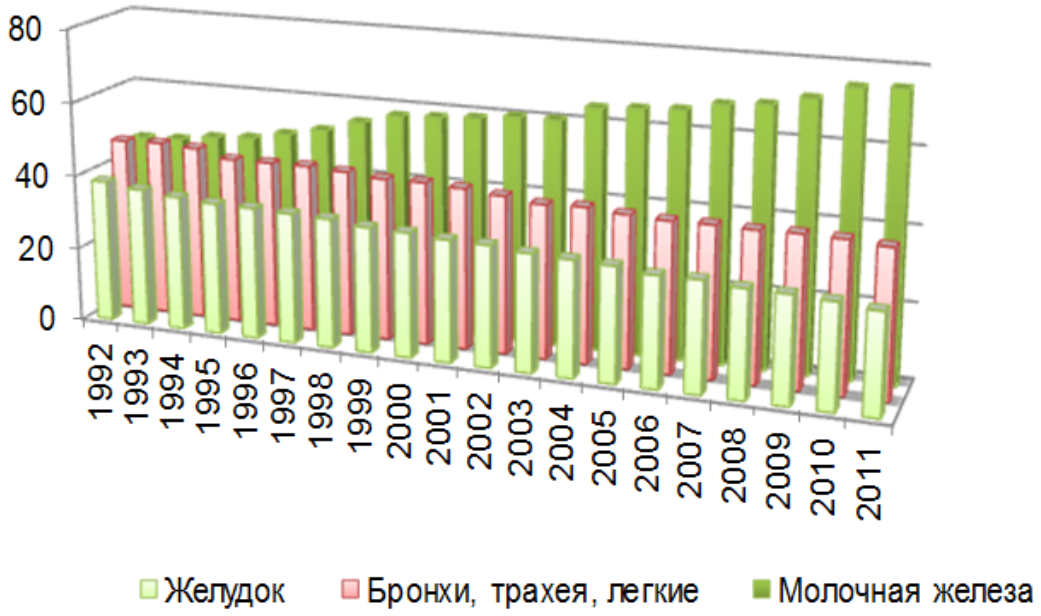

Рис. 1. Показатели заболеваемости самыми частыми злокачественными образованиями.

Целью данной работы является анализ и исследование алгоритмов распознавания для диагностики заболеваний молочной железы.

Распознавание – это отнесение конкретного объекта, представленного значениями его свойств (признаков), к одному из фиксированного перечня образов (классов) по определённому решающему правилу в соответствии с поставленной целью [2, 3, 4, 5].

Известно несколько алгоритмов распознавания новообразований, таких как метод опорных векторов (SVM), метод релевантных векторов (RVM) [6].

Формулировка SVM основана на принципе структурного риска минимизации. Вместо того, чтобы минимизировать объективную функцию, основанную на учебных образцах (такие как среднеквадратическая ошибка), SVM пытается минимизировать ограничения на обобщение ошибки (ошибка, сделанная машиной изучения на данных испытаниях, не используемых во время обучения). Как результат, SVM работают правильно, когда относятся к данным вне учебного набора.

Метод релевантных векторов основан на байесовской оценке для классификации проблемы. RVM работает быстрее метода опорных векторов, так как приводит к оптимальному решению с помощью меньшего количества обучающих образцов.

Однако, для успешного применения указанных алгоритмов необходимо одинаковое количество обучающих образцов в каждой группе, соответствующей типу новообразования. Если нет возможности обеспечить выполнение данного условия, то возможно применить дискриминантный анализ Фишера для распознавания новообразований [6, 7]. Данный метод заключается в вычислении среднего переменных в каждой группе и объединенной дисперсионной матрицы. На следующем этапе обращаем объединенную дисперсионную матрицу, затем вычисляем коэффициенты дискриминантных функций и оценки функций для каждого наблюдения (в отдельности).

Рассмотрим подробнее реализацию данного алгоритма.

Формирование обучающей выборки

Поскольку область молочной железы представляет собой пространственную организацию элементов в пределах некоторого ее участка с однородными статистическими характеристиками, то можно сказать, что область молочной железы – это сложное текстурное изображение. Поэтому для описания области молочной железы будут использоваться статистические моменты пространственных распределений второго порядка, которые называются также текстурными признаками Харалика [8, 9-11]. Были

выбраны 16 текстурных признаков Харалика, вычисленных на фрагментах маммографических снимков, где обнаружены подозрительные области.

Каждый признак вычисляется для матрицы GLCM, повернутой на 0°, 45°, 90° и 135°, соответственно количество признаков для 1 снимка будет равно 64. Для каждого типа новообразования все признаки заносятся в таблицу.

Обучающая выборка была сформирована из базы маммографических снимков MIAS совместно с врачом-рентгенологом. Все изображения были разделены на 3 группы: киста молочной железы, фиброаденома, рак молочной железы. В каждой группе было сформировано по 3 подгруппы: жировая инволюция, фиброзно-кистозная болезнь (ФКБ), аденоз, которые отражают степень сложности диагностики новообразований. Для каждого из эталонов были рассчитаны признаки Харалика, как указано выше.

Математическое описание алгоритма

Для каждой группы k=1,2,…,g вычисляют средние и суммы взаимных произведений отклонений от средних, как показано ниже.

Средние:

$$\overline{x}_{jk} = \frac{\sum_{i=1}^{n_k} x_{ijk}}{n_k}, \quad (1)$$

где nk - размер выборки в k-й группе, j=1,2,…,m - переменные.

Сумма взаимных произведений отклонений от средних:

$$S_k = \left\{ s_{jl}^k \right\} = \sum (x_{ijk} - \overline{x}_{jk}) \cdot (x_{ilk} - \overline{x}_{lk}), \quad (2)$$

где j=1,2,…,m; l=1,2,…,m.

Объединенная дисперсионная матрица вычисляется следующим образом:

$$D' = \frac{\sum_{k=1}^{g} S_k}{\sum_{k=1}^{g} n_k - g}, \quad (3)$$

где g - число групп.

На следующем этапе выполняется дискриминантный анализ при расчете совокупности линейных функций, которые служат указателями для классификации индивидуума в одну из k групп.

Для всех групп сочетаний получаем следующее.

Общие средние:

$$\overline{X}_j = \frac{\sum_{k=1}^{g} n_k \cdot \overline{x}_{jk}}{\sum_{k=1}^{g} n_k}, \qquad (4)$$

где g - число групп; j=1,2,…,m - переменные; nk - размер выборки в k-й группе; $\overline{x}_{jk}$ среднее j-й переменной в k-й группе.

Обобщенная D2 статистика рассчитывается на основании расстояния Махаланобиса V:

$$V = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} \cdot \sum_{k=1}^{g} n_k \cdot (\overline{x}_{ik} - \overline{X}_i) \cdot (\overline{x}_{jk} - \overline{X}_j), \qquad (5)$$

где dij - обратные элементы объединенной дисперсионной матрицы D.

Для каждой дискриминантной функции k*=1,2,…,g вычисляются следующие статистики:

Коэффициенты

$$C_{ik^*} = \sum_{j=1}^{m} d_{ij} \cdot \overline{x}_{jk}, \qquad (6)$$

где i=1,2,…,m; k=k*.

Константа

$$C_{0k^*} = \frac{-1}{2} \cdot \sum_{j=1}^{m}\sum_{l=1}^{m} d_{jl} \cdot \overline{x}_{jk} \cdot \overline{x}_{lk}. \qquad (7)$$

Для каждого i-го события в каждой k-й группе выполняются следующие вычисления.

Дискриминантные функции

$$f_{k^*} = \sum_{j=1}^{m} C_{jk^*} \cdot x_{ijk} + C_{0k^*}. \qquad (8)$$

где k*=1,2,…,g.

Вероятность, соответствующая наибольшей дискриминантной функции

$$P_L = \frac{1}{\sum_{k^*=1}^{g} e^{f_{k^*} - f_L}}, \qquad (9)$$

где fL - значение наибольшей дискриминантной функции; L-индекс наибольшей дискриминантной функции.

Исследование метода линейного дискриминантного анализа Фишера

Таблица 1

Средние для каждой группы

	Средние каждого признака в группе					
	1	**2**	**3**	**4**	**…**	**64**
Группа 1	0,012	1	1,861	0,67	…	0,137
Группа 2	0,021	1	1,984	0,605	…	0,16
Группа 3	0,011	1	1,355	0,673	…	0,125
Группа 4	0,004	1	1,397	0,633	…	0,064
Группа 5	0,01	1	7,247	0,608	…	0,083
Группа 6	0,024	1	2,19	0,603	…	0,133
Группа 7	0,003	1	2,308	0,564	…	0,056
Группа 8	0,013	1	1,757	0,644	…	0,114
Группа 9	0,244	1	16,49	0,904	…	0,983

Таблица 2

Сумма взаимных произведений отклонений от средних

	Сумма взаимных произведений отклонений от средних					
	1	**2**	**3**	**4**	**…**	**64**
Группа 1	0.00033	0.00033	-0.06500	-0.06000	…	-0.05600
	-0.05600	-0.05600	-0.05600	-0.05600	…	-0.05600
	-0.12200	-0.12200	24.57400	22.96500	…	22.20100
	22.20600	22.20600	20.59700	20.71000	…	20.77000
	…	…	…	…	…	…
	20.77300	20.77300	20.00900	20.06900	…	20.10500
Группа 2	0.00645	0.00645	-0.12100	-0.10600	…	-0.10700
	-0.10700	-0.10700	-0.10700	-0.10700	…	-0.10700
	-0.23400	-0.23400	3.63800	3.26400	…	3.34000
	3.35400	3.35400	2.98100	3.02000	…	3.01400
	…	…	…	…	…	…
	3.01300	3.01300	3.08800	3.08300		3.25800
…	…	…	…	…	…	…
	…	…	…	…	…	…
	…	…	…	…	…	…

	…	…	…	…	…	…
	…	…	…	…	…	…
Группа 9	0.00129	0.00129	-0.00890	-0.01100	…	-0.01200
	-0.01200	-0.01200	-0.01200	-0.01200	…	-0.01200
	-0.02200	-0.02200	215.07900	214.03500	…	211.51500
	211.51300	211.51300	210.46900	210.47700	…	210.49300
	…	…	…	…	…	…
	210.49200	210.49200	207.97200	207.98800	…	208.02200

Таблица 3

Дисперсионная матрица

0.000309	0.000309	0.004976	0.006011	…	0.006349
0.00635	0.00635	0.006349	0.006349	…	0.006349
0.011017	0.011017	22.863343	22.74401	…	22.672841
22.673876	22.673876	22.554543	22.565844	…	22.570717
…	…	…	…	…	…
22.571056	22.571056	22.499887	22.50476	…	22.513848

Таблица 4

Обратная дисперсионная матрица

-0.999999989232	-0.99999998365	1	1	…	-0.999999996643
1	1	-1.000000065392	-1.000000014272	…	1
0.000004317803	0.00000485339	-0.00000477634	-0.000005262897	…	0.000007932253
-0.00000712263	-0.000006907847	0.000007477431	0.000008538442	…	-0.00001916519
…	…	…	…	…	…
0.000002793939	0.000002038391	-0.00000263558	-0.000003261158		0.000011229484

Таблица 5

Общие средние признаков для всех групп

Признак	1	2	3	4	…	64
Значение	0.038	1	4.065	0.656	…	0.206

Расстояние Махаланобиса равно V = 13.502.

На следующем этапе вычисляются коэффициенты и константы для дискриминантных функций.

Таблица 6

Дискриминантные функции

Дискриминантная функция	Константа	Коэффициенты					
		1	2	3	4	…	64
1	0.684	1.383	-1.383	$-6{,}431*10^{-6}$	$1{,}003*10^{-5}$	…	$-3{,}487*10^{-6}$
2	0.689	1.408	-1.408	$-6{,}448*10^{-6}$	$9{,}878*10^{-6}$	…	$-3{,}308*10^{-6}$
3	0.441	0.891	-0.891	$-4{,}118*10^{-6}$	$6{,}489*10^{-6}$	…	$-2{,}29*10^{-6}$
4	0.479	0.963	-0.963	$-4{,}629*10^{-6}$	$7{,}695*10^{-6}$	…	$-2{,}982*10^{-6}$
5	3.346	6.763	-6.763	$-3{,}226*10^{-5}$	$5{,}082*10^{-5}$	…	$-1{,}809*10^{-5}$
6	0.798	1.635	-1.635	$-7{,}616*10^{-6}$	$1{,}188*10^{-5}$	…	$-4{,}133*10^{-6}$
7	0.903	1.813	-1.813	$-8{,}679*10^{-6}$	$1{,}407*10^{-5}$	…	$-5{,}244*10^{-6}$
8	0.63	1.275	-1.275	$-5{,}974*10^{-6}$	$9{,}464*10^{-6}$	…	$-3{,}382*10^{-6}$
9	5.736	15.168	-15.168	$-6{,}982*10^{-5}$	$1{,}035*10^{-4}$	…	$-3{,}266*10^{-5}$

Для каждого признака в каждой группе вычисляются дискриминантные функции (ДФ), затем выбираются номера функций имеющих наибольшие значения.

Таблица 7

Значения дискриминантных функций

		Значения дискриминантных функций									Вероятность наибольшей ДФ
		1	2	3	4	5	6	7	8	9	
Группа 1	1	**-0.435**	-0.676	-0.695	-0.467	-3.301	-0.809	-0.878	-0.624	-9.174	0,999748135
	2	**-0.429**	-0.666	-0.685	-0.460	-3.251	-0.797	-0.865	-0.614	-9.061	0,9999742147
	3	**-0.449**	-0.698	-0.717	-0.482	-3.408	-0.835	-0.907	-0.644	-9.413	0,9999995
	4	**-0.444**	-0.689	-0.708	-0.476	-3.364	-0.825	-0.895	-0.636	-9.315	0,999999247
	…	**…**	…	…	…	…	…	…	…	…	
	64	**-0.444**	-0.689	-0.708	-0.475	-3.363	-0.824	-0.895	-0.636	-9.312	0,999657421
Группа 2	1	-0.571	**-2.788**	-0.588	-0.368	-0.394	-0.685	-0.741	-0.527	-8.024	0,99999751463
	2	-0.696	**-3.396**	-0.715	-0.448	-0.480	-0.832	-0.904	-0.642	-9.386	0,99923541
	3	-0.696	**-3.396**	-0.715	-0.448	-0.480	-0.832	-0.904	-0.642	-9.387	0,99996857
	4	-0.697	**-3.401**	-0.716	-0.448	-0.481	-0.833	-0.905	-0.643	-9.396	0,99912745
	…										
	64	-0.694	**-3.390**	-0.714	-0.447	-0.479	-0.831	-0.902	-0.641	-9.374	0,9999996
…	1	…	…	…	…	…				…	
	2	…	…	…	…	…				…	
	3	…	…	…	…	…				…	
	4	…	…	…	…	…				…	
	…	…	…	…	…	…				…	
	64										

Группа 9	1	-0.399	-0.412	-0.273	-1.944	-0.481	-0.514	-0.368	-6.129	**-0.256**	0,99999654
	2	-0.339	-0.352	-0.232	-1.654	-0.411	-0.437	-0.313	-5.478	**-0.218**	0,996875965
	3	-0.355	-0.368	-0.243	-1.733	-0.430	-0.458	-0.328	-5.656	**-0.229**	0,9999555475
	4	-0.377	-0.390	-0.258	-1.838	-0.456	-0.486	-0.348	-5.892	**-0.243**	0,999365897
	…	…	…	…	…	…				**…**	
	64	-0.343	-0.356	-0.235	-1.673	-0.416	-0.442	-0.317	-5.522	**-0.221**	0,99999991

Таким образом, в ходе исследований было обнаружено, что с помощью дискриминантного анализа было правильно распознано все 9 экзаменационных объектов.

Литература

1. Состояние онкологической помощи населению России в 2011 году. Под ред. В.И. Чиссова, В.В. Старинского, Г.В. Петровой. – М.: ФГБУ «МНИОИ им. П.А. Герцена», Минздравсоцразвития России, 2012. ил. - 240 с.

2. Фомин А.А. Многомасштабный алгоритм обнаружения дефектов сварных соединений // Алгоритмы, методы и системы обработки данных. 2011. № 17. С. 15.

3. Фомин А.А., Данилов С.Д. Многомасштабный анализ объектов изображений// Алгоритмы, методы и системы обработки данных. 2008. № 13. С. 158-164.

4. Орлов А.А., Канунова Е.Е. Цифровая обработка текста на изображениях рукописей как линейчатых объектов // Информационные технологии, №1, 2008. С. 57-62.

5. Канунова Е.Е., Полякова Е.В. Особенности распознавания изображений старопечатных текстовых символов // Алгоритмы, методы и системы обработки данных. 2009. № 14. С. 55.

6. Садыков С.С. Автоматизированная обработка и анализ маммографических снимков: монография/ С.С. Садыков, Ю.А. Буланова, Е.А. Захарова; Владим. гос. Ун-т им. А.Г. и Н.Г. Столетовых.- Владимир: Изд-во ВлГУ, 2014. – 208 с.

7. Буланова Ю.А. Экспертно-аналитическая система обработки и анализа маммограмм // Прикаспийский журнал: управление и высокие технологии. 2014. №1. С. 092-102.

8. Luo S.-T., Cheng B.-W. Diagnosing Breast Masses in Digital Mammography Using Feature Selection and Ensemble Methods // J Med Syst, vol. 36, no.2, 2010. pp. 569-577

9. Cheng E., Xie N., Ling H., Bakic P.R Mammographic Image Classification Using Histogram Intersection // Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium, April 2010, pp. 197-200.

10. Садыков С.С. Регрессионные модели стенокардии и зависимость их информативности от количества параметров работы сердца/ С.С. Садыков, А.С. Белякова// Системы управления и информационные технологии. 2011. Т.45. №3.1. С.190-194.

11. Критерии выделения групп риска из лиц трудоспособного возраста при медицинских исследованиях на системе АСПО/О.И. Евстигнеева, С.С. Садыков, Е.Е. Суслова, А.С. Белякова// Алгоритмы, методы и системы обработки данных. 2012. №19. С.33-39.

12. Андрианов Д.Е. Математическая модель определения эмоционального состояния / Андрианов Д.Е., Ширабакина Т.А., Жолобов С. А. // Известия юго-западного государственного университета. 2012. №2 Часть 3. С.75-78

E-MAIL: YULIYABULANOVA@YANDEX.RU