# TECHNOLOGY OF CLASSIFICATION OF ELECTRONIC DOCUMENTS BASED ON THE THEORY OF DISTURBANCE OF PSEUDOINVERSE MATRICES

Volodymyr Donchenko, Viktoria Omardibirova

*Abstract: Technology of classification of electronic documents based on the theory of disturbance of pseudoinverse matrices was proposed.*

*Keywords: classification, training sample, a pseudoinverse matrix, Web Data Mining.*

*ACM Classification Keywords: I.5.2 Design Methodology, I.5.4 Applications, G.1.3 Numerical Linear Algebra*

## Introduction

One of the most important modern applied tasks of classification is classification of electronic documents. Applications may be different. For example, classification of E-Mail and elimination of so-called spam, i.e. letters which do not represent interest for the user; or classification of documents by subjects at their reception from such not structured warehouse as Internet. The given problems concern of a class of problems of extraction of the useful data from the Internet (Web Data Mining). In presented article the technology of classification of electronic documents by the given classes is described. As the mathematical instrument the theory of disturbance of pseudoinverse matrices is used. [1]

## Statement of the Task

There are some classes of electronic documents from some subject domain. Each class is characterized by a set of documents - templates. Whole subject domain is characterized by some given thesaurus identical to all classes of documents. It is necessary to construct and train the classifier with the help of training sample generated from documents - templates which will classify newly added documents to one of known classes. Standard way for decision of a task of recognition is from the one hand formation of significant attributes (Feature Extraction) [2], and from another - selecting of suitable distance function to formed classes.

## The Basic Designations

Let is present K classes of electronic documents, which we shall designate accordingly, $\Omega_k$, $k=\overline{1,K}$ for which the same thesaurus is fixed with quantity of terms that we shall designate L. Attributes on which classification will be made are relative frequencies of use of terms of the thesaurus. Thus, each document may be represented as a vector a: $a^T=(a_1,...,a_L)$ consisting of relative frequencies of occurrence of words from the thesaurus in given document. Let $n_k$, $k=\overline{1,K}$ is quantity of documents of the training sample concerning each of the classes.

Let's enter into consideration matrices $A_k$, $k=\overline{1,K}$, made of frequency vectors of each of classes. Evidently, each of matrices has dimension $L\times n_k$, $k=\overline{1,K}$.

Let's designate average of vectors of training sample on each of classes through $\overline{a}_k$, $k=\overline{1,K}$:

$$\overline{a}_k=\frac{1}{n_k}\sum_{a\in\Omega_k}a \quad , \; k=\overline{1,K}\,. \tag{1}$$

Let's shift each of vectors of training sample of this or that class by average on the same class, and the matrices constructed from received vectors as on vectors-columns.

We shall make the new matrices similar $A_k$, $k=\overline{1,K}$. We shall designate received matrices connected to each of classes through $\tilde{A}_k$, $k=\overline{1,K}$.

## Algorithm of Classification

The algorithm of classification is offered to be built on the basis of calculation of vectors $\bar{a}_k$, $k=\overline{1,K}$ and construction of singular decomposition [1] for matrices $\tilde{A}_k$, $k=\overline{1,K}$, describing the appropriate classes. As is known, according to singular decomposition matrices allow representation:

$$\tilde{A}_k=\sum_{i=1}^{r_k} y_i^{(k)}\left(\mathrm{x}_i^{(k)}\right)^T\lambda_i^{(k)}\ ,\quad r_k=rank\,\tilde{A}_k=rank\,A_k\ ,\quad k=\overline{1,K}\ , \tag{2}$$

where $\lambda_1^2\geq\ldots\geq\lambda_r^2$.

Eigen values $\lambda_i$ and eigen vectors $y_i^{(k)}\in R^L$, $x_i^{(k)}\in R^{n_k}$, $i=\overline{1,r_k}$, $k=\overline{1,K}$ in the given representation may be calculated, for example, by Jacobi method or a method of singular decomposition of matrix SVD [3].

Singular decomposition (2) of matrices of classes can be used for construction approximation of these matrices, which will be used for construction of belonging measures to each of classes. These approximations are constructing in two stages: highest members of singular decompositions are rejected on the first stage: items, answering to smaller values of modules of eigen values; on the second stage - received matrixes are used for construction of belonging measures to classes. An estimation of the error made after rejection of highest members of singular decomposition if it is left $s_k$ members: $s_k<r_k$ is described by the following inequality [1]:

$\left\|\Delta_{is_k}\right\|^2=\sum_{i=s_k+1}^{r_k}(\lambda_i^{(k)})^2\left|x_i^{(k)}\right|^2\leq(\lambda_{s_k+1}^{(k)})^2\sum_{i=s_k+1}^{r_k}\left|x_i^{(k)}\right|^2\leq(\lambda_{s_k+1}^{(k)})^2$, where

$$\tilde{A}_k=\sum_{i=1}^{s_k} y_i^{(k)}\left(\mathrm{x}_i^{(k)}\right)^T\lambda_i^{(k)}+\Delta_{is_k},\quad s_k<r_k\ ,\ k=\overline{1,K}$$

$$\Delta_{is_k}=\sum_{i=s_k+1}^{r_k} y_i^{(k)}\left(\mathrm{x}_i^{(k)}\right)^T\lambda_i^{(k)} \tag{3}$$

Parameter $s_k$, $s_k<r_k$, $k=\overline{1,K}$ is being chosen in the sense of smallness of the error, made at construction of suitable approximation $\tilde{A}_k$, $k=\overline{1,K}$, and as a rule $s_k$ may be chosen, that $\left|\lambda_{s_k+1}^{(k)}\right|$ answers several percents from module of the maximal eigen value. Such construction of approximation essentially simplifies computing procedure of construction of a belonging measure to each of classes: measures, which can be constructed either on the basis of initial matrices $\tilde{A}_k$, $k=\overline{1,K}$, or their approximation constructed according to the procedure described above.

Belonging measures are determined as square-law forms with matrices $R_k$, $k=\overline{1,K}$, which are being constructed on the basis of suitable approximations of matrices $\tilde{A}_k$, $k=\overline{1,K}$,– which we shall designate accordingly $\tilde{A}_{k,s_k}$, $k=\overline{1,K}$ – according to formulas:

$$R_k = \left(\tilde{A}^{+}_{k,s_k}\right)^T \cdot \tilde{A}^{+}_{k,s_k} = \sum_{i=1}^{s_k} y_i^{(k)} (y_i^{(k)})^T (\lambda_i^{(k)})^{-2} \quad (4)$$

Procedure of classifying of the electronic document described by its frequency vector $b^T = (b_1,...,b_L)$ to one of classes $\Omega_k$, $k = \overline{1,K}$ is made on the basis of calculation for each of them "distance" $m_k$, $k = \overline{1,K}$ up to a class, which is determined by expression:

$$m_k = \left(\left(b - \bar{a}^{(k)}\right)^T \cdot R_k \cdot \left(b - \bar{a}^{(k)}\right)\right), \; k = \overline{1,K}. \quad (5)$$

The classified document will belong to the class for which value of distance $m_k$, $k = \overline{1,K}$, determined according to (5) will accept the minimal value.

## Results of Experiments

For check of correctness of work of the offered technology the book in an electronic format «The Handbook of Data Mining» [2] with size of 689 pages and consisting of 3 parts was chosen. Training sample of three classes by 5 first documents - chapters in each class was accordingly generated. After training of the classifier on its input chapters of the book which were not used in training of the classifier were inputted to be classified to one of 3 classes.

For eighth chapter of the first part given on input of the classifier, values of functional (10) are equal:

0.01959
0.15240
0.09561

As the least value of functional is equal to 0.01959 the classified document concerns to the first class.

## Conclusions

In article the technology of classification of electronic documents on the given classes with use of the theory of disturbance of pseudoinverse matrices which has shown the efficiency at least in the considered modelling examples is described. The offered technology may be used for automatic classification of incoming E-Mail or for automatic extraction of interesting information from the Internet (Web Data Mining). Due to use of approximations for the pseudoinverse matrices, it is possible to essentially increase speed of work of algorithm.

## Bibliography

1. Кириченко Н.Ф., Куц Р., Лепеха Н.П. Распознавание трехмерных объектов по ультразвуковым эхо-сигналам // Проблемы управления и информатики.– 1999.– №5– С.110–122.
2. The Handbook of Data Mining / edited by Nong Ye, LAWRENCE ERLBAUM ASSOCIATES, London, 2003, 689p.
3. Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений.-М.:Мир,1980.-280 с.
4. Гантмахер Ф.Р. Теория матриц.-М.:Наука,1967.-287 с.

## Authors' Information

Volodymyr Donchenko – professor, Kiev National Taras Shevchenko University, Department of System Analysis and Decision Making Theory, e-mail: vsdon@unicyb.kiev.ua

Victoria Omardibirova – post-graduate course student, Kiev National Taras Shevchenko University, Department of System Analysis and Decision Making Theory, e-mail: sdp@unicyb.kiev.ua