Krzysztof SIWEK[1], Stanisław OSOWSKI[1,2], Tomasz MARKIEWICZ[1,3], Jacek KORYTKOWSKI[1]

Warsaw University of Technology (1), Military University of Technology (2), Military Institute of Medicine, Warsaw (3)

Analysis of medical data using dimensionality reduction techniques

Abstract. *The paper presents the application of dimensionality reduction methods for representation of the multidimensional medical data representing the images of the blood cells in leukemia. Different techniques of reduction belonging to linear and nonlinear methods will be applied and their efficiency compared. Their application to the visualization of different classes as well as clusterization and classification of data will be studied and discussed in the paper.*

Streszczenie *Praca przedstawia zastosowanie różnych metod redukcji wymiaru danych w reprezentacji numerycznej deskryptorów charakteryzujących klasy komórek krwiotwórczych w białaczce. Porównane zostaną różne podejścia do redukcji oparte na metodach liniowych i nieliniowych transformacji. W szczególności analizie poddane zostaną możliwości zastosowania tych metod w wizualizacji danych jak również klasteryzacji i klasyfikacji. W pracy pokazane zostaną wyniki przeprowadzonych eksperymentów dotyczących 11 klas komórek. (**Analiza wielowymiarowych danych medycznych z użyciem wybranych technik redukcji wymiarów**)*

Keywords: multidimensional reduction techniques, data visualization, data clustering and classification.
Słowa kluczowe: techniki redukcji wielowymiarowej, wizualizacja danych wielowymiarowych, grupowanie i klasyfikacja danych.

Introduction

Medical data usually contains very large amount of information hidden in the form of either images or signals. To uncover the kernel of this information we have to apply specialized tools allowing to detect the piece of information we actually need. The first step of such processing is the analysis of data usually associated with the optimal reduction of their size [2]. This step involves the reduction of either dimension of the observation vectors or representation of them by smaller representative population.

In this paper we apply both forms of reduction of medical data concerning the blood cell analysis. In the first step we reduce the dimensionality of the observation space and in the second we apply the clustering procedure to analyze the distribution of the centers of data. Few methods of reduction will be tried and compared: Principal Component Analysis (PCA), Kernel PCA (KPCA), Sammon mapping as well t-distributed Stochastic Neighbor Embedding (t-SNE) [3,6].

The analysis is aimed at discovering the intrinsic properties covered in the numerical descriptors of the family of the blood cells. We consider 8 types of the blood cells important in the recognition of the myelogenous leukaemia, which are distinguished in the image of the bone marrow smear. The data base contains the family of erythroblast cells (basophilic, polychromatic and ortochromatic), lymphocyte cells (prolymphocyte and lymphocyte), neutrophilic band and neutrophilic segmented gathered in one class, as well as myeloblasts, myelocytes, promyelocytes, metamyelocytes and plasmocytes.

The images of these cells have been characterized by 87 numerical descriptors on the basis of the geometry, texture and color distribution [4]. These numerical descriptors are subject to the analysis in this work.

Dimensionality reduction methods

The problem of dimensionality reduction is defined as follows. Let us assume we have a dataset represented by $p \times N$ matrix $\mathbf{X}$, containing p data vectors $\mathbf{x} = [x_1, x_2, ..., x_N]$ of zero mean value. Dimensionality reduction is understood as a transformation of the original data set $\mathbf{X}$ of dimensionality N into a new data set $\mathbf{Y}$ of dimensionality n, while retaining the geometry of data as much as possible. In general we know neither the geometry of data manifold, nor the intrinsic dimensionality of the original data set X. The original vectors of dimension N will be denoted by $\mathbf{x}_i$ and the transformed vectors of the reduced dimensionality by $\mathbf{y}_i$ (i=1, 2, ..., p).

The most popular transformation technique of this type is the PCA [3]. This transformation constructs the low-dimensional characterization of data searching for the direction where the variance of the data is maximum. This task is solved by finding the linear transformation

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{1}$$

in which the amount of variance in the data is maximal. In mathematical terms it can be written as the maximization of the determinant of the matrix

$$\max_{\mathbf{w}} J = \left|\mathbf{W}^T \mathbf{R}_{xx} \mathbf{W}\right| \tag{2}$$

This problem is solved by defining the auto-covariance matrix $\mathbf{R}_{xx}$ of the set, i.e., $\mathbf{R}_{xx}=E[\mathbf{x}\mathbf{x}^T]$ and performing the . eigenvalues decomposition

$$\mathbf{R}_{xx} = \sum_{k=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \tag{3}$$

On the basis of this description we define the PCA matrix $\mathbf{W}$, $\mathbf{W} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K]^T$, considering only the eigenvectors $\mathbf{v}_i$ corresponding to the highest eigenvalues λ_i.

In kernel PCA [6] we transform first the original data using the nonlinear kernel function and then perform the PCA on this transformed data set. KPCA computes first the kernel matrix $\mathbf{K}$ of the data points $\mathbf{x}_i$. The entries of this matrix are defined by $k_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)$, with k the kernel function adjusted in such a way that it gives rise to a positive semi-definite kernel matrix $\mathbf{K}$. This matrix is double centered using the following modification [6]

$$\widetilde{\mathbf{K}} = \mathbf{K} - \mathbf{1}_p \mathbf{K} - \mathbf{K}\mathbf{1}_p + \mathbf{1}_p \mathbf{K} \mathbf{1}_p \tag{4}$$

In this equation $\mathbf{1}_p$ means the quadratic $p \times p$ matrix, of the entries equal $1/p$. The eigenvalue decomposition is done on this kernel matrix and further steps identical as in PCA transformation.

The next transformation considered in the paper is the nonlinear Sammon mapping [3]. This transformation maps the original data points to the reduced dimension space in a way to preserve as much as possible the relations of distances between the points in the original and new space. Mathematically the problem is described as the minimization of the cost function defined in the following way

(5) $$\min\ E = \frac{1}{c}\sum_{i<j}\frac{\left(d_{ij}^{*}-d_{ij}\right)^{2}}{d_{ij}^{*}}$$

where d_{ij}^{*} is the Euclidean distance between the data points in the original space and d_{ij} in the reduced space. The parameter *c* is defined as $c=\sum_{i<j}^{p} d_{ij}^{*}$. The minimization of the cost function (7) is done by using Newton optimization method.

The last transformation method applied in this work is t-distributed Stochastic Neighbor Embedding [6]. This method starts by converting the high dimensional Euclidean distances between original data points into the conditional probability $p_{j|i}$ that $\mathbf{x}_i$ would pick $\mathbf{x}_j$ as its neighbor if neighbors were picked in proportion to their probability density under Gaussian centered at $\mathbf{x}_i$. To avoid the problem of non-symmetric distance measure the conditional probability is replaced in practice by a joint probability p_{ij}. The learning problem is transformed to the minimization of the Kullback-Leibler divergence between the joint probability distribution *P* in a high dimensional space and a joint probability distribution *Q* in a reduced dimensional space

(6) $$\min\ E = \sum_{i}\sum_{j} p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

where p_{ij} and q_{ij} are described by [6]

(7) $$p_{ij} = \frac{\exp\left(-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2/2\sigma^2\right)}{\sum_{k\neq l}\exp\left(-\left\|\mathbf{x}_k-\mathbf{x}_l\right\|^2/2\sigma^2\right)}$$

(8) $$q_{ij} = \frac{\left(1+\left\|\mathbf{y}_i-\mathbf{y}_j\right\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left\|\mathbf{y}_k-\mathbf{y}_l\right\|^2\right)^{-1}}$$

The parameter σ is the standard deviation of the Gaussian function (also subject to optimization). The original algorithm described in [6] applies the gradient descent method with momentum.

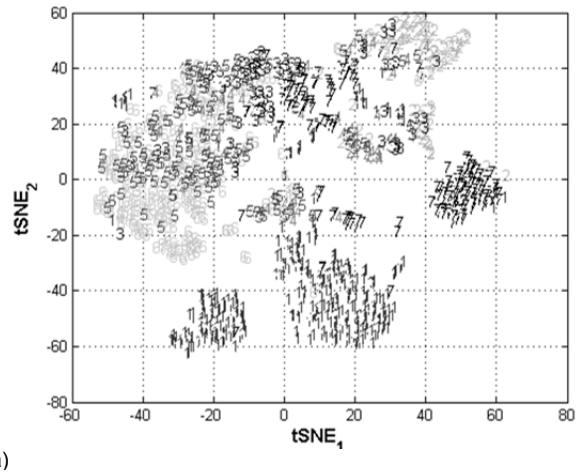
Data base

The medical data of blood cell images which are important in the recognition and treatment of leukemia have been obtained thanks to the cooperation with Hematology Institute of Warsaw [4]. The total number of cells considered in this research was equal 1329. In the introductory step the cell images have been converted to the numerical descriptors based on the geometry, texture and color description of the cells [4]. As a result 87 numerical descriptors have been generated and subject to further analysis in this work. The data have been arranged into 7 classes, depicting the development stages of these cells in bone marrow. This arrangement was done in the following way:

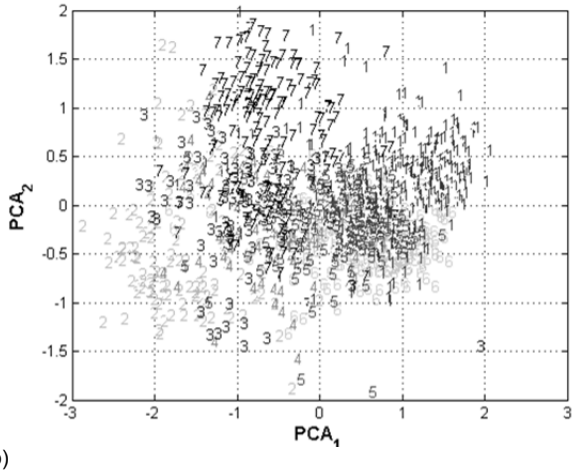
- class 1 – the erythroblast cells (basophilic, polychromatic and orthochromatic) – totally 374 records of data,
- class 2 – myeloblasts – 130 records,
- class 3 – promyelocytes – 93 records,
- class 4 – myelocytes – 102 records,
- class 5 – metamyelocytes – 142 records,
- class 6 – neutrophils (neutrophilic band and segmented) – totally 293 records,
- class 7 – lymphocyte cells (prolymphocyte and lymphocyte) – totally 195 records.

Visualization of the medical data

In the first step we have tried to apply the reduction techniques in the visualization of the data by keeping the size of transformed data equal 2. It is quite important step in the analysis, since it shows the potential tendency of clusterization of data and on the basis of this we may conclude of the difficulty of cell recognition in the next phase of analysis. This analysis has shown that the best results correspond to tSNE, while the other (PCA, KPCA, LDA, Sammon) were of inferior, although similar quality. The distributions of data mapped on the 2-dimensional space are presented in Fig. 1, where Fig. 1a depicts the results of tSNE and Fig. 1b – of PCA. The members of each class of data are denoted by the numbers (from 1 to 7).



a)



b)

Fig.1. The visualization of blood cell data by applying a) tSNE, b) PCA

It is evident that much better consistence of classes grouped in the clusters have been obtained at application of nonlinear mapping tSNE. We can see clear recognition between clusters, they are less interlaced and the uniformity of members of all clusters seems to be better.

Clusterization of the data

In the next step we have made the numerical characterization of clusterization results of the original and mapped data by applying the K-means algorithm [1,2,5] with the number of cluster equal to the number of classes. The aim is to discover what is the association of the clusters with the classes. In the first case we have clustered the original data of the dimension equal 87. Next we have mapped the original data to smaller dimension and performed the K-means clusterization procedure on these mapped data. We have tried different dimensions of transformed data. The best results of uniformity have been obtained at K=30.

Two different classification-oriented measures of clusterization have been taken into account. The first one is based on entropy, depicting the degree to which each

cluster consists of objects of a single class. Let p_{ij} denotes the probability that a member of cluster i belongs to class j, $p_{ij} = n_{ij} / n_i$, where n_{ij} is the number of objects of *j*th class in *i*th cluster and n_i is the number of objects in class *i*. Then the entropy of ith cluster is described by [5]

$$e_i = -\sum_{j=1}^{M} p_{ij} \log_2 p_{ij} \tag{9}$$

The total entropy for a set of clusters is calculated as the sum of entropies weighted by the relative size of each cluster, i.e.,

$$e = \sum_{i=1}^{K} \frac{n_i}{n} e_i \tag{10}$$

where *n* is the total number of data points. The smaller the total entropy the better is the clusterization.

Another measure of the quality of clusterization is the purity. It measures the extent to which a cluster contains objects of a single class. Applying the previous terminology the purity of *i*th cluster is $p_i = \max_j p_{ij}$ and the overall purity of the clustering is equal [5]

$$p = \sum_{i=1}^{K} \frac{n_i}{n} p_i \tag{11}$$

The higher is this measure the better clusterization results.

The clusterization have been performed for the original data and the transformed data mapped into 30 dimensions by applying such transformation techniques as PCA, KPCA, LDA, Sammon and tSNE. In table 1 we present the detailed results of membership of the different classes data to all clusters (special type of confusion matrix) at K-means clusterization of the original data.

Table 1. Class contents of clusters made for the original data

Cluster	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	e	p
1	*7*	*1*	*20*	*32*	*38*	*47*	*3*	*2,27*	*0,32*
2	*170*	*0*	*0*	*0*	*0*	*9*	*3*	*0,40*	*0,93*
3	*24*	*0*	*2*	*8*	*45*	*174*	*1*	*1,38*	*0,68*
4	*0*	*98*	*10*	*11*	*3*	*1*	*5*	*1,25*	*0,76*
5	*142*	*1*	*10*	*28*	*48*	*54*	*32*	*2,19*	*0,45*
6	*24*	*24*	*50*	*22*	*8*	*8*	*43*	*2,55*	*0,27*
7	*7*	*6*	*1*	*1*	*0*	*0*	*108*	*0,72*	*0,87*
Total	*374*	*130*	*93*	*102*	*142*	*293*	*195*	*1,77*	*0,52*

In the same way we have got the numerical results for the considered transformation techniques. In table 2 we present only the final entropy and purity measures obtained at application of the investigated reduction techniques: tSNE, PCA, KPCA, Sammon and LDA. In each case we have limited the representation to *K*=30.

Table 2. The clusterization measures of the data after application of different reduction techniques at K=30

Cluster	tSNE		PCA		KPCA		Sammon	
	e	p	e	p	e	p	e	p
1	*1.02*	*0.77*	*1.25*	*0.72*	*1.52*	*0.60*	*1.02*	*0.77*
2	*1.59*	*0.61*	*1.68*	*0.57*	*1.66*	*0.57*	*1.59*	*0.61*
3	*1.46*	*0.62*	*1.52*	*0.60*	*1.49*	*0.61*	*1.46*	*0.62*
4	*1.53*	*0.63*	*1.54*	*0.61*	*1.82*	*0.51*	*1.53*	*0.63*
5	*1.52*	*0.60*	*1.51*	*0.63*	*1.70*	*0.57*	*1.52*	*0.60*
6	*1.74*	*0.54*	*2.01*	*0.46*	*1.17*	*0.74*	*1.74*	*0.54*
7	*1.89*	*0.50*	*1.25*	*0.72*	*1.46*	*0.63*	*1.89*	*0.50*
Total	*1.49*	*0.63*	*1.53*	*0.62*	*1.49*	*0.62*	*1.49*	*0.63*

Classification of data

The last step of analysis is comparison of the efficiency of transformation techniques at classification of the data into 7 classes. In this phase of experiments we have mapped the original 87-dimensional data into smaller number of the most important components, using them as the input signals for classifier. As a classifier we have applied the Support Vector Machine (SVM) of the Gaussian kernel and applying the one-against-one classification approach [2,4].

The statistical results of classification of 7 classes of blood cells presented in the form of percentage mean ε of the misclassifications and standard deviation for the testing data not taking part in learning, at application of different reduction techniques are presented in Table 3. They refer to 10 cross validation runs at the number of features adjusted for each transformation method separately.

Table 3. Statistical mean percentage errors of cell classification

Method	No reduction	PCA (80 entries)	KPCA	tSNE	Sammon
ε±std [%]	18.59±1.10	18.51±1.95	21.67±1.30	25.71±2.1	28.82±1.35

It is seen that the transformed data not necessarily improve the classification accuracy. Moreover, observe that all nonlinear transformation techniques lead to deterioration of the accuracy of recognition.

Conclusions

The paper has shown that linear and nonlinear reduction techniques of the multidimensional medical data concerning the blood cells can find practical application in analysis of such data. We have shown their applicability in visualization, clusterization and classification of the data.

In the case of graphical visualization on the 2-dimensional space the best results have been obtained at application of tSNE method. Moreover, the experiments have shown that reduction of the data allows to cluster the multidimensional data with better quality measures than at original dimension of the input space. In the case of classification this is not necessarily true, especially in the case of nonlinear methods of transformation, where nonlinearity lead rather to the deterioration of recognition accuracy.

Acknowledgement

This research was financed by the Polish Ministry of Science and Higher Education as a research project within the years 2010-2012.

REFERENCES

[1] Duda, R.O., Hart, P.E., Stork, P., Pattern classification and scene analysis, 2003, Wiley, New York

[2] Jakubowski J., Ocena możliwości wykorzystania deskryptorów cech lokalnych obrazu twarzy w zadaniu automatycznej identyfikacji osób, *Przegląd Elektrotechniczny*, 2012, vol. 88, pp. 217-221

[3] Osowski S., Sieci neuronowe do przetwarzania informacji, 2006, Oficyna Wydawnicza PW

[4] Osowski S., Markiewicz T., Support vector machine for recognition of white blood cells in leukemia, (chapter in book of in G. Camps-Valls, J. L. Rojo-Alvarez, M. Martinez-Ramon, Kernel methods in bioengineering, signal and image processing, Idea Group Publishing, London, 2007), pp. 93-123.

[5] Tan P.N., Steinbach M., Kumar V., Introduction to data mining, 2006, *Pearson Education Inc*., Boston

[6] Van der Maaten L., Hinton G., Visualising data using t-SNE, *Journal MLR*, 2008, vol. 9, pp. 2579-2602

***Authors**: dr hab. inż. Krzysztof Siwek, Warsaw University of Technology, Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Email: ksiwek@iem.pw.edu.pl,*

prof. dr hab. inż. Stanisław Osowski, Warsaw University of Technology, Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Military University of Technology, Institute of Electronic Systems, Email: sto@iem.pw.edu.pl,

prof. PW, dr hab. inż. Tomasz Markiewicz, Warsaw University of Technology, Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Military Institute of Medicine, Warsaw, *Email: markiewt@iem.pw.edu.pl,*

dr inż. Jacek Korytkowski, Warsaw University of Technology, Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Email: jacek@iem.pw.edu.pl