ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 300 - 305 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

Nonlinear Approximations in Sigmoid Transfer Function for Improved Statistical Pattern Recognition Based On PNN Bayesian Approach

R. Murugadoss[1], Dr.M. Ramakrishnan[2]
[1]Sathyabama University, Research Scholar
Department of Computer Science and Engineering, Chennai,
murugadossphd@gmail.com.,mdossresearch@gmail.com
[2]Professor and Head, Department of Information Technology
Velammal Engineering College, Chennai. ramkrishod@gmail.com

Abstract: The exponential function neural networks are commonly used to replace the sigmoid activation function, and thus able to calculate the non-linear discriminant construct boundary probabilistic neural network (PNN), which determine the optimal Bayesian decision boundary is close to the surface. A similar nature were also discussed with other activation function. The proposed four-layer neural network that can put any input mode is mapped to multiple categories. If you can get the new data, you can use the new data in real time to modify determine boundaries, and can run in parallel using completely artificial "neurons" to fruition. Also estimate the probability of occurrence and reliability categories, and do judgment prepared. For the problem of back-propagation time to adapt to increase the substantial part of the total calculation time, this method shows the advantage of very rapid. PNN paradigm 200,000 times faster than the back-propagation.

Key words— Neural Grid, parallel processors, "Neuron", pattern recognition, Bayesian strategy

INTRODUCTION

Neural network based on the examples used to study the mode classification. Different neural grid paradigm (paradigm) using different learning rules, but in some way, according to the statistics of a set of training samples to determine the mode, then the new models are classified according to these statistics. Common methods such as back propagation, to obtain class-based statistics using heuristics. Heuristic system parameters typically contains many small improvements gradually improve system performance. In addition to the training needs of a long calculation time, but also show the increase in the minimum adaptation reverse spread approximation is sensitive to errors. In order to improve this method, find a classification method based on statistical theory has been established in the. Can be shown that, although the final structure obtained is similar to the network back-propagation, and the main difference is that the activation function derived by statistical means to activate an alternative sigmoid function, but with the characteristics of the network are: easy to satisfy certain conditions next, in order to determine the boundary PNN Bayesian approach to achieve progressively the best judgment surface.

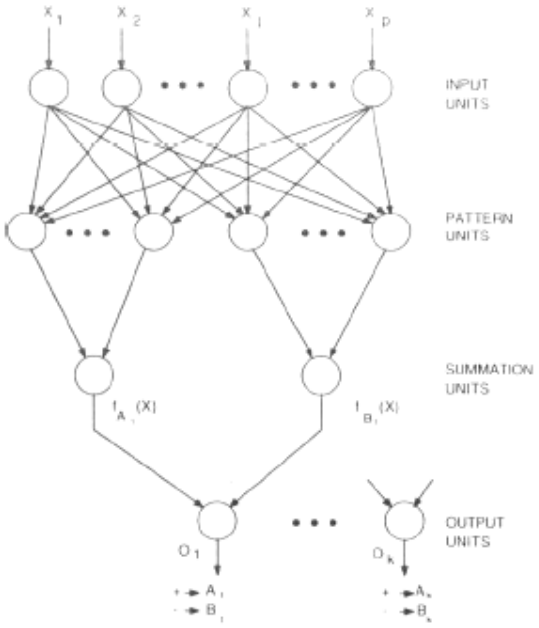

Fig 1. Pattern Classification Structure

To understand the basis for discussion PNN paradigm usually determine strategy and nonparametric Bayesian probability density function estimate from the beginning. Then you can show that the statistical method mapped to a feedforward neural network architecture, network architecture is based on a number of simple processors (neurons) on behalf of all the processors are running in parallel. Determination for pattern classification rules or policies recognized standard is: In a sense, so that the "expected risk" minimal. Such a strategy known as "Bayesian strategy," and applies to problems with many categories. Now examine the case of two types, wherein the state or the known class. If you want based on p-dimensional vector $X_T = [X_1 \ldots X_i \ldots X_p]$ described a set of measurements to determine = or =, Bayes decision rule becomes:

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 300 - 305 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

$d(X)=\theta_A$ In case $h_A l_A f_A(X) > h_B l_B f_B(X)$

$d(X)=\theta_B$ If $h_A l_A f_A(X) < h_B l_B f_B(X)$ (1)

In the formula, $f_A(X)$ and $f_B(X)$ Class A and B respectively, the probability density function l_A ; $\theta=\theta_A$ is determined as $d(X)=\theta_B$ the loss function; $\theta=\theta_B$ is determined as $d(X)=\theta_A$ the loss function (taking the correct determination of the loss is equal to 0); pattern from the prior probability for the class h_A A appears; $h_B=1-h_A$ = and $\theta=\theta_B$ for-priori probabilities.

Thus, the Bayes decision rule $d(X)=\theta_A$ regional boundaries between regions with Bayes decision rules $d(X)=\theta_B$ can be used the following formula

$$f_A(X)=Kf_B(X) \quad (2)$$

Where

$$K=h_B l_B / h_A l_A \quad (3)$$

Generally, by the formula (2) to determine the types of surfaces can be arbitrarily complex judgment, because there is no constraint on the density, those conditions are just all probability density function (PDF) must be satisfied that they are always non-negative, is plot in the whole space of points is equal to one. The same rule is applicable to many types of judgment problems. The key to using equation (2) is estimated based on the ability of PDF training mode. Typically, the a priori probability is known or can be estimated accurately, the loss of function requires subjective estimates. However, if the probability of the pattern to be divided into categories of unknown density, and is given a set of training pattern (training samples), then the only clue to provide the unknown underlying probability density that these samples. In this paper, he pointed out, as long as the basis of the density matrix is continuous, PDF estimator can asymptotically approaching categories based parent density.

LITERATURE SURVEY

Different regulatory approach has been developed to find significant intervals of continuous attributes. In D2 (Catlett) is a supervised discretization method is suitable for measuring the entropy potential division series to find cut-off values of two consecutive intervals. This is a static method, it requires the entire instance space. Instead of finding a single point cutting recursive binary ranges or sub ranges, until the stop criterion. Stopping criteria is essential in order to avoid over-segmentation. 1RD (Holt) is a discrete method binning supervised use. After successive values sorted, 1RD continuous range of values is divided into a number of disjoint intervals and adjust restrictions based on the value associated with the successive class marks. Cerquides and Mantaras have introduced the so-called Mantaras distance (Mantaras), to assess the distance measurement truncated. It also uses the minimum length principle (MDLP) Description stopping criteria to determine whether to enter multiple breakpoints. Ruiz et al propose a supervised discretization based on interval distances approach. This method is based on the separation distance, and a new concept in the vicinity. Furthermore, the proposed method can also be applied when using a variable output distribution a suitable distance class variables disorder exists. Pongaksorn et al proposed the so-called discrete class information DCR algorithm continuous attributes. Between the two types of information and the order of attributes in order to determine the optimal number and spacing of the minimum integration scheme is used. Property order is based on information gain with respect to each attribute of the class attribute. The number of discrete intervals of continuous attributes significant reduction of the best discrete algorithms, and maximize the accuracy and efficiency of the classification model. Although there are a variety of methods to establish a classification model, neural network classification as a tool, because of its resistance to high noise data and its classification model, a relationship between the properties and the ability to understand the class poorly. Feedforward neural network neural network is one of the most common types. Momentum back-propagation algorithm (Han Kamber) for training the network. The neural network has a good set of training the trainers to improve the performance speed of convergence and generalization of neural networks. Generalization level, that is, the new entries appropriately react to the capacity is very dependent on the quality of the training data. Many studies have been done to improve the versatility and reduce the convergence time. For classification problems Huyser and Horowitz has shown that a well-trained network boundary of the pattern, i.e. the pattern of the separation plane is closer than a trained over summarizes an example of the same number of

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 300 - 305 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

randomly selected network better . Distinguish between standard typical Wynn et al used the closest sample and confusing samples neighbors. Cohen et al, with a carefully selected dynamic training mode, you can get a better generalization performance. Ogawa proposed a dynamic selection strategy to reduce the convergence time, improve the generalization ability of candidates to form a pattern of neural networks training set. Many researchers (Cheung et al;. Engelbrecht and Cloete; Dasarathy, the Kelly and Davis) proposed to select a different standard of training patterns for classification. Particular the choice of method is based on the confidence metric, Euclidean distance, etc.

METHODOLOGY

Accuracy of the decision boundary is determined based on the accuracy of the estimated PDF. It discusses how to construct f (X) family of valuation

$$f_n(X) = \frac{1}{n\lambda}\sum_{i=1}^{n}\varpi\left(\frac{X - X_{Ai}}{\lambda}\right) \qquad (4)$$

PDF which all points in the continuous X on is the same. Independent random variables make X_{A1}, ... X_{Ai}, ... X_{An} for identically distributed, X because the distribution function of the random variable X f (X) = P [x ≤ X] is absolutely continuous. About weighting function $\varpi(y)$ the conditions

$$\sup_{-\infty<y<+\infty} |\varpi(y)| < \infty \qquad (5)$$

$$\int_{-\infty}^{+\infty} |\varpi(y)|\, dy < \infty \qquad (6)$$

$$\lim_{y\to\infty} |y\varpi(y)| = 0 \qquad (7)$$

$$\int_{-\infty}^{+\infty} \varpi(y)dy = 1 \qquad (8)$$

(4) In the formula, selected $\lambda = \lambda(n)$ as a function of n,

$$\lim_{n\to\infty}\lambda(n) = 0 \qquad (9)$$

$$\lim_{n\to\infty} n\lambda(n) = \infty \qquad (10)$$

$$E|f_n(X) - f(X)|^2 \to 0 \text{ with } n \to \infty \qquad (11)$$

Sense, f (X) consistent with the valuation of the mean square value. This definition is consistent, generally considered that when a large data set in accordance with the estimated expected error becomes small, which is particularly important, because it means that the real distribution can be approximated by a smooth manner. Also extends the results for multivariate case. In theorem indicate how to extend the results in this particular case to estimate multivariate nuclear core of the product is a single variable. In exceptional circumstances Gaussian kernel, multi-variable estimate can be expressed as

$$f_A(X) = \frac{1}{(2\pi)^{p/2}\sigma^p}\frac{1}{m}\sum_{i=1}^{m}\exp\left[-\frac{(X - X_{Ai})^T(X - X_{Ai})}{2\sigma^2}\right] \qquad (12)$$

Where, i = mode number,

m = total number of training patterns,

X_{Ai} = the i-th category training mode,

σ = "Smoothing parameter"

P = number of dimensions metric space.

$f_A(X)$ simply for each training center is located in a small sample of a multivariate Gaussian distribution. However, this and not limited to the Gaussian distribution. In fact, the density function can be approximated arbitrary smooth. The independent variable X for the next two-dimensional case, the effect of different smoothing parameter value pairs. Three different values in each case with the same training sample, according to equation (12) plotted density. Such that the smaller the estimated value of the density function corresponding to the parent position of training samples having different modes. Larger value, produce greater levels of the interpolation between the points. Here, the training samples near the X value, the estimated samples having approximately the same for a given probability of occurrence. The larger values, the interpolation to produce a larger level. A large value so that the estimated Gaussian density distribution, and has nothing to do with the actual underlying distribution. In "With σ→0 and as σ→∞ extreme conditions," a discussion to select the appropriate smoothing values. Decision rules can be used with a direct expression of the formula (1) of formula (12). To use these equations to perform pattern recognition tasks, has written a computer program, and on practical issues and achieved good results. However, to use the formula (12) the presence of two inherent limitations: It must be stored during testing and use of the entire training set, size, is divided into categories of the unknown point and the amount of computation required for the training set proportional. In this method, first proposed and applied to pattern recognition, these two factors severely limit the formula (12) directly for real-time or specialized applications. Approximation method must be used instead. Later, the computer memory into a compact and inexpensive enough to the training set and the memory is no longer an obstacle, but the calculation time of the computer is still connected in series is a point constraint. As a result of having a large massively parallel computing capability of neural networks, and the second restriction block (12) is about to lift directly.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 300 - 305 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

PROPOSED METHOD

In Figure 2, the distribution unit is an input unit, the same input value is provided to all the pattern unit. Each pattern unit (represented in more detail in Figure 3) to generate the input pattern vector and the weight vector W_i X scalar product $Z_i = X \bullet W_i$, and then, prior to its activation level is output to the summing means for computing a nonlinear Z_i. Instead of the back-propagation of the common S-type activation function, nonlinear operation here is $\exp\left[(Z_i - 1)/\sigma^2\right]$ X and W are assumed normalized to unit length, this is equivalent.

$$\exp\left[-\frac{(W_i - X)^T (W_i - X)}{2\sigma^2}\right] \qquad (13)$$

Form in formula (12). This scalar product is interconnected in natural finish, followed by the neuron activation function (index). Simply summation unit from the pattern input unit with accumulation of the pattern unit have the training patterns corresponding to the selected category. The determination unit is output or the two input neurons, as shown in Fig. Both single-generates a binary output. They have a single variable weights C_k,

$$C_k = -\frac{h_{B_k} l_{B_k}}{h_{A_k} l_{A_k}} \cdot \frac{n_{A_k}}{n_{B_k}} \qquad (13)$$

In the formula,

n_{A_k} = Number of training patterns from A_k class

n_{B_k} = Number of training patterns from B_k class.

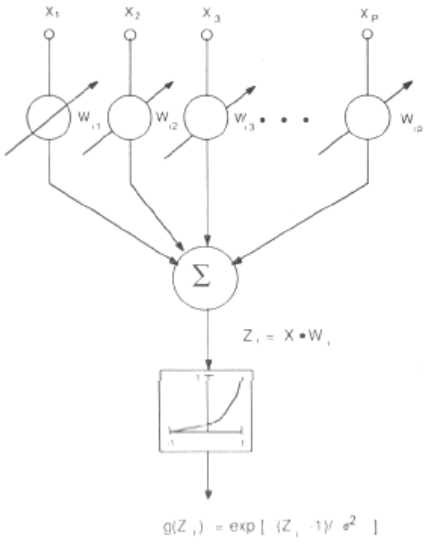

Fig 2. Mode unit

Please note, C_k priori probability ratio divided by the sample ratio and loss ratio multiplied. Any problems, and it can be in proportion to the a priori probabilities obtained from the category number of training samples A and B, which the variable weights $C_k = -l_{B_k} / l_{A_k}$. Not according to statistics of training samples, but only in accordance with the determination of significance to estimate the final ratio. If there is no particular reason for emphasis on judgment, simplified to -1 (converter). Network training method is: the weight vector of the specified pattern unit is equal to the training set within each of the X mode, and then, the output unit is connected to the appropriate mode of the summation unit. Each training mode requires a separate neurons (model unit). As shown above, the same pattern unit according to the different aggregation summation unit 2, to provide additional categories in the output vector of a binary code and additional information.

EXPERIMENTAL RESULT

Another form of (20) is simply the equation (12) the scalar product of formula estimator. No simplified form of the scalar product requires another network structure. Completely by the status quo, by calculating implement them. Had not been proved, which is a good estimate and should always be used. Because all estimators converge to correct the underlying distribution, it can be made based on simple or similar biological neural network computing model selection calculations. Among them, from the viewpoint of a simple calculation, especially attractive formula (21) (together with the formula (1) to (3)). When the measurement vector X restricted binary measurement of the formula (21) reduces to determination of the input vector and the Hamming distance between the storage amount Q, and then activated using the exponential function. Now, it is in itself make a final and very useful variation. If binary (+1 or -1) expressed in the form of input variables, all input vectors automatically have the same length, and do not have standardized. You can also use these patterns, as well as the network of Figure 2-4. In this case, in the range of + p to -p. By small changes in the activation function, you can adjust this change.

$$f_A(X) = -\frac{1}{n(2\lambda)^p} \sum_{i=1}^{n} 1 \qquad (14)$$

when,

$$| X_j - X_{Aij} | \leq \lambda \qquad (15)$$

$$f_A(X) = \frac{1}{n\lambda^p} \sum_{i=1}^{n} \prod_{j=1}^{p} \left[1 - \frac{X_j - X_{Aij}}{\lambda}\right] \qquad (16)$$

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 300 - 305 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

$$| X_j - X_{Aij} | \leq \lambda \quad (17)$$

$$f_A(X) = \frac{1}{n(2\pi)^{p/2}\lambda^p} \sum_{i=1}^{n} \prod_{j=1}^{p} e^{-1} \frac{\left(X_j - X_{Aij}\right)^2}{\lambda^2} \quad (18)$$

$$= \frac{1}{n(2\pi)^{p/2}\lambda^p} \sum_{i=1}^{n} \exp\left[\frac{-\sum_{j=1}^{p}\left(X_j - X_{Aij}\right)^2}{2\lambda^2} \right] \quad (19)$$

$$f_A(X) = \frac{1}{n(2\lambda)^p} \sum_{i=1}^{n} \prod_{j=1}^{p} e^{-|X_j - X_{Aij}|/\lambda} \quad (20)$$

$$= \frac{1}{n(2\lambda)^p} \sum_{i=1}^{n} \exp\left[-\frac{1}{\lambda} \sum_{j=1}^{p} | X_j - X_{Aij} | \right] \quad (21)$$

$$f_A(X) = \frac{1}{n(\pi\lambda)^p} \sum_{i=1}^{n} \prod_{j=1}^{p} \left[1 + \frac{\left(X_j - X_{Aij}\right)^2}{\lambda^2} \right]^{-1} \quad (22)$$

$$f_A(X) = \frac{1}{n(2\pi\lambda)^p} \sum_{i=1}^{n} \prod_{j=1}^{p} \left[\frac{\sin\frac{\left(X_j - X_{Aij}\right)}{2\lambda}}{\frac{X_j - X_{AIJ}}{2\lambda}} \right]^2 \quad (23)$$

Under the circumstances the best attributes of the network without abandon Bayesian asymptotically optimal, allowing shown to activate a function table changes shape. Even when the measured value of the input original as a continuous, preferably converting them into a binary representation suitable for massively parallel because some of the hardware technology lends itself to the Hamming distance calculation. Continuous observation can be in binary form, a coding scheme (sometimes known as "thermometer code"), where the characteristics of each of n-bit binary coded representation, i.e., a series of binary coded +1, followed by a series of -1. To +1 and represents the value of features. The upper surface of the invalid encoding has the following advantages:

1. Hamming distance can be the absolute value of the difference between the amount of the eigenvalues of the training vectors to be classified and stored training pattern characteristic values.
2. By the feature vector (n × p) bit length, a large Hamming distance calculation, the characteristic P can be operated (21) required for the entire sum.

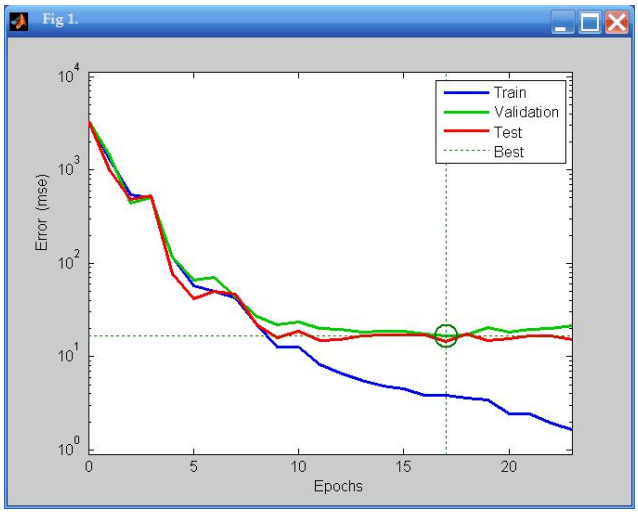

Fig 3. Neural network pattern recognition

On the run, the most important advantage of probabilistic neural network is trained easily and instantaneously. You can use it in real time, because just finished a representative of a pattern observed for each category, the network began to be extended to the new model. When viewed and stored in the additional mode to the network, will increase the universality, may become more complex decision boundary. Other advantages of PNN are: (a) by selecting the appropriate smoothing parameter values, you can make the necessary determination of complex surface shape, or simply desired; (b) determining the surface can approximate Bayesian optimization; (c) allow the error sample ; (d) insufficient sample applied to the network performance; (e) in the case without the repeated training, when n becomes larger, the smaller can; (f) for the statistics change with time, with a new body mode rewrite the old model.

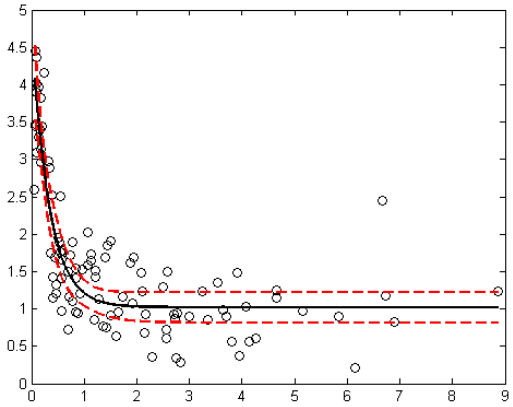

Fig 4. Nonlinear regression prediction

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 300 - 305 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

Another practical advantage of probabilistic neural network are: the majority of the network is different, it does not require input from each neuron to neuron feedback, can be completed in parallel computing. The system contains thousands of neurons (too much into a single semiconductor chip), such feedback path rapidly over the pins on the chip available to teach. However, the proposed use of the network, if only by summing unit can parallel computing chip part and any number of chips connected to the same input. Only two such portions and, each output bit a. It has been indicated that the precise form of the activation function is not critical to network performance. In the simulation of neural networks or hybrid neural network design in order to perform the activation function analog components, it is very important. Proposed here can be used to map probabilistic neural network classification, associative memory, or directly estimate the posterior probability.

DISCUSSION AND CONCLUSION

The best neural network topology can be determined by the pruning process. Classification of knowledge can easily extract the classification rules in the form of a simple, its performance is achieved by pre-treatment and/or trim improved. The focus of this paper is an important way to promote the effective use of neural network classifier building. New algorithms have been proposed to quantify and select the mode used to determine the optimal neural network topology. The paper also impact on the black-box nature of neural networks disclosure, provides an ideal classification rule extraction from neural classification table. Finally, we also studied the problem of artificial neural network, which is their poor explanatory. Nerve internal knowledge network rule extraction algorithm translated into a set of symbolic rules. . Similar properties are also discussions with other activation functions. This may be given in any of the four input patterns are assigned to the neural network of the plurality of categories. If you can get the new data, you can use the new real-time data to determine changes in boundaries, and can be used in parallel entirely artificial "neurons" can be achieved running. We also estimate the probability of occurrence and reliability categories, and prepare for trial.

REFERENCES

[1] Bashir, Z. A., & El-Hawary, M. E. (2009). Applying wavelets to short-term load forecasting using PSO-based neural networks. Power Systems, IEEE Transactions on, 24(1), 20-27.

[2] Karlik, B., & Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, *1*(4), 111-122.

[3] Yilmaz, A. S., & Özer, Z. (2009). Pitch angle control in wind turbines above the rated wind speed by multi-layer perceptron and radial basis function neural networks. *Expert Systems with Applications*, *36*(6), 9767-9775.

[4] LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K. R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade* (pp. 9-48). Springer Berlin Heidelberg.

[5] Yonaba, H., Anctil, F., & Fortin, V. (2010). Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting. *Journal of Hydrologic Engineering*, *15*(4), 275-283.

[6] Beale, M. H., Hagan, M. T., & Demuth, H. B. (2010). Neural Network Toolbox 7.*User's Guide, MathWorks*.

[7] LukošEvičIus, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*(3), 127-149.

[8] Tong, D. L., & Mintram, R. (2010). Genetic Algorithm-Neural Network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection. *International Journal of Machine Learning and Cybernetics*, *1*(1-4), 75-87.

[9] Ahmadlou, M., & Adeli, H. (2010). Enhanced probabilistic neural network with local decision circles: A robust classifier. *Integrated Computer-Aided Engineering*, *17*(3), 197-210.

[10] Cheng, J. H., Chen, H. P., & Lin, Y. M. (2010). A hybrid forecast marketing timing model based on probabilistic neural network, rough set and C4. 5. *Expert systems with Applications*, *37*(3), 1814-1820.

[11] Tripathy, M., Maheshwari, R. P., & Verma, H. K. (2010). Power transformer differential protection based on optimal probabilistic neural network. *Power Delivery, IEEE Transactions on*, *25*(1), 102-112.

[12] Karlaftis, M. G., & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights.*Transportation Research Part C: Emerging Technologies*, *19*(3), 387-399.

[13] Wang, H., & Chen, P. (2011). Intelligent diagnosis method for rolling element bearing faults using possibility theory and neural network. *Computers & Industrial Engineering*, *60*(4), 511-518.

[14] Manjunath Aradhya, V. N., Niranjan, S. K., & Hemantha Kumar, G. (2010). Probabilistic neural network based approach for handwritten character recognition. *International Journal of Computer & Communication Technology*,*1*(2, 3, 4.), 9-13.

[15] BOLAT, B., & Yildirim, T. (2011). A data selection method for probabilistic neural networks. IU-Journal of Electrical & Electronics Engineering, 4(2), 1137-1140.

[16] Simard, P. Y., LeCun, Y. A., Denker, J. S., & Victorri, B. (2012). Transformation invariance in pattern recognition–tangent distance and tangent propagation. In Neural networks: tricks of the trade (pp. 235-269). Springer Berlin Heidelberg.