Bull. Nov. Comp. Center, Comp. Science, 33 (2012), 79–84

Data space dimensionality reduction in the problem of diagnosing a thyroid disease

M.S. Tarkov, E.A. Chiglintsev

Abstract. Analysis of a set of data space dimensionality reduction methods in image recognition problems is carried out. A problem of diagnosis of thyroid diseases with the use of images of cytological preparation is investigated. It is shown that the morphological image analysis combined with the method of diffusion maps makes it possible to obtain a higher recognition accuracy in the problem of diagnosis of a thyroid disease than with the previously proposed method based on the Fourier correction spectrum and selection of principal components.

1. Introduction

Currently, the image recognition is widely used in many areas, such as automatic monitoring system, electronic locks using retina and finger-print images instead of a key, medical and engineering diagnostics, etc. The research into the image recognition collides with a problem of "dimensionality damnation", that is, a quickly increasing data dimensionality. The data dimensionality reduction becomes an immediate problem for image recognition techniques.

2. Dimensionality reduction

In the general case [1], the dimensionality reduction problem is stated as follows. Let us have a data set X in a multidimensional space E^n. The data presented by a set of k vectors is really a matrix with $n \times k$ dimensionality. We suppose the data set to have an internal dimensionality $m < n$, and in practice, we have often situations with $m \ll n$. The conception of internal dimensionality implies that the elements of X belong to some manifold with dimensionality m in the space E^n. Dimensionality reduction algorithms implement a mapping of the set X with dimensionality n into a set Y with dimensionality m with a possible retention of the initial geometry of the set X. In the general case, geometrical properties of the set X and its internal dimensionality are not known. Thus, the problem of dimensionality reduction is, in general, improperly posed and cannot be solved without additional assumptions about initial data. In particular, we can make a priori assumption about internal dimensionality of the set X.

Let us consider the dimensionality reduction problem as applied to the image recognition. Let us have a set of images, and we need to reduce their

dimensionality for the following recognition. It is impossible to essentially diminish the data dimensionality without loss of information. The question arises: what sort of information we want to save in images when we reduce their dimensionality. The reply is not obvious. Moreover, when mapping the data onto a space with a lesser dimensionality we can obtain additional properties or we can distinguish properties that are not brightly expressed in the original space. After all, this depends upon a recognition problem under consideration.

Let us consider a classification problem. As the objective of the dimensionality reduction we consider organization of mapping which will provide the greatest separability of classes in the output space. A statistical sampling parameter known as separability [2], characterizes the complexity of the classification. A characteristic feature of the image classification problem is that the number of samples is usually much less than the dimension of the problem. The difference can be several orders of magnitude.

An approach based on the analysis of morphologic functions of the binary images characterizing the connectivity, the number and the shape of objects is proposed in [3]. The integral geometry provides mathematical fundamentals for determination of such functions called the Minkowski functionals. In a space R^2, there are three Minkowski functionals: an area M_0, a perimeter M_1, and an Euler characteristic M_2. The Euler characteristic is a variable-sign sum of numbers of simple elements (symplexes) with different dimensionalities resulting in the decomposition of a body. A simple algorithm for evaluation of the Euler characteristic is based on this definition. By counting the number of edges, nodes and facets, entirely owned by a region bounded by two contours, it is easy to see that

$$M_2 = \textit{number-of - “islands”} - \textit{number-of- “holes”}.$$

We can assume every black pixel of a binary image a node of the body, two black adjacent pixels form the edge, and four black adjacent pixels to form the body facet. Thus, it is possible to evaluate the Euler characteristic of the binary image.

Let the gray picture be subject to the threshold processing with the threshold values from 0 up to 255. For every resulting binary image evaluate the Euler characteristic. As a result, we have a vector with dimensionality 255. In the analysis of images of slices of biological tissues, the morphological properties play the key role [3].

The diffusion maps [1] provide a nonlinear approach to mapping of multidimensional data onto spaces with lesser dimensionality. They demonstrate good results in the solution of some practical problems. This approach is based on representation of data as nodes of a weighted graph.

Assume it is required to construct a mapping of the data set $X = \{x^1, x^2, \ldots, x^k\} \subset R^n$, where $x^i = \{x_1^i, x_2^i, \ldots, x_n^i\}$, $i = 1, 2, \ldots, k$, onto

a data set $Y = \{y^1, y^2, \ldots, y^k\} \subset R^m$, where $y^i = \{y_1^i, y_2^i, \ldots, y_m^i\}$, $i = 1, 2, \ldots, k$, and $m \ll n$. Suppose that $X \subset M$, where M is a manifold embedded in R^n. Let us consider k points as nodes of a complete weighted graph with special edge weights. We are interested in the matrix of probabilities of transitions between the nodes of the data graph. Eigenvectors of this matrix can be considered as axes of the resulting map. Thus, constructing of the diffusion map consists of the following four stages [1]:

1. Construction of the adjacency matrix W of the considered graph. The entries of the matrix are the edge weights. The edges are weighted as follows:

$$W_{ij} = \exp\left(-\frac{\|x^i - x^j\|^2}{2\sigma^2}\right).$$

Here $\|\cdot\|$ is the Euclidean norm, σ is a parameter chosen in terms of an experiment, and W is a symmetric matrix of $k \times k$ size.

2. Normalization of the matrix W:

$$P_{ij}^{(1)} = \frac{W_{ij}}{\sum_k W_{ik}}.$$

The matrix $P^{(1)}$ is considered to be a matrix of probabilities of transitions between nodes of the data graph. Then the transition matrix for t steps is $P^{(t)} = (P^{(1)})^t$.

3. Evaluation of eigenvalues and eigenvectors of the spectral problem

$$P^{(t)} y = \lambda y,$$

where $y \in R^m$ are columns of the matrix Y.

4. The largest eigenvalue $\lambda_1 = 1$ and the corresponding eigenvector y^1 are not considered. The eigenvectors corresponding to other m largest eigenvalues are used for the representation of the resulting space by the diffusion map:

$$\Psi_m : x^i \to (y_2^i, y_3^i, \ldots, y_{m+1}^i).$$

3. Cytological diagnosis of follicular thyroid tumors

The cytological diagnosis of a thyroid disease is usually a determining factor when choosing of conservative or a surgical treatment. The greatest difficulties of a differential diagnosis is follicular thyroid tumors, in which there are no distinct differential diagnostic features that distinguish follicular adenoma from a highly differentiated follicular cancer (Figure 1).

Neural networks are conventionally used in the medical diagnosis. In most cases, data are preprocessed before making the direct analysis and classification. In [4], a principal component method with transformation of

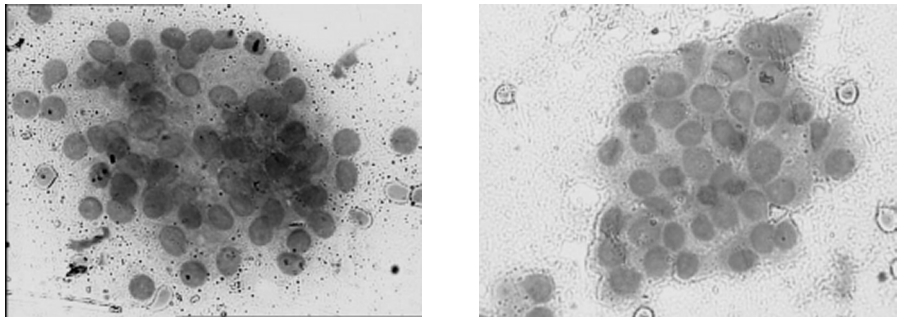

Figure 1. Images of cytological preparations: follicular adenoma (left) and follicular cancer (right)

the Fourier spectrum of images is used to increase the image separability and dimensionality reduction.

Application of the morphological approach does not give a good separation of classes without additional data transformation. In this approach, the Euler characteristics of the sample data vectors are mapped onto a plane using multidimensional scaling and the Euclidean distance [1, 5]. The mapping result is shown in Figure 2. As can be seen from Figure 2, the classes are significantly mixed.

A data sample obtained from the source images by calculating the Euler characteristics was mapped onto a plane by a number of different generic

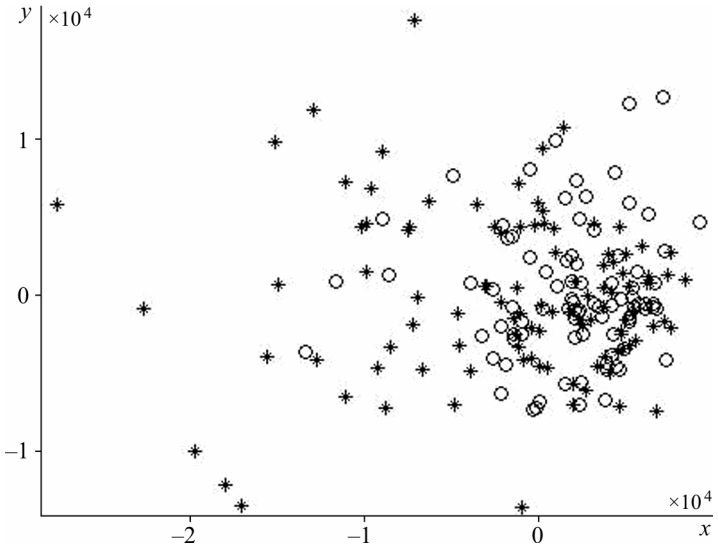

Figure 2. Mapping of images of cytological preparations of a thyroid gland onto a plane with the morphological approach and multidimensional scaling ("stars" — adenoma, "circles" — cancer)

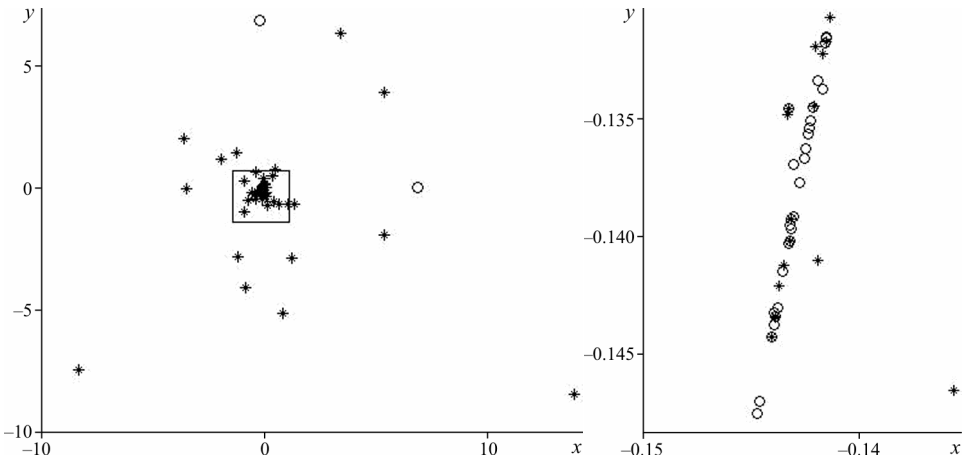

Figure 3. Mapping of images of thyroid cytological preparations onto a plane by morphological analysis and diffusion maps with $t = 32$ and $\sigma = 1$ ("stars" — adenoma, "circles" — cancer)

methods [1]. Ultimately, the diffusion maps gave the best mapping. The result of applying diffusion maps with $t = 32$ and $\sigma = 1$ to the result of morphological analysis of the sample is shown in Figure 3 (left).

In Figure 3 (right), the central site from the left of Figure 3 is several-fold increased. As can be seen, the mapping turned out to be good. Despite the fact that the points corresponding to the cancer samples are somewhat mixed with samples of adenoma, many of them are quite compact and isolated from most points of adenoma. Thus, the classes are easily distinguished visually.

The image mapping by the morphological analysis and diffusion maps allow making a good classification of cytological data. As in [4], we use test images on 45 patients with adenoma and 39 patients with follicular cancer. Other data were considered as a training set with known classes. The results from [4] are shown in Table 1.

Here N^+ equals the number of images classified as adenoma, and N^- equals the number of images classified as cancer. The authenticity of classification is evaluated as the ratio of the number of correctly recognized images to the total number of images classified as images of the same type.

Table 1

Class	Adenoma	Cancer
N^+	37	16
N^-	8	23
Authenticity	0.70	0.74
Sensitivity	0.82	0.59

Table 2

Class	Adenoma	Cancer
N^+	37	8
N^-	8	31
Authenticity	0.82	0.79
Sensitivity	0.82	0.79

The sensitivity is the ratio of the number of correctly recognized images of a considered class to the total number of images of this class. In this paper, for recognition we use a method of k nearest neighbors with a parameter $k = 5$. The results are shown in Table 2.

As can be seen from Tables 1 and 2, the morphological analysis in conjunction with diffusion maps allow us to obtain slightly better results in terms of authenticity and sensitivity.

4. Conclusion

The choice of a method of data dimensionality reduction plays an important role in obtaining high quality solutions of the problems of image recognition. In this paper, we show that the morphological image analysis in conjunction with the diffusion maps makes it possible to obtain a higher quality of recognition in the complex problem of diagnosing diseases of the thyroid gland by the images of cytological preparations as compared to the previously proposed method based on the correction of the Fourier spectrum of images and distinguishing principal components.

References

[1] Van der Maaten L., Postma E., van den Herik J. Dimensionality Reduction: A Comparative Review, October 2009 / Tilburg Center for Creative Computing, Tilburg University, the Netherlands. — http://www.uvt.nl/ticc.

[2] Jing X.-Y., Tang Y.-Y., Zhang D. A Fourier-LDA approach for image recognition // Pattern Recognition. — 2005. — Vol. 38. — P. 453–457.

[3] Makarenko G., Mekler A.A., Zabrodskaja Yu.M., Knjazeva I.S., Diagnostics of brain tumors by intelligent analysis of histological data // Proc. XIII Russian Scientific-eng. Conf. Neuroinformatika-2011, Part 3. — Moscow, 2011. — P. 71–77 (In Russian).

[4] Shapiro N.A., Poloz T.L., Shkurupij V.A., Tarkov M.S., Poloz V.V., Demin A.V. Application of artificial neural network for classification of thyroid follicular tumors // Analytical and Quantitative Cytology and Histology. — 2007. — Vol. 29, No. 2. — P. 87–94.

[5] Kohonen T. Self-Organizing Maps. — 3rd ed. — Springer, 2001. — (Springer series in information sciences; 30).