SOUTHEAST EUROPE JOURNAL OF SOFT COMPUTING
Available online at www.scjournal.com.ba

Neural Networks to Diagnose the Parkinson's Disease

Mehmet Can
mcan@ius.edu.ba

International University of Sarajevo
Faculty of Engineering and Natural Sciences
Hrasnicka Cesta 15, 71000 Sarajevo
Bosnia and Herzegovina

Abstract

To identify the presence of Parkinson's disease, a neural network system with back propagation together with a majority voting scheme is presented in this paper. The data used has an imparity of the ratio 3:1. Previous research with regards to predict the presence of the disease has shown accuracy rates up to 92.9% [1] but it comes with a cost of reduced prediction accuracy of the small class. The designed neural network system is boosted by filtering, and this causes a significant increase of robustness. It is also shown that by majority voting of eleven parallel networks, recognition rates reached to > 90 in spite of 3:1 imbalanced class distribution of the Parkinson's disease data set.

Keywords—Machine learning, Parallel neural networks, boosting by filtering, Parkinson's Disease

1. INTRODUCTION

The cause of Parkinson's disease is unknown, however research has shown that a degradation of the dopaminergic neurons affect the dopamine production to decline [2]. Dopamine is used by the body to control movement, hence the less dopamine that is in circulation the more difficult the person to control the movements and may experience tremors and numbness in extremities. As a direct cause of reduced control of motor-neurons in the central nervous system, the ability of articulating vocal phonetics is reduced. In this case the symptom, the inability to articulate words, is related to the presence of Parkinson's disease and is described as Dysphonia, a reduced functionality of the vocal cords. One of the immediate effects of vocal Dysphonia is that the sound of the words is hardly recognizable [3].

Although the field of speech processing and development of speech recognition systems have received considerable attention during the last decades, scientific researches on vocal recordings of patients that suffer from Parkinson's disease are not abundant. With the availability of portable phones and analyzing methods involving traditional digital signal processing approaches such as hidden Markov models, Kalman filter, short-time frequency analysis and wavelet transforms are successfully used for both speech enhancement and speech recognition applications [4, 5, 6, 7, 8, 9, 10, 11].

Scientific research on vocal recordings of patients that suffer from Parkinson's disease are not abundant. The

data set used in this study was collected by M. A Little et. al. [12] who used support vector machine techniques to distinguish between the people who have normal vocal signals and who suffer from Parkinson's disease. They achieve a classification accuracy of 91.4% but they do not report single class true positive rates. This is noteworthy because of the highly imbalanced sick to healthy ratio (3:1) data class distribution of the Parkinson's disease data set [13].

R. Das [1] has made a comparative study on this data set making the use of the neural networks, DMNeural analysis, regression analysis, and decision trees presented results of classification accuracy of 92.9%, 84.3%, 88.6% and 84.3% respectively. The analysis was carried out on data exploration of SAS software. Another study by M. Lee et. al. [14] on the imbalanced data problem in biomedical data uses a sampling scheme in collaboration with a naive Bayes classifier to deal with the imbalanced data problem. The sampling pattern starts with a small portion of the data to train the classifier, and then successively to increase the number of training samples regardless of the initial class distribution. This method results in positive predictive rates of 66.2% for normal subjects and 90.0% for subjects with Parkinson's disease.

Neural networks are the tools that should be recalled for any classification job. They are developed enormously since the first attempts made modeling the perceptron architecture six decades ago [15].

The massive parallel computational structure of neural networks is what has contributed to its success in predictive tasks. It has been shown that the approach of using parallel networks is successful with respect to increasing the predictive accuracy of neural networks in robotics [16] and in speech recognition [17].

This work presents a parallel networks system which is bound together with a majority voting system in order to further increase the predictive accuracy of a Parkinson's Disease data set based on vocal recordings.

For the proposed system it is shown with a case study of Parkinson's disease that some of the difficulties with imbalanced data sets are resolved. The type of network used is the standard feedforward back-propagation neural network, since they have proven useful in biomedical classification tasks [18]. The performance of the trained neural networks is evaluated according to the true positive, and true negative rate of the prediction.

The paper is organized as follows; first, the data used in this work is introduced in section 2. The neural network that is boosted by filtering is illustrated in section 3. Results of the research are shown in section 4 which followed by a conclusion.

2. DATA SET OF PARKINSON'S DISEASE

The data used in this study is a voice recording originally done at University of Oxford by M.A. Little [12]. In the same study a detailed presentation is made on the specificities of the recording equipment as well as in what environment the experiment was conducted. The data consists of 195 recordings extracted from 31 people whom 23 are suffering of Parkinson's disease. The time since first diagnosis of Parkinson's disease was done 0 to 28 years ago and the age of the subjects ranged from 46 to 85 years and a total of 6 vocal sounds were recorded from each subject. For more information on the data set refer to ref. [12]. Furthermore, the data set consists of 22 attributes. Little et al. apply a correlation filter and of these 22 attributes 12 are removed after applying the filter. Each correlation coefficient, which is less than 0.95 is considered not to contribute to classification accuracy, thus the attribute is removed. A total of 11 attributes are kept after the correlation filter has been applied. Table 1 gives a brief explanation of meaning of the attributes; references [19, 12, 13] should be consulted for details on how the attributes are derived and what they indicate.

Table 1: Table describing the attributes that are not removed after applying the correlation filter or by other reasons mentioned in Little et. al [12] where the exact computations of each measurement is described.

No	**Attribute name**	**Description**
1	MDVP:Jitter(Abs)	Variation in fundamental frequency
2	Jitter:DDP	Variation in fundamental frequency
3	MDVP:APQ	Measures of variation in amplitude
4	Shimmer:DDA	Measures of variation in amplitude
5	NHR	Ratio of noise to tonal components
6	HNR	Ratio of noise to tonal components
7	status	(1)-Parkinson'sDisease, (0)-Healthy
8	RPDE	Dynamic complex measurement
9	DFA	Signal fractal scaling exponent
10	D2	Dynamic complex measurement
11	PPE	Non-linear measure of fundamental frequency

3. ARTIFICIAL NEURAL NETWORKS

Nervous systems existing in biological organism for years have been the subject of studies for mathematicians who tried to develop some models describing such systems and all their complexities. Artificial neural networks emerged as generalizations of these concepts with mathematical model of artificial neuron due to McCuloch and Pitts [20] described in 1943 definition of unsupervised learning rule by Hebb [21] in 1949, and the first ever implementation of Rosenblatt's perceptron [22] in 1958. The efficiency and applicability of artificial neural networks to computational tasks have been questioned many times, especially at the very beginning of their history the book "Perceptrons" by Minsky and Papert [23], published in 1969, caused dissipation of initial interest and enthusiasm in applications of neural networks.

It was not until 1970s and 80s, when the backpropagation algorithm for supervised learning was documented that artificial neural networks regained their status and proved beyond doubt to be sufficiently good approach to many problems. Artificial Neural Network can be looked upon as a parallel computing system comprised of some number of rather simple processing units (neurons) and their interconnections. They follow inherent organizational principles such as the ability to learn and adapt, generalization, distributed knowledge representation, and fault tolerance. Neural network specification comprises definitions of the set of neurons (not only their number but also their organization), activation states for all neurons expressed by their activation functions and offsets specifying when they fire, connections between neurons which by their weights determine the effect the output signal of a neuron has on other neurons it is connected with, and a method for gathering information by the network that is its learning or training rule.

3.1. Architecture

From architecture point of view neural networks can be divided into two categories: feed-forward and recurrent networks. In feed-forward networks the flow of data is strictly from input to output cells that can be grouped into layers but no feedback interconnections can exist. On the other hand, recurrent networks contain feedback loops and their dynamical properties are very important.

The most popularly used type of neural networks employed in pattern classification tasks is the feedforward network which is constructed from layers and possesses unidirectional weighted connections between neurons. The common examples of this category are Multilayer Perceptron or Radial Basis Function networks, and committee machines.

Multilayer perceptron type is more closely defined by establishing the number of neurons from which it is built, and this process can be divided into three parts, the two of which, finding the number of input and output units, are quite simple, whereas the third, specification of the number of hidden neurons can become crucial to accuracy of obtained classification results.

The number of input and output neurons can be actually seen as external specification of the network and these parameters are rather found by trial. For classification purposes as many distinct features are defined for objects which are analyzed that many input nodes are required. The only way to better adapt the network to the problem is in consideration of chosen data types for each of selected features. For example instead of using the absolute value of some feature for each sample it can be more advantageous to calculate its change as this relative value should be smaller than the whole range of possible values and thus variations could be more easily picked up by artificial neural network. The number of network outputs typically reflects the number of classification classes.

The third factor in specification of the multilayer perceptron is the number of hidden neurons and layers and it is essential to classification ability and accuracy. With no hidden layer the network is able to properly solve only linearly separable problems with the output neuron dividing the input space by a hyperplane. Since not many problems to be solved are within this category, usually some hidden layer is necessary.

With a single hidden layer the network can classify objects in the input space that are sometimes and not quite formally referred to as simplexes, single convex objects that can be created by partitioning out from the space by some number of hyperplanes, whereas with two hidden layers the network can classify any objects since they can always be represented as a sum or difference of some such simplexes classified by the second hidden layer.

Apart from the number of layers there is another issue of the number of neurons in these layers. When

the number of neurons is unnecessarily high the network easily learns but poorly generalizes on new data. This situation reminds auto-associative property: too many neurons keep too much information about training set rather "remembering" than "learning" its characteristics. This is not enough to ensure good generalization that is needed.

On the other hand, when there are too few hidden neurons the network may never learn the relationships amongst the input data. Since there is no precise indicator how many neurons should be used in the construction of a network, it is a common practice to build a network with some initial number of units and when it learns poorly this number is either increased or decreased as required. Obtained solutions are usually task-dependant.

3.2 Activation Functions

Activation or transfer function of a neuron is a rule that defines how it reacts to data received through its inputs that all have certain weights.

Among the most frequently used activation functions are linear or semi-linear function, a hard limiting threshold function or a smoothly limiting threshold such as a sigmoid or a hyperbolic tangent. Due to their inherent properties, whether they are linear, continuous or differentiable, different activation functions perform with different efficiency in task-specific solutions.

For classification tasks antisymmetric sigmoid tangent hyperbolic function is the most popularly used activation function:

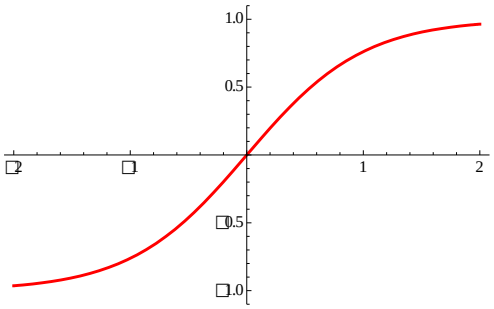

Fig. 1. Antisymmetric sigmoid tangent hyperbolic activation function

3.3 Learning Rules

In order to produce the desired set of output states whenever a set of inputs is presented to a neural network it has to be configured by setting the strengths of the interconnections and this step corresponds to the network learning procedure. Learning rules are roughly divided into three categories of supervised, unsupervised and reinforcement learning methods.

The term supervised indicates an external teacher who provides information about the desired answer for each input sample. Thus in case of supervised learning the training data is specified in forms of pairs of input values and expected outputs. By comparing the expected outcomes with the ones actually obtained from the network the error function is calculated and its minimization leads to modification of connection weights in such a way as to obtain the output values closest to expected for each training sample and to the whole training set.

In unsupervised learning no answer is specified as expected of the neural network and it is left somewhat to itself to discover such self-organization which yields the same values at an output neuron for new samples as there are for the nearest sample of the training set.

Reinforcement learning relies on constant interaction between the network and its environment. The network has no indication what is expected of it but it can induce it by discovering which actions bring the highest reward even if this reward is not immediate but delayed. Basing on these rewards it performs such re-organization that is most advantageous in the long run [22].

The modification of weights associated with network interconnections can be performed either after each of the training samples or after finished iteration of the whole training set.

The important factor in this algorithm is the learning rate η whose value when too high can cause oscillations around the local minima of the error function and when too low results in slow convergence. This locality is considered the drawback of the backpropagation method but its universality is the advantage.

3.4 Architecture of artificial neural networks, Committee Machines

As the base topology of artificial neural network committee machines [25] with the feed-forward multilayer perceptron with sigmoid activation function trained by backpropagation algorithm is used.

In committee machines approach, a complex computational task is solved by dividing it into a number of computationally simple tasks and then combining the solutions to those tasks. In supervised learning, computational simplicity is achieved by distributing the learning task among a number of experts, which in turn divides the input space into a set of subspaces. The combination of experts is said to constitute a committee machine. Basically, it fuses knowledge acquired by experts to arrive at an overall decision that is supposedly superior to that attainable by anyone of them acting alone. The idea of a committee machine may be traced back to Nilsson [24] (1965); the network structure considered therein consisted of a layer of elementary perceptrons followed by a vote-taking perceptron in the second layer.

Committee machines are universal approximators. They may be classified into two major categories:

1. *Static structures.* In this class of committee machines, the responses of several predictors (experts) are combined by means of a mechanism that does not involve the input signal, hence the designation "static." This category includes the following methods:

• Ensemble averaging, where the outputs of different predictors are linearly combined to produce an overall output.

• Boosting, where a weak learning algorithm is converted into one that achieves arbitrarily high accuracy.

2. *Dynamic structures.* In this second class of committee machines, the input signal is directly involved in actuating the mechanism that integrates the outputs of the individual experts into an overall output, hence the designation "dynamic."

In this research ensemble averaging category of committee machines will be used.

Ensemble averaging

Figure 1 shows a number of differently trained neural networks (i.e., experts), which share a common input and whose individual outputs are somehow combined to produce an overall output *y*. In this research the outputs of the experts are scalar-valued. Such a technique is referred to as an ensemble averaging method. The motivation for its use is two-fold:

• lf the combination of experts in Fig. 2 were replaced by a single neural network, we would have a network with a correspondingly large number of adjustable parameters. The training time for such a large network is likely to be longer than for the case of a set of experts trained in parallel.

• The risk of overfitting the data increases when the number of adjustable parameters is large compared to cardinality (i.e., size of the set) of the training data.

In any event, in using a committee machine as depicted in Fig. 2, the expectation is that the differently trained experts converge to different local minima on the error surface, and overall performance is improved by combining the outputs in some way.

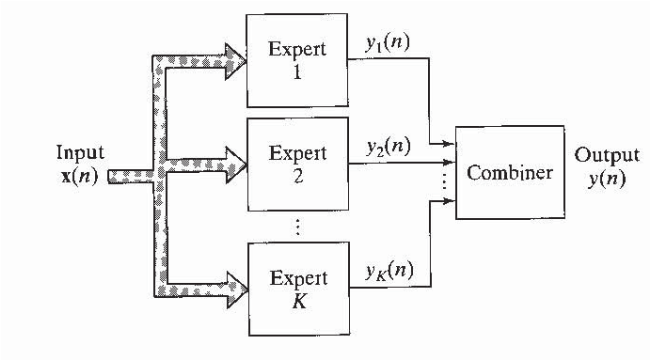

Fig. 2. Block diagram of a committee machine based on ensemble-averaging.

The number of input terminals equaled the number of attributes in the human voice data, thus it is eleven. There are two hidden layers with eleven neurons within each of eleven neural networks in the committee machine for preserving generalization properties but achieving as shown in the signal flow graph in Figure 3.

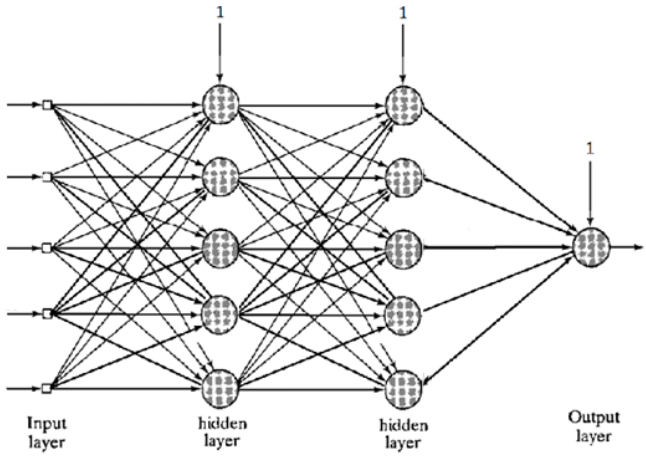

Fig. 3. Signal flow graph of an expert neural network

In this research two hidden layer, feed forward, back propagation artificial neural networks are used as the eleven committee machines. The Parkinson data was 11 dimensional. Therefore eleven input ports equaled the number of eleven attributes used, thus it is eleven. There are two hidden layers with eleven neurons within each of eleven neural networks in the committee machines for preserving generalization properties but achieving convergence during training with tolerance at most 0.14 for all training samples recognized properly.

For all structures of artificial neural networks, only one output is produced. Actually, it was possible to use a single output and by interpretation of its active state as one class and inactive output state the second class the task would have been solved as well.

4. RESULTS AND DISCUSSION

To demonstrate the increased robustness of the system and to justify forward propagation of untrained data samples, experiments are conducted. Results from the experiments can be seen in table 1.

Table 1: Performance measurements of eleven committee machines and majority vote. The high false positive is a result of imbalanced data ratio 1:3 of healthy subjects to the ones with Parkinson's disease.

	Committee Machines											
%	1	2	3	4	5	6	7	8	9	10	11	MV
TP	80	93	91	91	91	93	93	92	83	92	91	92
TN	50	73	84	84	84	83	71	73	75	73	73	73
FP	50	27	16	16	16	17	29	27	25	27	27	27
FN	20	7	9	9	9	7	7	8	17	8	9	8

It has been shown in this study that parallel neural networks in combination with a majority rule based system increase performance of true recognition rates in an imbalanced data set. In conducted experiments all measurement parameters are improved compared to single network predictions. From the experiments it is proven the parallel system with forward propagation of untrained data samples increases the robustness and decrease the variability as seen in the system which does not have this feature.

Despite the advantages of having an accurate system prediction, the training time and complexity of the parallel network algorithm do increase as the number of parallel networks increases [26-27]. The data set is very unbalanced with regard to the class distribution. This, in combination with the small sample size, makes it difficult to train any type of classifier to predict the presence of Parkinson's disease. Out of 195 samples, 75.4% are Parkinson's disease type and the remainder is of healthy character.

It implies that the baseline prediction is 75.4% and any prediction accuracy less than the baseline is not relevant. A common problem with imbalanced data sets is that they can increase to high false positive rates. Traditionally, the problem with false positive predictions is dealt with over- or undersampling [28]. However techniques to adjust the sample distribution sometimes overweight the benefits of generalising the classifier. Any modification to the data set is merely artificial alternatives to the problem of inadequate training data. In this paper, it has been demonstrated that parallel neural networks are strong at adjusting the imbalanced data set problem.

False positive rates up to 25 - 30% of the positive class have been reported [29] in the literature. It has been demonstrated in this study that a true positive rate up to 90% of positive class is achieved by using eleven parallel networks. This is a significant improvement compared to previously demonstrated results. It has also evident that networks with forward propagation of untrained data do increase the robustness of the parallel system. For the case of forward propagation of untrained data, this threshold is after 7 networks.

5. CONCLUSIONS

A new system has been presented consisting of parallel distributed neural networks and a majority voting system. An empirical investigation demonstrates that it is possible to achieve >90% true positive rate for each class in a Parkinson's Disease data set with class distribution of 3:1 ratio.

REFERENCES

1. R. Das, A comparison of multiple classification methods for diagnosis of Parkinson disease,
Expert Systems with Applications 37 (2) (2010) 1568 – 1572.

2. D.M. R, J. J. L, *Harrison's Principles of Interna Medicine*, 17th Edition, The McGraw-Hill Company, Inc,
3. http://www.accessmedicine.com/content.aspx?aID=2905868, 2009, Ch. 366. Parkinson's Disease and Other Extrapyramidal Movement Disorders.
4. A. H. Ropper, M. A. Samuels, *Adams and Victor' Principles of Neurology*, 9th Edition, The McGraw-Hill Companies, Inc,
5. http://www.accessmedicine.com/content.aspx?aID =3633872, 2009, Ch. 23. Disorders of Speech and Language.
6. L. A. da Silva, M. B. Joaquim, Noise reduction in biomedical speech signal processing based on time and frequency Kalman filtering combined with spectral subtraction, *Computers and Electrical Engineering* 34 (2008) 154 – 164.
7. Q. Yan, S. Vaseghi, E. Zavarehei, B. Milner, J. Darch, P. White, I. Andrianakis, Kalman tracking of linear predictor and harmonic noise models for noisy speech enhancement, *Computer Speech and Language* 22 (1) (2008) 69 – 83.
8. J. J. Sroka, L. D. Braida, Human and machine consonant recognition, *Speech Communication* 45 (4) (2005) 401 – 423.
9. M. D. Skowronski, J. G. Harris, Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments, *Speech Communication* 48 (5) (2006) 549 – 558.
10. A. Esposito, M. Marinaro, *Nonlinear Speech Modeling and Applications*, Vol. 3445 of Lecture Notes in Computer Science, Springer Berlin /Heidelberg, 2005, Ch. Some Notes on Nonlinearities of Speech, pp. 1–14.
11. A. Hussain, M. Chetouani, S. Squartini, A. Bastari, F. Piazza, *Progress in Nonlinear Speech Processing*, Vol. 4391, Springer Berlin / Heidelberg, 2007, Ch. Nonlinear Speech Enhancement: An Overview, pp. 217–248.
12. J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, D. Klakow, *Machine Learning for Multimodal Interaction*, Vol. 4892/2008, Springer-Verlag Berlin Heidenberg, 2007, Ch. To Separete Speech, A System For Recognizing Simultaneous Speech, pp. 283 – 294.
13. M. J. F. Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. thesis, Gonville and Caius College (September 1995).
14. M. A. Little, P. E. McSharry, E. J. Hunter, L. O. Ramig, Suitability of dysphonia measurements for telemonitoring of parkinson's disease, *IEEE Transactions on Biomedical Engineering* 56 (4) (2009) 1015–1022.
15. Freddie Åström, Rasit Koker, "A parallel neural network approcah to prediction of Parkinson´s Disease", Expert systems with applications, 38(10): 12470-12474, 2011.
16. M. S. Lee, J.-K. Rhee, B.-H. Kim, B.-T. Zhang, *Aesnb: Active example selection with naive Bayes classifier or learning from imbalanced biomedical data*, 2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering (2009) 15–21.
17. M. L. Minsky, and S. A. Papert, (1988) *Perceptrons*, Expanded Edition. Cambridge, MA: MIT Press. Original edition, 1969.
18. R. Koker, Reliability-based approach to the inverse kinematics solution of robots using Elman's networks, *Engineering Applications of Artificial Intelligence* (18) (2008) 685 – 693.
19. A. J. Lee, Parallel neural networks for speech recognition, *Neural Networks*,1997., International Conference on 4 (9-12) (1997) 2093–2097.
20. M. A. Mazurowskia, P. A. Habasa, J. M. Zuradaa, J. Y. Lob, J. A. Bakerb, G. D. Tourassib, *Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, Advances in Neural Networks Research*: IJCNN '07, 2007 International Joint Conference on Neural Networks IJCNN '07 21 (2-3) 2008, 427–436.
21. M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, I.M. Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, *BioMedical Engineering OnLine* 2007 6 (23).
22. W. S. Mcculloch, and W.Pill's. (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics*, 5:115-133. Reprinted in Anderson& Rosenfeld [1988], pp. 18-28.

23. B. O. Hebb, (1949). *The Organization of Behavior*. New York: John Wiley & Sons. Introduction and Chapter 4 reprinted in Anderson & Rosenfeld 1988, pp. 45-56.
24. Rosenblatt, E, 1958.The Perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Review*, vol. 65, pp. 386-408.
25. H. Tang, K. C. Tan, and Z. Yi, *Neural Networks: Computational Models and Applications*, Springer-Verlag Berlin Heidelberg 2007.
26. N. J. Nilsson, *Learning Machines: Foundations of Trainable Pattern-Classifying Systems*, New York: Macgraw-Hill, 1965.
27. S. Haykin, *Neural Networks A Comprehensive Foundation*, Second Edition, Prentice-Hall, Inc., Simon & Schuster,A Viacom Company Upper Saddle River, New Jersey 07458, 1999.
33. Introduction and Chapter 4 reprinted in Anderson & Rosenfeld, 1988, pp. 45-56.
28. R. E. Schapire, R.E, The strength of weak learnability, *Machine Learning*, vol. 5, pp.197-227, 1990.
29. R. E. Schapire, 1997. Using output codes to boost multiclass learning problems, Machine Learning: *Proceedings of the Fourteenth International Conference*, Nashville, TN.
30. H. Drucker, C. Cortes, L.D. Jackel, and Y. LeCun, Boosting and other ensemble methods. Neural Computation, vol. 6, 1994, pp.1289-1301
31. H. Drucker, R.E. Schapire, and P. Simard, Improving performance in neural networks using a boosting algorithm, *Advances in Neural Information Processing Systems*, vol. 5, 1993, pp. 42--49, Cambridge, MA: MIT Press.
32. B. O. Hebb, *The Organization of Behavior*. New York: John Wiley & Sons, 1949.