Liliana BYCZKOWSKA-LIPIŃSKA [1], Agnieszka WOSIAK[2]

University of Computer Sciences and Skills (1)
Institute of Information Technology, Lodz University of Technology (2)

doi:10.15199/48.2016.12.22

Fuzzy Classification of Medical Data Derived from Diagnostic Devices

Abstract. *The research described in this paper concerns fuzzy classification of medical datasets obtained from diagnostic devices. Experimental studies were performed with use of fuzzy c-means algorithm. It was shown that despite the low accuracy of the results, fuzzy classification reduce the risks associated with the loss of internal relationships in the characteristics of the data, and thus increases the chances of finding the pathological cases, as well as taking preventive actions or therapy.*

Streszczenie*. W ramach niniejszej pracy przeprowadzona została klasyfikacja rozmyta w odniesieniu do medycznych zbiorów danych pozyskanych z urządzeń diagnostycznych. Zastosowana została rozmyta metoda k-średnich. Badania wykazały, że pomimo niskiej dokładności rezultatów, klasyfikacja rozmyta zmniejsza ryzyko związane z utratą wewnętrznych zależności w charakterystyce danych, a tym samym zwiększa szanse na stwierdzenie ryzyka patologii i tym samym szybsze podjęcie działań zapobiegawczych lub terapeutycznych (**Rozmyta klasyfikacja danych medycznych pozyskanych za pomocą urządzeń diagnostyki medycznej**).*

Słowa kluczowe: eksploracyjna analiza danych, klasyfikacja rozmyta, rozmyta metoda k-średnich, dane medyczne
Keywords: fuzzy classification, data mining, fuzzy c-means algorithm, medical data

Introduction

Classification techniques enable automated analysis and diagnostic process of data. It is an extremely important and difficult issue as regards to medical cases, due to the need to achieve the highest rates of accuracy for the results of classification.

Numerous research studies have been undertaken to point the most accurate classification method. Nevertheless, there is no universal approach that could be successfully applied to a variety of scientific, industrial or medical problems [1]. For this reason, there is a constant need for further research.

The main difficulty, that emerges in automated diagnosis arise from a degree of ambiguity and uncertainty [2]. In a hard classification each instance must be assign to a particular group or cluster. This restriction does not reflect the reality since many cases may have the characteristics similar to instances assign to several, different categories.

It is worth mentioning, that while considering an automated classification process, it is necessary to include also the cases, where it is not possible to strictly point out their full belonging to one or many categories. Fuzzy classification indicates degrees of membership of a particular medical instance to many classes instead of strict pointing one of them. As a result it protects from losing some information, that may be important for further medical analysis, for example approaching the pathologic characteristics. Moreover, difficult or outlying cases of a dataset are better recognized since the degree of membership is continuous rather than all-or-none [3].

The aim of this research was to constitute an independent contribution to the relevant literature in terms of fuzzy classification process as well as a try to prove that introducing fuzzy approach into classification is recommended and may bring benefits for results of medical inference and future diagnosis.

The remainder of this paper is organized as follows. Section 2 (Related Works) presents literature review concerning fuzzy classification applied in the diagnosis of medical data. Next section (Materials and Methods) concerns the description of the proposed methodology. In Section 4 (Experimental Analysis and Results) we describe the studies that were conducted. We introduce data collected for this application and discuss the results. Finally, in Section 5 (Conclusions) we summarize our research and describe further works.

Related Works

Large number of researches concerning data classification [4, 5, 6] as well as implementation of fuzzy techniques were discussed in the literature during the last years [7]. They confirm that these techniques may be successfully applied in medical domain. At the same time it is noticeable that very few methods are used for treatment.

The fuzzy classification algorithms are successfully applied in economic and industrial areas of interest [8, 9]. However their implementation for medical data analysis is still controversial. It is mainly due to the fact that medical diagnosis (including automated one) is supposed to give a precise answer. Nevertheless, for some diseases, such as heart rate disorder [10, 11] or intrauterine growth restriction [12], it is not possible to fully classify the medical cases based on data derived from medical equipment.

Further scientific investigations, regarding fuzzy classification and including this research, may increase the chances to implement fuzzy approach in practice. Consequently, in the future medical staff will be able to make usage from estimated prompts, indicating degrees of membership for the particular medical case to the selected class of diagnosis, and as a result it will be possible to avoid difficult, manual analysis of numerous parameters obtained from medical studies.

Materials and Methods

Algorithms of fuzzy classification belong to the unsupervised machine learning group together with clustering techniques.

Three most popular fuzzy classification methods can be distinguished:

- fuzzy c-means clustering (FCM) [13],
- Gustafson-Kessel clustering [14],
- Gath-Geva clustering [15].

The goal of fuzzy c-means clustering is to minimize the objective function (1):

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} \mu_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad (1)$$

where: m - any real number greater than 1, μ_{ij} - the degree of membership of x_i in the cluster j, x_i - the ith of d-dimensional measured data, c_j - the d-dimension center of the cluster, and $\|*\|$ is any norm expressing the similarity between measured data and the center.
The fuzzy c-means algorithm can be described by the following four steps [13]:

1. Initialize $U = [\mu_{ij}]$ matrix: $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)} = [c_j]$ with $U^{(k)}$ according to the equation (2):

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} \mu_{ij}} \tag{2}$$

3. Update $U^{(k)}$ and $U^{(k+1)}$ according to the equation (3):

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_i\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{3}$$

4. If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$, then STOP, otherwise return to Step 2.

The Gustafson-Kessel fuzzy clustering extends the standard fuzzy c-means algorithm by changing the way the distance is calculated [15]. Instead of the Euclidean distance used by FCM, Mahalanobis distance is used in the form of the equation (4):

$$d_{GK}^2 = (x_k - c_i)^T A_i (x_k - c_i) \tag{4}$$

The matrix A_i is derived from the inversed fuzzy covariance C_i matrix represented by the equation (5):

(5)

$$A_i = (\rho_i |C_i|)^{1/d} C_i^{-1}, C_i = \frac{\sum_{k=1}^{N} \mu_{ik}^{m} (x_k - c_i)(x_k - c_i)^T}{\sum_{k=1}^{N} \mu_{ik}^{m}}$$

The Gath-Geva algorithm is also known as Gaussian Mixture Decomposition [16]. It also resembles FCM algorithm, however it uses Gauss distance instead of Euclidean distance denoted as follows in the equation (6):

(6)

$$d_{GD}^2 = \frac{1}{P_k} \sqrt{|A_i|} e^{\left(\frac{1}{2} (x_i - c_k)^T A^{-1} (x_i - c_k) \right)}$$

where: P_k – the probability that an element x_i belongs to the cluster c_k according to the equation (7):

$$P_k = \frac{\sum_{i=1}^{N} \mu_{ik}^{m}}{\sum_{i=1}^{N} \sum_{k=1}^{C} \mu_{ik}^{m}} \tag{7}$$

To evaluate the quality of the results, cluster validity should be performed. In literature (inter alia [17, 18]) different validity indices have been proposed and they can divided into two main groups:

- external evaluation - to compare the results with known labels, if possible,
- internal evaluation - based on the data being subjected to classification to assess the compactness, connectedness and the separation.

External validity is based on additional parameters – they are usually labels defined by experts for separate subgroups. This kind of evaluation determines the degree of accuracy for obtained results to the predefined, standardized classes.
In some cases it is not possible to get class names in advanced. Therefore a validity based on data used for clustering may be performed. Such an approach – called an internal measure – assigns a high value to the models for which there is strong similarity between objects within clusters and weak similarity between object belonging to different clusters [19]. The assessment is based on a compactness, a connectedness and a separation between clusters.
Compactness (also called homogeneity) usually measures intra-cluster variance by within-sum-of-squared-errors [20].
The connectedness determines a degree of local densities [20].
The separation quantifies the degree of disconnection between clusters [20].
There are also measures that combine the above mentioned factors: Dunn Index [21], Davies–Bouldin Index [22] or Silhouette Width [23] to name a few.

Experimental Analysis and Results

Data for the analysis were obtained from the research servers providing their resources for scientific purposes [24, 25]. The following medical studies with the corresponding data sets were analyzed:

- cardiotocographic data (CTG),
- cardiac SPECT images data.

The CTG dataset refers to cardiotocographic data. It consists of 2126 measurements and classifications of foetal heart rate (FHR) signals. Each instance is described by 21 parameters and one label that classifies all cases into three categories: normal, suspect and pathological.
The SPECT dataset was gathered as a result of Single Proton Emission Computed Tomography images and their diagnosis. It consisted of 267 instances described by 44 attributes. Each data case was categorized by appropriate binary label: normal and abnormal.

The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images.

As a result, 44 continuous feature pattern was created for each patient.

The pattern was further processed to obtain 22 binary feature patterns.
The experimental studies consisted of three main steps:

1. data preparation,
2. performing fuzzy clustering on all datasets,
3. validation of results.

The first step – data preparation - referred to all actions that were necessary to perform further classification and included verification of missing data and removing labels form the parameter list.
As it has been already proved that the improvements over the standard fuzzy c-means algorithm (including Gustafson-Kessel clustering and Gath-Geva method) do not produce significant differences for most datasets being investigated [15], only the results of FCM are introduced in this research.
The experiments were performed with use of WEKA data mining tool (Waikato Environment for Knowledge Analysis) [26], Matlab environment [27] and Statistica (Statsoft, USA) [28].
The second step - fuzzy clustering – returned the results in the form of distances between each case and the centroids of clusters. The outcomes were verified with the predefined categories assigned to each instance. The results of classification in terms of accuracy, precision and sensitivity for each dataset are summarized in Table 1.

Table 1. The results of cluster analysis.

Dataset	Accuracy	Precision	Sensitivity
CTG	0.44	1.00	0.44
SPECT	0.50	0.71	0.63

One can noticed that obtained accuracies are not satisfactory enough to use this approach as a hard classification.
Nevertheless, as is was stated before in the introduction of this research, fuzzy classification of medical data enables discovering some hidden information on how "close" are particular cases to another kind of diagnosis. In our research fuzzy clustering allows to state, that basing on the values of parameter, the case classified by the expert as still normal is far close to abnormal or suspect.
Tables 2 and 3 shows the degrees of membership for the most fuzzified cases of CTG and SPECT datasets. By "the most fuzzified" cases we mean instances for which the differences between two adjacent clusters are the smallest.

Table 2. The degrees of membership for the most fuzzified cases of CTG dataset using fuzzy c-means clustering.

μ_{1j}	μ_{2j}	μ_{3j}
Class: Normal		
0.6430	0.2355	0.1215
0.5988	0.2623	0.1389
0.4999	0.3708	0.1293
Class: Suspect		
0.1268	0.6667	0.2065
0.1103	0.6253	0.2643
0.0987	0.4690	0.4324
Class: Pathological		
0.0720	0.2631	0.6650
0.0988	0.2840	0.6175
0.1256	0.3833	0.4911

Table 3. The degrees of membership for the most fuzzified cases of CTG dataset using fuzzy c-means clustering.

μ_{1j}	μ_{2j}
Class: Normal	
0.5119	0.4881
0.5135	0.4865
0.5201	0.4799
Class: Abnormal	
0.4960	0.5040
0.4960	0.5040
0.4973	0.5027

One can see, that for each dataset and for each assigned cluster, there are medical cases, where the distances between two adjacent classes are comparable. For example the case from CTG dataset classified as "Normal" differs in less the 5% in the degree of membership from being classified as "Abnormal". This information may suggest more careful observation of patients as their medical characteristics is close to the pathologic cases. Such a conclusion proves that fuzzy (also called soft) classification is required and should be included into automated classification of medical data.

Conclusions

Medical diagnosis, including automated inference, is expected to give a precise answer. However, for certain types of diseases such as cardiac arrhythmias, or intrauterine fetal growth, full classification on the basis of data obtained from diagnostic equipment it is not possible. Moreover, hard classification may cause losing some important information on data characteristics, as its process is based on "all-or-nothing" rule, whereas there might be slight differences in parameters between medical cases assigned to different classes.

The studies on fuzzy classification increase the chances of its application in practice. As a result the medical staff will be able to use the estimated suggestions defining the degree of membership of a clinical case to a particular class of diagnosis, and thus avoid a difficult, manual analysis of a number of parameters obtained during the investigation.
Further studies should be primarily associated with the use of other strategies of membership function selection. Moreover, techniques for accuracy improvement should be considered, such as Artificial Immune System (AIS) method for fuzzy rules mining [29] or Similarity Adjustment Model (SAM) using adjusted Fuzzy Similarity Functions (FSF) presented in [30].

***Authors**: prof. dr hab. inż. Liliana Byczkowska-Lipińska,, University of Computer Sciences and Skills, ul. Rzgowska 17 a, 93-008 Lodz, Poland e-mail: liliana.byczkowska-lipinska@p.lodz.pl*
dr inż. Agnieszka Wosiak, Lodz University of Technology, Institute of Information Technology, ul. Wólczańska 215, 90-924 Lodz, Poland, e-mail: agnieszka.wosiak@p.lodz.pl

REFERENCES

[1] ESFANDIARI, N., BABAVALIAN, M. R., MOGHADAM, A. M. E., AND TABAR, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications, 41(9), 4434-4463.
[2] MAHMOODABADI, S. Z., ALIREZAIE, J., BABYN, P., KASSNER, A. AND WIDJAJA, E. (2010). Wavelets and fuzzy relational classifiers: A novel spectroscopy analysis system for pediatric metabolic brain diseases. Fuzzy sets and systems, 161(1), 75-95.
[3] GUSTAFSON, D. E., AND KESSEL, W. C. (1978). Fuzzy clustering with a fuzzy covariance matrix." Scientific Systems. Inc., Cambridge, MA.
[4] KORZENIEWSKA E., DURAJ A., KRAWCZYK A.: Detection of local changes in resistance by means of data mining algorithms. Przegląd Elektrotechniczny, 2014, 90.12: 229-232.
[5] LAKSHMI JEETHA, SARAVAN KUMAR, A. SURESH: A novel hybrid medical diagnosis system based on genetic data adaptation decision tree and clustering, ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 16, 2015
[6] WU, C. H., LAI, C. C., CHEN, C. Y., CHEN, Y. H. (2015). Automated clustering by support vector machines with a local-search strategy and its application to image segmentation. Optik-International Journal for Light and Electron Optics, 126(24), 4964-4970.
[7] PANDEY B., MISHRA R. B.: Knowledge and intelligent computing system in medicine. Computers in biology and medicine, 2009, 39.3, pp. 215-230.
[8] JEFMAŃSKI, B. (2009). Rozmyte metody klasyfikacji w analizie segmentów rynkowych na przykładzie rynku motoryzacyjnego.
[9] WIDJAJA M, DARMAWAN A, MULYONO S, Fuzzy classifier of paddy growth stages based on synthetic MODIS data. In Advanced Computer Science and Information Systems (ICACSIS), 2012 IEEE International Conference on; 239-244.
[10] USHER, J., CAMPBELL, D., VOHRA, J. AND CAMERON, J. (1996). Fuzzy classification of intra-cardiac arrhythmias. In Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine. Proceedings of the 18th Annual International Conference of the IEEE (Vol. 3, pp. 997-998). IEEE.
[11] ZAMOJSKA J., NIEWIADOMSKA-JAROSIK K., WOSIAK A., LIPIEC P., STAŃCZYK J.: Myocardial dysfunction measured by tissue Doppler echocardiography in children with primary arterial hypertension, Kardiologia Polska 2015, DOI: 10.5603/KP.a2014.0189
[12] ZAMECZNIK, A., NIEWIADOMSKA-JAROSIK, K., WOSIAK, A., ZAMOJSKA, J., MOLL, J. AND STAŃCZYK J. (2014) Intra-uterine growth restriction as a risk factor for hypertension in children six to 10 years old, Cardiovascular Journal of Africa, 2014, pp. 73-77
[13] BEZDEK, J. C., EHRLICH, R. AND FULL, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers and Geosciences, 10(2), 191-203.

[14] GATH, I., GEVA, A. B. (1989). Fuzzy clustering for the estimation of the parameters of the components of mixtures of normal distributions. Pattern Recognition Letters, 9(2), 77-86.
[15] GRAVES D., PEDRYCZ W.: Fuzzy c-means, Gustafson-Kessel FCM, and kernel-based FCM: A comparative study. Analysis and Design of Intelligent Systems using Soft Computing Techniques. Springer Berlin Heidelberg, 2007. pp. 140-149.
[16] WANG N., LIU X., YIN J.: Improved Gath–Geva clustering for fuzzy segmentation of hydrometeorological time series. Stochastic environmental research and risk assessment, 2012, 26.1: 139-155.
[17] REZAEE B.: A cluster validity index for fuzzy clustering. Fuzzy Sets and Systems, 2010, vol.161(23), pp. 3014-3025.
[18] JASZUK M., MROCZEK T.: FRYC, Barbara. Testy porównawcze metod klasteryzacji jako narzędzia identyfikacji grup studenckich oraz tworzenia klas pytań ankietowych. Projekt współfinansowany ze środków Unii Europejskiej z Europejskiego Funduszu Rozwoju Regionalnego oraz z budżetu Państwa w ramach Regionalnego Programu Operacyjnego Województwa Podkarpackiego na lata 2007 – 2013. Inwestujemy w rozwój województwa podkarpackiego.
[19] MANNING C.D., RAGHAVAN P., SCHÜTZE H.: Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008
[20] HANDL J., KNOWLES J., KELL D.B., Computational cluster validation in postgenomic data analysis. Bioinformatics, 21(15) 2005, pp. 3201–3212.
[21] DUNN J.C.: Well separated clusters and fuzzy partitions. J. Cybernet. (1974) 4:95–104.
[22] DAVIES D.L., BOULDIN D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell. (1979) 1:224–227.
[23] ROUSSEEUW P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. (1987) 20:53–65.
[24] BERNARDES J., Faculdade de Medicina, Universidade do Porto, Porto, Portugal, Reference: D AYRES DE CAMPOS ET AL. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318
[25] CIOS K.J., KURGAN L., University of Colorado at Denver, Denver, CO 80217, U.S.A, Reference: KURGAN, L.A., CIOS, K.J., TADEUSIEWICZ, R., OGIELA, M., GOODENDAY, L.S.: Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis, Artificial Intelligence in Medicine, vol. 23:2, pp. 149-169, 2001
[26] WITTEN I.H., FANK E.: Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
[27] NATICK M.A.: Fuzzy logic toolbox for use with Matlab, The MathWorks Inc., 1998
[28] STANISZ A., TADEUSIEWICZ R.: Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. StatSoft, 2007.
[29] MĘŻYK E., UNOLD O.: Mining fuzzy rules using an Artificial Immune System with fuzzy partition learning. Applied Soft Computing, 2011, 11.2: 1965-1974.
[30] MEGED A., GELBARD R.: Adjusting Fuzzy Similarity Functions for use with standard data mining tools. Journal of Systems and Software, 2011, 84.12: 2374-2383.