The Big Data mining to improve medical diagnostics quality

Ilyasova N.Yu., Kupriyanov A.V.

Samara State Aerospace University,
Image Processing Systems Institute, Russian Academy of Sciences

Abstract. The paper offers a method of the big data mining to solve problems of identification of cause-and-effect relationships in changing diagnostic information on medical images with different kinds of diseases. As integrated indices of the fundus vessels and coronary heart blood vessels we have used a global set of geometric features which is supposed to be a rather complete characteristic of diagnostic images and allows to make a successful diagnosis of vascular malformations. To evaluate informativity of vascular diagnostic features based on a classification efficiency criterion and in order to form new features required to improve a diagnostics quality a method of discriminative analysis of sample data has been considered. A filtration method of invalid data is proposed using a clustering algorithm to improve a performance quality of the developed algorithm of the discriminative analysis of feature vectors.

Keywords: human vascular system, image processing, diagnostic features, discriminative analysis

Citation: Ilyasova N.Yu., Kupriyanov A.V. The Big Data mining to improve medical diagnostics quality. Proceedings of Information Technology and Nanotechnology (ITNT-2015), CEUR Workshop Proceedings, 2015; 1490: 346-354. DOI: 10.18287/1613-0073-2015-1490-346-354

1. Introduction

The very large data ('big data') mining is a key problem of modern information technologies. In accordance with the Forecast of the Scientific and Technology Development of the Russian Federation for the period until 2030 approved by the Chairman of the Government of the Russian Federation Dmitry A. Medvedev promising research areas include the "Technologies of data processing and analysis" including methods and techniques of collection, processing, analysis and storage of very large volumes of information. The purpose of the paper is to develop methods and algorithms of the bid data mining to solve problems of identification of cause-and-effect relationships in changing diagnostic information on medical images with different kinds of diseases. It is offered to use a single approach to the analysis of different image classes based on evaluation of combined geometric parameters of selected areas of interest which are considered to be a basic feature set for further diagnostic analysis [1-2]. In order to identify images based on the big data mining using methods of the discriminative analysis a technique of efficient feature space formation has been developed [3-5]. As integrated indices of the fundus vessels and coronary heart blood vessels it is proposed to use the global set of geometric features which allow to make the successful diagnosis of vascular malformations [6-7]. Based on the specified methods new distributed technologies and software are developed to provide the remote image processing, analysis and understanding which are intended for implementation in automated telemedicine systems. The methods being developed are aimed at improvement of the medical diagnostics quality due to obtaining objective numerical estimates of biomedicine image parameters using large volumes of accessible information.

2. Diagnostic Image Mining Information Technology

The diagnostic image mining information technology includes a method of formation of the efficient feature space to classify a predefined image set.

A highlighting technique of diagnostically significant information on blood vessel images is based on a new generalized mathematical model of blood vessels for two classes of diagnostic images, i.e. the fundus blood-vessel system and coronary blood vessels, characterized by a set of geometric parameters [6-8].

A geometric approach to formation of diagnostic features, which, unlike traditional abstract spectral-and-correlation features are well accustomed and understandable by medical professionals, demonstrate a good visual effect and take into account an object’s specific character, allows to finally increase a diagnostics efficiency.

In order to select the most successful features we use their correlation with results of expert evaluations, as well as a dispersion analysis of learning samples or the diagnostic error analysis using particular characteristics. The effectiveness of different features is evaluated for automated diagnosis problems and proper recommendations are articulated on how to use different groups of features in medical practice.

The diagnostic image mining information technology includes the following advanced techniques and algorithms:

- the technique and the algorithm of increasing a degree of the feature informativity based on the discriminative analysis and formation of an optimal learning sample to learn a disease diagnosis expert system;
- an estimation method of a class separatability which is not influenced by distribution of objects through classes and is independent from a used classifier;
- the algorithm of decrease of feature space dimensions and formation of new informative features that maximize a separatability criterion based on methods of the discriminative analysis and allow to increase a diagnosis accuracy of a pathology degree;
- a technique of optimum learning sample formation to learn a diagnostic system based on exemption of anomalous observations that will also enable to increase a disease diagnosis accuracy.

Problem-oriented and distributed software solutions of the analysis of medical and diagnostic images are developed to detect pathological changes including software tools for quantitative estimates of the pathology degree based on expert findings and the proposed methods of classification. They are intended to ensure the user with an opportunity to manage a decision-making and analysis process [9]. Automated systems of the quantitative feature analysis make it possible to standardize a diagnosing process, considerably reduce an examination time and decrease its cost. The systems allow to carry out the analysis of subclinical morphological changes of pathomorphological components, computerize diagnosis stages and make a quantitative monitoring of pathological changes in diagnostic images. The peculiarity is the use of expert system components, i.e. a database of diagnostic features, the correlative, discriminative and cluster analysis of the feature space, and a prognosis of the pathology degree based on expert estimates.

A system of classification and diagnostic researches [1] provides tools for the correlative and discriminative analysis to form the informative feature space, resources to form an optimal feature sample based on the criterion of separation efficiency by pathology groups, and facilities for the cluster analysis to filtrate the learning sample in order to eliminate invalid data and to obtain standard feature values in accordance with pathology groups. The data mining system allows the user to obtain a proper pathology degree, standard feature values for each degree of disease pathology and a predicted disease development, and will provide proper diagnostic decisions.

3. The Discriminative Analysis to Form the Informative Feature Space

Proper researches have been performed together with medical professionals from the Medical and Stomatological University (Moscow), Ophthalmology Academic Department, based on a digital image analysis of the fundus. Diagnostics techniques for ophthalmic diseases have been developed based on evaluation of global vascular characteristics (features). The paper considers geometric characteristics proposed in [6-10]. These characteristics involve the following: the mean diameter, straightness, beads-looking shape, amplitude of thickness variations, frequency of thickness variations, thickness tortuosity, amplitude of path variations, frequency of path variations and path tortuosity, which correspond to diagnostic features of the fundus vessels. If two or more classes available (in our case there are 5 classes including a normal feature and 4 degrees of diabetic retinopathy and diabetes), the objective of feature selection is to select those features which are the most efficient in accordance with the class separability [5,8]. In the discriminative analysis the class separability criteria are formed using scatter matrixes inside classes and scatter matrixes between classes [11,12].

The scatter matrix inside classes demonstrates a variety of objects with respect to mean vectors of classes: $\mathbf{W} = \sum_{k=1}^{g}(\mathbf{X}_k - \bar{\mathbf{x}}_k)(\mathbf{X}_k - \bar{\mathbf{x}}_k)'$, where k – class data will correspond to the mean vector $\bar{\mathbf{x}}_k = [\bar{x}_{1k}\bar{x}_{21k}\ldots\bar{x}_{pk}]$, and g – is a total number of classes.

Elements of the scatter matrix between classes B is counted by the formula: $b_{ij} = \sum_{k=1}^{g} n_k \left(\bar{x}_{ik} - \bar{x}_i\right)\left(\bar{x}_{jk} - \bar{x}_j\right), i, j = \overline{1, p}$, $\bar{x}_i = \left(1/n\right)\sum_{k=1}^{g} n_k \bar{x}_{ik}$ – is a mean feature value of i - feature in all classes, n_k – is a number of objects in k - class, $\overline{x}_{ik} = 1/n_k \sum_{m=1}^{n_k} \overline{x}_{ikm}$ – is the mean feature value in class k, and x_{ikm} – is a value of i - feature for m-object in k - class. The matrixes W and B contain all basic information about interrelationships inside and between classes. In order to obtain the class separability criterion some number is to be associated to these matrixes. This number should be increased with the increase of scattering between classes or with the decrease of scattering inside classes. For this purpose the following criterion is more frequently used [11]: $J_1 = tr(\mathbf{T}^{-1}\mathbf{B})$, where $\mathbf{T} = \mathbf{B} + \mathbf{W}$. The greater the value of the criterion - the more separability of classes. The following algorithm presented in Figure 1 has been developed to form new features.

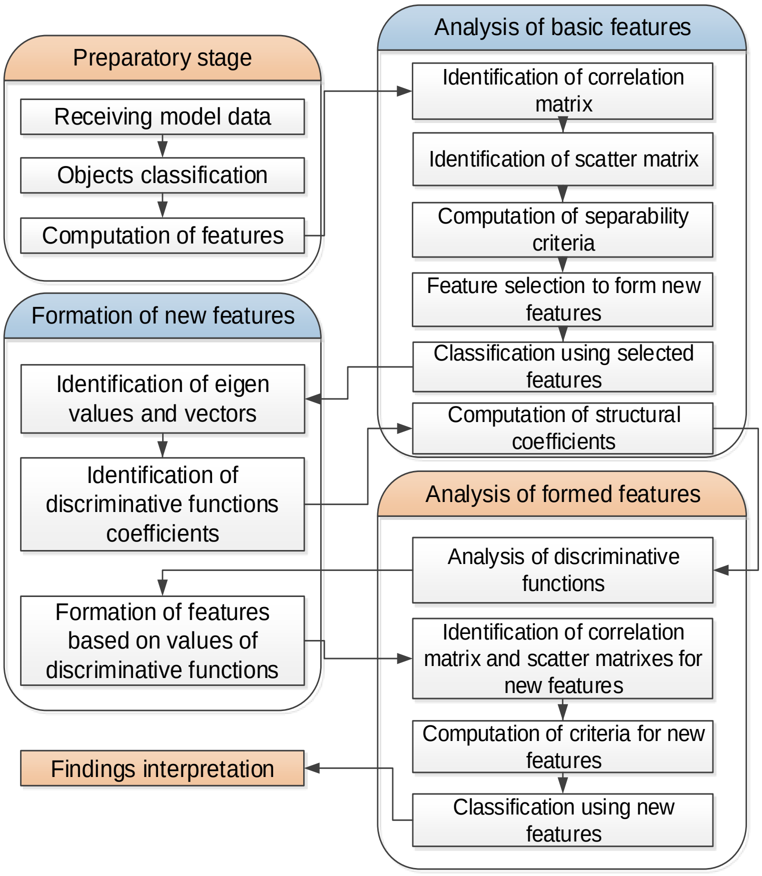
Preparatory stage
Receiving model data
Objects classification
Computation of features
Analysis of basic features
Identification of correlation matrix
Identification of scatter matrix
Computation of separability criteria
Feature selection to form new features
Classification using selected features
Computation of structural coefficients
Formation of new features
Identification of eigen values and vectors
Identification of discriminative functions coefficients
Formation of features based on values of discriminative functions
Analysis of formed features
Analysis of discriminative functions
Identification of correlation matrix and scatter matrixes for new features
Computation of criteria for new features
Classification using new features
Findings interpretation

Fig. 1. – Algorithm of the discriminative feature analysis

4. Experimental Studies

A number of studies of vascular malformations have been undertaken for diabetic retinopathy (DR) of 151 patients suffered with diabetes (D) based on the digital image analysis of the fundus. After processing of images the sample amounted to 8175 measurements including 1490 arteriols of the 1st-order, 2345 arteriols of the 2nd-order, 1960 venules of the 1st-order, and 2380 venules of the 2nd-order. Medical professionals separately consider venules and arteriolas of the 1st and 2nd order (GROUP 1 – GROUP 4) since different tendencies of vessel changes can be observed in these GROUPS at different pathology stages. Figure 2 shows some examples of diagnostic images of the fundus for different stages of diabetes (1D, 4D) and a measurement sketch of the vascular system that gives an order of the vessel examination accepted by ophthalmologists.

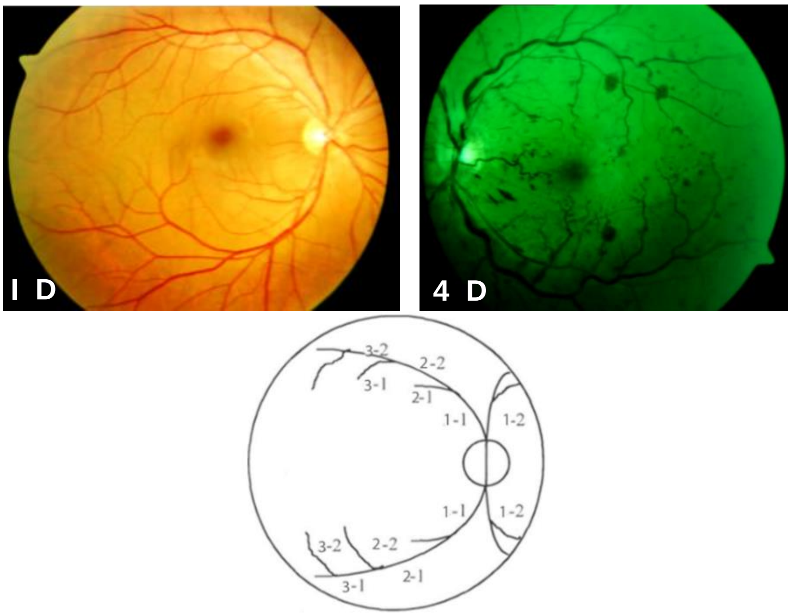

Fig. 2. – Examples of diagnostic images of the fundus for different stages of diabetes (1 D, 4 D) and a measurement sketch of the vascular system

At the examination of features it was concluded that there are two high-correlated characteristic groups. The first group includes those features which describe path parameters (e.g. the path straightness and tortuosity), and the second group includes the features characterizing a vessel radius function (e.g. the radius tortuosity and a beads-factor). Figure 3 gives values of the separability criteria of single features for each of the four GROUPS (venules and arteriols of the 1st and 2nd orders).

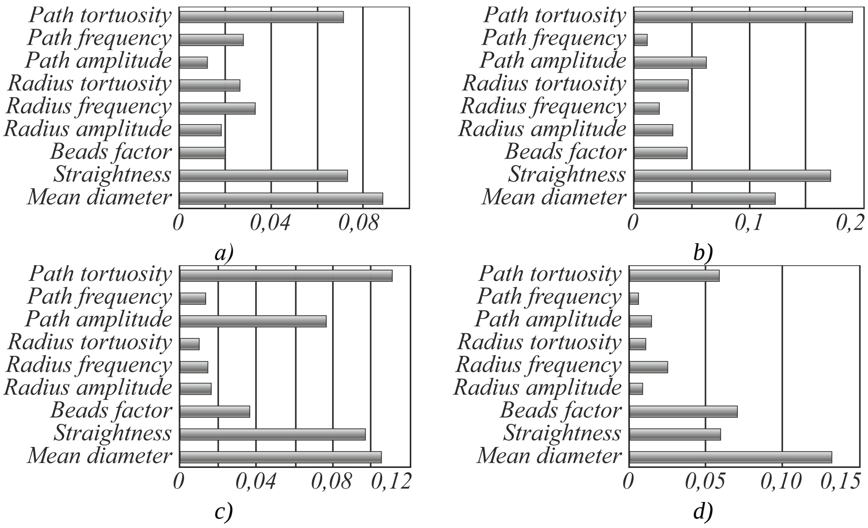

Fig. 3. – Values of feature separability criteria for each of the four GROUPS of vessels before filtration: a) and b) arteriols of the 1st and 2nd orders, and c) and d) venules of the 1st and 2nd orders

Figure 3 shows that maximum feature values belong to the vessel features which are not particularly significant for medical professionals in diagnosing. Besides, in proper pathological conditions each group may have vessels which do not comply with the given pathology (for instance, the norm feature). It may be thus concluded that the sample may contain some noisy data.

In order to eliminate noise in the sample it is proposed to filter down an original sample using the clustering approach of k-means. Each GROUP was divided into 5 clusters in accordance with original clusters inside each of the GROUPS. Feature vectors which were not included into a correct cluster were removed from the sample. Values of individual separability criteria for features inside the GROUPS after filtration are given in Figure 4.

Analyzing Figure 4 we may conclude that in the sample obtained after filtration the features with the greatest separability criterion are the characteristics which are considered by medical professionals as specifically diagnostically significant in visual diagnosing of pathology that corresponds to an information letter [13]. The discriminative analysis algorithm has been applied for original and filtered samples. In order to form new features we have completely enumerated original features to search a combination of new features that has maximized the separability criterion. The result was that we have obtained a set of four features for all GROUPS of either sample.

The classification error was evaluated for the GROUPS obtained thereby. The error was estimated by means of the U-approach [11]. Two samples were formed, i.e. the learning and test samples. The classifier was tuned using the learning sample based on the SVM approach (Support Vector Machines), and the test sample was thereby classified. Objects of the learning sample which are not contained within the test sample are only used for the classifier synthesis. There are many opportunities to implement the U-approach. In order to evaluate probability of the classification error in researches a one-object-elimination method was used. The findings are illustrated in Table 1 where criterion values before the discriminative analysis are presented for the combination of original sample quads with the best separability values inside the GROUPS.

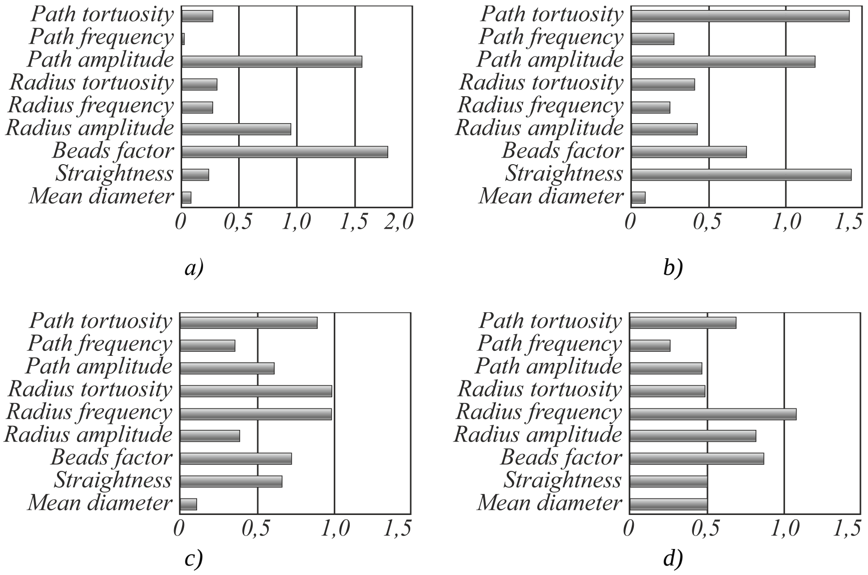

Fig. 4. – Values of feature separability criteria for four GROUPS of vessels after filtration: a) and b) arteriols of the 1st and 2nd orders, and c) and d) venules of the 1st and 2nd orders

Analyzing the obtained findings it may be concluded that filtration allows not only to increase feature separability criteria and decrease the classification error, but also to identify diagnostically significant features. Researches made on four GROUPS of blood vessels have shown that it is important for each GROUP to have its own set of diagnostic features that is proved by clinical research studies of medical practitioners. For example, the mean diameter of the blood vessel acts differently for venules and arteriols under pathological changes. The research results have showed that the application of the proposed feature formation technique allowed not only to eliminate invalid data, but also led to a reduction in errors of classification. This resulted to the increase of the separability criterion by 39% for GROUP 1, by 42% - for GROUP 2, by 24% - for GROUP 3 and by 39% - for GROUP 4. Besides, we have obtained some additional information based on the used features such as their informativity and have identified relationships between some features.

Table 1. Findings of the feature space discriminative analysis

Group		Original sample			Filtered sample		
		J	Increasing criterion	Error	J	Increasing criterion	Error
Arteriols of the 1st order	befor	0.21	40%	0.185	1.69	9%	0.071
	after	0.31		0.104	1.85		0.024
Arteriols of the 2nd order	befor	0.35	19%	0.144	2.48	6%	0.083
	after	0.42		0.090	2.62		0.028
Venules of the 1st order	befor	0.29	50%	0.128	2.07	17%	0.045
	after	0.44		0.096	2.41		0.035
Venules of the 2nd order	befor	0.28	27%	0.162	2.21	12%	0.072
	after	0.36		0.113	2.47		0.023

5. Conclusion

A new technology has been developed for the data mining to identify cause-and-effect relationships of changing diagnostic information on medical images with different kinds of diseases including the filtration of invalid data and the discriminative analysis based on maximizing the separability criterion.

The algorithm based on the selection of features with the biggest separability criterion and on a complete enumeration followed by new features maximizing this criterion has been developed. Separability criteria are formed using the scattering matrixes between and inside classes. As a result of the discriminative analysis the best features have been defined for each group of vessels based on the separability criterion. It is shown that each of the four GROUPS of vessels has efficiently used its own set of global geometric features that has been proved by clinical researches. The classification error was calculated for each group of vessels before and after the algorithm performance. It is shown that the proposed technology of the feature space analysis by groups, including the algorithm of filtration of sampled data and the algorithm of formation of the efficient feature space, allowed to increase the classification efficiency of vessels by “normal-feature” classes and different degrees of “pathology”. The classification error was hereby reduced to 2.3%-3.5% for different pathology GROUPS.

Acknowledgements

This work was partially supported by the Ministry of education and science of the Russian Federation in the framework of the implementation of the Program of increasing the competitiveness of SSAU among the world’s leading scientific and educational centers for 2013-2020 years; by the Russian Foundation for Basic

Research grants (# 14-01-00369-a, # 14-07-97040-p_ povolzh'e_a, # 15-29-03823, # 15-29-07077); by the ONIT RAS program # 6 “Bioinformatics, modern information technologies and mathematical methods in medicine” 2015.

References

1. **Ilyasova NYu.** Diagnostic computer complex for vascular fundus image analysis, Biotehnosfera, 2014; 33(3): 20-24. [in Russian]
2. **Ilyasova NYu.** Methods for digital analysis of human vascular system. Literature review. Computer Optics, 2013; 37(4): 517-541. [in Russian]
3. **Simcher, VM.** Methods of multivariate statistical analysis. M.: Finance and statistics, 2008. [in Russian]
4. **Mookiaha MRK, Acharyaa UR, Lima CM, Petznickb A, Jasjit S.** Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features. Knowledge-Based Systems, 2012; 33: 73-82.
5. **Ilyasova NYu, Kupriyanov AV, Paringer RA.** Formation features for improving the quality of medical diagnosis based on the discriminant analysis methods. Computer Optics, 2014; 38(4): 851-855. [in Russian]
6. **Ilyasova NYu.** Estimating the geometric features of a 3D vascular structure. Computer Optics, 2014; 38(3): 529-538. [in Russian]
7. **Ilyasova NYu.** Methods to Evaluate the Three-Dimensional Features of Blood Vessels. Optical Memory and Neural Networks (Information Optics), 2015; 24 (1): 36-41.
8. **Ilyasova NYu, Kupriyanov AV, Khramov AG.** Information technologies of image analysis in medical diagnostics. M: Radio and Communication, 2012. [in Russian]
9. **Ilyasova N.** Computer Systems for Geometrical Analysis of Blood Vessels Diagnostic Images. Optical Memory and Neural Networks (Information Optics), 2014; 23 (4): 278-286.
10. **Ilyasova NYu, Kupriyanov AV, Ananin MA.** Measurement of the biomechanical vessels parameter for the diagnostics of the early stages of the retina vascular pathology. Computer Optics, 2005; 27: 165-169. [in Russian]
11. **Fukunaga K.** Introduction to statistical pattern recognition. New York and London: Academic Press, 1972.
12. **Kim JA, Myuller ChU, Klekka WR.** Factor, discriminant and cluster analysis. M.: Finance and Statistics, 1989. [in Russian]
13. **Moshetova LK, Yushchuk ND, Tsyganov DI, Sister VG, Branchevsky SL, Ilyasova NYu, Pavlova Y.** Method for digital image processing fundus. Information letter, 2004; 1.