Implementation, Testing, and Evaluation of Center Selection Methods for Supervised Radial Basis Networks to Enhance Breast Cancer Analysis

Manan Nanavati*, Boris R. Bracio

Anhalt University of Applied Sciences, Koethen, Germany; *manan.nanavati@live.com

Simulation Notes Europe SNE 25(3-4), 2015, 133 - 139
DOI: 10.11128/sne.25.tn.10303
Received: January 158, 2015 (Selected ASIM SST 2014 Postconf. Publ.); Accepted: August 15, 2015;

Abstract. Breast cancer is a chronic disease which has been classified as a cancer type having one of the highest mortality rates. An early and accurate diagnosis of any chronic disease plays an important role and can be life-saving. Numerous research articles state that the role of computerized diagnostic tools supporting the decision making in diagnosis of chronic diseases has increased over the past decade. The presented study implements and evaluates three different artificial neural networks in form of supervised radial basis networks (RBN). The performance of the RBN's in regards to different center selection methods using different clustering algorithm are evaluated with the help of the Wisconsin breast cancer dataset (WBCD) by UCI machine learning repository.

Introduction

Cancer is a disease where the neoplasm or the tumor cells show uncontrolled death which leads to formation of a mass or lumps [1], [2]. Today clinicians face the challenge to screen almost 200 different types of cancer [3]. In regards to breast cancer the World Cancer fact-sheet [4] lists it as the second most diagnosed cancer across the globe along with the second highest mortality rate. The regain of a healthy state of patients in case of chronic illness depends on an early detection as well as on a proper treatment. The most common methods for breast cancer diagnosis are: (1) surgical biopsy and (2) fine needle aspiration cytology (FNAC).

The physician experience and analytical skills are part of the subset for accurate diagnosis of breast cancer. One of the earliest computerized implementation for cancer diagnosis dates back to 1995 using a linear programming approach [5]. During the following decade several machine learning and data mining implementation followed. The positive results of those implementations led to an increase in usage of such computerized tools and underlined the promising nature of computer assisted diagnosis for chronic disease.

Machine learning like artificial neural networks (ANN's) and support vector machine (SVM) have been proven accurate and fast enough for the disease diagnosis. ANN is a soft computing approach, which processes the input data in an adaptive way, i.e. the designed algorithm involves 'learning' from the past information. After learning, the designed ANN can be specifically used for classification of patterns or prediction or forecasting. ANN has become an accurate method for analysis of clinical data for diagnosis purposes in the linear and non-linear domain [6], [7], [8], [9].

In this research study, RBNs were used for the analysis of breast cancer. The fast learning rate and unique design in its own class makes RBN more dominant in some of the applications than conventional multilayer perceptron networks (MLP). This study involves implementation and evaluation of three different RBN's using supervised learning methods in regards to their centers selection methods:

1. Fixed selection of centers at random [10]
2. Selection of centers using the default k-means algorithm of MATLAB – proposed method
3. Selection of centers using 'customized' k-means algorithm – proposed method

The main reason behind choosing the supervised learning algorithm is the accuracy obtained in the end stage when compared to that of an unsupervised method. One of the examples representing dominance of supervised RBN over unsupervised RBN was described in [11].

The breast cancer data used for this study is the Wisconsin Breast Cancer Dataset (WBCD) from the University of California Irvin (UCI) Machine Learning Repository [12].

This paper is organized as follows: Section 1 gives the background information on radial basis networks, WBCD data and the previous work using it. Section 2 describes the methods used for this study with subsections explaining all three designs and its implementation. The results obtained are discussed in Section 3 followed by summary of this study in Section 4.

1 Background

1.1 Wisconsin breast cancer dataset:

In this research study, the Wisconsin breast cancer dataset available on [12] was used for breast cancer analysis using a RBN. The original data consists of recordings from 699 patients towards their FNAC findings which accumulates 9 different attributes on a scale of $0-10$. In each of the data 10 was classified as most abnormal value and 1 as most normal value, also the class labeled for diagnosis was assigned, '2' stands for benign and '4' for malignant breast cancer. According to WBCD original data out of 699 patients, 458 patients were classified into the benign class cancer and 241 as malignant. The FNAC recorded attributes are as follows:

- Clump thickness
- Uniformity of cell size
- Uniformity of cell shape
- Marginal adhesion
- Single epithelial cell size
- Bar nucleoli
- Bland chromatin
- Normal nucleoli
- Mitoses

In the pre-processing part of the data, the original WBCD data available on [12] contains a total of 16 instances having missing attributes value, thus they have been eliminated from the database before implementing of the RBN. The output class of '4' as malignant breast cancer was changed to '1', which states a new set of class values i.e. '1' representing malignant and '2' representing benign. The same data was also normalized from $0-1$ scale using a minmax method.

1.2 Radial Basis Networks:

A radial basis function is a function whose output value depends on its distance from a center 'c'.

The activation function of RBN is given as:

$$\varphi = f\|x - c\| \tag{1}$$

φ = activation function
x = input value
c = center of the radial neuron

Radial basis function networks (RBFN) possess a radial symmetric property in regards to their own centers and are a subset network of MLP's. RBFN have a different design and algorithm, it works on analysis of the data during a learning process and applying a 'best fit' approach during a testing phase.

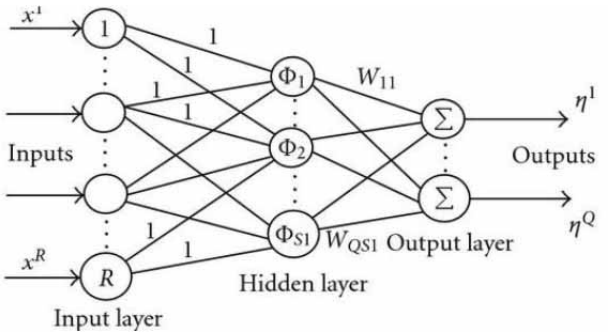

Figure 1: Basic schematic of Radial basis function network [13].

As shown in Figure 1 the construction of a basic RBFN consist of the three different layers:

1. Input Layer: The layer which acts as the source node for the input data.
2. Hidden Layer: It is layer having high number of neurons with a basis function working as the network activation function.
3. Output Layer: It acts as a sink in the network, giving an output response for the given input data.

Considering all above design constraints, the output response of the radial basis network is given as:

$$y = \sum w_{ij}.\varphi_i(x) \quad (2)$$

y = output response of network

φ = activation function

x = input value

c = center of radial neuron

w_{ij} = weight of output layer

For a Gaussian RBF the activation function is given for a spread value σ of the radial neuron as:

$$\varphi = e^{-\frac{\|x-c\|^2}{2\sigma^2}} \quad (3)$$

The most versatile characteristic of RBFN is that the mapping between input to hidden layer is nonlinear whereas it is linear between the hidden-output layer [10].

1.3 Previous works on WBCD data:

The breast cancer diagnosis on WBCD data performed at the University Of Sydney, Australia achieved a 94.74% accuracy using a C4.5 decision tree algorithm [14]. Researchers at University of Veszprem, Hungary have obtained an overall accuracy of 95.57% using a fuzzy clustering method [15] whereas researchers at Swiss Federal Institute of Technology, Switzerland have obtained accuracy 97.36% using a fuzzy genetic algorithm [16]. Researcher at the Tobb University of Economics and Technology, Turkey has demonstrated a SVM machine learning approach which yielded an overall accuracy of 99.54% [17] where on other side a 100% accuracy was obtained using a Rough subset theory by SVM at University Changchun, China [18]. When focusing on RBN methods, researcher from the Indian Institute of Management and Technology, India obtained an accuracy of 49.79% [19], on the other hand researchers at the Yildiz Technical University, Turkey achieved an accuracy of 96.18% [20].

2 Methods and Implementation

As mentioned in Section 1.2, radial basis networks are symmetric to its own center(s), and the output value depends on the distance between input value and the respective center.

Thus, the determination of the center value plays a critical role in the performance of supervised RBN's. In this research study three methods regarding the selection of centers of radial neurons were implemented and their obtained results were compared to each other.

2.1 Algorithm 1: Fixed selection of centers at random

In this supervised learning method the random selection of center(s) method described in [10] was used. This was implemented in the form of a radial basis network supervised algorithm present in MATLAB neural network toolbox. The basic steps involved in this method are summarized as follows.

1. Selection of RBN architecture
2. Initialization of network
3. Training of network
4. Validation of the network
5. Test of network
6. Presentation of results

Here the WBCD data for supervised learning were randomly divided for analysis into 70% training, 15% validation and 15% test dataset.

2.2 Algorithm 2: Selection of centers by default k-means algorithm

The second RBN implementation was also done in MATLAB, but without the neural network toolbox. First a statistical test was used for the determination of a 'Silhouette Index' to calculate the best possible 'k' value for a k-means clustering on the WBCD data. For this study '$k = 2$' was obtained, so two clusters of data will be implemented for this algorithm.

The main aim of using the k-means algorithm in this method is to determine the cluster centers for each category input. In this method the default k-means algorithm of MATLAB to determine center values is used. It was followed by the design of a supervised RBN for the WBCD analysis. The steps involved in this method are listed as follows:

1. Determination of Silhouette index
2. Perform k-means clustering to determine the centers of input value
3. Calculation of $1/2\sigma^2$ term in activation function, where σ represents the spread value of radial part
4. Calculate the output activation of radial neurons over inputs
5. Determination of output weights using pseudoinverse method
6. Evaluation of RBN response
7. Presentation of results.

2.3 Algorithm 3: Selection of centers by modified k-means algorithm

This last RBN was also implemented using MATLAB without using the neural network toolbox. The main aim of using a k-means algorithmic approach in this method was to determine the centers for the input categorical data. Furthermore, after a successful determination of a center value using a modified kmeans approach, a supervised RBN was implemented.

The difference between 'Algorithm 2' and 'Algorithm 3' is the method used to calculate the k-means. In Algorithm 2, the given algorithm of MATLAB was used, whereas in 'Algorithm 3' a customized k-means approach was designed for the determination of the centers value. The methodological step number 2 to 7 presented in Algorithm 2 in Section 2.2 are same to analyze the WBCD cancer data. The following steps represent the customized k-means algorithm for determination of centers:

1. Determination of number of unique category in target for determining value for k-means
2. Selection of Centers per categorial data
3. Selection of intial centroids
4. Perform k-means over iterative loop
5. Remove, if any empty clusters present
6. Find the closest centroid to determine membership class.

The performance measure for all three implemented supervised RBN algorithms was the overall accuracy obtained in regards to the total classification rate. The overall accuracy obtained for all three supervised algorithms/methods for centers' determination were compared.

3 Results and Discussion

3.1 Results of Algorithm 1: Fixed selection of centers at random

The used RBN algorithm in this method for selection of centers is described in [10]. The fixed centers from the categorical input data at random are selected in this method. The algorithm was implemented in MATLAB 7.10 with use of neural network toolbox. The nine different categorical inputs presented in Section 1.1 were taken as the input to the RBN model. The key factor which influences the performance of RBN model is the 'spread' value and centers of the given input data. Here in this method, we have selected centers randomly as mentioned and best spread value was determined using 'spread over loop' method. The following response was obtained when network was simulated with 'loop over spread' method.

The results of implemented model obtained highest overall accuracy rate at spread value of 0.7 represented in Figure 2.

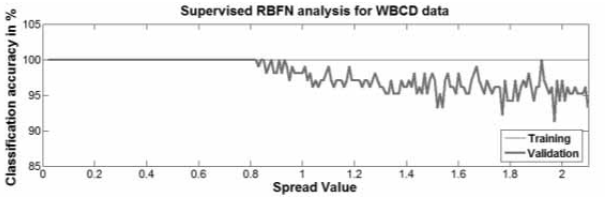

Figure 2: Classification analysis v/s spread values of RBN model.

The following tables, Table 1 and Table 2 represent the confusion matrix obtained through Algorithm 1during training and testing phase.

Training Dataset

	Actual Class	Predicted Class
	Positive (Malignant) = 195	**Negative (Benign) = 285**
Positive (Malignant) = 195	195	0
Negative (Benign) = 285	0	285

Table 1: Results of Training phase- Algorithm 1.

Test Dataset

	Actual Class	Predicted Class
	Positive (Malignant) = 195	**Negative (Benign) = 285**
Positive (Malignant) = 21	20	1
Negative (Benign) = 79	1	78

Table 2: Results of Testing phase- Algorithm 1.

The total accuracy for the analysis of results obtained is calculated as:

$$T.A = \frac{TP + TN}{TP + FP + FN + TN} * 100\% \quad (4)$$

TA = Total accuracy
TP = True Positive value
TN = True Negative value
FP = False Positive value
FN = False Negative value

The overall accuracy of 100% for the training and validation and 98% for the testing data set was obtained.

3.2 Results of Algorithm 2: Selection of centers using default k-means algorithm of MATLAB

In this second method, the RBN based diagnostic tool for breast cancer was also implemented and tested on MATLAB 7.10, but here the neural network toolbox was not used. The same set of input parameters as taken in Algorithm 1 were taken as end point and to compare and analyze these algorithms. But here no partition of data was done prior before giving input (i.e. 100% training dataset). Prior to the design and implementation of supervised RBN model for breast cancer two steps were done: (1) Determination of Silhouette index (The silhouette value for input data is a measure of similarity of points within in its own cluster, compared to the other points in the cluster [21].) (2) k-means clustering for determination of centers. The maximum silhouette index of '0.75' was obtained at '$k = 2$'. Later on the obtained 'k' value was adapted for default k-means algorithm. Then in the same way an implementation of the RBN model and activation function was determined based on the input values.

Finally the learning weights were determined using pseudoinverse method. The result of the same method were implemented and analyzed by calculation of the total accuracy obtained as mentioned in above method. The following table represents the results obtained through implementation of 'Algorithm 2'.

The overall accuracy of 'Algorithm 2' was found to be 96.77%. Here in this method, 2 centers per categorical data were used as the outcome of the default k-means algorithm.

Training Dataset

	Number of predicted value (Benign + Malignant)
Right predicted values	661
Wrong predicted values	22

Table 3: Results of Algorithm 2.

3.3 Algorithm 3: Selection of centers using 'centralized' k-means algorithm

In this last method, the RBN model for breast cancer diagnosis was also implemented on MATLAB 7.10 but here also no neural network toolbox is used. The same input dataset used in 'Algorithm 1' and 'Algorithm 2' is also used in this method, but likewise no partition of data as mentioned in Section 4.2 is used i.e. 100% training dataset. Prior before implementation of RBN model for diagnosis, two steps were done: (1) Determination of number of unique categories in target data (2) Perform 'customized' k-means for determination of centers. The only difference between both k-means approach is the involved substeps of it.

Here in this method, there is feature added to decide the 'number of centers per category'. Also, empty clusters were also removed, which may lead to higher accuracy, was also implemented in this method. In this method, the total accuracy was recorded over different values of 'number of centers per category'. The other steps like determination of activation function and determination of output weights were same as used in 'Algorithm 2'. The following Table 4 describes the accuracy analysis over different values of 'number of category'. Here in this method numbers of centers ranges from 1 to 10 were evaluated for accuracy analysis.

At the end of all implementation involved in this research study following few things can be summarized when all results are compared in regards to their overall accuracy. All the methods implemented in this study have acheived noticeable performance of 90% and higher accuracy in the end results obtained. One of our proposed method i.e. Algorithm 3 i.e. 'Selection of centers using customized k-means algorithm' has achieved accuracy of 97.07% at 3 selected centers per category. On other side no significant difference in overall accuracy was found in higher number of centers per category. Likewise on other side, by use of sophisticated neural network toolbox gave almost same level of accuracy.

Number of centers per category	Total overall accuracy	Number of centers per category	Total overall accuracy
1	96.48%	6	96.92%
2	96.92%	7	96.92%
3	**97.10%**	8	96.92%
4	96.77%	9	96.92%
5	96.92%	10	96.92%

Table 4: Accuracy outcome obtained in Algorithm 3.

	Number of predicted value (Benign + Malignant)
Right predicted values	663
Wrong predicted values	20

Table 5: Results for $number\ of\ selected\ centers = 3$ (Algorithm 3).

4 Conclusions

The main purpose of this research work was to compare performance of several centers selection methods for developing RBN models. The overall classification rate obtained at the end of this study stated that, all the implemented supervised RBN models show higher accuracy rate of 90% and above.

The noticeable performance of 97.07% was obtained in last method based on 'number of centers per category' choice. On the other side the other two methods i.e. basic supervised RBN algorithm i.e. method 1 showed total accuracy of 98% and method 2 with basic k-means showed total accuracy of 96.77%.When comparing all above results obtained, significant differences were found. This states that there is dominance of centers selection for performance of RBN model. The noticeable result obtained in this study also states: the algorithm without neural network toolbox showed almost the same results as that of the algorithm involving sophisticated toolbox.

Thus, in the end to summarize the research work we conclude that there is a critical role of centers determination in performance of RBN when accuracies of all methods are compared to each other, as it is clear from the definition that RBN are radially symmetric in regards to their own centers. In future sophisticated algorithms to determine centers can be developed for evaluation of RBN which further can be compared to standard supervised RBN algorithm of MATLAB which may produce noticeable results.

References

[1] American Cancer Society. *Cancer Glossary*. cancer.org. Retrieved September 11, 2013.

[2] National Cancer Institute. *What is cancer?*. cancer.gov. Retrieved September 11, 2013.

[3] Cancer Research UK: CancerHelp UK. *How many different types of cancer are there?*. Retrieved: 11 May 2012.

[4] International Agency for Research on Cancer and Cancer Research UK. *World Cancer Factsheet*. Cancer Research UK, London, 2014.

[5] Mangasarian OL, Wolberg WH. Cancer diagnosis via linear programming. *SIAM News*. 1990; vol. 23(5): 1 - 18.

[6] Cichocki A, Unbehauen R. *Neural Networks for optimisation and signal processing*. J. Wiley, Sons Ltd. And B.G.Teubner, Stuttgart: 1993.

[7] Ahmed A, Medhat M, Muhamed FW. Using data mining for assessing diagnosis of breast cancer. *Proc. International multiconference on computer science and information Technology*, 2010, 11-17.

[8] Burke HB et al. Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction. *Cancer*. 1997, vol.79, pp.857-862.

[9] Revett K, Gorunescu F, Gorunescu M, El-Darzi E, Marius. A Breast Cancer Diagnosis System: A CombinedApproach Using Rough Sets and Probabilistic Neural Networks. *EUROCON 2005*, 2005 Nov, Serbia & Montenegro, Belgrade, 22- 24, 1124-1127.

[10] Haykin, S. *Neural networks: acomprehensive foundation.* New York:Macmillan. 1994.

[11] Wilamowski BM, Vieira K. Clustering of patterns using radial base function networks. *In Artificial Neural Networks in Engineering ANNIE'95*. 1995; 109-115.

[12] *UCI Repository of Machine Learning Databases*. www.archive.ics.uci.edu/ml/machine-learningdatabases/breast-cancer-wisconsin/

[13] Ab Malek MN, Mohamed Ali MS. Evolutionary Tuning Method for PID Controller Parameters of a Cruise Control System Using Metamodeling. *Modelling and Simulation in Engineering*. 2009; vol. 2009: 8 pages. doi:10.1155/2009/234529.

[14] Quinlan JR. Improved use of continuous attributes in C4.5. *J. Artif. Intell. Res.* 1996; 4: pp. 77–90.

[15] Abonyi J, Szeifert F. Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Lett*. 2003; 24: 2195–2207.

[16] Pena-Reyes CA, Sipper M. A fuzzy-genetic approach to breast cancer diagnosis. *Artif. Intell. Med*. 1999;17: 131–155.

[17] Übeyli, ED. Implementing automated diagnostic systems for breast cancer detection. *Expert Systems with Applications.* 2007; 33(4): 1054-1062.

[18] Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications.* 2011; 38(7): 9014-9022.

[19] Janghel, R. R., Anupam Shukla, Ritu Tiwari, and Rahul Kala. Breast cancer diagnosis using artificial neural network models. In *3rd International Conference on Information Sciences and Interaction Sciences (ICIS)*, 2010; 89-94. IEEE.

[20] Kiyan, Tüba, and Tülay Yildirim. Breast cancer diagnosis using statistical neural networks. *IUJournal of Electrical & Electronics Engineering.* 2011; 4(2): 1149-1153.

[21] Kaufman L, P. J. Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis.* Hoboken, NJ: John Wiley & Sons, Inc.; 1990.

EUROSIM 2016

9th EUROSIM Congress on Modelling and Simulation

City of Oulu, Finland, September 12 – 16, 2016

EUROSIM Congresses are the most important modelling and simulation events in Europe. For EUROSIM 2016, we are soliciting original submissions describing novel research and developments in the following (and related) areas of interest: Continuous, discrete (event) and hybrid modelling, simulation, identification and optimization approaches. Two basic contribution motivations are expected: M&S Methods and Technologies and M&S Applications. Contributions from both technical and non-technical areas are welcome.

Congress Topics The EUROSIM 2016 Congress will include invited talks, parallel, special and poster sessions, exhibition and versatile technical and social tours. The Congress topics of interest include, but are not limited to:

- Intelligent Systems and Applications
- Hybrid and Soft Computing
- Data & Semantic Mining
- Neural Networks, Fuzzy Systems & Evolutionary Computation
- Image, Speech & Signal Processing
- Systems Intelligence and Intelligence Systems
- Autonomous Systems
- Energy and Power Systems
- Mining and Metal Industry
- Forest Industry
- Buildings and Construction
- Communication Systems
- Circuits, Sensors and Devices
- Security Modelling and Simulation
- Bioinformatics, Medicine, Pharmacy and Bioengineering
- Water and Wastewater Treatment, Sludge Management and Biogas Production
- Condition monitoring, Mechatronics and maintenance
- Automotive applications
- e-Science and e-Systems
- Industry, Business, Management, Human Factors and Social Issues
- Virtual Reality, Visualization, Computer Art and Games
- Internet Modelling, Semantic Web and Ontologies
- Computational Finance & Economics
- Simulation Methodologies and Tools
- Parallel and Distributed Architectures and Systems
- Operations Research
- Discrete Event Systems
- Manufacturing and Workflows
- Adaptive Dynamic Programming and Reinforcement Learning
- Mobile/Ad hoc wireless networks, mobicast, sensor placement, target tracking
- Control of Intelligent Systems
- Robotics, Cybernetics, Control Engineering, & Manufacturing
- Transport, Logistics, Harbour, Shipping and Marine Simulation

Congress Venue / Social Events The Congress will be held in the City of Oulu, Capital of Northern Scandinavia. The main venue and the exhibition site is the Oulu City Theatre in the city centre. Pre and Post Congress Tours include Arctic Circle, Santa Claus visits and hiking on the unique routes in Oulanka National Park.

Congress Team: The Congress is organised by SIMS - Scandinavian Simulation Society, FinSim - Finnish Simulation Forum, Finnish Society of Automation, and University of Oulu.

Esko Juuso EUROSIM President, Erik Dahlquist SIMS President, Kauko Leiviskä EUROSIM 2016 Chair

Info: eurosim2016.automaatioseura.fi, office@automaatioseura.fi