RESEARCH ARTICLE

European Journal of Artificial Intelligence and Machine Learning
www.ej-ai.org

A New Procedure for Unsupervised Clustering Based on Combination of Artificial Neural Networks

Yaroslava Pushkarova and Paul Kholodniuk

ABSTRACT

Classification methods have become one of the main tools for extracting essential information from multivariate data. New classification algorithms are continuously being proposed and created. This paper presents a classification procedure based on a combination of Kohonen and probabilistic neural networks. Its applicability and efficiency are estimated using model data sets (iris flowers data set, wine data set, data with a two-hierarchical structure), then compared with the traditional clustering algorithms (hierarchical clustering, k-means clustering, fuzzy k-means clustering). The algorithm was designed as M-script in Matlab 7.11b software. It was shown that the proposed classification procedure has a great advantage over traditional clustering methods.

Keywords: artificial neural network, clustering, cross-validation, multivariate data.

Published Online: Septemberr 18, 2023

ISSN: 2796-0072

DOI : 10.24018/ejai.2023.2.4.31

Y. Pushkarova*
Bogomolets National Medical University, Ukraine.
(e-mail: yaroslava.pushkarova@gmail.com)
P. Kholodniuk
Production Laboratory and Research Department, LLC “Firm Soyuz, Ltd”, Ukraine.
(e-mail: sofir.lab@gmail.com)

Corresponding Author

I. Introduction

The classification of objects is widely used in any field where present processing of multivariate data. Classification of objects can be considered as identification (assigning the sample to a certain class of objects), discrimination (separating the studied samples into groups), as well as clustering (identifying homogeneous groups of objects) [1], [2].

Supervised classification methods (discriminant analysis, support vector machine, soft independent modeling for class analogies, artificial neural networks) require a set of samples with the known class membership (training set) to develop classification rules. The training set must be representative and contain all information about the task. Unsupervised classification methods (hierarchical clustering, k-means method) require information about the number of classes. Thus, the application of both supervised and unsupervised classification algorithms is impossible without information about the number of classes. But the number of classes is not always known (for example, novel experimental data) [3]–[5].

This paper aims to realize and verify the new clustering algorithm based on a combination of unsupervised Kohonen neural network and supervised probabilistic neural network. The proposed classification procedure allows to assign the objects to certain classes based on a set of their characteristics without a priori information about the number of classes and without a training set.

II. Theory and Data Processing Details

A. *Proposed Clustering Algorithm*

For the first time, the proposed clustering algorithm was described in [2] and used for solvent classification.

This paper presents the results of a comparison of the proposed clustering algorithm with traditional clustering algorithms using model data sets.

The core of the idea is the combination of unsupervised Kohonen neural network [6] with supervised learning probabilistic neural network [7]. The role of Kohonen neural network is to determine the first training data set for probabilistic neural network. Kohonen neural network was created with different numbers of input neurons. We took the number of neurons from 3 and above. The first training data set for probabilistic neural network included the objects which were assigned to the same classes independently to the number of neurons. The remaining objects were divided into testing sets and classified by means of probabilistic neural network. A leave-one-out cross-validation [8] was used for verification and correction of the obtained classification.

The algorithm was designed as M-script in Matlab 7.11b software.

The proposed algorithm was tested in the classification of different data sets with a variable parameter of the algorithm (the number of neurons for Kohonen network).

B. *Data Sets*

Two well-known multivariate model data sets were used in this work for testing the proposed classification algorithm:

1) The data set of iris flowers contains information on 150 samples, each of which is characterized by four numerical

features (sepal length and width, petal length and width). The iris flower samples are divided into three classes, each containing 50 samples [9].

2) The wine data set contains the results of determining 13 features of 178 Italian wine samples belonging to three classes (59 wines of the first class, 71 wines of the second class, 48 wines of the third class) [10].

And one more data set. Model data with a two-hierarchical structure represents a set of 41 samples, and these samples can be assigned both to three classes and to nine classes (Fig. 1).

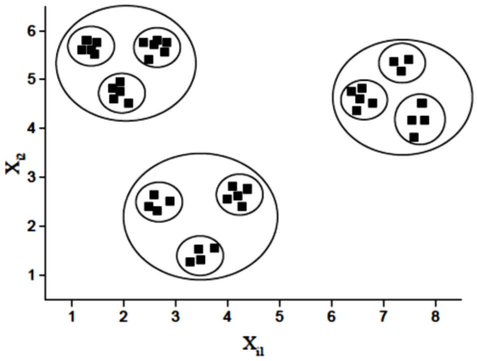

Fig. 1. Model data with a two-hierarchical structure.

C. Traditional Clustering Methods

Hierarchical clustering is a common method in data analysis [11]. Its essence is that at each step the object is considered as a separate cluster. The process of merging clusters is an identification of the two closest clusters by use of an appropriate distance. There are a variety of possible metrics: single-linkage (nearest neighbor), complete-linkage (farthest neighbor), median-linkage, centroid-linkage, Ward-linkage.

K-means clustering [12] and fuzzy k-means clustering [13] are also common methods for solving classification problems. Action of K-means clustering is to minimize the total squared deviation of cluster points from the centers of these clusters. Fuzzy clustering methods allow the same object to belong to several (or even all) clusters at the same time but with different degrees of membership. Fuzzy clustering is more appropriate for objects located on the border of clusters.

III. Results and Discussion

Auto-scaling transformation of data sets was performed due to the large range of initial data [14]:

$$x_i^{norm} = \frac{x_i - \bar{x}}{std(x)}, \; i = 1, 2, \dots, N \qquad (1)$$

where x^{norm} is modified values of parameters, $\bar{x}$ is mean value of the corresponding parameter, x_i is initial values of parameters, *std (x)* is standard deviation of the corresponding parameter.

The classification error was estimated as the proportion of incorrectly classified objects [7]:

$$P = \frac{n}{N}, \qquad (2)$$

where *n* is the number of objects classified wrongly, and *N* is the total number of objects.

The results of clustering iris flowers data set using different methods are shown in Table I. The minimum classification error is observed as the result of using proposed classification procedure with the number of neurons for Kohonen neural network from 3 to 10. The classification errors for the proposed procedure with the other numbers of neurons are also less than for traditional clustering algorithms. One can observe the decreasing of proportion of incorrectly classified objects with the increasing number of neurons for Kohonen neural network.

TABLE I: Results of Clustering Iris Data Set

Method	*P*, %
Single-linkage clustering	32.7
Complete-linkage clustering	15.3
Median-linkage clustering	10.0
Centroid-linkage clustering	32.0
Ward-linkage clustering	9.3
K-means clustering	11.3
Fuzzy k-means clustering	10.7
Proposed algorithm (number of neurons for Kohonen network from 3 to 6)	6.0
Proposed algorithm (number of neurons for Kohonen network from 3 to 7)	4.0
Proposed algorithm (number of neurons for Kohonen network from 3 to 8)	4.0
Proposed algorithm (number of neurons for Kohonen network from 3 to 9)	2.6
Proposed algorithm (number of neurons for Kohonen network from 3 to 10)	1.3

The results of clustering wine data set using different methods are shown in Table II. The minimum classification errors is observed as the result of using the proposed procedure with the number of neurons for Kohonen network from 3 to 7 and from 3 to 9.

For the wine samples, one can see some variation in the error, which reaches minimum values at odd values of the number of neurons. This can be explained by the high complexity of data on wine samples for classification, because each sample has a large number of characteristics, compared to other data sets.

TABLE II: Results of Clustering Wine Data Set

Method	*P*, %
Single-linkage clustering	69.1
Complete-linkage clustering	32.6
Median-linkage clustering	33.1
Centroid-linkage clustering	49.4
Ward-linkage clustering	32.6
K-means clustering	36.0
Fuzzy k-means clustering	32.6
Proposed algorithm (number of neurons for Kohonen network from 3 to 6)	41.6
Proposed algorithm (number of neurons for Kohonen network from 3 to 7)	28.7
Proposed algorithm (number of neurons for Kohonen network from 3 to 8)	35.4
Proposed algorithm (number of neurons for Kohonen network from 3 to 9)	29.2
Proposed algorithm (number of neurons for Kohonen network from 3 to 10)	33.1

Results of clustering model data set with a two-hierarchical structure into 3 and 9 classes using different methods are given in Table III. The zero classification error is observed for all traditional clustering algorithms except fuzzy k-means clustering.

The proposed procedure corrects the classified model data set with a two-hierarchical structure into 3 classes using number of neurons for Kohonen network from 3 to 10. Results obtained by using of proposed procedure with other numbers of neurons for Kohonen network are characterized by high values of classification error.

The proposed procedure correctly classified the model data set with a two-hierarchical structure into 9 classes using the number of neurons for Kohonen network from 3 to 7 and from 3 to 9.

TABLE III: RESULTS OF CLUSTERING MODEL DATA SET WITH A TWO-HIERARCHICAL STRUCTURE

Method	*P*, %	
	Three classes	Nine classes
Single-linkage clustering	0.0	0.0
Complete-linkage clustering	0.0	0.0
Median-linkage clustering	0.0	0.0
Centroid-linkage clustering	0.0	0.0
Ward-linkage clustering	0.0	0.0
K-means clustering	0.0	0.0
Fuzzy k-means clustering	13.4	13.4
Proposed algorithm (number of neurons for Kohonen network from 3 to 6)	34.1	7.3
Proposed algorithm (number of neurons for Kohonen network from 3 to 7)	36.6	0.0
Proposed algorithm (number of neurons for Kohonen network from 3 to 8)	34.2	7.3
Proposed algorithm (number of neurons for Kohonen network from 3 to 9)	39.0	0.0
Proposed algorithm (number of neurons for Kohonen network from 3 to 10)	0.0	19.5

IV. CONCLUSIONS

An applicability to solving complex classification tasks of the new clustering approach has been studied. The proposed clustering algorithm is based on the combination of Kohonen and probabilistic neural networks, and does not use a priori information about the number of classes and training set. The proposed clustering algorithm is quite competitive compared to traditional methods. The optimal number of neurons for Kohonen neural network for the realization of the proposed classification procedure is from 3 to 7 or from 3 to 9.

The proposed clustering algorithm can be used as an exploratory analysis of multidimensional data sets.

CONFLICT OF INTEREST

The authors declare that they do not have any conflict of interest.

REFERENCES

[1] Li N, Martin A, Estival R. Heterogeneous information fusion: combination of multiple supervised and unsupervised classification methods based on belief functions. *Inf. Sci*. 2021; 544: 238–265. doi: 10.1016/j.ins.2020.07.039.

[2] Pushkarova Y, Kholin Y. A procedure for meaningful unsupervised clustering and it’s application for solvent classification. *Centr. Eur. J. Chem.* 2014; 12(5): 594−603. doi: 10.2478/s11532-014-0514-6.

[3] Banerjee P, Chattopadhyay T, Chattopadhyay AK. Comparison among different Clustering and Classification Techniques: Astronomical data-dependent study. *New Astron*. 2023; 100: 101973. doi: 10.1016/j.newast.2022.101973.

[4] Guan C, Yuen KKF, Coenen F. Particle swarm optimized density-based clustering and classification: Supervised and unsupervised learning approaches. *Swarm Evol. Comput*. 2019; 44: 876–896. doi: 10.1016/j.swevo.2018.09.008.

[5] Mutihac L, Mutihac R. Mining in chemometrics. *Anal. Chim. Acta*. 2008; 612(1): 1–18. doi: 10.1016/j.aca.2008.02.025.

[6] Marini F, Zupan J, Magrì AL. Class-modeling using Kohonen artificial neural networks. *Anal. Chim. Acta*. 2005; 544(1-2): 306–314. doi: 10.1016/j.aca.2004.12.026.

[7] Pushkarova Y, Kholin Y. The classification of solvents based on solvatochromic characteristics: the choice of optimal parameters for artificial neural networks. *Centr. Eur. J. Chem*. 2012; 10(4): 1318−1327. doi: 10.2478/s11532-012-0060-z.

[8] Wong Tzu-Ts. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit*. 2015; 48(9): 2839–2846. doi: 10.1016/j.patcog.2015.03.009.

[9] UCI Machine Learning Repository. Center for Machine Learning and Intelligent Systems. *Iris Data Set* [Internet]. 2021 [cited 2023 August 11]. Available from: http://archive.ics.uci.edu/ml/datasets/Iris.

[10] UCI Machine Learning Repository. Center for MachineLearning and Intelligent Systems. *Wine Data Set* [Internet]. 1994 [cited 2023 August 11]. Available from: https://archive.ics.uci.edu/ml/datasets/wine.

[11] Darányi A, T. Czvetkó T, Kummer A, Ruppert T, Abonyi J. Multi-objective hierarchical clustering for tool assignment. *CIRP J. Manuf. Sci. Technol*. 2023; 42: 47–54. doi: 10.1016/j.cirpj.2023.02.002.

[12] Hu H, Liu J, Zhang X, Fang M. An Effective and Adaptable K-means Algorithm for Big Data Cluster Analysis. *Pattern Recognit*. 2023; 139: 109404. doi: 10.1016/j.patcog.2023.109404.

[13] Zhao X, Nie F, Wang R, Li X. Improving projected fuzzy K-means clustering via robust learning. *Neurocomputing*. 2022; 491: 34–43. doi: 10.1016/j.neucom.2022.03.043.

[14] Pushkarova YN, Sledzevskaya AB, Panteleimonov AV, Titova NP, Yurchenko OI, Ivanov VV, *et al*. Identification of water samples from different springs and rivers of Kharkiv: Comparison of methods for multivariate data analysis. *Mosc. Univ. Chem. Bull*. 2013; 68(1): 60−66. doi: 10.3103/S0027131412060077.