Personal data segmentation based on conjugation index usage

P V Hripunov¹ and D A Zherdev²

¹Pension Fund of the Russian Federation, Shabolovka str. 4, Moscow, Russia, 119991
²Samara 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 in processing of huge amount data. Different methods of pattern recognition, classification, images clustering and others have found 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:

\[ X_k = [x_1(k), x_2(k), ..., x_j(k), ..., x_M(k)], k = 1, K, \]

where \( x_j = [x_1, x_2, ..., x_i, ..., x_N] \) is a \( N \times 1 \) feature vector.

The conjugation index can be presented as:

\[ R_k(x_j) = \frac{x_j^T Q_k x_j}{x_j^T x_j}, \quad k = 1, K, \]

where \( K \) is a class count,

\[ Q_k = X_k [X_k^T X_k]^{-1} X_k^T, \quad k = 1, K, \]

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 $x_1$, $x_M$.

Then the algorithm from the remain set of vectors adds two new vectors $(x_2, x_{M-1})$. Each one of these vectors must have the maximum of the correlation ratio:

$$ R_{1,2} = \frac{(x_1^T x_2)^2}{\|x_1\| \|x_2\|}, $$

$$ R_{M-1,M} = \frac{(x_{M-1}^T x_M)^2}{\|x_{M-1}\| \|x_M\|}, $$

with one vector was obtained at the first step. In the result received pairs of vectors $x_1$, $x_2$ and $x_{M-1}$, $x_M$ form the subspaces that were formed by matrices $X_{1,2}$ and $X_{M-1,M}$ correspondingly. Then using the remaining set next two vectors $x_3$, $x_{M-2}$ that are closest to the subspaces are joined to previously formed subspaces using computation of conjugation 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 $X_1$, $X_M$ 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 the described 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 comparison of the initial database. Therefore, after the classification of the current vector, it will be easy to analyze data in a subclass and determine is it 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”.

![Figure 1](image)

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