Personal data segmentation based on conjugation index usage

P V Hripunov[1] and D A Zherdev[2]

[1]Pension Fund of the Russian Federation, Shabolovka str. 4, Moscow, Russia, 119991
[2]Samara National Research University, Moskovskoe Shosse 34, Samara, Russia, 443086

Abstract. The paper proposes a method for processing personal data that allows them to be divided into many segments or classes. The customer database is used as the source data. We use the indicator of conjugacy that has already proved the effectiveness in both recognition and clustering of data problems.

1. Introduction

Data mining problem is a primary problem inprocessing of huge amount data. Different methods of pattern recognition, classification, images clustering and others have foundthe implementation in the data mining. Many works [1-4] study clusterization processes of big data. In this study, we research the recognition ability of some personal data that was presented by a digit vector. For classification within some probability, it is necessary to find out is a similar element consist in the database. To achieve this, we use conjugation index. The index effectiveness was shown in study of face recognition problems [5], as well as objects recognition in radar images [6], [7].

2. Segmentation and classification of personal data

In this section the clustering approach based on conjugation index usage is described. We can decide that a vector belongs to a class by calculation of conjugation index value. The higher probability of conjugation index shows that a vector has similarity to the vectors that form a class:

$$\mathbf{X}_k = \left[\mathbf{x}_1(k), \mathbf{x}_2(k), ..., \mathbf{x}_j(k), ..., \mathbf{x}_M(k)\right], k = \overline{1,K},$$

where $\mathbf{x}_j = \left[x_1, x_2, ..., x_i, ..., x_N\right]$ is a $N\times 1$ feature vector.

The conjugation index can be presented as:

$$R_k\left(\mathbf{x}_j\right) = \frac{\mathbf{x}_j^T \mathbf{Q}_k \mathbf{x}_j}{\mathbf{x}_j^T \mathbf{x}_j}, \quad k = \overline{1,K},$$

where K is a class count,

$$\mathbf{Q}_k = \mathbf{X}_k \left[\mathbf{X}_k^T \mathbf{X}_k\right]^{-1} \mathbf{X}_k^T, \quad k = \overline{1,K},$$

is a $N \times N$ matrix of k -class.

Each letter in a single categorical data is coded by some index thus digital vector of a string can be formed. Each letter of Russian alphabet "А-Я" coded by numbers1-33. In the result the new database of vectors can be formed.

There are three fields in the database: first name, second name and middle name. For convenience, the maximum number of possible symbols in the database was chosen to be 100. As the result, all

vectors in the new dataset contain 300 features. All vectors in dataset must be the same size. For example, if the first field size equals 12 symbols then the first 12 features of a vector contain the field value and other 88 features are filled by zero value.When we processed the database, all personal data was encrypted by summing up with some digital key.

We use the similar procedure for clustering which was used in work [6] for clustering of radar images. At first step of the whole set we choose the two most "distanced" vectors. These vectors have the minimal value of the correlation ratio and we can be labeled them as $\mathbf{x}_1$, $\mathbf{x}_M$.

Then the algorithm from the remain set of vectors adds two new vectors ($\mathbf{x}_2$, $\mathbf{x}_{M-1}$). Each one of these vectors must have the maximum of the correlation ratio:

$$R_{1,2} = \frac{(\mathbf{x}_1^T \mathbf{x}_2)^2}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|},$$

$$R_{M-1,M} = \frac{(\mathbf{x}_{M-1}^T \mathbf{x}_M)^2}{\|\mathbf{x}_{M-1}\|\|\mathbf{x}_M\|},$$

with one vector was obtained at the first step. In the result received pairs of vectors $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_{M-1}$, $\mathbf{x}_M$ form the subspaces that were formed by matrices $\mathbf{X}_{1,2}$ and $\mathbf{X}_{M-1,M}$ correspondingly. Then using the remaining set next two vectors $\mathbf{x}_3$, $\mathbf{x}_{M-2}$ that are closest to the subspaces are joined to previously formed subspaces using computation ofconjugation index with a maximum value.

Since the database contains a large number of vectors, process continues due finding of specific number of vectors in both subspaces. For example, for such dataset the resulted subspace contains 15 vectors in both matrices $\mathbf{X}_k$, $\mathbf{X}_l$, which correspond to two subclasses.

The procedure described above is repeated iteratively with all unlabeled vectors. Clustering is continued until all the vectors will be specified to any of the subspaces.At the recognition stage with a certain decision rule, the vector closest to one of the subclasses formed in thedescribed manner is considered to belong to the class.

3. Results and discussion

In this paper, the problem of the determining possibility whether there is some given record in the database is examined. After clustering process we can figure out the belonging of a vector to some class. The subclass stores a small number of vectors in comparisonof the initial database.Therefore, after the classification of the current vector, it will be easy to analyze data in a subclass and determine isit possible to add a new value into the database.

Thus, to verify the above assumption, we performed the experiment. From the database of 1041100 records there was performed the random selection of 1040 records five times. Each selection was divided onto 80 subclasses, a subclass consists of 13 vectors. After the clustering procedure, the generated vectors were classified.

The testing vectors were formed using the existed in the dataset records with some modifications. For example, there was simulated situation of incorrect handwritten letters conversion when the personal data was filled in some document. As it was shown in the work [8] the problem of text recognition is a difficult and can have many solutions. Figure 1 a, b presents the images of two handwritten words. The word in Figure 1 a was correct converted by some letters recognition software into “ADAM” as opposed to word in Figure 1 b that led to incorrect result: “ADRM”.

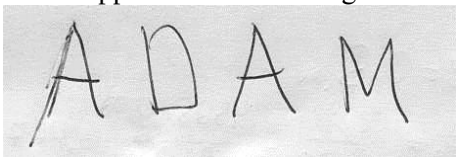
a)  b) 

Figure 1. Examples of a) correct and b) incorrect handwritten letters conversion.

In the classification experiment, 20 vectors of the type described above were tested. All vectors were successfully classified based on the many-to-many approach [9] extending the possibilities of the

binary classification of the support subspaces algorithm [7]. Moreover the average value of conjugation index was 0.95 for true defined class. This fact undoubtedly indicates the reliability of using the conjugation index in problems of this kind. This is an advantage for following research of such kind both with databases of a more complex type, with a larger field number, and for classification using a whole database of one million or more records.

4. References

[1] Yang Y and Guan J 2002 CLOPE: a fast and effective clustering algorithm for transactional data *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* 682-687

[2] Zhang T, Ramakrishnan R and Livny M 1996 BIRCH: an efficient data clustering method for very large databases *ACM Sigmod Record* **25(2)** 103-114

[3] He Z, Xu X and Deng S 2005 A cluster ensemble method for clustering categorical data *Information Fusion* **6(2)** 143-151

[4] Huang Z 1997 A fast clustering algorithm to cluster very large categorical data sets in data mining *DMKD* **3(8)** 34-39

[5] Fursov V and Kozin N 2007 Recognition through constructing the eigenface classifiers using conjugation indices *IEEE Conference on Advanced Video and Signal Based Surveillance* 465-469

[6] Minaev E and Fursov V 2016 Support subspaces method for fractal images recognition *CEUR Workshop Proceedings* **1638** 379-385

[7] Zherdev D A, Kazanskiy N L and Fursov V A 2015 Object recognition in radar images using conjugation indices and support subspaces *Computer Optics* **39(2)** 255-264 DOI: 10.18287/0134-2452-2015-39-2-255-264

[8] Bolotova Y A, Spitsyn V G and Osina P M 2017 A review of algorithms for text detection in images and videos *Computer Optics* **41(3)** 441-452 DOI: 10.18287/2412-6179-2017-41-3-441-452

[9] Bishop Ch M 2006 *Pattern Recognition and Machine Learning* (New York: Springer) p 738