Artificial neural network,
granular computing, data fusion,
medical data analysis

Marcin BERNAS[1]

ANN AS JUSTIFIED GRANULAR COMPUTING MECHANISM FOR MEDICAL DATA CLASSIFICATION

The medical data and its classification have to be treated in particular way. The data should not be modified or altered, because this could lead to false decisions. Most state-of-the-art classifiers are using random factors to produce higher overall accuracy of diagnosis, however the stability of classification can vary significantly. Medical support systems should be trustworthy and reliable, therefore this paper proposes fusion of multiple classifiers based on artificial Neural Network (ANN). The structure selection of ANN is performed using granular paradigm, where granulation level is defined by ANN complexity. The classification results are merged using voting procedure. Accuracy of the proposed solution was compared with state-of-the-art classifiers using real medical data coming from two medical datasets. Finally, some remarks and further research directions have been discussed.

1. INTRODUCTION

Medical diagnosis support systems gained acceptance among physicians and are commonly used in medical daily routines. Nowadays, these systems are often integrated into medical devices. Despite the active research in this topic, even well-known and reliable classifiers still tend to fail when faced with atypical data. Therefore, there is a continuous search for new methods to tackle with arising challenges and to improve the quality of decision support systems. In this paper a method for medical data classification is proposed, with fusion of granules [22] produced using artificial neuron network (ANN) [12]. Multi-layer perceptron (MLP) ANN proved to be a vital tool in classification and based on newly develop faster learning method [12] its possibilities extended significantly. Therefore, the paper proposes to use this method in connection with granular computing paradigm GC [23]. Using GC paradigm, the data can be aggregated into many formal representations of information granules: intervals [11], fuzzy sets [16], rough sets [7], shadowed sets [20], or probabilistic sets [10]. There are several works, which prove the usefulness of this concept [18], [8], [13], [19], [17]. The granulation method tends to generalize the information at the cost of losing the details. Therefore, multiple levels (specificity) of granulation were proposed [5] to perform correct classification. This solution allows using general classification and also takes into consideration particular data sets. The accuracy of proposed method was compared with various classifiers: k-NN classifier [3], fuzzy rules classifier [4], random trees [6] and Bayes network classifier [14]. The proposed

[1]Institute of Computer Science, University of Silesia,Bedzinska 39, 41-200 Sosnowiec, Poland

method was verified using two archival medical datasets. The remainder of the work is organized as follows. Section 2 presents a proposed method. The classification results are presented in Section 3. Section 4 includes discussion with remarks on possibilities of further development of the method.

2. PROPOSED METHOD

The proposed method bases on previous research [1], which proved that using the information granules of different specificity (level) allows to increase the classification accuracy. Additionally, the accuracy increase can be achieved by fusion of classifiers [2]. The medical diagnosis is often taken based on the patients parameter analysis. In classical approach an information representation is constructed, e. g. intervals. If the observed parameter is in appropriate interval then given diagnosis is favored. Unfortunately, based on simple interval representation and one granulation level some information is lost.

Therefore, justified granulation paradigm [21] approach is used with various representations to tackle this problem. Instead of using intervals [11], fuzzy sets [16], rough sets [7] or other representation, the weights of the neural network are used as granular model.

In contrast to previous works [1] [2], the ANN of various structures was used as granulation mechanism. Intuitively, if the structure of ANN is simple (e. g. one neuron), the obtained granule level is high. On the other hand, if the ANN architecture contains multiple layer and multiple neurons per layer then the granule level is low. Therefore, the granule specificity level is defined as the structure complexity of ANN defined by α = [0, 1]. The proposed method as a parameter obtain the number of granulation levels (p=[1,...,n]). Using these levels the p different ANN architectures are created. The number of neurons (n) and layers (l) for given level α_j j=1,...,p is defined as:

$$\alpha_j = \begin{cases} \frac{j-1}{p-1} & : \quad 1 < j \leq p \\ 0 & : \quad p = 1 \end{cases} \tag{1}$$

$$n_j = \alpha_j * (n_{max} - 1) + 1 \tag{2}$$

$$l_j = \alpha_j * (l_{max} - 1) + 1 \tag{3}$$

where:n_{max} - maximal number of neurons in layer, l_{max} - maximal number of layers.

The generated architectures (n_j neurons in l_j layers) are used as the knowledge at given j-th specificity level. The example of ANN structures generated for p parameter equal to 3 was ilustrated in Fig. 1

The final diagnosis is performed as majority voting of all specificity levels. If a tie occurred while voting, the decision of ANN with middle specificity layer $\lfloor \alpha_{p/2} \rfloor$ is selected. The method diagram is presented in Fig. 2.

The proposed method was verified using two datasets containing real medical data. The first database contains heart disease diagnosis [9]. The second database contains electrical impedance measurements of freshly excised breast tissue samples [15]. These datasets have no missing values. The datasets mentioned above can be characterized as follows:

- Heart disease: 13 parameters, 270 cases, 2 groups of diagnoses (angiographic disease status), group distribution: ">50% narrowing" – 44%, "<50% narrowing" – 56%;
- Breast tissue: 9 parameters, 106 cases, 6 groups of diagnoses (tissue type), group distribution: "Connective" – 13%, "Fibro–adenoma" – 14%, "Glandular" – 15%, "Mastopathy" – 17%, "Carcinoma" – 20%, "Adipose" – 21%;

The aforementioned datasets were used in the calibration and classification process described in next section.

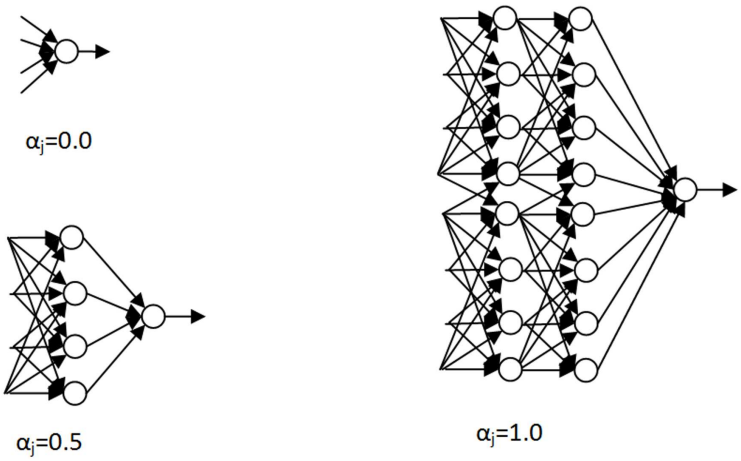

Fig. 1. The complexity of ANN network based specificity α_j.

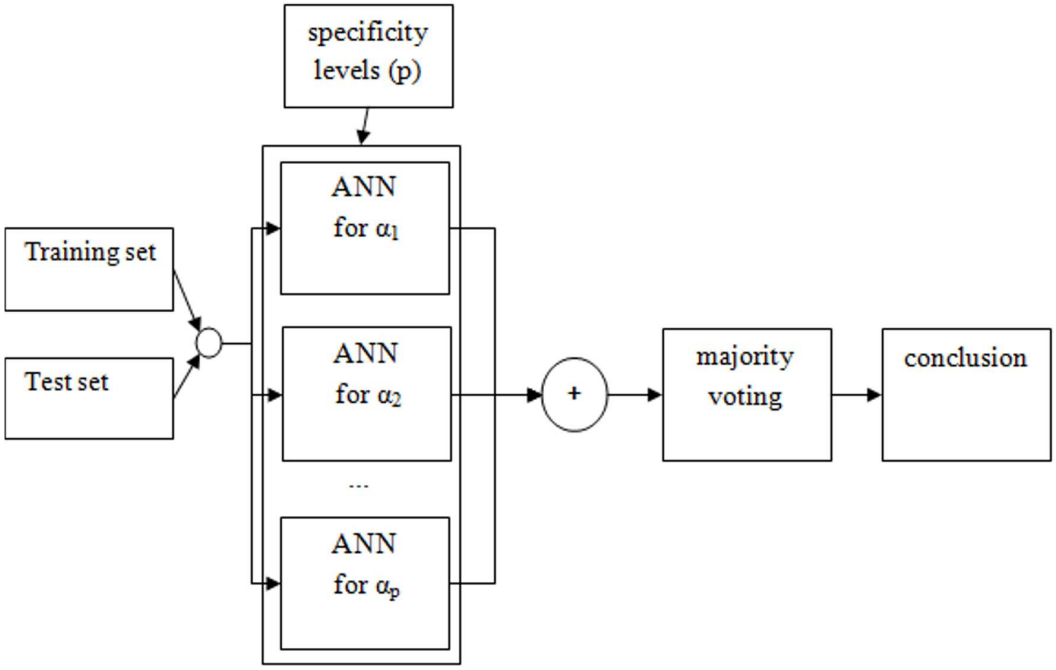

Fig. 2. The overview of proposed method.

3. OBTAINED RESULTS

The advantage of proposed method is only one parameter to tune, that allows defining a number of ANN and two constant values: maximal number of neurons in layers, and number of layers. Maximal ANN structure was limited by hardware limitations. The initial research was performed on a standard PC computer with 8 GB RAM, therefore the border values of network are $l_{max} = 4$ for number of layers and $n_{max} = 30$ for number of neurons. The calibration process was performed on training dataset. Using 10 fold cross-validation the optimal number of specificity levels was found (p parameter). As a test value the overall accuracy was used. The results obtained for various p values was illustrated in Fig. 3.

The calibration shows that the best results are obtained using relatively small granulation levels. In case of heart disease database it is p=5, while in case of breast tissue databases p=3 gives best results. Often, ties cause the decrease of overall accuracy in case of even number

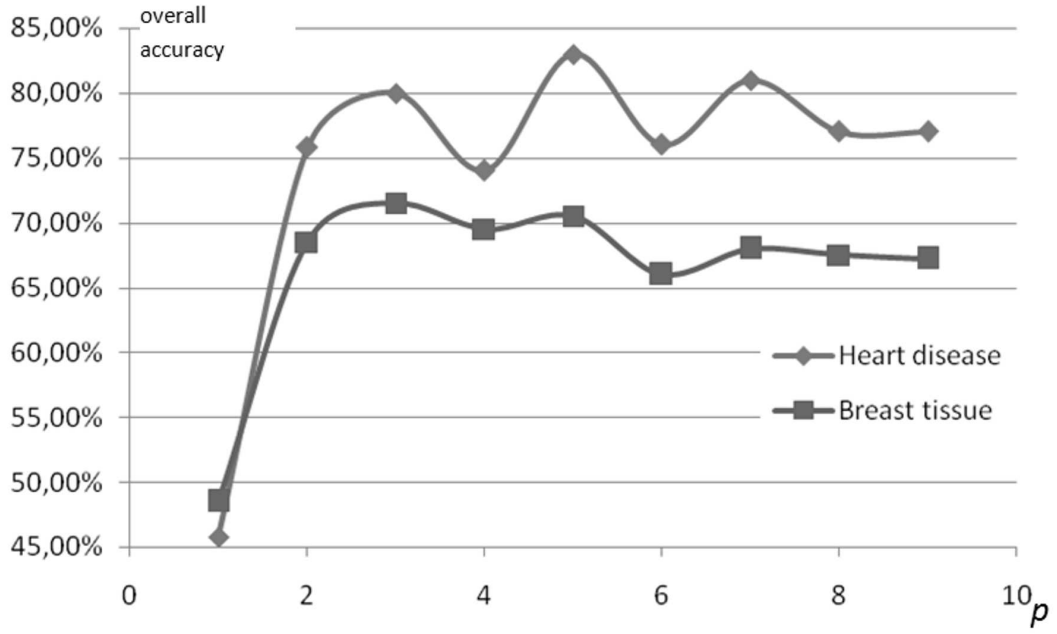

Fig. 3. The classifiers overall accuracy changes for different p values.

of classifiers. In case of the tie, the one of classification results are used as the diagnosis.

The calibrated method was verified using 10-fold cross validation, separately for two aforementioned medical datasets. The method was run twenty times to reduce the impact of random factors on obtained results. Proposed method using voting was than compared with well-known classification methods. All compared methods were tuned using training set. Table 1 presents the classifiers comparison result, where quality of classification was measured using the mean overall accuracy measured based on twenty attempts.

Table 1. The overall accuracy of various analyzed classifiers.

Database / Method	Proposed method	Fusion kNN + interval granules	Random Forest	k-NN	Bayes Network	PNN
Heart disease database	**82.6%**	81.6%	79.8%	80.1%	**83.1%**	63.8%
Breast tissue database	**74.9%**	**75.5%**	69.8%	69.1%	67%	53.8%

Then t-student test (t-test) was applied to compare the means of two classifiers results. The t-test compares if obtained results from proposed method varies significantly from other benchmark classifiers. The t-test results, using 95% confidence interval, showed that difference between proposed method and benchmark classifiers are considered to be statistically significant, with exceptions bolded in Table 1.

The obtained results encourage further research of this method. In general, the method proved to be most reliable and obtained results are closed to the results obtained by the best classifiers. In case of heart disease database the method was second to Bayes network. In case of breast tissue database, it was second to a method proposed by author in [2]. It is worth mentioning, that in case of [2] method two parameters have to be tuned. Additionally, the Bayes network gave slightly better result for heart disease dataset, however it provided one of the worst results in case of breast tissue dataset. The proposed method, despite using ANN as classification mechanism, offers a stable classification results. Stability was obtained by applying different

granulation levels, which allows to avoid overestimation of algorithm. The drawback of the method is simple voting mechanism. In future the weights for each classifier can be applied based on the calibration process.

4. CONCLUSIONS

The paper proposes a fusion of multiple levels of specificity of granules constructed using ANN. The specificity level is set by the structure of ANN. In this introductory research, the simplest multilayer perceptron network was used. The level of specificity influences the number of layers and neurons in each layer. As results the classification results offers generalized decisions as well as specific ones. As it was proved, classifiers fusion offers higher overall accuracy compared to a single classifiers. The results also shows that three and five specificity levels are enough for correct classification. The results are consistent with previous research using interval granules [2]. Obtained results are promising and proposed method can be considered to apply in specialized medical expert systems. In the future the proposed solution will be extended by improving the voting mechanisms by introducing the weighting of specific granulation level. Furthermore, the method will be tested using arbitrary and fully connected neural networks.

BIBLIOGRAPHY

[1] BERNAS M., ORCZYK T., MUSIALIK J., HARTLEB M., BOSKA-FAJFROWSKA B. Justified granulation aided noninvasive liver fibrosis classification system. Bmc Medical Informatics And Decision Making, 2015, Vol. 15(64).

[2] BERNAS M., ORCZYK T., PORWIK P. Fusion of granular computing and k-nn classifiers for medical data support system. The Series Lecture Notes In Computer Science, 2015, Vol. 9012. pp. 62–71.

[3] BERNAS M., PLACZEK B., PORWIK P., PAMULA T. Segmentation of vehicle detector data for improved k-nearest neighbours-based traffic flow prediction. Iet Intelligent Transport Systems Doi: 10.1049/Iet-Its.2013.0164, 2015.

[4] BERTHOLD M. Mixed fuzzy rule formation. International Journal Of Approximate Reasoning, 2003, Vol. 32 (2-3). pp. 67–84.

[5] BERTHOLD M., DIAMOND J. Constructive training of probabilistic neural networks. Neurocomputing, 1998, Vol. 19 (1-3). pp. 167–183.

[6] BREIMAN L. Random forests. Machine Learning, 2001, Vol. 45 (1). pp. 5–32.

[7] CAO Y., LIU S., ZHANG L., QIN J., WANG J., TANG K. Prediction of protein structural class with rough sets. Bmc Bioinformatics, 2006, Vol. 7:20.

[8] EMAM K., DANKAR F., NEISA A., JONKER E. Evaluating the risk of patient re-identification from adverse drug event reports. Bmc Medical Informatics And Decision Making, 2013, Vol. 13(114).

[9] FENG C., SUTHERLAND A., KING R., MUGGLETON S., HENERY F. Comparison of machine learning classifiers to statistics and neural networks. Proceedings Of The Third International Workshop In Artificial Intelligence And Statistics, 1993. pp. 41–52.

[10] HIROTA K. Concepts of probabilistic sets. Fuzzy Sets And Systems, 1981, Vol. 5 (1). pp. 31–46.

[11] HUANG B., ZHUANG Y., LI H. Information granulation and uncertainty measures in interval-valued intuitionist fuzzy information systems. European Journal Of Operational Research, 2013, Vol. 231. pp. 162–170.

[12] HUNTER D., YU H., PUKISH I M. S., KOLBUSZ J., WILAMOWSKI B. M. Selection of proper neural network sizes and architecturesa comparative study. Ieee Transactions OnIndustrial Informatics, 2012, Vol. 8(2). pp. 228–240.

[13] JIN T., SUN B., ZHANG Y. Granular support vector machines for medical binary classification problems. Proc. On Computational Intelligence In Bioinformatics And Computational Biology, 2004. pp. 73 – 78.

[14] JOHN G., LANGLEY P. Estimating continuous distributions in bayesian classifiers. In Proceedings Of The Eleventh Conference On Uncertainty In Artificial Intelligence, 1995. pp. 338–345.

[15] JOSSINET J. Variability of impedivity in normal and pathological breast tissue. Med. Biol. Eng. & Comput, 1996, Vol. 34. pp. 346–350.

[16] KUDLACIK P., PORWIK P. A new approach to signature recognition using the fuzzy method. Pattern Analysis And Applications, 2014, Vol. 17(3). pp. 451–463.

[17] LI B., WANG K., ZHANG D. On-line signature verification based on pca (principal component analysis) and mca (minor component analysis). In: Proc. Of First International Conference On Biometric Authentication Icba04, 2004. pp. 540–546.

[18] MAGO V., MORDEN H., FRITZ C., TIANKUANG W., NAMAZI S., GERANMAYEH P., CHATTOPADHYAY R., DABBAGHIAN V. The impact of social factors on homelessness: A fuzzy cognitive map approach. Bmc Medical Informatics And Decision Making, 2013, Vol. 13(94).

[19] PANTAZI S., AROCHA J., R M. Case-based medical informatics. Bmc Medical Informatics And Decision Making, 2004, Vol. 4(19).

[20] PEDRYCZ W. Interpretation of clusters in the framework of shadowed sets. Pattern Recognition Letters, 2005, Vol. 26 (15). pp. 2439–24493.
[21] PEDRYCZ W., GOMIDE F. Fuzzy systems engineering: Toward human-centric computing. John Wiley press, 2007.
[22] SONG M., WANG Y. Human centricity and information granularity in the agenda of theories and applications of soft computing. Applied Soft Computing Doi: 10.1016/J.Asoc.2014.04.040, 2014. pp. 610–613.
[23] ZHANG Y., ZHANG L., XU C. The property of different granule and granular methods based on quotient space. Information Granularity, Big Data, And Computational Intelligence Studies In Big Data, 2012, Vol. 8. pp. 171–190.