COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES ON PIMA INDIAN DIABETES DATA

Farhana Afroz and Rashedur M. Rahman
Department of Electrical Engineering and Computer Science, North South University, Bashundhara, Dhaka, Bangladesh

Keywords: Classification, Neural network, Decision tree, Rule based classifier, Fuzzy lattice, Fuzzy inference system, ANFIS.

Abstract: The development of data-mining applications such as classification and clustering has been applied to large scale data. In this research, we present comparative study of different classification techniques using three data mining tools named WEKA, TANAGRA and MATLAB. The aim of this paper is to analyze the performance of different classification techniques for a set of large data. The algorithm or classifiers tested are Multilayer Perceptron, Bayes Network, J48graft (c4.5), Fuzzy Lattice Reasoning (FLR), NaiveBayes, JRip (RIPPER), Fuzzy Inference System (FIS), Adaptive Neuro-Fuzzy Inference Systems(ANFIS). A fundamental review on the selected technique is presented for introduction purposes. The diabetes data with a total instance of 768 and 9 attributes (8 for input and 1 for output) will be used to test and justify the differences between the classification methods or algorithms. Subsequently, the classification technique that has the potential to significantly improve the common or conventional methods will be suggested for use in large scale data, bioinformatics or other general applications.

1 INTRODUCTION

The aim of this study is to investigate the performance of different classification methods using WEKA, TANAGRA and MATLAB for PIMA Indian Diabetes Dataset (PIDD). A major problem in bioinformatics analysis or medical science is in attaining the correct diagnosis for certain important information. For the ultimate diagnosis, a large number of tests generally involve the clustering or classification of large scale data. All of these test procedures are said to be necessary in order to reach the final diagnosis. On the other hand, huge amount of tests could complicate the main diagnosis process and lead to the difficulty in obtaining the end results, particularly in the case where many tests are performed. This kind of difficulty could be resolved with the aid of machine learning. It could be used to obtain the end result with the aid of several artificial intelligent algorithms which perform the role as classifiers. Machine learning covers such a broad range of processes that it is difficult to define precisely. A dictionary definition includes phrases such as to gain knowledge or understanding of or skill by studying the instruction or experience and modification of a behavioural tendency by experienced zoologists and psychologists study learning in animals and humans (Nilson, 2011). The extraction of important information from a large pile of data and its correlations is often the advantage of using machine learning. New knowledge about tasks is constantly being discovered by humans and vocabulary changes. There is a constant stream of new events in the world and continuing redesign of Artificial Intelligent systems to conform to new knowledge is impractical but machine learning methods might be able to track much of it (Han and Kamber, 2000).

There is a substantial amount of research with machine learning algorithms such as Bayes network, Multilayer Perceptron, Decision tree and pruning like J48graft, C4.5, Single Conjunctive Rule Learner like FLR, JRip and Fuzzy Inference System and Adaptive Neuro-Fuzzy Inference System.

2 DATA SET DESCRIPTION

The characteristics of the data set used in this research are summarized in Table 1. The detailed descriptions of the data set are available at UCI repository (UCI, 2011).

Afroz F. and M. Rahman R..
COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES ON PIMA INDIAN DIABETES DATA.
DOI: 10.5220/0003496803650368
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (ICEIS-2011), pages 365-368
ISBN: 978-989-8425-53-9

Table 1: Characteristics of PIMA Indian Dataset.

Data Set	No. of Example	Input Attributes	Output Classes	Number of Attributes
Pima Indian Diabetes	768	8	2	9

The objective of this data set was diagnosis of diabetes of Pima Indians. Based on personal data, such as age, number of times pregnant, and the results of medical examinations e.g., blood pressure, body mass index, result of glucose tolerance test, etc., try to decide whether a Pima Indian individual was diabetes positive or not. The attributes are given below:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/ (height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

3 METHODOLOGY

In this research we deploy various classification techniques. Those techniques are described briefly below:

3.1 Multilayer Perceptron (MLP)

The architecture used for the MLP (Werbos, 1974) during simulations with PIDD dataset consisted of a three layer feed-forward neural network: one input, one hidden, and one output layer. Selected parameters for the model are: learningRate = 0.3/0.15; momentum = 0.2; randomSeed = 0; validationThreshold = 20, number. of epochs = 500.

3.2 BayesNet

BayesNet (John and Langley, 1995) (learns Bayesian networks under the presumptions: nominal attributes (numeric one are pre descretized) and no missing values (any such values are replaced globally). There are two different parts for estimating the conditional probability tables of the network. In this study run BayesNet with the SimpleEstimator and K2 search algorithm without using ADTree.

3.3 NaiveBayes

The NaiveBayes (John and Langley, 1995) classifier provides a simple approach, with clear semantics, to representing and learning probabilistic knowledge. It is termed naïve because is relies on two important simplifying assumes that the predictive attributes are conditionally independent given the class, and it posits that no hidden or latent attributes influence the prediction process.

3.4 J48graft (C4.5 Decision Tree Revision 8)

Perhaps C4.5 algorithm which was developed by Quinlan (Quinlan, 1993) is the most popular tree classifier. Weka classifier package has its own version of C4.5 known as J48 or J48graf. For this study, C4.5 classifier used in TANAGRA platform and J48graft classifier used in WEKA platform. J48graft is an optimized implementation of C4.5 rev. 8. J48graft is experimented is this study with the parameters: confidenceFactor = 0.25; minNumObj = 2; subtreeRaising = True; unpruned = False. C4.5 is experimented in this study with the parameters: Min size of leaves = 5; Confidence-level for pessimistic = 0.25.

3.5 Fuzzy Lattice Reasoning (FLR) Classifier

The Fuzzy Lattice Reasoning (FLR) classifier is presented for inducing descriptive, decision-making knowledge (rules) in a mathematical lattice data domain including space R^N. Tunable generalization is possible based on non-linear (sigmoid) positive valuation functions; moreover, the FLR classifier can deal with missing data. Learning is carried out both incrementally and fast by computing disjunctions of join-lattice interval conjunctions, where a join-lattice interval conjunction corresponds to a hyperbox in R^N. In this study evaluated FLR classifier in WEKA with the parameters: Rhoa = 0.5; Number of Rules = 2.

3.6 JRip (RIPPER)

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) (Witten and Frank, 2005) is one of the basic and most popular algorithms.

Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced-error pruning. In this study evaluated RIPPER through JRip, an implementation of RIPPER in WEKA with the parameters: folds = 10; minNo = 2; optimizations = 2; seed = 1; usePruning = true.

3.7 Fuzzy Inference System (FIS)

Fuzzy Inference Systems (FISs) is a technology developed for granular rule induction and generalization based on fuzzy logic. Note that since a data cluster can be interpreted as a (fuzzy) granule, data clustering may be closely related to fuzzy rule induction. Neural implementations have provided conventional FISs a capacity for parallel implementation.

3.8 Adaptive Neuro-Fuzzy Inference Systems (ANFIS)

In this work uses ANFIS (Adaptive Neuro-Fuzzy Inference Systems), a fuzzy classifier that is part of the MATLAB Fuzzy Logic Toolbox (FLT, 2011). ANFIS is a fuzzy inference system implemented under the framework of adaptive networks (Jyh and Roger, 1993).

4 RESULT ANALYSIS

In this study, we examine the performance of different classification methods. We use accuracy estimate and error estimates of those classifiers. We get highest accuracy is 81.33% belongs to J48graft and lowest accuracy is 51.43% that belongs to FLR. Based on Figure 3 and Table 3, we could compare various error metrics among different classifiers in WEKA. We find out that J48graft is best, second best is Bayes Net and MLP & JRip is moderate but FLR is arguable.

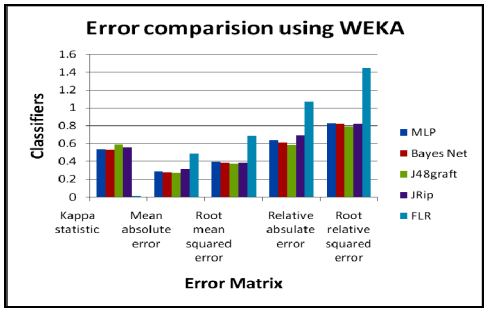

Figure 1: Error comparing for WEKA.

An algorithm which has a lower error rate will be preferred as it has a more powerful classification capability. The total time required to build the model is also a crucial parameter in comparing the classification algorithm. In this experiment, FLR classifier requires the shortest time which is around 0.025 seconds compared to the others. MLP algorithm requires the longest model building time which is around 63.13 seconds. The second on the list is Bayes network with 0.04 seconds. And J48graft takes 0.135 seconds.

Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity (Kappa, 2011). The average Kappa score from the selected algorithm is around 0.01-0.59. Based on the Kappa Statistic criteria, the accuracy of this classification purposes is substantial. So according to best average kappa statistic the J48graft classifier is best among others.

Rule accuracy is 71.51% and 78.79% for FIS and ANFIS respectively for different network and architectures. This is shown in Table 2. IF – THEN rules are used for adaptive classifiers. We use 7 IF – THEN fuzzy rules and mamdani operator for FIS and sugeno operators for ANFIS membership function. The rules are presented in Table 4.

We also measure our performance with True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F-measure and area under ROC curve. Those results are shown in Table 3.

Table 2: Performance measuring in rule based fuzzy approach using MATLAB.

Learning systems	Training/test epochs	Avg. Error after training/test	No. of Extracted Rules	Rules Accuracy (%)
FIS	500	7.6358	7	71.51
ANFIS	500	7.6358	7	78.79

5 CONCLUSIONS

We use WEKA, Tanagra and MATLAB to bring out an extensive performance comparison among the most popular classifier algorithms. In the absence of medical diagnosis evidences, it is difficult for the experts to opine about the grade of disease with affirmation. There is a need to undertake diagnostic studies medically to construct more realistic fuzzy numbers for characterizing the imprecision and thereby fuzzily describing the patient's disease nature. First, the misclassification cost is not considered explicitly here. In future, cost-sensitive

Table 3: Different Performance Matrix in the Training and Test Data Set using WEKA.

Classifier	Phase	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
MLP	Training	0.806	0.191	0.819	0.806	0.809	0.872
	Testing	0.778	0.306	0.774	0.778	0.776	0.813
Bayes Net	Training	0.783	0.26	0.783	0.783	0.783	0.851
	Testing	0.797	0.253	0.799	0.797	0.798	0.848
J48graft	Training	0.841	0.241	0.842	0.841	0.836	0.888
	Testing	0.785	0.189	0.816	0.785	0.792	0.803
JRip	Training	0.794	0.257	0.792	0.794	0.793	0.785
	Testing	0.824	0.294	0.821	0.824	0.816	0.766
FLR	Training	0.358	0.344	0.774	0.358	0.2	0.507
	Testing	0.67	0.662	0.582	0.67	0.572	0.504

Table 4: Sample rules framed for the proposed FIS and ANFIS.

	IF								THEN	
Rule No.	preg.	plas	bp	skin	insl	bmi	dpf	age	Class 0 (Weight)	Class1 (Weight)
1	0	<=103	>40	<=26	<=156	<=35.3	<=0.179	<=34	0.955	0.5
2	<=3	NDF	NDF	<=35	>156	<=35.3	<=0.787	NDF	0.5	0.928
3	NDF	NDF	NDF	NDF	NDF	NDF	<=0.179	<=34	0.955	0.5
4	NDF	<=103	NDF	NDF	NDF	NDF	<=0.787	NDF	0.944	0.5
5	NDF	NDF	NDF	NDF	<=156	<=35.3	NDF	>34 or <=37	0.912	0.5
6	NDF	>135	NDF	NDF	<=185	>33.7	<=1.096	>37	0.5	0.928
7	6	>103	NDF	NDF	NDF	>35.3	<=1.096	>34	0.5	0.909

learning might make the study more practical and valuable. Second, in this survey used only 7 rules for FIS and ANFIS but if increase the rules then might be got more accurate diagnosis result.

REFERENCES

FLT, 2011. The mathworks - fuzzy logic toolbox, from http://www.mathworks.ch/access/helpdesk r13/help/ toolbox/fuzzy/fuzzy.html

Han J., Kamber, M., 2000. *Data Mining Concept and Techniques*, Morgan Kaufmann Publishers

John, G. H., Langley, P., 1995. *Estimating Continuous Distributions in Bayesian Classifiers*. In: Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence.

Jyh, S., Roger, J., 1993.*Anfis: Adaptive-network-based fuzzy inference system*, IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, pp. 665–685.

Kappa Statistic, 2011. Link http://www.dmi.columbia. edu/homepages/chuangj/kappa.

Nilsson, N. J., 2011. *Introduction to Machine Learning*, http://ai.stanford.edu/~nilsson/mlbook.html.

Quinlan, J., 1993. C4.*5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo.

UCI machine learning repository, 2011. Link: http://www. ics.uci.edu/mlearn/MLRepository.html

Werbos, P., 1974. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioural Sciences*, PhD Thesis, Harvard University, 1974.

Witten, I. H., Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco, USA.