М. Ю. Сидоров, С. Г. Заблотский., Е. С. Семенкин., В. Минкер

ЭВОЛЮЦИОННОЕ ФОРМИРОВАНИЕ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ ПРОГНОЗИРОВАНИЯ ФИНАНСОВЫХ ВРЕМЕННЫХ РЯДОВ

Прогнозирование в различных технических, экономических и др. системах является важнейшей задачей современности. Методы искусственного интеллекта и машинного обучения являются эффективными средствами анализа в том числе и финансовых данных. Основной проблемой использования таких методов остается сложность настройки параметров моделей. Предлагается эволюционный способ формирования нейросетевых технологий, не требующий экспертных знаний в области нейронных сетей и теории оптимизации от конечного пользователя. Произведен сравнительный анализ показателей качества прогнозирования предложенной модели с другими методами искусственного интеллекта на исторических данных 13 валютных пар рынка FOREX, более чем за 12 лет. Предложенный алгоритм показал наилучшую результативность более чем на половине задач. На остальных задачах, алгоритм незначительно уступил многослойному перцептрону, обученному стайным алгоритмом.

Ключевые слова: нейронные сети, эволюционные алгоритмы, стайный алгоритм оптимизации, прогнозирование на FOREX.

UDK 004.89 (681.3)

Yu. V. Smeshko, T. O. Gasanova

MODIFICATION OF FUZZY C-MEANS ALGORITHM WITH AUTOMATIC SELECTION OF THE NUMBER OF CLUSTERS FOR SPEECH UTTERANCE CATEGORIZATION

In this paper we propose a fuzzy clustering algorithm, which is able to find the clusters in a data set without the number of clusters as a user input parameter. The algorithm is based on the standard fuzzy c-means method and consists of two parts: 1) detecting the number of clusters c^*; 2) calculating the cluster partition with the obtained c^*. We apply this method to the preprocessed database which was provided by Speech Cycle Company. The proposed algorithm has been tested with optimal parameters which we have calculated on the test data.

Keywords: unsupervised fuzzy classification.

Cluster analysis consists of methods used to find the group structure in a certain data set. These algorithms can be applied to many different problems, such as image segmentation, data mining, the analysis of genomic and sensorial data, among others. A lot of different clustering techniques have been appeared in the last few years. Clustering algorithms do not need knowledge about the object's group labels, thus they use unsupervised machine learning models.

However, many clustering algorithms require parameters which should be selected by user, i. e., the obtained clustering depends on some user input parameters which should be chosen for the certain dataset. The target number of clusters or equivalents of it (such as density indicators in density models) are usually required. In this respect, we should consider and develop approaches, which can automatically detect the true number of clusters in a given database with no prior information about the structure and group labels.

With the appearance of the classical clustering approaches in the 1970s, researchers like J. A. Hartigan (k-means) were very conscious about the problem of detecting the correct number of clusters and proposed some metrics for automatically determining this value. The general approach is to evaluate the quality of solutions which were obtained with the different number of clusters and select the value of number of clusters that originates the optimum partition according to a quality criterion.

In this work we considered the fuzzy c-means clustering algorithm and developed its modification which is able to discover the number of clusters automatically. In contrast to hard clustering methods where each object can be unequivocally assigned to only one cluster, in soft clustering (fuzzy clustering), objects are associated to all possible clusters. There is a so-called membership matrix, whose elements are degrees of certain object's membership to each cluster. Fuzzy approaches are more appropriate to deal with the existence of polysemous words and phrases.

We have chosen fuzzy c-means algorithm and developed its modification in order to solve the clustering problem on the database provided by Speech Cycle. This database consists of utterances in text form and some phrases and words have different meaning in another context.

This paper is organized as follows: Section 2 and Section 3 introduce the standard fuzzy c-means algorithm and its modification with automated choice of cluster's number, respectively. The value of fuzzyfier parameter is described in Section 4. We present the corpus employed in this study in Section 5. Results are summarized and discussed in Section 6. Finally, in Section 7, we draw conclusions and discuss the future directions.

Standard algorithm fuzzy-c-means. In the work [1] fuzzy-c-means algorithm has been proposed. The algorithm computes the fuzzy membership matrix starting from an initial choice of cluster medoids. The elements of the membership matrix μ_i^l denote the degree of membership of each object x_i to each cluster medoid τ^l. The objective function $Q(P)$ of fuzzy-c-means is

$$Q(P) = \sum_{l=1}^{c} \sum_{i=1}^{m} (\mu_i^l)^\gamma d(x_i, \tau^l), \qquad (1)$$

where c is the number of clusters; m is the quantity of objects; $d(x_i, \tau^l)$ is the dissimilarity between the object x_i and the medoid τ^l; γ is a fuzzyfier factor denoting the smoothness of the clustering solution.

The algorithm of fuzzy-c-means consists of an iterative repetition of two steps [1; 2]: (i) (re)computation of the cluster medoids; (ii) (re)calculation of objects membership to the classes. The main procedure is iterated until either the updated medoids remain the same or the maximum number of iteration is reached. The final solution is obtained when the deviation of objects from the medoids reaches the minimum.

Modified version of algorithm. The main disadvantage of the standard fuzzy-c-means algorithm is that the number of clusters must be necessarily known in advance. In this work we present the modification of fuzzy-c-means algorithm which automates the choice of the number of clusters. Modified algorithm must select the most "natural" number of clusters c^* from some range $[cn, ck]$. This range is defined by the user. An algorithm for solving this problem consists of two parts. In the first part, the database is partitioned into an integer number of fuzzy clusters from range $[cn, ck]$. The data partitioning is carried out using the iterative procedure of the standard fuzzy-c-means algorithm. For each partition the value of the additional functional $Q(x, c)$ for determining the number of clusters is estimated. c^* is taken to be the value of c that maximizes the functional $Q(x, c)$.

The number of clusters c^* when the functional reaches an extreme value is taken as the "most natural" number of fuzzy clusters. In the second part of the algorithm, the database is partitioned on c^* fuzzy clusters using the iterative procedure of the standard fuzzy-c-means algorithm.

The functional for determining the number of fuzzy clusters c^* is constructed using the compactness hypothesis [3]. The algorithm for calculating the functional value consists of three steps:

1) Computation of the index of "closeness", S, of objects within a cluster:

$$S = \frac{1}{c} \sum_{l=1}^{c} \frac{1}{m} \sum_{i=1}^{m} \mu_i^l d(x_i, \tau^l). \qquad (2)$$

2) Computation of the index of "remoteness", D, of clusters from each other:

$$D = \frac{1}{c} \sum_{l=1}^{c} \frac{1}{c-1} \sum_{l^{'}, l^{'} \neq l}^{c} d(\tau^l, \tau^{l'}), \qquad (3)$$

where $d(\tau^{'}, \tau^{l'})$ is the dissimilarity between the medoid τ^l and the medoid $\tau^{l'}$, $l \neq l^{'}$.

3) Computation of the value of functional [4]:

$$Q(x, c) = \ln \frac{D^a}{S^b + \varphi}, \qquad (4)$$

where a, b, φ are the parameters that define the importance of the appropriate indices. The number of clusters c^* corresponds to the partition when the functional reaches a maximum value.

The fuzzyfier parameter. The parameter γ controls the influence of the membership matrix on the clustering. The parameter value γ can be selected in the range $1 \leq \gamma \leq \infty$ [5]. There are no recommendations about the selection of this parameter. Many researches use $\gamma = 2$ in their works.

In this work the criterion for a quantitative estimation of the degree of partition smoothness is offered. The initial information for the estimation of smoothness is given by the membership matrix. The algorithm for computing the estimation consists of the following steps: (1) to define the maximum degree of membership, k_l, to clusters for all objects; (2) to determine how many objects are assigned to each cluster on the basis of k_l; (3) to calculate for each cluster its fuzzy weight, w_l, which is the sum of the membership degrees of all objects to this cluster; (4) to calculate the estimation of the smoothness partition using

$$W = \frac{1}{c} \sum_{l=1}^{c} |k_l - w_l|. \qquad (5)$$

Figure 1 show 3 examples of the partitioning of a test set obtained by using different values of the fuzzyfier parameter (3 clusters, 20 binary objects in each cluster, the dimension of attributes space is 100). For the graphical presentation of the fuzzy classification results in a multidimensional attributes space the linear diagram is used [5]. On figure 1, *a* the "very fuzzy" partition is presented. Each object is assigned to all clusters with the equal membership degree. A result of this classification is unsatisfactory because there is very high level of uncertainty associated with the assignment an object to a cluster. The fuzzyfier parameter is equal to 3. The estimation of smoothness partition $W = 11.333$. On figure 1, *b* the fuzzy partition is presented. Some objects with high membership degree are assigned to only one cluster. Some objects with the equal membership degree are assigned to more than one cluster. Level of uncertainty of this classification is medium. The fuzzyfier parameter is equal to 2. The estimation of smoothness partition $W = 0.542$. On figure 1, *c* the unambiguous partition is presented. All objects are assigned to its cluster with a high membership degree. In results of this classification the level of uncertainty is smallest. The fuzzyfier parameter is equal to 1.1. The estimation of smoothness partition $W = 0.026$.

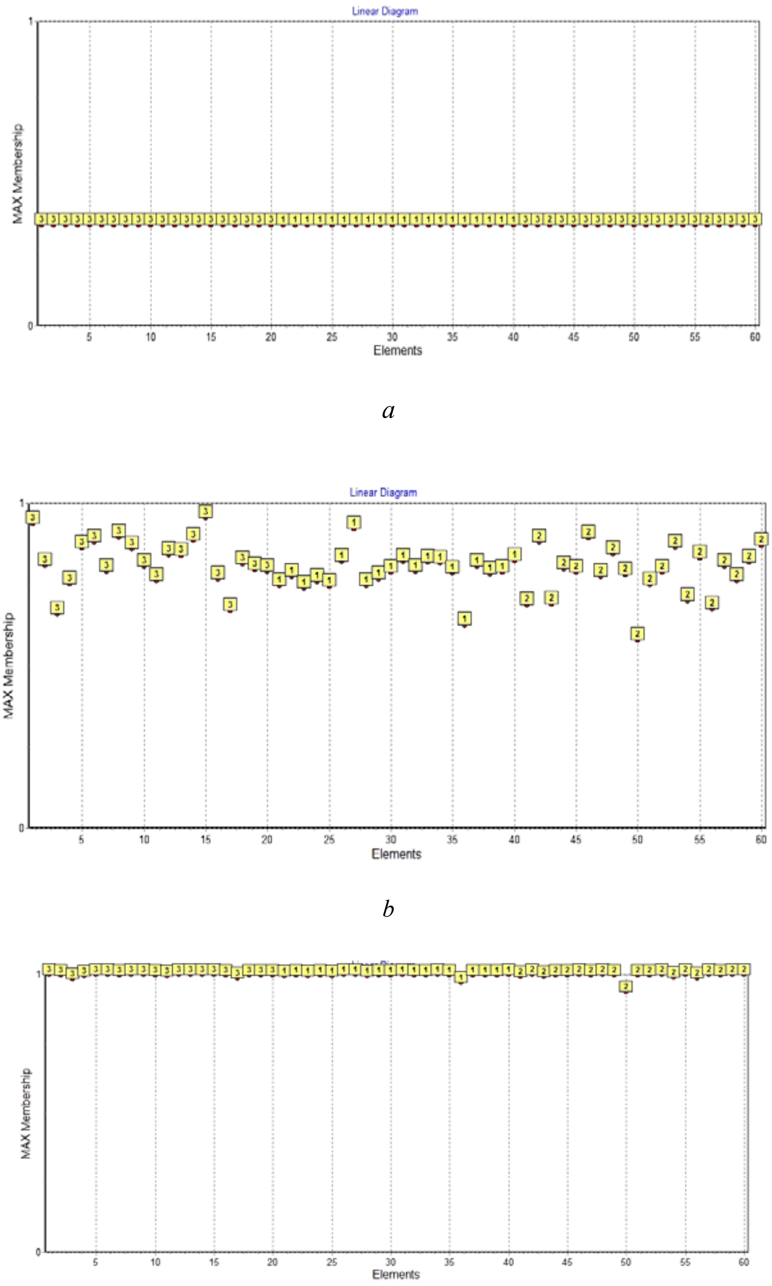
The value of the estimation W depends on the fuzzyfier parameter γ. In the process of solving fuzzy classification problem the value of parameter γ we recommend to select in the range $1 < \gamma \leq 1.2$. In this case the value estimation $W < 1$. This means that each object has a high membership degree to only one cluster.

Corpus Description. For the experiments we used corpora data set collected from spoken language dialogue system (SLDSs). SLDS is an important form of human-machine communication [6].

The initial database consists of utterances gathered from user interactions of a commercial troubleshooting agent of the Cable TV domains.

The initial database was preprocessed (the details of preprocessing of the initial database are described in the work [7]). The preprocessing module consists of part-of-speech (POS) tagging, morphological analysis, stop-word filtering, and bag-of-words representation.

As an intermediate result they obtained the vocabulary which is used in all utterances.



a

b

c

Fig. 1:
a – the very fuzzy partition of test objects; *b* – the fuzzy partition of test objects;
c – the unambiguous partition of test objects

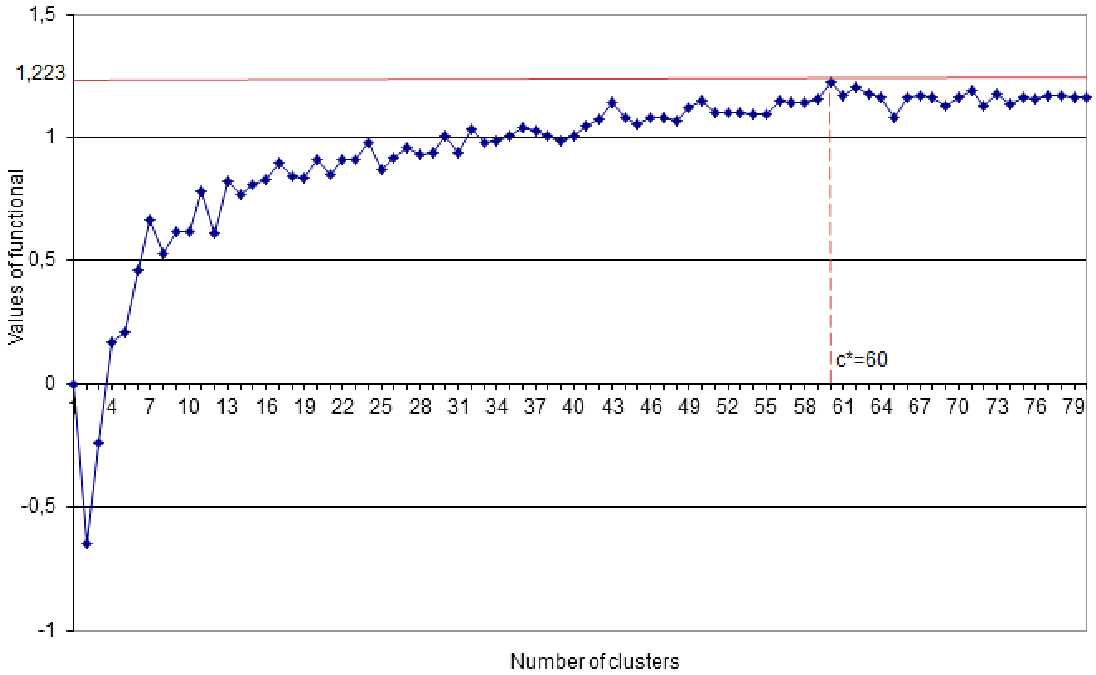


Fig. 2. The values of functional for determining the number of fuzzy clusters after implementation the algorithm on corpora data base

The processed data was then collected in an $m \times n$ dimensional matrix of binary elements. Each row of the matrix is an n-dimensional vector that represents an utterance form one corpora caller. The elements of the vector represent the presence or absence of the respective vocabulary elements. The quantity of utterances was $m = 2940$ and the vocabulary dimension was $n = 554$.

In order to find the optimal parameters of the proposed algorithm we have generated the binary test datasets. The objects of test data were chosen to be similar to the initial database. We use the vector of words distribution through all utterances to do this. At first is necessary to choose the values of parameters for procedure of generating test objects, such as cw is the number of new clusters; s is the number of objects in each cluster; v is the threshold value of the words appearance probability, where $0 < v < 1$; ok is the threshold value of the neighborhood objects.

The procedure of generating the test objects consists of the following steps: (i) computation the vector of distribution words as the sum of the appropriate words among the all utterances of the initial database; (ii) formation the binary matrix of the centers of cw new clusters with using the vector of distribution words and the random number generator by following rule: if the random number is less than the appropriate value of the vector of distribution words then appropriate element of matrix is equal to 1, otherwise the element is equal to 0; (iii) verification that the equivalent centers not were generated. If the matches are found then is necessary go back and repeat step (ii), otherwise go to following step; (iiii) creation the test object with using the random number generator on the basis of the center of new cluster by following rule: if the random number is more than value of parameter v then the current value of binary object is equal to appropriate value of center on the basis of which the given object is generated, otherwise the value of binary object is replaced by the opposite value; (iiiii) computation the neighborhood of test object and comparison its value with the threshold value ok (the neighborhood of object is the sum of discrepancies between the appropriate values of generated object and the center of cluster on the basis of which the given object was created). If the value of neighborhood test object is more than the threshold value than is necessary go back and repeat step (iiii), otherwise the given object is saved in new database. The steps (iiii) and (iiiii) are repeated until for all cw clusters the s test objects with the given neighborhood will be generated.

As a result of generating process the researcher has information about a number of clusters and object memberships. This information is necessary for choosing the appropriate algorithm parameters.

The modified algorithm was applied on corpora data base. The following values of algorithm parameters are chosen: the boundaries range $cn = 1$, $ck = 80$; the fuzzyfier parameter $\gamma = 1.1$; the stop parameter of algorithm procedure $\varepsilon = 0.0001$. The values of functional for determining the number of clusters are presented on figure 2.

The number of fuzzy clusters $c^* = 60$ because when the initial data base partitioned on sixty fuzzy clusters the maximum value of the criterion (1.223) on the entire range is obtained. The value of criterion W for the estimation of smoothness partition is equal to 0.193. The biggest fuzzy cluster consists of 160 utterances. The smallest fuzzy cluster includes 7 utterances.

There is labeled data which were manually marked by experts. They partitioned the initial utterances base into 77 clusters. However, in their partition there are 22 clusters which consist of less than 20 utterances. The 17 smallest clusters are jointed with the biggest clusters on the basis of the maximum similarity. This clusters are compared with the 60 clusters which obtained by the proposed algorithm. For the estimation of algorithm's results the following index is computed

$$E_r = \sum_{l=1}^{c^*} \frac{1}{m} \sum_{i=1}^{m} \left\| (\mu_i^l)^* - (\mu_i^l)^e \right\|, \qquad (6)$$

where $(\mu_i^l)^*$ are results obtained by proposed algorithm, $(\mu_i^l)^e$ are results labeled by experts manually. For example, the Euclidean distance can be used as a norm. The results obtained by the proposed algorithm agree with the manually labeled data by 60 %.

Conclusion and Future Directions. In this paper we proposed a modification of fuzzy c-means clustering algorithm which does not need a number of clusters as an input parameter. This approach represents the way of an automated detecting the number of clusters. The algorithm is designed for solving the clustering task in the database which was provided by Speech Cycle company. This approach was also tested on specially created data to find optimal parameters of the algorithm (such as the fuzzyfier parameter γ and the estimation of smoothness partition W). With those parameters, we show that the initial data should be divided into 60 clusters. Furthermore, this method does not require the same sizes of clusters and can detect very small or very big groups of objects. The proposed approach can be extended not only to the given database but to any other clustering problem as well. Although for the better results parameters of approach should be given using the a priori information of the data structure (if there is such information). Otherwise, the observed parameters can be applied.

These results will be used in an ensemble of unsupervised classifiers as one of algorithms. For the future work, we will also apply the proposed approach to other databases of Speech Cycle.

Acknowledgements. We would like to thank Speech Cycle company for providing the TV Cable database. This work was supported by Leonhard-Euler-Programm which is funded by Deutscher Akademischer Austausch Dienst (DAAD).

References

1. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York : Plenum Press, 1981.

2. Oliveira J. V., Pedrycz W. Advances in Fuzzy clustering and its Applications. England, West Sussex, John Wiley & Sons Inc. Pub, 1 edition, 2007.

3. Zagoruyko N. G. Application Methods of Data Analysis and Knowledges. Novosibirsk : Mathematic institute, 1999.

4. Zagoruyko N. G. Recognition Methods and its Application. Moscow : The Soviet Radio, 1972.

5. Everitt B. S., Landau S., Leese M., Stahl D. Optimization Clustering Techniques in Cluster Analysis. Chichester, UK, John Wiley & Sons Ltd., 2011.

6. Minker W., Albalate A., B¨uhle D., Pittermann A., Pittermann J., Strauss P., Zaykovskiy D. Recent Trends in Spoken Language Dialogue Systems. Egypt, Cairo, In Proc. of the ICIT, 2006.

7. Amparo G. New Strategies on Semi-Supervised and Unsupervised Machine Learning. Germany, Ulm, Ph.D. Thesis, 2010.

Ю. В. Смешко, Т. О. Гасанова

МОДИФИКАЦИЯ АЛГОРИТМА НЕЧЕТКИХ С-СРЕДНИХ С АВТОМАТИЧЕСКИМ ВЫБОРОМ КОЛИЧЕСТВА КЛАССОВ ДЛЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ БАЗЫ ЗАПРОСОВ

Предложен алгоритм нечеткой кластеризации с возможностью автоматического выбора количества кластеров, на которое может быть разделен набор исходных данных. Разработанный алгоритм основан на стандартной процедуре алгоритма нечетких С-средних и состоит из двух частей: 1) выбирается «наиболее подходящее» количество нечетких кластеров c^*; 2) набор данных разбивается на c^* нечетких кластеров. Алгоритм применен к обработанной базе данных, которая предоставлена компанией Speech Cycle. Значения параметров алгоритма подобраны на тестовых задачах.

Ключевые слова: нечеткая кластеризация с самообучением.