A COMPARISON STUDY OF TWO KNOWLEDGE ACQUISITION TECHNIQUES APPLIED TO THYROID MEDICAL DIAGNOSIS DOMAIN

Abdulhamed Mohamed Abdulkafi and Aiman Subhi Gannous
Faculty of Information Technology, Garyounis University, Benghazi, Libya

Keywords: ID3, C4.5, Backpropagation.

Abstract: This study compares the performance of two famous methods used in knowledge acquisition and machine learning; the C4.5 (Quinlan 1986) algorithm for building the decision tree and the Backpropagation algorithm for training Multi layer feed forward neural network. This comparison will be based on the task of classifying thyroid diagnosis dataset. Both methods will be applied on the same data set and then study and discuss the results obtained from the experiments.

1 INTRODUCTION

Knowledge acquisition from databases has long been recognized as an essential feature of artificial intelligence. When the data formatted, one or more algorithms must be applied to extract patterns, regularities or general laws from the data. It's a challenge with extracting knowledge when the expert is not consciously aware of the knowledge used. Artificial Intelligence (AI) has been working on problems of Knowledge Acquisition and has made a great contribution to the understanding and application of neural nets and decision trees. Techniques include relatively recent algorithms like neural nets and decision trees increased computer power to work on the huge volumes of available data. (Levia, 2002).

1.1 Research Problem

The process of choosing method for acquiring knowledge from a set of data is still vague and need more experiments to do. The comparison process can not be done in general but in specific domain because the differences of the data types, and according to (Touretzky, 1989, 1990), "we still lack an understanding of the situations for which method is appropriate"(Thomas, Hild, and Ghulum, 1995). This lack is the motivation for conducting the comparison of C4.5 with Backpropagation in the field of Thyroid medical diagnosis and to discover the differences between these two algorithms that may or may not be enough to decide which is better to chose.

1.2 Objectives

Compare the test results in knowledge acquisition of the medical diagnosis applied in Thyroid data set and shows the advantages and disadvantages of each method when used in this domain.

1.3 Related Work

Scientists paid a great attention to Artificial Intelligence field and its applications, one of the most application is the knowledge acquisition from a large set of data bases; many of them had done researches about comparing and studying the differences between various methods used in this application.

(Thomas, Hermann and Ghulum, 1995) conduct a comparative study of ID3Algorithm with Backpropagation in the task of mapping English text to phonemes and stresses their experimental comparison shows that Backpropagation consistently out-performs ID3 on this task by several percentage points.

(Berkman, Lubomir, Ping, Chuan, Wei, 2005) compare the performance of prediction accuracy of a learning classifier system based on Wilson's XCS with Decision trees, Artificial Neural Networks and support Vector Machines. The experiments are

Mohamed Abdulkafi A. and Subhi Gannous A. (2008).
A COMPARISON STUDY OF TWO KNOWLEDGE ACQUISITION TECHNIQUES APPLIED TO THYROID MEDICAL DIAGNOSIS DOMAIN.
In *Proceedings of the Third International Conference on Software and Data Technologies - PL/DPS/KE*, pages 360-365
DOI: 10.5220/0001880503600365

performed on the forest Cover type database. They find that C5 Decision trees perform significantly better than other techniques.

(Bagnall, Cawley, 2000) conducted a Comparison of decision tree classifier with neural network and linear discriminant analysis(LDA) classifiers for computer-aided diagnosis and studied the comparison of performance of decision tree (DT) classifiers with artificial neural network (ANN) and linear discriminant analysis classifiers under different conditions for the class distributions, feature space dimensionality, and training sample size using a Monte Carlo simulation study. Three types of feature space distributions were studied: the Gaussian feature space, a mixture of Gaussians, and a mixture of uniform distributions. The results indicated that, in the Gaussian feature space, the (LDA) outperformed the other two classifiers. The authors conclude that a (DT) can be a viable alternative to (ANN) and (LDA) classifiers in certain feature spaces.

2 THYROID DATASET

Thyroid disease records supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Sydney, Australia. 1987. Experiments are performed on the Thyroid datasets that available at the university of California archive (UCI KDD archive, www.uci.edu, 2007) for researchers to study and carry on such experiments in the artificial intelligence field.

The data contains 2800 cases with 29 Attributes, 7 attributes are continues and 22 are binary.

The Class attribute(Outputs) of the dataset:
Negative(class 1),
Increased binding protein(class 2),
and decreased binding protein(class 3).

3 PRACTICAL WORK

3.1 Applying C4.5 Algorithm on Thyroid Dataset

C4.5 is a program for inducing classification rules in the form of decision trees from a set of given examples. The program expects to find at least two files: a names file defining class, attribute and attribute value names, and a data file containing a set of objects, each of which is described by its values of each of the attributes and its class. All trees generated in the process are saved. After each tree is generated, it is *pruned* in an attempt to simplify it. All trees produced, both pre- and post-simplification, are evaluated on the training data.

3.2 Experiment on Decision Tree

When we apply the algorithm to Construct Decision Tree for predicting the bending protein in the Thyroid the results were as follows:

Table 1: Evaluation on training data (2800 items).

Before Pruning				After Pruning			
S	C	MC	E	S	C	MC	E
127	2774	26	0.9%	51	2762	38	1.4%
S-Size E- Error. C- Classified. M C.- MisClassified							

Table 2: Evaluation on Test data (972 items).

Before Pruning				After Pruning			
S	P	MP	E	S	P	MP	E
127	941	31	2.3%	51	951	21	2.2%
S-Size E- Error. P- Predicted. M P.- MisPredicted.							

As shows in figure (1) the percentage in prediction is not equal in all classes. class1 has a better ratio of error than class2 and class3 because the allowance in the number of records in each classes, class1 had(2667) records, class2 had (124) and class3 had (9) records. Figure(5.1) shows the relation in predicting and diagnosed cases.

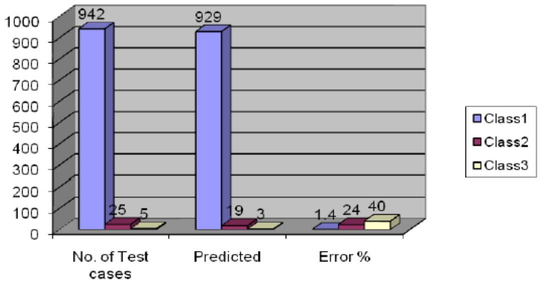

Figure 1: Shows the prediction percentage with 972 Test cases.

As noticed from Table (3) C4.5 Decision tree algorithm gives a better results in prediction phase on unseen data set when pruning is applied on the constructed tree that decrease the error ratio from 3.2% to 2.2% , also the algorithm was fast and easy to apply.

3.3 Experiment on Artificial Neural Network

3.3.1 Preparing the Thyroid Dataset

The data must be processed before applying the (ANN) approach because the data set used contains missing values and from the observations we noticed some modifications we can do on the data set to perform better in the training phase.

Reduction of Attributes. The attribute (TBG measured) has only one value for all records that it can not be useful for learning because it will not affect or help the (ANN) for better classification so we can delete the attribute from the data set and also we can delete the attribute (TBG) because it correlated to the value of (TBG measured) attribute that if the (TBG measured) value is (f) then it leads that the patient did not do this test so the values of the (TBG) attribute will considered as missing value and not useful because all values of (TBG measured) attribute have one single (f) value.

Coding Attributes. The (referral source) attribute had six categories (WEST, STMW, SVHC, SVI, SVHD, other) and these attributes are coded as follows (1, 2, 3, 4, 5, 0). The class attribute of the bending protein dataset are coded for each value as follows: **(**negative 1, Increased binding protein 2, Decreased binding protein 3).

3.3.2 Designing the Network

The network was designed according to feed forward structure for the experiments as the figure below:

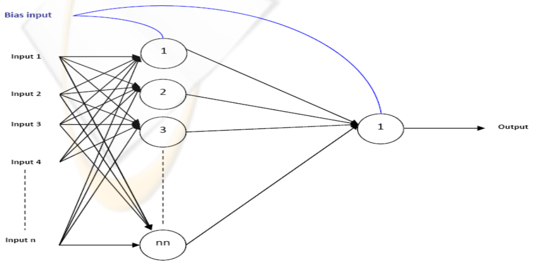

Figure 2: Feed forward neural network design.

3.3.3 Train a Feed forward Neural Network for Predicting the Bending Protein in the Thyroid

The network was designed according to a multi layer feed forward structure, it contains a two layers with linear function in the hidden layer and nonlinear (sigmoid) function in the output layer with considering the first layer as an input layer, the training is done using Back propagation algorithm programmed with the scientific programming language 'Mat lab'.

The first experiment used all records of the data set considering that any missing data are replaced by a value of -1. After many training tests, learning rate was determined by 0.00001 as a best value to train the network and number of iterations determined by 20000 iterations. The best results were at 20 neurons with 12.5% of error in classification accuracy and 12.2% of error in prediction accuracy when applied on unseen data set.

In the second experiment, the outputs were changed to binary output to simplify the training process and it shows an improvement in the results at 10 neurons that gives 5.4% of error in classification accuracy and 3.7% of error in prediction accuracy, but in the third experiment all missing data are deleted and this effects the results in a good way that at 10 neurons the classification accuracy was 3.8% of error and 3.1% of error in prediction accuracy.

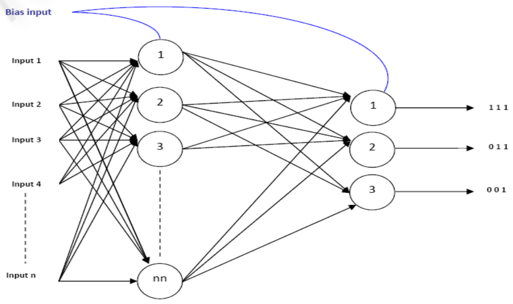

Figure 3: Designed network for training with binary outputs.

Table 3: Training after removing all missing data.

No. of Experiment	Number of Neurons	Learning rate	Allowed Error	Reached Error	Number of iterations in thousands	M(ssec) in the training set	Miss classified in the training set		M(ssec) in the test set	Miss classified in the test set	
								error %			error %
1	10	0.00001	0.001	0.0017	20	0.0018	106	3.8	0.0020	30	3.1
2	20	0.00001	0.001	0.0017	20	0.0018	106	3.8	0.0020	30	3.1
3	30	0.00001	0.001	0.0017	20	0.0018	106	3.8	0.0020	30	3.1
4	40	0.00001	0.001	0.0017	20	0.0018	106	3.8	0.0020	30	3.1
5	50	0.00001	0.001	0.0017	20	0.0018	106	3.8	0.0020	30	3.1
6	60	0.00001	0.001	0.0017	20	0.0018	106	3.8	0.0020	30	3.1
7	70	0.00001	0.001	0.0017	20	0.0018	106	3.8	0.0020	30	3.1
8	80	0.00001	0.001	0.0017	20	0.0018	106	3.8	0.0020	30	3.1

Also experiments are carried out on the data set after ranging the data set between 0 and 1, and that did not shows any improvement in the results, also using High Order Inputs with data set attributes did not show any enhancement on the classification or prediction accuracy.

In the forth experiment the inputs changed by calculating the mean for these inputs and train the network to learn this process using the mean of inputs as an outputs for the network, the results were very poor at this side, but when using the mean of inputs(output) as an inputs for the network to classify the bending protein cases the results in classification were perfect with 0% of error but a 100% of error in prediction accuracy, which led that the network was overfitting the training set.

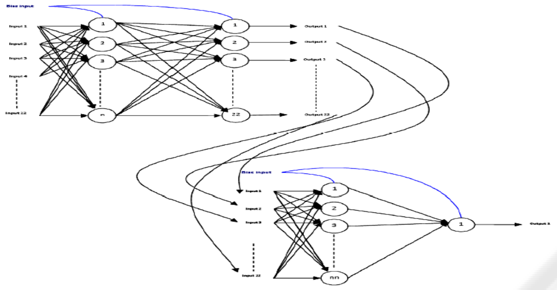

Figure 4: The structure of training network using mean of input data set.

The last step of the experimental work was an explanation of the weakness in learning process in Artificial Neural Network, from the observations the network was good in learning how to predict class (negative) in thyroid binding protein with a less value of error rather than the other classes, because class1 contains a plenty of examples were enough to learn this case well.

4 COMPARISONS

4.1 Classification Accuracy

The classification accuracy is tested by both models constructed, the tree that constructed using C4.5 algorithm have a better results as shown in table (8.1) C4.5 have a 0.9% of error before pruning the decision tree and 1.4% of error after pruning that beat ANN 3.8% of error in classifying the dataset results.

Table 4: Shows the classification percentage of both methods.

Training dataset			
	ANN	C4.5 before Pruning	C4.5 after pruning
Misclassified	106/1947	26/2800	38/2800
Error%	3.8	0.9	1.4

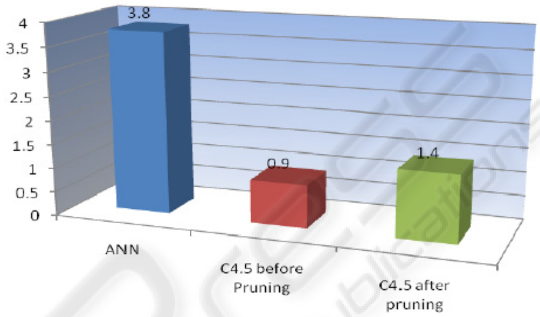

Figure 5: Shows the classification percentage of both methods.

However, the affecting of the number of examples in constructing the classification model using ANN method should not be ignored. The most correct classification was in class1, and also very accurate because it have a rich and plenty of examples that made the ANN model classify it very well, but in classifying class2 and class3 there were not as much as necessary examples to the ANN model to make a good classification. In general C4.5 decision tree model was better and more accurate in classification results, but in comparing the classification accuracy of each class individually, ANN model is better in classifying class 1, but C4.5 model is better in classifying class 2 and class3 that ANN model misclassified 106 patterns of 1947 and all of them in class2 with 99 patterns, and in class3 with 7 patterns, but C4.5 model misclassified 26 of 2800 patterns and obviously it has a correct classification in both classes 2 and 3.

4.2 Prediction Accuracy

After conducting the ability of prediction test on both methods under study using unseen dataset contains 972 records, ANN have 3.1% of error and C4.5 have 3.2% of error before pruning, both results is almost equal with a little bit difference for the ANN model, but when testing the C4.5 decision tree model after pruning process it gives a 2.2% of error that overcome the ANN test results as shown in table (5) and figure (6).

Table 5: Shows the prediction percentage of both methods.

Test dataset			
	ANN	C4.5 before Pruning	C4.5 after pruning
Mispredicted	30	31	21
Error%	3.1	3.2	2.2

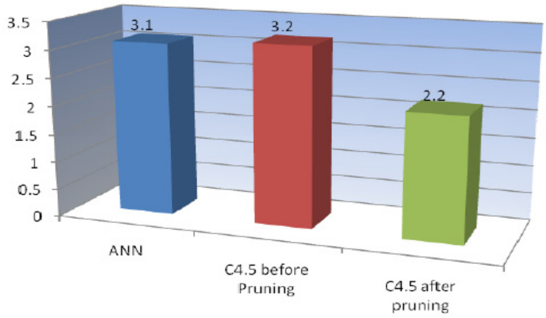

Figure 6: Shows the prediction percentage of both methods.

It is worth saying that the pruning process which available in C4.5 algorithm has a clear effect on the model constructed to predict or diagnose the patients cases. It was expected that C4.5 overcome the ANN model according to its overcome in constructed the basic classification tree model for the fact that there is a shortage in the number of examples.

In general C4.5 is overcome on ANN, but in the prediction of class 1 ANN is overcome because class1 is available in a plenty of numbers of examples. ANN model is better in predicting class 1, but C4.5 model is better in predicting class 2 and class3 that ANN model miss predicted 30 cases of 972 and all of them in class2 with 25 cases and in class3 with 5 cases, but C4.5 model miss predicted 21 of 972 cases and obviously it has a 22 correct prediction of 30 cases in both classes 2 and 3.

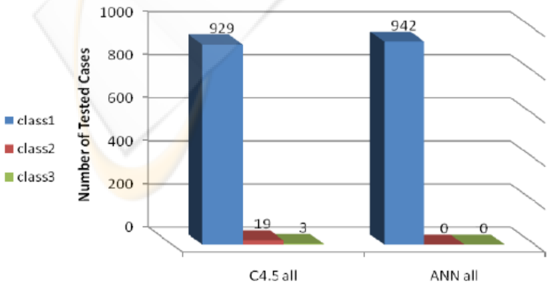

Figure 7: Shows differences in ability to predict each class.

4.3 Speed of Attainment

From the experiments conducted using both methods, C4.5 was less time consuming than ANN for the reason that C4.5 counts on selecting the best attribute to divide the dataset using Entropy Gain theory, but ANN counts on training the network using all the examples and reach the smallest error as could as possible by altering the weights and that what makes it very slow when compared with C4.5. In fact speed of implementation is not important in this type of application as the accuracy of prediction that if the model constructed gives us a very accurate diagnosing with a very small error ratio it does not care too much about which one is more speedy than the other.

4.4 Dealing with Missing Data

C4.5 algorithm has the ability to use the dataset records in constructing the decision tree directly even when it contains missing data. The missing values is filled by (?) symbol so the algorithm can recognized it as a missing value and process it using a certain function, but ANN method needs to preprocess the data separately and remove all the missing values before using the data set in training that makes C4.5 overcome on ANN in dealing with missing data.

4.5 Knowledge Expanding

Expanding the knowledge base in both methods can be done by restart applying the methods from the beginning if we get more records or examples or changing the dataset with another good quality one. However in this criteria of comparison if the speed and simplicity of implementation are concerned, then C4.5 algorithm will overcome on the Back propagation algorithm.

5 CONCLUSIONS AND FUTURE WORK

The prediction accuracy of C4.5 (DT) was better than Back propagation neural network but also not accurate. In fact it has a large value of error, the reason is the shortage of examples in class2 and class3 in the dataset. The missing data effects the process of training the neural network that forced to remove it completely which led to reduce the number of training examples, while C4.5 can handle the missing data problem using processing function included in the algorithm. The pruning process has a great effect on the decision tree that improved the results of prediction after applying it on the tree by reducing the error of prediction from 3.2% to 2.2%.

The learning rate 0.00001 leads to the best training results in ANN and 20000 iterations was enough to run the training that the error is decreased very slowly after 10000 iterations. Changing the output to binary offered improvement on the classification and prediction process using ANN that reduced the error of prediction the state of bending protein from 12.2% to 3.7%. Removing the missing data reduced the number of training examples from 2800 example to 1947 examples made the network faster in training phase and improved the error of prediction from 3.7% to 3.1%. Finding the mean of data set for each class as applied in section 6.3.3 was not helpful in any way but also have a poor results in prediction, also rearranging the input values between 0 and 1 shows no improvement on the results or the speed of learning. Increasing the number of neurons more than 20 neurons effects the results of prediction in a bad way that decrease the speed of learning and increase the error of prediction.

In this study it was obvious that the dataset quality is not good, it have a lot of missing values and there are a shortage of examples in class 2 and class 3, but the comparison in this data circumstances shows that C4.5 decision tree is more reliable method to use but In future it is worth trying to run these experiments again using another data sets in many aspects of medical diagnosis domain that have better quality examples with less numbers of missing values and rich examples in each class attributes and then run the comparison again to see which method will overcome in the field of medical diagnoses domain.

REFERENCES

Leiva, H., 2002. A multi-relational decision tree learning algorithm ,Msc, Iowa State University Ames.

Thomas, D., Hild, H., and Ghulum Bakiri, G., 1995. A Comparison of ID3 and Backpropagation for English Text-to-Speech Mapping Machine Learning, Kluwer Academic Publishers, Boston.

Berkman, S., Lubomir, H., Ping, C., Chuan, Z., Wei Jun, W,. 2005. Comparison of decision tree classifiers with neural network and linear discriminant analysis classifiers for computer-aided diagnosis: a Monte Carlo simulation study, Medical Imaging: Image Processing, Volume 5747.

Bagnall, A., Cawley, G., 2000. Learning Classifier Systems for Data Mining: A Comparison of XCS with Other Classifiers for the Forest Cover Data Set, University of East Anglia, England.

The UCI KDD archive. Irvine, University of California, Department of Information and Computer Science, http://kdd.ics.uci.edu. Last access September 2007.

Mitra, S., Tinkuacharya ., 2003, Data mining multimedia, soft computing, and Bioinformatics, John Wiley & Sons, Inc.

Ye, N., 2003. Hand book of data mining, Arizona State University, Lawrence Erlbaum Associates, Inc, New Jersey.

Kantardzic, M., 2003. Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons.Inc.

Larose, D., 2005, Discovering Knowledge in data, an introduction to data mining, John Wiley & Sons.Inc.

Michael, A., Berry, S., 2004. Data mining Techniques, John Wiley & Sons.Inc, 2nd edition.

Paplinski, A., 2004. Basic structures and properties of Artificial Neural Networks.retrivedfrom: lsc.fie.umich.mx/~juan/Materias/ANN/Papers/basic structures-and-properties.pdf, last access June 2006.

Zurada, J., 1992. Introduction to Artificial Neural Systems, West publishing Company, Singapore.