Data Mining Techniques for the Detection Nonlinear Mapping of Data Sets Using Neural Networks

G. Sankara Subramaniam[1], Dr. Ashish Chaturvedi[2]
Research Scholar, Department of Computer Science and Engineering,
Himalayan University, Arunachal Pradesh, India,
Professor and Associate Director, Arni School of Computer Science and Applications,
Arni University, Indora (Kathgarh), Himachal Pradesh, India.

Abstract— This paper deals with methods for finding optimal neural network architectures learn particular problems. A genetic algorithm is used to find the appropriate architectures, this evolutionary algorithm uses direct encoding and error using trained as performance measure to guide the network evolution. The training of the network takes place by retro-propagation algorithm or back-propagation; techniques such as repetition training, early stopping and regularization of complexity to improve the performance of the evolutionary process. The evaluation criterion is based on the skills of both learning and generalization of specific domain architectures generated. The classification results are also compared with architectures found by other methods. Genetic algorithm is an optimization technique based on Darwin evolution theory. The basic principles and some further modifications implemented to improve its performance are presented, as well as a historical development. A numerical example of a function optimization is also shown to demonstrate how the algorithm works in an optimization process. Finally several chemistry applications realized until now is commented to serve as parameter to future applications in this field.

Index Terms— data mining; neural networks; symbolic rules; weight freezing; machine learning

I. INTRODUCTION

Artificial intelligence has been defined as how to design processes that exhibit characteristics commonly associated with intelligent human behavior. His address modeling approaches, based on different architectures, different own human thought processes such as decision making, reasoning and learning. One of the architectures that have emerged to emulate the behavior of the neural network is learning, modeled on the human brain. Artificial neural networks provide an attractive paradigm for the design and analysis of intelligent adaptive systems for a wide range of applications in artificial intelligence for many reasons including: flexibility for adaptation and learning (by modifying the computational structures used) robustness in the presence of noise (errors or omissions), ability to generalize, resilience to failures, potential for massive parallel computing, and likeness (albeit superficial) with biological neural networks. The performance (and cost) of a neural network on problems is critically dependent on, among other things, the choice of processing elements (neurons), the network architecture and the learning algorithm used. For example, many of the learning algorithms used in neural networks essentially looking for a proper setting of the adjustable

parameters (also called weights) within a specified network topology a priori, under the guidance of input samples (training examples) from the environment of the task. Clearly, for this approach to be successful, the desired parameter setting must be done in the space where they are looking (which in turn is constrained by the choice of network topology) and the search algorithm used must be able to find. Dynamic programming is a mathematical optimization technique widely used in a variety of disciplines, however, suffers from several problems among which the dimensionality. The genetic algorithm meanwhile, is a more general optimization technique that can not only solve the kinds of problems that are usually left to dynamic programming, but also is able to remain stable even in situations where the search spaces are very big. This paper shows how to transform a dynamic programming problem into one that can be solved by a genetic algorithm.

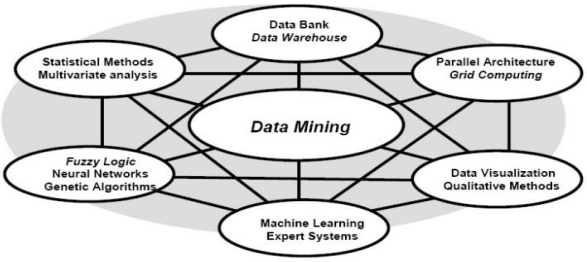

Fig 1 Data Mining Process

Even when you can find a suitable adjustment of parameters using this approach, the ability of the resulting network to generalize on unseen data during learning, or the cost of using the network (measured by its size, power consumption, hardware implementation, etc.) may be far from optimal. These factors make it difficult to process design of neural networks. Additionally, the lack of design principles is a major obstacle in developing systems for large-scale neural networks for a wide variety of practical problems. Therefore of great interest are techniques to automate the design of neural architectures for particular classes of problems under different design constraints and performance.

II. DATA MINING BACKPROPAGATION TRAINING ALGORITHM

The number of input nodes equals the dimensionality of input patterns and the required problem output desired the number of nodes in the output desired the number of nodes in the output layer. The number of nodes in the hidden layer depends on the problem complexity. Each hidden node and output node applies a sigmoid function to its net input. Its formula it is continuous, monotonically increasing, and invertible and everywhere differentiable function. Training set contains set of input output patterns that are used to train the network. Testing set contains collection of input output patterns that are used to access network performance. Learning rate is used to set the rate of weight adjustments in the time of network training.

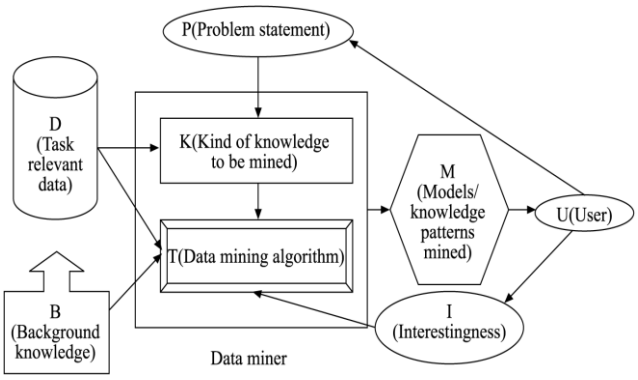

Fig 2 Sequential Pattern Data Mining And Visualization

The back propagation algorithm trains a given feed forward multiplayer neural network for a given set of input pattern with known classifications. When each entry of the sample set is presented to the network, the network examines its output response to the sample input compared to the known and desired output and the error value is calculated. Based on the error, the connection weights are adjusted. The back propagation algorithm is based on delta learning rule in which the weight adjustment is done through mean square error of the output response to the sample input. The set of these sample patterns are repeatedly presented to the network until the error value is minimized.

III. METHODOLOGY

In the most general case, a genotype can be thought of as an array (a string) of genes, where each gene takes values from a suitably defined (also known as alleles) domain. Each encodes a phenotype or genotype candidate solution in the domain of interest - in this case a kind of neural architecture. Such encoding may employ genes that take numeric values to represent a few parameters or complex structures which become symbols phenotypes (in this case, neural networks) via the appropriate decoding process. This process can be very simple or very complex. The resulting neural networks (phenotypes) They can also be equipped with learning algorithms that train using the environmental stimulus or simply be evaluated on the given task (assuming the weight of the network are also determined by the mechanism for encoding/decoding). This assessment determines a phenotype corresponding genotype fitness. The evolutionary process operates on a population of such genotypes, preferably selecting genotypes that encode high fitness phenotypes, and reproducing. The genetic operators such as mutation, crosses, etc., are used to introduce variety within the population and test variants of candidate solutions represented at the current population. Thus, over several generations, the population will evolve gradually genotypes to phenotypes that correspond to high fitness.

Table 1 Confusion matrix representing network performance

		Predicted	
		Positive	***Negative***
Actual	***Positive***	**5**	**5**
	Negative	**1**	**9**

Table shows the classification performance matrix of the created neural network. The created artificial neural network system has classified 14 cases correctly out of 20 given input patient cases. It has given 70% correct classification and 30% miss classification result.

The topology of a neural network, ie the number of nodes and the location and number of connections between them, has a significant impact on network performance and ability to generalize. The density of connections in a neural network determines its ability to store information. If a network does not have enough connections between nodes, the training algorithm may never converge; the neural network is not able to approximate the function. On the other hand, in a densely connected network, it may be over fitting (over fitting). The over fitting is a problem of too many parameters where statistical models are presented. This is a bad situation because instead of learning to approximate the function present in the data, the network can simply memorize each training example. The noise in the training data is then learned as part of the function, often destroying the ability of the network to generalize.

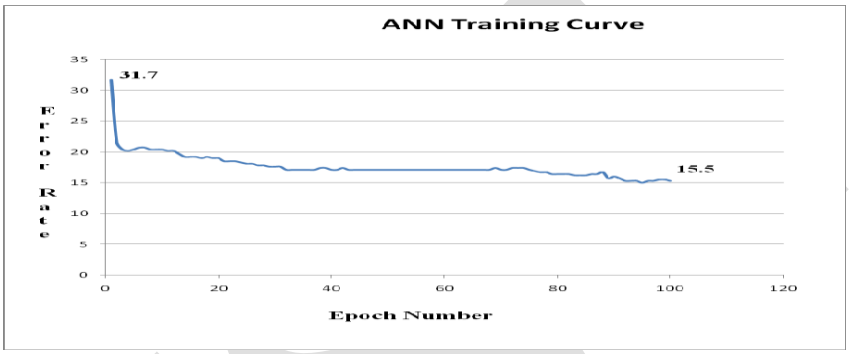

Fig 3 Figure ANN Training Curve for Test Case

IV. RESULT

Case based reasoning has been applied to solve different types of tasks like decision assistance and diagnosis, electronic commerce, customer support, medicine, and tutoring and help systems. Decision support, or assistance, helps a user by providing useful information or advice without necessarily providing a complete solution [13]. CASE Line is used by British Airways to diagnose and repair Boeing 747- 00 jets. Cognitive Systems implemented it using the CBR software development tool- ReMind. CASELine uses nearest neighbor to look up repair and maintenance information, using a two-digit fault number, error message, or a variable-length fault description. CASELine retrieves solution descriptions that have a high probability of fitting the current problem situation. CASELine is important in assisting airplane technicians in finding the quickest and best possible solution during the brief time the airplane is on the ground between flights.

CASEY combined case-based and rule-based reasoning to aid in the diagnosis of heart failure. CASEY attempted to determine the differences between similar cases and use the applicable diagnosis. If the differences between the two similar cases were very significant, it tried to adapt the case and explain it. CBR is being applied in medical image research. One such system uses two-layer CBR architecture for interpreting computed tomography images. The first CBR layer performs lower level segment identification, and the second CBR layer performs an overall interpretation of the image. SCINA is another CBR-based image understanding system. It automatically performs a diagnosis of coronary artery disease from a stored case base of image data.

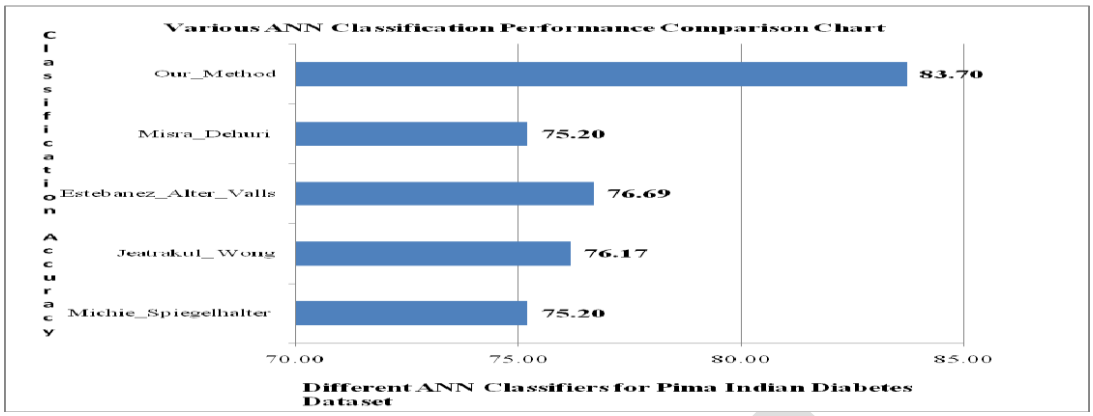

Figure 4 Various ANN Classifiers Performance Comparisons

A Top Management Fraud Diagnostic Tool (TMFDT) was developed by the multinational accounting firm of Deloitte and Touch. TMFDT is a CBR tool that assists auditors in determining the likelihood of top management fraud happening within a company. TMFDT was implemented using Cognitive Systems' Remind CBR tool. Several experienced auditors were questioned to help identify the characteristics of top management, which could indicate fraud is occurring. They identified about 160 case features, which were important in assessing potential fraud situations. After these features were identified, the case base was populated with several hundred real world cases. Deloitte and Touch used precision and noise to evaluate TMFDT, using both nearest-neighbor indexing and inductive indexing. Precision is defined as the number of correct fraud cases retrieved per total number of cases retrieved. Noise is defined as the number of incorrect fraud cases retrieved per total number of cases retrieved.

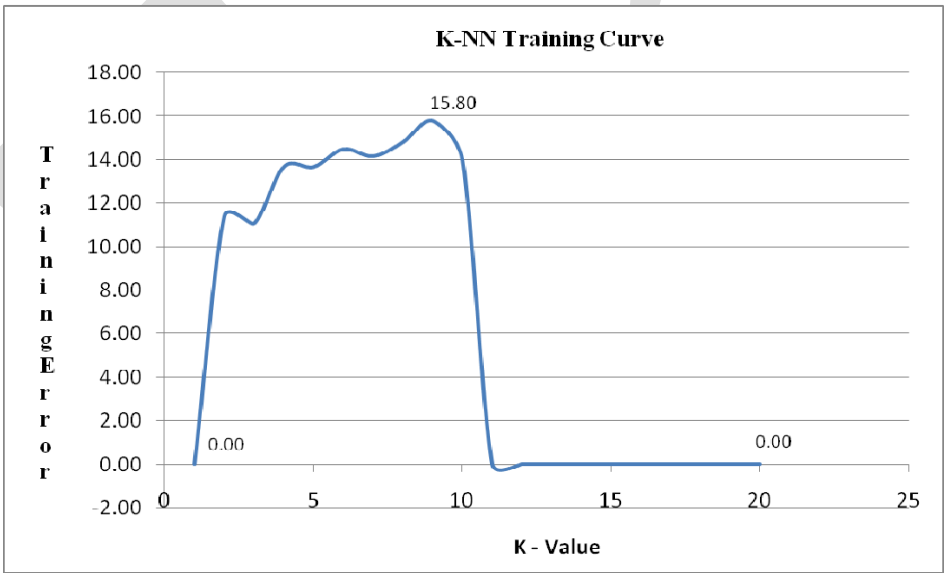

Figure 5 K-NN Training Curve

V. DISCUSSION AND CONCLUSION

We used only 3 different machine learning methods like Artificial Neural Network, Case Based Reasoning and Classification Tree methods. Lot of other machine learning methods are available in the market like Support Vector Machine and Rough Set

Theories. If we apply the same hybrid technique SVM, Rough Set theories may give some better classification performance and other required features in the medical data mining.. The Pima Indian Diabetes dataset we used in the research have only the numeric parameters. Some times medicine dataset may contain different data formats like text, images (x-ray, EGC report) and dates and time. Artificial neural network accepts only numeric data formats. The other machine learning algorithms can be used for handling the text and image data types. Redundancy, Outliers affects the Classifier's Classification performance. We want to find different ways to remove the outliers, duplicate data from the Pima Indian Diabetes Dataset. It may help to increase the classification accuracies of the machine learning methods like ANN, KNN and CT methods.

The algorithm presented in this thesis uses genetic algorithms for prediction as well as reduce the number of features. Hence it improves the performance of the classifiers. For data mining. It does this by performing feature selection to replace the original attributes of the dataset with a novel set of features. The initial hypothesis, that the algorithm could be used to sufficiently improve the accuracy of a simple classifier to make it competitive with a more complex classifier such as SVM, Regression Tree and Random forest while making explicit the feature construction implicitly performed by the more complex classifier. The performance of the Random Forest is considerably greater because it generates a model with more number of trees. It is clear that the classifiers SVM, Regression Tree and the Random Forest has been improved for the reduced features.

REFERENCES

[1] Alavi, A. H., & Gandomi, A. H. (2011). A robust data mining approach for formulation of geotechnical engineering systems. Engineering Computations,28(3), 242-274.

[2] Helbing, D., & Balietti, S. (2011). From social data mining to forecasting socio-economic crises. The European Physical Journal Special Topics, 195(1), 3-68.

[3] Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

[4] Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. In Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on(pp. 173-177). IEEE.

[5] Suh, S. (2012). Practical applications of data mining. Jones & Bartlett Publishers.

[6] Kusiak, A., & Verma, A. (2012). A data-mining approach to monitoring wind turbines. Sustainable Energy, IEEE Transactions on, 3(1), 150-157.

[7] Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems, 52(2), 464-473.

[8] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229.

[9] Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical Pharmacology & Therapeutics, 91(6), 1010-1021.

[10] Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. arXiv preprint arXiv:1202.4815.

[11] Hüllermeier, E. (2011). Fuzzy sets in machine learning and data mining.Applied Soft Computing, 11(2), 1493-1505.

[12] De Bie, T. (2011, August). An information theoretic framework for data mining. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 564-572). ACM.

[13] Köksal, G., Batmaz, İ., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. Expert systems with Applications, 38(10), 13448-13467.

[14] Sufi, F., & Khalil, I. (2011). Diagnosis of cardiovascular abnormalities from compressed ECG: A data mining-based approach. Information Technology in Biomedicine, IEEE Transactions on, 15(1), 33-39.

[15] Amatriain, X., Jaimes, A., Oliver, N., & Pujol, J. M. (2011). Data mining methods for recommender systems. In Recommender Systems Handbook (pp. 39-71). Springer US.

[16] Vaclavik, L., Lacina, O., Hajslova, J., & Zweigenbaum, J. (2011). The use of high performance liquid chromatography–quadrupole time-of-flight mass spectrometry coupled to advanced data mining and chemometric tools for discrimination and classification of red wines according to their variety.Analytica chimica acta, 685(1), 45-51.

[17] Yeh, J. Y., Wu, T. H., & Tsao, C. W. (2011). Using data mining techniques to predict hospitalization of hemodialysis patients. Decision Support Systems,50(2), 439-448.