Zibaldone. Estudios Italianos - Vol. X, Issue 2 (2023): 502-511

ISSN: 2255-3576

RESEARCH ON THE CHOICE OF ALGORITHMS FOR RECOGNIZING NEOPLASMS ON MAMMOGRAMS

Yu.A.Bulanova
Candidate of Technical Sciences, Associate Professor,

S.S.Sadykov
Doctor of Technical Sciences, Professor,

I.R.Samandarov
Candidate of Technical Sciences, Associate Professor

M.N. Malikov
Candidate of Technical Sciences, Associate Professor,

Mashurov Sh.T.
senior lecturer.

1Murmansk Institute of Vladimir State University, Murmansk, Russia
2Almalyk branch of Tashkent State University, Almalyk, Uzbekistan

Annotation

Studies of methods have been conducted to determine the possibility of their use for the recognition of neoplasms on mammographic images.

Key words: *Recognition, mammogram, neoplasm, image,*

Introduction

Recognition is the process of assigning a specific object, represented by its property values (features), to one of the fixed set of patterns (classes) based on a specific decision rule in accordance with the set goal [1, 4-6]. Since there are many different methods and algorithms, there is a need to choose an algorithm that best solves the task of recognizing new formations in mammographic images.

1. Analysis of Recognition Algorithms Features

1.1 The method of building standards

For each class, a standard is built based on the training sample, which has the values of the signs

Recibido: 27 October 2023 / aceptado: 23 November 2023 / publicado: 08 December 2023

$$\bar{x}^0 = \{x_1^0, x_2^0, ..., x_N^0\}, \quad (1)$$

где $x_i^0 = \frac{1}{K}\sum_{k=1}^{K} x_{ik}$, K – the number of objects of this image in the training sample, i is the number of the attribute.

In essence, the standard is an abstract object averaged over the training sample. It is called abstract because it may not coincide not only with any object of the training sample, but also with any object of the general population [7-9].

Recognition is performed as follows. An object is received at the input of the system $\overline{x*}$, whose belonging to this or that image is unknown to the system. From this object, the distances to the standards of all images are measured, and $\overline{x*}$ the system refers to the image whose distance to the standard is minimal. The distance is measured in the metric that is introduced to solve a specific recognition problem.

Distance measures [1]:

== Euclidean distance is the most common type of distance, which is a geometric distance in a multi-dimensional space:

$$r(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \quad (2$$

where r(x,y) – distance, x, y – objects (or features).

== The square of the Euclidean distance is used to assign greater weights to objects that are farther apart from each other:

$$r(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2, \quad (3)$$

==Manhattan distance is the average of differences along the coordinates. In most cases, this distance measure yields the same results as Euclidean distance. However, for this measure, the impact of individual large differences (outliers) is reduced since they are not squared. Manhattan distance is calculated using the formula:

$$r(x,y) = \sum_{i=1}^{n}|x_i - y_i|, \quad (4)$$

== Chebyshev distance is a measure used to determine 'different' objects based on a single coordinate (or single dimension).

$$r(x,y) = \max(|x_i - y_i|), \quad (5)$$

== Mahalanobis distance is applied in the case of dependent components x_1, x_2, ..., x_k vectors y_1, y_2, ..., y_k observations and their different significance in solving the classification issue:

$$r(x,y) = (x_i - y_i)^T \times C^{-1} \times (x_i - y_i), \quad (6)$$

Since the breast region is a spatial organization of elements within a certain part of it with homogeneous statistical characteristics, it can be said that the breast region is a complex textural image. Therefore, statistical moments of second-order spatial distributions will be used to describe the breast area, which are also called textural features of Haralik [2,3]. 16 textural features of Haralik were selected [2,3], calculated on fragments of mammographic images, where suspicious areas were found.

Each feature is calculated for the GLCM matrix rotated by 0□, 45□, 90□ and 135, respectively, the number of features for 1 image will be 64. For each type of neoplasm, all signs are entered in the table [10-11].

1.2 The k-nearest neighbors method.

The k-nearest neighbors algorithm is a metric classification method based on estimating the similarity of objects. An unknown object is assigned to the class to which the nearest objects in the training sample belong. Then, the majority class among the nearest objects in the training sample is determined, and the unknown object is assigned to that class. The optimal number of neighbors, K, is usually chosen experimentally. Increasing K reduces the influence of random errors in the data, but it may lead to less distinct class boundaries.

Let $X \in R^n$ – the set of objects, Y is the set of valid answers.

For an arbitrary object $x \in X$ there are objects of the training sample x_i in ascending order of distances up to x.

In general , the nearest neighbor algorithm looks like this:

$$a(x) = \arg\max_{y \in Y} \sum_{i=1}^{m} [x_{i;x} = y] \times w(i,x), \qquad (7)$$

where w(i,x) – a given weight function that evaluates the degree of importance of the i-th neighbor for the classification of the object u.

1.3 Fisher's linear discriminant classification

This method consists in the following sequence of actions:

Step 1. Calculation of the mean of variables in each group and the combined variance matrix;

Step 2. Reversal of the combined dispersion matrix;

Step 3. Calculation of coefficients of discriminant functions and evaluation of functions for each observation (separately).

For each group k=1,2,…,g calculate the averages and sums of mutual products of deviations from the averages.

2. Research of recognition algorithms

Since it is required to choose a specific algorithm for recognizing neoplasms, and the classification stage is the final one (after preprocessing and textural segmentation), at this point, a study of recognition methods is carried out in advance to select a specific identification algorithm.

2.1 Formation of Test (Training) Samples

The training sample was created from the MIAS database of mammographic images provided by the World Health Organization in collaboration with a radiologist. All images were categorized into 3 groups: breast cyst, fibroadenoma, and breast cancer. Within each group, 3 subgroups were formed: fatty involution, fibrocystic breast disease (FBD), adenosis, representing different levels of complexity in diagnosing new formations. For each of these prototypes, Haralick features were calculated.

Formation of Texture Features for Breast Cyst Prototypes (Fatty involution) (Group 1); (FBD) (Group 2); (Adenosis) (Group 3);

Formation of Texture Features for Fibroadenoma Prototypes (Fatty involution) (Group 4); (FBD) (Group 5); (Adenosis) (Group 6);

Formation of Texture Features for Breast Cancer Prototypes (Fatty involution) (Group 7); (FBD) (Group 8); (Adenosis) (Group 9).

Table 1

Examination Objects and Their Features"

Sign An object	Sign 1	Sign 2	Sign 3	Sign 4	…	Sign 64
Exam 1	0,001287	0,9999967	6,267989	0,393412	…	0,012568
Exam 2	0,00304	0,99999727	1,424784	0,623059	…	0,045511
Exam 3	0,005929	0,999999909	0,597948	0,761518	…	0,213413
Exam 4	0,002484	0,999999636	1,742708	0,595836	…	0,041931
Exam 5	0,004147	1	2,453902	0,552431	…	0,047062
Exam 6	0,007849	1	1,915754	0,421547	…	0,092451
Exam 7	0,002374	0,999999727	1,839062	0,584507	…	0,037071
Exam 8	0,04486	1	0,634222	0,78384	…	0,277906
Exam 9	0,25786	0,999999909	12,946662	0,920896	…	1,034229

2.2 "Study of Prototype Construction Method

For each group of objects, we construct a prototype - an abstract object averaged over the training sample, which does not coincide with any object in the training sample or in the entire population.

Table 2

(Groups 1 - 9)

Sign An object	Sign 1	Sign 2	Sign 3	Sign 4	…	Sign 64
reference 1	0.0116858	0.99999934	1.961346	0.6704056	…	0.1365834
reference 2	0.0043148	0.9999988	1.984176	0.6054398	…	0.1602186
reference 3	0.0109604	0.99999942	1.354812	0.6726674	…	0.1251634
reference 4	0.0037974	0.99999991	1.3970552	0.633027	…	0.063736
reference 5	0.0104546	0.99999938	7.247219	0.608357	…	0.0825366

reference 6	0.0239768	0.9999998	2.1896648	0.6027914	…	0.1334004
reference 7	0.0031646	0.99999956	2.3075112	0.5664378	…	0.056217
reference 8	0.012582	0.99999996	1.7574318	0.6441644	…	0.1137528
reference 9	0.2437976	0.99999998	16.4897616	0.9044898	…	0.9825436

At the next stage, the Euclidean distance (clause 1 of expression (2)) is calculated from the object of the examination sample (Table 11) to the reference objects (Table 12):

Table 3

Calculation of the minimum distance to the standards of Groups 1-9

Standards	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6	Exam 7	Exam 8	Exam 9
1	*4.317*	0.546	1.369	0.250	0.514	0.257	0.180	1.340	11.028
2	4.292	*0.571*	1.396	0.269	0.486	**0.208**	0.191	1.367	11.005
3	4.922	0.117	***0.767***	0.404	1.108	0.615	0.500	**0.746**	11.633
4	4.877	**0.035**	0.823	*0.348*	1.060	0.561	0.445	0.808	11.597
5	**1.005**	5.823	6.652	5.505	*4.794*	5.335	5.408	6.618	5.792
6	4.086	0.770	1.602	0.457	0.283	*0.331*	0.365	1.573	10.802
7	3.964	0.885	1.728	0.566	**0.147**	0.419	*0.469*	1.702	10.693
8	4.519	0.340	1.170	**0.088**	0.706	0.274	**0.127**	*1.144*	11.233
9	10.283	15.099	15.913	14.782	14.073	14.611	14.687	15.873	***3.544***

In the table.3 the minimum distances for each group are highlighted in bold; the affiliation of each examination object to the standard group is highlighted in italics; the correct recognition of the examination object to the standard is highlighted in bold italics. It turned out that using the method of building standards, only 2 objects out of 9 were correctly recognized (22.2%).

2.3. *Study of the k-Nearest Neighbors Method*

In the first step of this algorithm, we calculate the distance (d) from the examination objects (Table 11) to each object in the training samples:

-Distance from examination objects 1, 2, 3, 4, 5, 6, 7, 8, 9 to objects in the training samples of Groups 1 - 9.

Table 4

Recognition Results using the k-Nearest Neighbors Method

Groups	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6	Exam 7	Exam 8	Exam 9
1	***0.000***	3.216	2.711	4.530	3.471	3.371	1.978	3.392	1.434
2	0.211	***0.000***	0.283	*0.042*	0.247	0.294	0.348	0.349	4.299

3	*0.132*	*0.079*	***0.000***	0.290	0.317	1.100	0.508	*0.087*	5.062
4	0.526	0.319	0.601	***0.000***	*0.073*	0.148	0.084	0.198	3.997
5	1.235	0.012	1.107	0.713	***0.000***	0.208	*0.051*	0.224	*3.325*
6	0.731	0.533	0.804	0.251	0.312	***0.000***	0.189	0.159	3.838
7	0.623	0.416	0.698	0.097	0.170	*0.189*	***0.000***	0.103	3.907
8	0.171	0.156	*0.087*	0.302	0.316	1.072	0.500	***0.000***	5.012
9	6.782	9.952	9.457	11.255	10.203	10.095	8.730	10.126	***0.000***

In Table 4, the minimum distances for each group are highlighted in bold italics. It was found that using the k-Nearest Neighbors method, all 9 objects were correctly recognized (100%).

2.4. *Study of Fisher's Linear Discriminant Analysis Method*

We find: the means for each group; the sum of mutual deviations from the means; the dispersion matrix; the inverse dispersion matrix.

Table 5

Overall Means of Features for All Groups"

Sign	Meaning
1	0.038
2	1
3	4.065
4	0.656
…	…
64	0.206

The Mahalanobis distance is V = 13.502.

At the next stage, coefficients and constants for discriminant functions are calculated.

Table 6

Discriminant functions

Discriminant function	Constant	Coefficients					
		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

Discriminant functions (DF) are calculated for each feature in each group, then the numbers of functions with the highest values are selected.

Table 7

Values of discriminant functions

		Values of discriminant functions									The probability of the greatest DF
		1	2	3	4	5	6	7	8	9	
Group 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
Group 2	1	-0.571	**-2.788**	-0.588	-0.368	-0.394	-0.685	-0.741	-0.527	-8.024	0,999997514 63
	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										
Group 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**	0,999955547 5
	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

Thus, during the research, it was found that all 9 examination objects were correctly recognized using discriminant analysis.

Conclusion:

In the course of the research, it has been established that the k-Nearest Neighbors method and Fisher's discriminant classifier are the best solutions for the task of classifying new formations."

References:

1. Haralick R. M., Shanmugan K., Dinstein I. Textural Features for Image Classification // IEEE Transactions on Systems, Man, and Cybernetics, 1973. – Vol. SMC-3, No. 6. – P. 610-621.
2. Gaydel A. V., Pervushkin S. S. Research on Texture Features for the Diagnosis of Bone Tissue Diseases from X-ray Images // Computer Optics – 2013. – Vol. 37, No. 1. – P. 113-119.
3. Vapnik V. The Nature of Statistical Learning Theory, - Springer, 2000. – 314 p.
4. D. Orlov, Ya. Yu. Kulkov, S. S. Sadykov, Samandarov I. R. Algorithm for Determining the Side of a Flat Object with Symmetrical Shape. Optics, Electronic Devices, and Instruments in Pattern Recognition and Image Processing Systems. Kursk, 2021. P. 281-283
5. Sadykov S. S., Savicheva S. V., Samandarov I. R. Recognition of Individual and Overlapping Real Flat Objects Based on the Curvature of Contour Points in Their Binary Images. 2022, Gospodarka i Innowacje, 22, pp. 383-398.
6. Sadykov S. S., Savicheva S. V., Samandarov I. R. Recognition of Individual and Overlapping Real Flat Objects Based on the Curvature of Contour Points in Their Binary Images. GOSPODARKA I INNOWACJE Volume: 22 | 2022 ISSN: 2545-0573, pp. 383-398. [Link: https://www.gospodarkainnowacje.pl/index.php/poland/article/view/246]
7. Bulanova Y. A., Sadykov S. S., Samandarov I. R., Dushatov N. T., Miratoyev Z. M. Research on Methods to Enhance Contrast in Mammographic Images. Oriental Renaissance: Innovative, Educational, Natural, and Social Sciences. Vol. 2, No. 10, pp. 304-315." https://cyberleninka.ru/article/n/issledovaniya-metodov-povysheniya-kontrasta-mammograficheskih-snimkov/viewer
8. Bulanova.Yu.A, Sadykov.With.S., Samandarov I.R. , Dushatov N.T., Miratаev.Z.M. Investigation of noise filtering methods on mammographic images. Eastern Renaissance: innovative, educational, natural and social sciences. Volume 2. No. 10. pp. 177-191. https://cyberleninka.ru/article/n/issledovanie-metodov-filtratsii-shuma-na-mammograficheskih-snimkah/viewer
9. Samandarov I.R., Matsurov V. T., Mukhatov N.T., Mirzoev Z.M. Image processing in C++ using the OpenCV library. THE UNIVERSE. Technical sciences. No. 5 (110) , May, 2023

10. Samandarov I.R., Suleymanova D.B. Recognition of separate and superimposed real flat objects by the curvature of contour points of their binary images using C++. BioGecko, Journal of Herpetology of New Zealand. volume 12, issue 03.2023 . ISSN No. 2230-5807, pp. 4740-4751.
11. Y.A.Bulanova, S.S.Sadykov, I.R.Samandarov, N.T.Dushatov, Z.M.Muratov. A study of algorithms for isolating neoplasms on mammographic images was conducted. ECB. Eur.Chem.Bull.2023, 12 (Special Issue 8), pp.998-1016.