Indian Journal of Science and Technology, Vol 9(S1), DOI: 10.17485/ijst/2016/v9iS1/94094 , December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Enhanced Fuzzy K-NN Approach for Handling Missing Values in Medical Data Mining

R. Naveen Kumar* and M. Anand Kumar

Department of Computer Science, Karpagam University, Coimbatore - 641021, Tamil Nadu, India;
rnaveenkumarooty@gmail.com, anand2kumarm@gmail.com

Abstract

Objectives: Exploratory data study is regularly indispensable to evaluate a potential premise for the subsequent studies such as grouping the data in clusters or diversifying the data in classification. Very common incident in the real data is incompleteness. **Methods/Statistical Analysis:** This problem can result in the biased treatment comparisons and also impacts the overall statistical power of the study. Missing data are proposed in several methods. The central idea of this proposed method is to handle the uncertainty of the missing values due to the vagueness arises in the real world datasets. This research work overcomes the inconsistency of the missing datasets and the proposed method tolerates the missing values using the fuzzy based K-NN. Three different well known datasets are used in this paper. **Findings:** The results demonstrate that the proposed method is capable of imputing the missing values even with high presence of missing values and overwhelmed the problem of uncertainty precisely. **Application/Improvements:** As compared to other techniques the proposed fuzzy based K-NN gives high inconsistency. In future it can be improved in concentrating with big dataset and more effectively and efficiently result could be substituted by applying the expected value.

Keywords: Data Mining, Fuzzy, Imputation, K-NN, Missing Values, Uncertainty

1. Introduction

A general problem in all type of dataset is senesced as the missing data. Various assortments of methods are used in handling the missing values[1]. The missing–indicator method and complete case analysis are the frequently used method in finding the missing[2] value. Traditional approach can also be used in finding missing value but it may lead to biased estimate data and either reduces statistical power[3,4]. It is understood easily which is kept hidden and unpredicted. Now a days missing data is a common problem in data mining and knowledge discovery database. Even in the large database or dataset many of the attributes value are incorrect. In political survey 50% data having missing value[5,6]. Family journalists are having a serious problem in this issue[7,8]. The common process for analyzing the data and knowledge is explicable in nature it can be easily detected, for the previously hidden data.

Noise data is a common problem in data mining and knowledge discovery[9]. Many attribute values are missing or incorrect because of the erroneous instrument measuring property or by human registering[10]. The uncertainty in handling the problem is not much focused in missing value handling[11]. When there is a vagueness in handling the missing values using traditional imputation[12] methods which leads to misclassification and misconception of the pattern analysis that may leads to false alarms. This paper introduces the concept of fuzzy

*Author for correspondence

based K-NN which tackles the problem of incompleteness in the dataset.

2. A Study on Missing value and Uncertainty in Medical Data Mining

Real-life data are frequently imperfect and erroneous. This may lead to incomplete and vagueness. Missing attribute value is considered as one of the form of data incompleteness[13]. Eradicating erroneous data is an important step when heterogeneous data sources are combined. In the data warehouse the dirty data files are seen prevalently. Incomplete information and naming attribute conventions are the two common problems in the data warehouse. In supervised classification of the missing value treatment of LDA classifier are consider only as two classes in multivariate mode[14]. K-NN imputation techniques are introduced for the missing value using the supervised classification techniques. Data mining[16] and Microarrays[17] are the two statical applications dealing with missing value. These applications include supervised classification as well as unsupervised classification (clustering)[18]. Missing values are replaced by zero in the Micro arrays[19]. Data analysis techniques are mainly based on the assumption using intelligent techniques[20]. Trails and tribulations could not derive conclusion on incomplete knowledge. In the medical data analysis the most common intelligent techniques[21] used are the Bayesian classifier[22,23], Genetic Algorithms[24], Decision Trees[25], Fuzzy theory[26–28] and Neural Network[29–31]. This paper handles the problem of uncertainty while imputing the missing values.

3. Dataset Description

3.1 Breast Cancer Dataset

This breast cancer databases is obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolbergset. In this dataset 699 instances and 10 plus the class attribute are used.

Information regarding Attributes:

- Diagnosis (M = Malignant, B = Benign) Ten real-valued features are computed for the each cell nucleus.
- ID number
- Mean of distances from center to points on the perimeter - Radius.
- Standard deviation of gray-scale values - Texture.
- Finding - perimeter and Area's.
- Local variation in radius lengths – Smoothness.
- Perimeter^2 / area - 1.0 - Compactness.
- Severity of concave portions of the contour – Concavity.
- Number of concave portions of the contour – Concave point.
- Calculating - symmetry.
- "coastline approximation" – 1 (fractal dimension).

3.2 Hepatitis Dataset

Number of Instances: 155

This dataset consist of 155 instances and the attributes including the class label is 20[7].

Information regarding Attributes:

- CLASS: DIE, LIVE
- AGE: 10, 20, 30, 40, 50, 60, 70, 80
- SEX: male, female
- STEROID: no, yes
- ANTIVIRALS: no, yes
- FATIGUE: no, yes
- MALAISE: no, yes
- ANOREXIA: no, yes
- LIVER BIG: no, yes
- LIVER FIRM: no, yes
- SPLEEN PALPABLE: no, yes
- SPIDERS: no, yes
- ASCITES: no, yes
- VARICES: no, yes
- BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
- ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
- SGOT: 13, 100, 200, 300, 400, 500
- ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
- PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
- HISTOLOGY: no, yes

3.3 Lung Cancer Dataset

Three types of pathological lung cancers dataset are described in this section. 32 instances and 56 one class attribute are totally used. The author does not give the source of individual variables or their originality of data.

Information regarding Attributes:

- 32 Number of instances.
- 57 Number of Attributes.
- 1 class Attributes, 56 predictive attributes.
- Class label is Attribute 1.
- All predictive attributes are nominal.
- 0-3 are integer values.
- Attribute 5 and 30(*) are Missing values.
- Fourth value from fifth attribute is -1 in original values.
- These values have been changed to ? (Unknown)(*).

In the original data 1 value for the 39 attribute was 4.

- Class Distribution: - 3 classes,
  - 9 observations
  - 13"
  - 10"

3.4 Materials and Methods

The original (historical) dataset is cleaned from all kind of inconsistencies. All tuples with the same conditional and different classification attributes are removed. This will clearly improve the efficiency because it removes all the suspected cases. The experiments are then designed to check the best model to fill the missing values that generates the highest coverage of the data set. In this section the detailed description of four different techniques used for the missing value handling are discussed.

4. Mean Imputation

Missing values are replaced with the estimates derived by applying the statistics methods to the available data. Impute Missing Values allows the estimation of the missing data for all or selected attributes. The substitution with a measure of central tendency which is commonly known as the mean substitution is one of the most frequently used method. The missing data for the giving attribute is replaced by the means of all known feature (attribute) in the class where the instance with the missing attribute belongs. Let us consider that the value Z_{ij} of the t-th class, C_t, is missing then it will be replaced by:

$$\bar{z}_{ij} = \sum_{i:z_{ij} \in C_t} \frac{z_{ij}}{n_t}$$

Algorithm:
Input: Data set, DS.
Output: Data set, DS, contains instances with no missing values.
Method:
for each selected Attribute Att in DS
Calculate the Mean Value (MV) of Att
{end for}
for each selected Attribute Att in DS
for each case C of Att
if the value of Att is null
fill the value of Att as MV
{end if}
{endfor}
{endfor}

4.1 Crisp K-NN

In the crisp Nearest Neighbor based imputation,[32,33] an attribute att with missing value is imputed by finding its k^{th} Nearest Neighbor and assigning its value to the attribute att.

Algorithm
Input :
Split the input Dataset DS into two:
DS_m – Dataset containing the instances in which at least one of the attribute value is missing
DS_c – dataset containing complete attribute information
Output: Dataset DS containing no missing values
Method

1. For each vector x in D_m:
2. Observe the instance vector before divide.
3. Calculate all the instance vectors from the set DC and also from the distance between x0.
4. Only features in instances are used from the complete set x0.
5. Which are observed in the vector x.
6. Use the K closest instances vectors (K-nearest neighbors).
7. Perform a majority voting to estimate the missing values for categorical attributes.
8. For the continuous attributes, replace the missing value using the mean value of the attribute in the k-nearest neighborhood.

4.2 Weighted K-NN

The weighted KNN utilizes the contribution of attributes information[34,35]. The algorithm of the weighted K-NN as follows:

Read the training data.
Read the testing data.
Set K to some value.
Set the learning rate α.
Set the value of N for the number of folds in the cross validation.
Using Standard deviation Normalize the attribute.
For each attribute Ai assign random weight wi.
Divide the number of training examples into N sets.
Train the weights by the cross-validation.
For every set N_k in N, do
Set N_k = Validation Set
For every example xi in N such that xi does not belong to Nk do
Find the K nearest neighbors based on the Euclidean distance
Return the class that represents the maximum of the k instances
If the actual class != predicted class then apply gradient descent
Error = Actual Class – Predicted Class
For every Wk
Wk = Wk + α * Error * Vk (where Vk is the query attribute value)
Calculate the accuracy as
Accuracy = (# of correctly classified examples / # of examples in Nk) X 100

4.3 Imputation based on Enhanced Fuzzy K-NN

The main drawback of the conventional K-Nearest Neighbor is that it looks for the most similar instances, the algorithm searches through all the data sets. The main objective of analyzing the large dataset in KDD is very critical for the researchers. In this proposed work the usage of Fuzzy K-NN is adopted for handling the missing values instead of classification. The proposed method reduces the search space of the K-NN by finding the similar instances of same class to which it belongs. Based on the membership of each classes the instances to which the highest membership value of the missing instances belong is alone considered for the similarity measurement and among them K-Nearest Neighbor is selected and depending on the type of the attribute the values are assigned to the discrete or continuous the values .

Logical assignments are not made by the algorithm because the class membership is not directly assign to a particular class rather sample vectors are used. Data normalization is used to get a better result for a large value.

Given a set of sample vectors {xl,x2 . . . . . x,}, a fuzzy c-partition of these vectors specifies the membership of each vector in each of c classes. This is denoted by the c by n matrix U, where $u_{ij} = u_i(x_j)$ is the membership of xj in class i for i = 1, 2 . . . . . c, and j = 1, 2 .... ,n. The following properties must be held for U to be a fuzzy c-partition:

$$\sum_{I=1}^{c} u_{ij} = 1,$$

$$0 < \sum_{j=1}^{n} u_{ij} < n,$$

$$u_{ij} \epsilon [0,1].$$

assign membership as a function of the vector's distance from its K-nearest neighbors and those neighbors' memberships are in the possible classes. Let W = {xl, x2 .... ,xn} be the set of n labeled samples and uij is the membership of the ith class of the jth vector of labeled sample set. The algorithm as follows:

Consider W={w_1, w_2, ..., w_t} a set of *t* labeled data. Each object w_i is defined by *l* characteristics *wi* = (w_{i1}, w_{i2}, ..., *wil*).

Input element y with missing value.

k the number of the closest neighbors of *y*.

E the set of *k*- Nearest Neighbors (NN).

$\mu_i(y)$ is the membership of *y* in the class *i*

μ_{ij} is the membership in the *i*th class of the *j*th vector of the labelled set (labelled *wj* in class *i*) .

BEGIN:

Let *t* be the number of elements that identify the classes. Let c be the number of classes.

Let *W* be the set that contain the *t* elements

Each cluster is represented by a subset of elements from *W*.

for i = 1 to c // number of classes

Calculate $\mu_i(y)$ using

$$\mu_i(y) = \frac{\sum_{j=1}^{k} \mu_{ij} \left(\frac{1}{\|y - x_j\|^{2/(m-1)}} \right)}{\sum_{j=1}^{k} \left(\frac{1}{\|y - x_j\|^{2/(m-1)}} \right)}$$

end loop

Let s be the sorted set of instances based on the highest class membership value.

Select the K-Nearest Neighbor of the highest membership class. It computes the similarity measure of the instances with the corresponding attributes of the instances under observation. If the type of the missing attribute is categorical then fill the missing value of y with the most frequently occurring value of the attribute in the K-Nearest Neighborhood else if the type of the missing attribute is continuous then replace the missing value of y with mean value of the attribute in the K-Nearest Neighborhood.

The fuzzy based distance metric is as follows: The distance metric similarity measures, the W (*F, G*) in which, for two fuzzy sets *F* and *G*,

$$W(F, G) = 1 - \frac{\sum_{i=1}^{n} |\mu Fi - \mu Gi|}{n}$$

In the case, if f is the instance with the missing value and g is the instance with the complete value. Then find the similarity among the two instances using the membership value of corresponding attribute instances expect the missing value attribute.

5. Experimental Result

To evaluate the performance of the proposed approach the three different benchmark dataset of UCI machine learning repository is used. The Table 1 shows the dataset characteristics such as the number of instances, number of features used in the dataset. The percentage of the missing value with respect to the whole dataset and the percentage of missing instances with at least one missing value.

The Table 2 shows the performance of the various imputation methods on the missing value estimation. The work is constructed as a complete matrix by removing all the rows containing the missing values and randomly creating 5% to 25% of the matrix entries. The artificial missing value is introduced by based on the row and by randomly selecting specific percentage of the entries in the complete matrix and the values between 5% – 25% are removed in each row. From the Table 2 it reveals that the normalized root mean square's error estimation for evaluating the missing value estimation for the proposed approach, fuzzy based K-NN performs the remaining techniques in the entire four different missing ratios [Figure 1]. Next Weighted K-NN works better [Figure 2] and the worst case is mean based imputation [Figure 3].

Table 1. Characteristics of the three dataset

Dataset	No. of Instances	No. of. Attributes	Missing Value (%)	Missing Instances (%)
Breast Cancer dataset	699	9	0.25	2.28
Hepatitis dataset	155	19	5.67	48.38
Lung Cancer dataset	32	56	0.15	1.45

Table 2. Performance of Imputation methods on missing value estimation.

Missing Ratio	Dataset	Mean Imputation	K-NN	Weighted-NN	Fuzzy K-NN
5%	Breast Cancer	0.84665	0.81768	0.78908	0.73911
	Hepatitis	0.82748	0.78677	0.71547	0.66732
	Lung Cancer	0.19858	0.19548	0.19173	0.19076
10%	Breast Cancer	0.85722	0.82813	0.79351	0.75606
	Hepatitis	0.83075	0.79108	0.72328	0.67124
	Lung Cancer	0.20554	0.19640	0.19249	0.19087
15%	Breast Cancer	0.86148	0/83257	0.80762	0.83091
	Hepatitis	0.83512	0.79659	0.73461	0.69786
	Lung Cancer	0.20817	0. 19786	0.19306	0.19138
20%	Breast Cancer	0.87753	0.83697	0.80934	0.76790
	Hepatitis	0.84754	0.79901	0.73895	0.69547
	Lung Cancer	0.20974	0. 19874	0.19695	0.19548
25%	Breast Cancer	0.88925	0.83997	0.81173	0.85455
	Hepatitis	0.85266	0.80023	0.72979	0.70941
	Lung Cancer	0.21054	0.19982	0.19884	0.19766

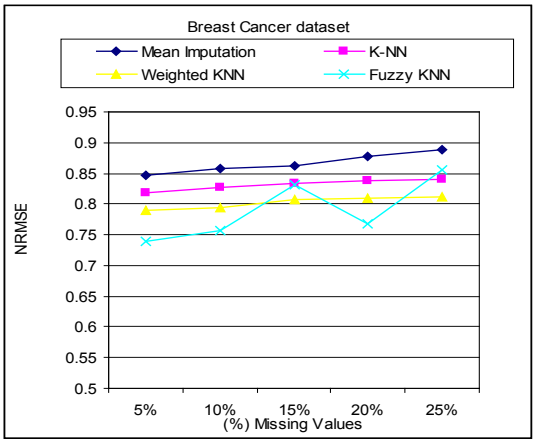

Figure 1. Breast cancer dataset.

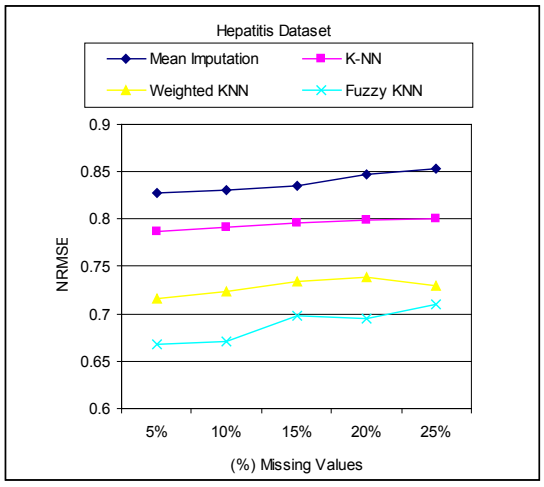

Figure 2. Hepatitis dataset.

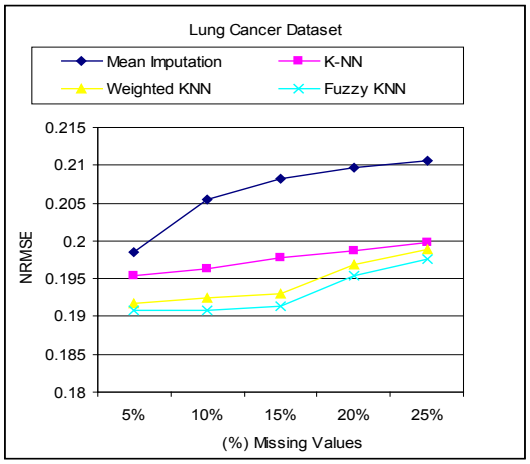

Figure 3. Lung cancer dataset.

6. Conclusion

This paper focuses on handling the problem of uncertainty while imputing missing values to the dataset. The proposed method utilizes the concept of fuzzy based K-NN in order to fill the missing values in case of high inconsistency in the given dataset. The result shows that by introducing different levels of missing ratio in the three different datasets it is identified that the proposed method outperforms the remaining algorithm.

7. References

1. Verma A, Kaur I, Arora N. Comparative analysis of information extraction techniques for data mining. Indian Journal of Science and Technology. 2016 Mar; 9(11). DOI: 10.17485/ijst/2016/v9i11/80464.
2. Lohita K, Amitha Sree A , Poojitha D , Renuga Devi T, Umamakeswari A. Performance analysis of various data mining techniques in the prediction of heart disease. Indian Journal of Science and Technology. 2015 Dec; 8(35). DOI: 10.17485/ijst/2015/v8i35/87458.
3. Kuppusamy V, Paramasivam I. A study of impact on missing categorical data - A qualitative review. Indian Journal of Science and Technology. 2016 Aug; 9(32). DOI: 10.17485/ijst/2016/v9i32/83088.
4. Grund S, Ludtke O, Robitzsch A. Multiple imputation of multilevel missing data: An introduction to the R Package pan. SAGE Open; 2016 Oct-Dec. p. 1–17. DOI: 10.1177/2158244016668220.
5. Sarab AlMuhaideba S, El Bachir Menaib M. An individualized preprocessing for medical data classification. Symposium on Data Mining Applications, SDMA2016. Riyadh, Saudi Arabia. 2016 Mar; 82:35–42.
6. Hulse JV, Khoshgoftaar TM. Incomplete case Nearest Neighbor imputation in software measurement data. Proceedings of the IEEE International Conference on Information Reuse and Integration; Japan. 2007. p. 630–7.
7. King G, Hopnaker J, Joseph A, Scheve K. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American Political Science Review. 2001. 95(1):49–69.
8. Chan P, Dunn OJ. The treatment of missing values in discriminant analysis. Journal of the American Statistical Association. 1972; 67(338):473–7.
9. Gil M, Sarabia EG, Llata JR, Oria JP. Fuzzy c-means clustering for noise reduction, enhancement and reconstruction of 3D ultrasonic images. Proceedings of the Emerging Technologies and Factory Automation. 1999; 1:465–72.

10. Li D, Gu H, Zhang LY. A fuzzy c-means clustering algorithm based on Nearest Neighbor intervals for incomplete data. Expert System Applications. 2010; 37(10):6942–7.
11. Agrawal R. Design and development of data classification methodology for Uncertain data. Indian Journal of Science and Technology. 2016 Jan; 9(3). DOI: 10.17485/ijst/2016/v9i3/72262.
12. Aydilek IB, Ahmet Arslan A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a Genetic Algorithm. Information Sciences. 2013 Jun; 233:25–35.
13. Manjula KR, Keshari AK, Pahlazani A. An approach to perform uncertainty analysis on a spatial dataset using clustering and distance based outlier detection technique. Indian Journal of Science and Technology. 2015 Dec; 8(35). DOI: 10.17485/ijst/2015/v8i35/71972.
14. Tresp V, Neuneier R, Ahmad S. Efficient methods for dealing with missing data in supervised learning. Editors, Advances in NIPS 7. MIT Press; 1995.
15. Magnani M. Techniques for dealing with missing data in knowledge discovery tasks. 40127 Bologna, Italy; 2015.
16. Edgar Acuna1 E, Rodriguez C. The treatment of missing values and its effect in the classifier accuracy. Proceedings of the Meeting of the International Federation of Classification Societies (IFCS); Illinois Institute of Technology, Chicago. 2004 Jul. p. 639–47.
17. Scheel I, Aldrin M, Glad IK, Sorum R, Lyng H, Frigessi A. The influence of missing value imputation on detection of differentially expressed genes from microarray data. Bioinformatics. 2005; 21(23):4272–9.
18. Sharma S, Srivastava SK. Review on text mining algorithms. International Journal of Computer Applications. 2016 Jan; 134(8):39–43.
19. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Bostein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17(6):520–5.
20. Maksood FZ, Achuthan G. Analysis of data mining techniques and its applications. International Journal of Computer Applications. 2016 Apr; 140(3):6–14.
21. Naveen Kumar R, Anand Kumar M. Medical data mining techniques for health care systems. IJESC; 6(4):3498–503. DOI: 10.4010/2016.811. ISSN 2321 3361.
22. Goel A, Srivastava, SK. Study of data mining algorithms in the context of performance enhancement of classification. International Journal of Computer Applications. 2016 Jan; 134(9):1–5.
23. Kharya S, Soni S. Weighted naive bayes classifier: A predictive model for breast cancer detection. International Journal of Computer Applications. 2016 Jan; 133(9):32–7.
24. Li D, Gu H, Zhang L . A fuzzy c-means clustering algorithm based on Nearest Neighbor intervals. Expert System Application. 2010; 37(10):6942–7.
25. Madadipouya1K. A new decision tree method for data mining in medicine. Advanced Computational Intelligence: An International Journal (ACII). 2015 Jul; 2(3).
26. Liao Z, Lu X, Yang T, Wang H. Missing data imputation: A fuzzy K means clustering algorithm over sliding window. Fuzzy Systems and Knowledge Discovery. 2009; 3(1):133–7.
27. Bezdek JC, Keller J, Krishnapuram R, Pal NR. Fuzzy models and algorithms for pattern recognition and image processing. Kluwer Academic Publishers. 1999; 4:89.
28. Krishnapuram R, Keller JM. A possibilistic approach to clustering, IEEE Transactions on Fuzzy Systems.1999; 1:98–110.
29. Lim CP, Leong JH, Kuan MM. A hybrid neural network system for pattern classification tasks with missing features. IEEE Transactions Pattern Analysis. 2005; 27(4):648–53.
30. Ho Tim TN. Predicting HIV status using neural networks and demographic factors. [Unpublished]. Johannesburg, South Africa: University of Witwatersrand; 2006 Apr.
31. Abdella M, Marwala T. Treatment of missing data using neural networks and genetic algorithms. Proceedings of the 2005 IEEE International Joint Conference on Neural Networks (IJCNN'05);. 2005; 1:598–603.
32. Jena PK, Chattopadhyay S. Comparative study of fuzzy K-Nearest Neighbor and fuzzy c-means algorithms. International Journal of Computer Applications. 2012 Nov; 57(7).
33. Karegowda AG, Kishore B. Enhancing performance of KNN classifier by means of Genetic Algorithm and Particle Swarm Optimization. IJAFRC, 2014 May; 1(5). ISSN 2348 – 4853.
34. Duch W, Grudzinski K. The weighted K–NN with selection of features and its neural realization.
35. Duch W, Adamczak R, Grabczewski K, Zal G, Hayashi Y. Fuzzy and crisp logical rule extraction methods in application to medical data. Fuzzy Systems in Medicine, Springer. 1999; 41:593–616.