RBF Neural Network and Filter Methods for Feature Selection in Medical Classification Problems

Jasmina Đ. Novakovic [1] Alempije Veljovic [2] and Sinisa S. Ilic [3]

Abstract – **In this paper we evaluate classification accuracy of radial basis function (RBF) neural network and filter methods for feature selection in medical data sets. To improve diagnostic process in everyday routine and avoid misdiagnosis, machine learning methods can be employed. Diagnosis of tumor, heart diseases, hepatitis, liver and Parkinson diseases are some of the medical problems in which we used artificial neural networks. The main objective of this paper is to show that it is possible to improve the performance of the system for inductive learning rules with RBF neural network for medical classification problems, using the filter methods for feature selections. The goal of this research is also to present and compare different algorithmic approaches for constructing and evaluating systems that learn from experience to make the decisions and predictions and minimize the expected number or proportion of mistakes.**

Keywords – **Classification problems, Feature selection, Filter methods, RBF.**

I. INTRODUCTION

A lot of information is available to medical specialists, ranging from details of clinical symptoms to various types of biochemical data and outputs of imaging devices. To modernize the diagnostic process in daily routine and avoid misdiagnosis, machine learning methods can be employed. These algorithms can handle diverse types of medical data and integrate them into categorized outputs. In machine learning, a wide range of algorithms for classification is available, each with their own strengths and weaknesses. One of the most commonly used algorithms for diagnosing the disease is artificial neural networks [1, 2, 3, 4]. Coronary heart disease diagnosis [5, 6], tumor diagnosis [7], diagnosis of hypoglycemic episodes [8], prognosis of chronic myeloid leukemia [9], tuberculosis disease diagnosis [10], predict thyroid bending protein diagnosis [11], prediction approach of carcinoma patients [12], control of blood glucose [13], colorectal cancer metastases [14], detection of ovarian cancer [15] are some of the medical problems in which the used artificial neural networks.

In this paper, we used RBF neural network as artificial neural networks. RBF neural network offers a number of advantages, including requiring less formal statistical training, the ability to implicitly detect complex nonlinear relationship between the dependent and independent variables, the ability to detect all possible interactions between predictor variables and the availability of multiple algorithms for training.

The main aim of this paper was to experimentally verify the impact of filter methods as one of the feature selection technique on classification accuracy with RBF neural network. A process that chooses a minimum subset of M features from the original set of N features, so that the feature space is optimally reduced according to a certain evaluation criterion can be defined as feature selection. Finding the best feature subset is usually intractable and many problems related to feature selection have been shown to be NP-hard.

For this purpose we have organized the paper in the following way. In the second part of this paper we present a model of RBF neural network, in the third part of the paper we present description of data sets. The fourth part of the paper describes the methodology of the experiment and also we present the results of experimental study. In the fifth part of the paper, we discuss the obtained results and give directions for further research.

II. RBF NEURAL NETWORK

The classification of neural networks proved to be very good just for serious classification problems, problems where is difficult or impossible to use the classical technique. Besides, neural networks are well suited to work in conditions of noise in the data. The above listed reasons are conditioned to use neural networks.

From a structural point of view, depending on the model used to build neural networks, neural networks can be divided into static and dynamic. From the point of view of a layered mode of organization of neurons in the network, the network can be divided into single-and multi-layered. The first layer is called the input layer, the last layer output, and all other layers are called hidden layers. As a rule, each layer receives inputs from the previous layer and sends their outputs on a next layer.

The main characteristic of static neural networks is that the neurons are organized beforehand, so that neurons connected in a way with no form of feedback. These networks can not contain dynamic members, making them structurally stable. Since there are no dynamic members, static response of the neural network depends only on the current state and input values of the network parameters. Static neural networks are commonly used in the identification process, process management, and signal processing and pattern recognition. The most common types of static neural networks are MLP

[1]Jasmina Đ. Novakovic, professor of applied studies, Belgrade business school, Higher education institution for applied science, Kraljice Marije 73, 11000 Belgrade, Serbia, jnovakovic@sbb.rs.

[2]Alempije Veljovic, full professor, Faculty of technical science Cacak, University of Kragujevac, Svetog Save 65, 32000 Cacak, Serbia.

[3]Sinisa S. Ilic, associate professor, Faculty of technical science in Kosovska Mitrovica, University of Pristina, Kneza Milosa 7, 38220 Kosovska Mitrovica, Serbia.

and RBF neural networks. In this paper we are used the RBF neural network.

Dynamic neural networks are used much less frequently than static neural networks because their stability is not guaranteed. The learning process of the dynamic neural networks is much more demanding than learning static. Another shortcoming is that the output dynamic neural network depends on its initial conditions making it difficult to study because of poorly selected initial conditions may give poor results.

In the case where the application of static neural network does not give good results it is reasonable to use the dynamic neural network. The most common types of dynamic neural networks are the Hopfield, Elman and NARX dynamic neural network. Pseudo code for RBF training [16] is shown in Fig. 1.

```
trainRBF (in, out, width, MaxError, data) {
hidden = 0;
net = initRBFNetwork (in, out, hidden);
do {
        / / find the vector data that produces the highest error
        i = findMaxNetworkError (data, net); / / i = index
        vectors
                / / add a layer of RBF neurons in the same
                place where the vector data addRBFNeuron
                (net, width, data (i)); / / data (i) = midpoint
        / / find the overall network error
        NetError = trainOutputWeights (net, data);
} while (NetError> MaxError);
}
```

Figure 1. Pseudocode for RBF training

III. Description of Data Sets

Nine real medical data sets were used for tests, taken from the UCI repository of machine learning databases [17]. We used these data sets to compare results of classification accuracy with RBF network. In the following, we provide the details for the benchmark data sets we have used from UCI repository of machine learning databases.

Breast cancer (bc): The task of this data set is to predict whether or not there is recurrence of breast cancer. In this data set there is a single instance of the class 201 (no recurrence of breast cancer), and 85 instances of other classes (a recurrence of breast cancer).

Cardiography (ct): This set of data consists of attribute measurement of fetal heart rate and uterine contractions attributes on ultrasound that are classified doctors [18]. This data set contains 2126 instances and 23 attributes.

Hepatitis (he): The main aim of this data set is to predict whether hepatitis patients will die or not. In this data set, there are two classes: live (123 instances) and die (32 instances).

Liver (li): In this data set, the first five variables are all blood tests, which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each row in this data set constitutes the record of a single male individual.

Lung cancer (lc): A set of data for the cancer of the lung contains data describing the three kinds of pathological forms of lung cancer. There are 32 instances and 56 attributes.

Mammographic mass (ma): The task is to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS features and the patient's age [19].

Parkinson (pa): This data set consists of a range of biomedical voice measurements in 31 persons, 23 of them suffering from Parkinson's disease [20]. Each column in the table is a distinctive feature of a person's voice, and each row corresponds to one of the 195 recordings of person's voice. The main goal of this dataset is to separate healthy people from those people who are suffering from Parkinson's.

Pima Indians diabetes (pi): In this data set the diagnostic is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care).

Statlog Heart (sh): The task is to predict absence or presence of heart disease. This data set contains 13 features (which have been extracted from a larger set of 74).

TABLE I
REPRESENTATION OF DATA SETS

Data set	Attributes			Number of class	Size of training	Default accuracy
	all	categorical	numerical			
bc	9	9	0	2	286	70.30
ct	23	0	23	3	2126	95.00
he	19	13	6	2	155	78.10
li	6	0	6	2	345	58.10
lc	56	0	56	3	32	26.80
ma	5	0	5	2	961	84.00
pa	23	0	23	2	195	76.00
pi	8	0	8	2	768	65.10
sh	13	3	10	2	270	55.00

Table I shows the comparative characteristics of the observed data sets. There are 9 real data sets relating to diagnosis and prognosis of medical diseases. They were obtained by collecting data from real systems that exist. The default accuracy is the accuracy of always predicting the majority class on the whole dataset.

In addition, the observed total number of attributes in each set of data, as well as the number of attributes which belong to the category of categorical and numerical attributes. Three sets of data has more than 20 attributes, *lc* with 56 attributes, *ct* and *pa* with 23 attributes. Data sets with minimum attributes are *ma* with 5 attributes and *li* with 6 attributes. We conclude that the observed data sets are the sets with a very large number of attributes, as well as those sets that have a small number of attributes, which is good from the perspective of research. The observed data sets are balanced because there are sets that contain only or categorical or numerical attributes, as well as datasets containing both categorical and numerical data.

Regarding the number of classes in the analyzed data sets, 7 sets of data has two classes; only two data sets have three classes, namely *ct* and *lc*. The reason for this is the fact that in most cases in the classification problems, instances classify in two, possibly three classes, and rarely in a larger number of classes.

In Table I, shows that the number of instances provided for training varies from a small number of collected instances, which is the case with *lc* which has only 32 instances, to events that have a much greater number of instances such as *ct* which has 2126 training instance. Regarding the size of the testing set, initially in all real datasets, we had prepared a set of data for testing using 10-fold cross-validation.

IV. Results of Experimental Research

The experiment was performed using WEKA (Waikato Environment for Knowledge Analysis) tools for data preparation and research developed at the University of Waikato in New Zealand. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

When searching for the model that best approximates the target function, it is necessary to provide measures of quality models and learning. Different measures can be used depending on the problem, in our experimental studies; we used the classification accuracy as a measure of the quality of the model. In experimental research we used filter method to reduce the dimensionality of data. In all experiments is selected the solution with the number of attributes that will be used further in the study, which gives the highest classification accuracy. Our results provide the accuracy that is obtained as the average of ten repetitions each time with a 10-fold cross-validation.

In our experimental research, we used Paired *t*-test, where the level of significance was set to a value of 0.05. We use Paired *t*-test to determine whether the value obtained by different methods differs significantly. Paired *t*-test test the significance of the mean differences pairs *d* according to the following equation: $t = \frac{\bar{d}\sqrt{N}}{s_d}$, where s_d is standard deviation of the obtained differences. If the calculated value of the parameter *t* is greater than tabular (critical value), the null hypothesis is rejected and it is said that *d* is significantly different from zero, or that the difference in pairs statistically significant.

In the table of classification accuracy "+" indicates a significantly higher value for classification accuracy, while "-" indicates a significantly lower value for classification accuracy. In our experimental research, whenever we compare two or more algorithms, we give a table of classification accuracy. Comparison is such that the second algorithm is an algorithm in which was performed pre-selection attributes, and the first algorithm is a standard algorithm without pre-selection of attributes.

In this paper, we use the following filtering methods for ranking attributes that are statistically and entropy-based, and show good performance in various domains: IG, GR, SU, RF, OR and CS. All methods of filtering, IG, GR, SU, RF, OR and CS have done the ranking of attributes for each data set. Considering that the method of ranking lists all the attributes in the order as to their importance for classification problem, this method does not perform automatic reduction of the number of attributes. In order to realize the reduction of the number of attributes, there are two possibilities: (1) the use of a threshold, or (2) the use of an appropriate number of attributes for each data set and each of the filtering methods. In our experiment we use second possibility.

TABLE II
THE NUMBER OF ATTRIBUTES IN THE ORIGINAL DATASET AND NUMBER OF ATTRIBUTES SELECTED BY THE FILTER METHOD

Data set	Original data set	IG	GR	SU	RF	OR	CS
bc	9	8	3	3	2	8	3
ct	23	18	12	8	21	19	6
he	19	1	1	5	6	2	1
li	6	4	4	4	4	5	4
lc	56	5	17	2	4	9	4
ma	5	3	2	2	4	2	2
pa	23	21	21	21	13	22	21
pi	8	1	4	1	1	4	1
sh	13	3	9	3	6	3	3

Table II shows the optimal number of attributes for the purposes of classification, after searching the set of all possible solutions for each method. The table shows the original size of the set, in order to compare the effects of the reduction of the dimensionality of data. Using filter methods, the six data sets, from 9 observed, reduce the number of attributes exactly half or more than half compare with the original data set. These data sets are *bc*, *he*, *lc*, *ma*, *pi* and *sh*.

The greatest benefit of reducing the dimensionality of the data set has *lc*, with 56 attributes. These methods for *lc* data set selected a small number of attributes, even less than one-sixth, for each method, except GR method. For *he* data set, which originally has 19 attributes, all filtering methods show that the most important attribute for this studied phenomenon is 6. Filtering methods for data collection *pi*, show that up to 4 attributes are important for the classification problem.

TABLE III
CLASSIFICATION ACCURACY OF RBF ALGORITHM FOR ORIGINAL AND REDUCED DATA SET

Data set	RBF	RBF_IG	RBF_GR	RBF_SU	RBF_RF	RBF_OR	RBF_CS
bc	71.41	71.34	74.32	74.46	71.00	71.20	73.62
ct	97.93	98.35	98.41	97.65	98.13	96.90	96.27 -
he	85.29	81.31	83.45	83.05	80.49	82.69	81.25
li	65.06	58.16 -	58.16 -	58.16 -	57.33 -	60.96	58.16 -
lc	76.00	73.58	79.75	79.00	76.75	72.92	74.92
ma	77.31	77.66	79.67	79.24	77.07	77.51	79.16
pa	81.22	80.92	80.92	80.92	83.39	81.98	80.67
pi	74.04	73.84	76.28	73.84	73.84	75.32	73.84
sh	83.11	78.44	83.44	78.15	81.56	78.44	78.52

Further experimental research, the optimal number of selected attributes for each data set and filtering method, checked the accuracy of the classification algorithm using RBF network.

In none data set, we have significantly worse data for all filtering methods, which means that we can always choose the

method for a given set of data that has better results or results that approximate the original dataset (table III). Using RBF classifier, we can conclude that the GR method of filtering in most cases led to better results in the observed data sets (in six cases have better results). Only one data set (ma), with all implemented filter methods, achieves less than the default accuracy for this data set (84.0%). In all other data sets, values of classification accuracy with at least one filter method are greater than the default accuracy.

TABLE IV
THE STANDARD DEVIATION FOR THE ACCURACY OF RBF ALGORITHM WITH ORIGINAL AND REDUCED DATA SET

Data set	RBF	RBF_IG	RBF_GR	RBF_SU	RBF_RF	RBF_OR	RBF_CS
bc	7.88	7.66	6.41	6.16	6.81	8.34	6.28
ct	1.02	1.02	0.94	1.21	1.00	2.19	1.29
he	8.29	7.38	8.54	7.90	9.44	8.25	7.51
li	8.80	8.10	8.10	8.10	7.66	9.62	8.10
lc	22.91	22.91	21.10	20.70	23.31	22.17	22.52
ma	3.31	3.67	4.14	4.51	3.83	4.35	4.50
pa	7.37	7.49	7.42	7.53	7.39	7.24	7.35
pi	4.91	4.65	5.18	4.65	4.65	5.31	4.65
sh	6.50	7.28	6.44	7.25	7.29	7.13	7.28

Table IV shows the standard deviation for the classification accuracy of RBF algorithm with original and reduced data set using filter methods. From the table it can be seen that the standard deviations generally do not differ much between the standard algorithm and algorithms that use reduced data set.

V. DISCUSSION OF RESULTS AND FUTURE RESEARCH

According to the obtained results, we can conclude that the basic hypothesis was proved - it is possible to improve the system performance of inductive learning rules in the medical classification problems, using the filter method for reducing the dimensionality of the data. To prove the hypothesis, have been implemented and empirically tested filter methods for reducing the dimensionality of the data.

In further research it would be interesting to apply other techniques to solve the problem of dimensionality reduction of data, such as wrapper methods and extraction of attributes and analyze and compare the effects of their implementation. These techniques could also improve the performance of classification learning algorithms.

ACKNOWLEDGEMENT

The authors are grateful to the Ministry of Science and Technological Development of the Republic of Serbia for the support (project TR 34009).

REFERENCES

[1] F. Amato, A. López, P. Vaňhara, A. Hampl, J. Havel, "Artificial neural networks in medical diagnosis", Jornal of Applied Biomedicine 11: 47–58, 2013.

[2] E. Alkim, E. Gürbüz, E. Kiliç, "A fast and adaptive automated disease diagnosis method with an innovative neural network model", Neur Networks. 33: 88–96, 2012.

[3] D. Brougham, G. Ivanova, M. Gottschalk, D. Collins, A. Eustace, R. O'Connor, J. Havel, "Artificial neural networks for classification in metabolomic studies of whole cells using 1H nuclear magnetic resonance", J Biomed Biotechnol. 2011: 158094, 2011.

[4] M. Catalogna, E. Cohen, S. Fishman, Z. Halpern, U. Nevo, E. Ben-Jacob, "Artificial neural networks based controller for glucose monitoring during clamp test", PloS One. 7: e44587, 2012.

[5] O. Atkov, S. Gorokhova, A. Sboev, E. Generozov, E. Muraseyeva, S. Moroshkina S, N. Cherniy, "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters", J Cardiol. 59: 190–194, 2012.

[6] H. Uğuz, "A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases", J Med Syst. 36: 61–72, 2012.

[7] D. Barbosa, D. Roupar, J. Ramos, A. Tavares, C. Lima, "Automatic small bowel tumor diagnosis by using multi-scale wavelet-based analysis in wireless capsule endoscopy images", Biomed Eng Online. 11:3, 2012.

[8] K. Chan, S. Ling, T. Dillon, H. Nguyen, "Diagnosis of hypoglycemic episodes using a neural network based rule discovery system", Expert Syst Appl. 38: 9799–9808, 2011.

[9] P. Dey, A. Lamba, S. Kumari, N. Marwaha, "Application of an artificial neural network in the prognosis of chronic myeloid leukemia", Anal Quant Cytol Histol. 33: 335–339, 2012.

[10] E. Elveren, N. Yumuşak, "Tuberculosis disease diagnosis using artificial neural network trained with genetic algorithm", J Med Syst. 35: 329–332, 2011.

[11] AS. Gannous, YR. Elhaddad, "Improving an Artificial Neural Network Model to Predict Thyroid Bending Protein Diagnosis Using Preprocessing Techniques", WASET. 50: 124–128, 2011.

[12] Ho W-H, Lee K-T, Chen H-Y, Ho T-W, Chiu H-C., "Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network", PloS One. 7: e29179, 2012.

[13] BS. Leon, AY. Alanis, E. Sanchez, F. Ornelas-Tellez, E. Ruiz-Velazquez, "Inverse optimal neural control of blood glucose level for type 1 diabetes mellitus patients", J Franklin I. 349: 1851–1870, 2012.

[14] L. Spelt, B. Andersson, J. Nilsson, R. Andersson, "Prognostic models for outcome following liver resection for colorectal cancer metastases: A systematic review", Eur J Surg Oncol. 38: 16–24, 2012.

[15] A. Thakur, V. Mishra, S. Jain, "Feed forward artificial neural network: tool for early detection of ovarian cancer", Sci Pharm. 79: 493–505, 2011.

[16] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Second edition, Prentice Hall, 2003.

[17] A. Frank, A. Asuncion, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2010, [http://archive.ics.uci.edu/ml].

[18] D. Ayres de Campos, et al., "SisPorto 2.0 A Program for automated analysis of cardiotocograms", J Matern Fetal Med 5:311-318, 2000.

[19] M. Elter, R. Schulz-Wendtland, T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process", Medical Physics 34(11), pp. 4164-4172, 2007.

[20] M.A. Little, P.E. McSharry, S.J. Roberts, D.A.E. Costello, I.M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection", BioMedical Engineering OnLine, 6:23, 2007.