Applying rule-based classification techniques to medical databases: an empirical study

R.P. Datta*

Department of Information Technology,
Indian Institute of Foreign Trade,
Kolkata Campus, Premise No. 1583,
Madurdaha, Kolkata-700107, West Bengal, India
Email: rpdatta@gmail.com
*Corresponding author

Sanjib Saha

Department of Computer Science and Engineering,
Dr. B.C. Roy Engineering College,
Jemua Road, Fuljhore, Durgapur, West Bengal, 713206, India
Email: sanjibsaha.ju@gmail.com

Abstract: In the process of analysing and interpreting medical data, various classification techniques have been widely applied with a lot of success. A number of classification algorithms are available for this purpose and many researchers face the problem of choosing the best method for a particular dataset. In this paper, we apply five well-known, rule-based classification techniques on different medical datasets and compare their relative merits and demerits. Subsequently, we interpret their applicability in classifying patients into groups.

Keywords: classification; data mining; knowledge discovery; rule-based classification.

Reference to this paper should be made as follows: Datta, R.P. and Saha, S. (2016) 'Applying rule-based classification techniques to medical databases: an empirical study', *Int. J. Business Intelligence and Systems Engineering*, Vol. 1, No. 1, pp.32–48.

Biographical notes: R.P. Datta is a Professor of Information Technology and Information Management at Indian Institute of Foreign Trade, Kolkata campus. He holds a Bachelor degree from Indian Statistical Institute, Kolkata, India, He also holds MSc degree from Indian Institute of Technology, Kharagpur, India and MS degree from Colorado State University, USA. He is a PhD from University of Texas at Arlington, USA.

Sanjib Saha is an Assistant Professor at Dr. B.C. Roy Engineering College, Durgapur, India. He holds a Bachelor in Engineering from Burdwan University and post-graduate degree in Engineering from Jadavpur University, Kolkata, India.

1 Introduction

The management and analysis of information, and using existing data for prediction of the future has been an important and challenging research area for many years. Information can be analysed in various ways, one of which is classification. Classification of information is an important part of business decision making tasks. Classification enables us to classify records in a large database into a predefined set of classes. The classes are defined before studying or examining records in the database. It also enables us to predict the future behaviour of that sample data. Many decision making tasks are instances of classification problems or can be formulated into a classification problem. Examples are prediction and forecasting problems, diagnosis, and pattern recognition. Classification of information can be done either by statistical methods or by data mining methods.

Data mining may be considered as the automated discovery of non-trivial, previously unknown, and potentially useful patterns inherent in databases (Frawley et al., 1991). Also known as knowledge discovery in databases (KDD), data mining aims to find useful information from large collections of data. Data mining is the technique of extracting meaningful information from a large and mostly unorganised database. It is the process of performing automated extraction and generating predictive information from large databases. The discovered knowledge may be rules that describe properties of the data, patterns that occur frequently and objects that are found to be in clusters in the database etc. (Heikki, 1997).

Classification can be looked upon as a key data mining technique, whereby database tuples, acting as training samples, are analysed in order to produce a model of the given data (Fayyad et al., 1996a, 1996b). It is assumed that each tuple belongs to a predefined class, as determined by one of the attributes, called the classifying attribute. Once the classification model is derived, it can be used to categorise future data samples, as well as provide a better understanding of the database contents.

In data mining and knowledge discovery, an important task is to construct fast and accurate classifiers for large datasets. There are numerous applications of classification that include credit approval, product marketing, and medical diagnosis. Data mining practitioners often are confronted with the problem of selecting the most accurate algorithm for their classification tasks. It is an accepted fact that each algorithm performs well only on a subset of classification tasks and in general there are no clear winners. This may be looked upon as a direct consequence of the no free lunch (NFL) theorem which basically says that given any classifier, a dataset can be always constructed that beats it, i.e., the classifier will perform very poorly (Schaffer, 1994; Wolpert, 2001). In other words, if algorithm A outperforms algorithm B on some cost functions, then there must exist exactly as many other functions where B outperforms A. Thus, we can say that it is the ‘data’ that is important, not the classifier per se. There are a growing number of available algorithms, and therefore, finding the best algorithm for a particular classification task is a challenging task. It becomes important to ask the question “from the vast and ever increasing array of classification algorithms, which one should be the first choice for my present classification problem?”

This is one of the major problems in the analysis of bioinformatics data, where the aim is to arrive at the correct diagnosis of a certain illness based on certain important attributes that are available. Many tests that generally involve the clustering or classification of large scale data are used for the ultimate diagnosis. It is commonly

assumed that all these test procedures are necessary in order to reach the final diagnosis. However, it is equally likely that too many tests can complicate the main process of diagnosis and this may lead to difficulty in obtaining the end results. Machine learning can be used to resolve this kind of difficulty by directly obtaining the end result with the help of different artificial intelligence algorithms, which perform the role of classifiers.

A machine learning algorithm is an algorithm that can be implemented on a computer, which can learn from past experience (observed instances) with respect to some category of tasks and some measure of performance (Mitchel, 1997). Machine learning methods are suitable for a variety of data, such as, transactional data, financial data, molecular biology related data etc. The learning ability of the algorithm can construct classifiers/hypotheses that can explain complex relationships in the data which are not visible otherwise. The classifiers or hypotheses thus constructed can be further verified by domain experts or subject matter specialists, who can suggest some real lab experiments, if needed, to validate or refute the hypotheses.

The main aim of this research is to assist in the selection of an appropriate classification algorithm without the need for trial-and-error testing of the vast array of available algorithms. There are many studies that propose new classification algorithms. They attempt to produce empirical evidence of the superiority of one algorithm over another based on different datasets. The NFL theorem on the other hand suggests, that a more useful strategy would be to increase our understanding of the dataset characteristics that enable different learning algorithms to perform better, and to use this knowledge to help determine which learning algorithm to select based on the characteristics of the given dataset. The objective of our study is to try to suggest some answers to the following questions faced by researchers:

1 How does one choose the algorithm that is best suited to the particular dataset under consideration?

2 How does one compare the effectiveness of a particular algorithm with that of another?

In this paper, we study the performance of five different rule-based classification techniques on some commonly available medical databases and look at their ability to segment patients into groups. This paper is organised as follows: In Section 2, we give a brief survey of the existing work, in Section 3, we describe our method, in Section 4, we discuss the results, and in Section 5, we summarise and conclude.

2 Background

Classification and Association rules play a major role in data mining. Classification is the process of dividing a dataset into mutually exclusive groups. Association rules provide the means to find relationships among data items in a given dataset. Comparison of classification techniques have been the subject of many some previous studies. A comparison of rule-based and association rule mining algorithms is dealt with in Mazid et al. (2009). A comparison of fuzzy-based classification with neural network approaches for medical diagnosis is given in Herrmann et al. (1995). Again, in the field of medical detection, the paper Frame et al. (1998) gives a comparison of computer-based classification methods applied to the detection of microaneurysms in ophthalmic

fluorescein angiograms. In a study by Nosofsky et al. (1994), the authors have partially replicated and extended the Shepard et al.'s (1961) classic study involving the task difficulty for learning six fundamental types of rule-based categorisation problems. A comparison of various classification methods for predicting deception in computer-mediated communication is presented in Zhou et al. (2004).

A variety of statistical methods and heuristics have been used in the past for the purpose of classification. Work done in the field of decision science also show many different data mining techniques used to classify and predict data. Data mining techniques have also been used primarily for pattern recognition in large volumes of data. If we look at the available research, we can see that statistical and data mining techniques have been used for purposes like bankruptcy prediction (Wilson and Sharda, 1994), educational placement of students (Lin et al., 2004), supporting marketing decisions for target marketing of individual mailings (Levin et al., 1995; Kim and Street, 2004), assessing consumer credit risk (Henley and Hand, 1996) and helping in customer credit scoring (Hand and Henley, 1997). Different authors like Kiang (2003), Chiang et al. (2006) and Asparoukhov and Krzanowski (2001) have studied data mining and statistical classification methods and have analysed the results for a comparative assessment of classification methods.

A study by Finch and Schneider (2007) compares the classification accuracy of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression (LR), and classification and regression trees (CART) under a variety of data conditions. Morris and Meshbane (1995) have looked at the effect on predictive accuracy assuming equal versus unequal prior probabilities of group membership in discriminant analysis. Some of the other commonly used approaches for classification used to extract knowledge from data are statistical (John and Langley, 1995) which use Bayesian classifiers to look at continuous distributions. Furnkranz (1996) gives an overview of a large family of symbolic rule-learning algorithms, the so-called separate-and-conquer or covering algorithms. All members of this family share the same top-level loop. This method basically uses a separate-and-conquer algorithm that searches for a rule that explains a part of its training instances, separates these examples, and recursively conquers the remaining examples by learning more rules until no examples remain. This ensures that each instance of the original training set is covered by at least one rule.

Cendrowska (1987) describes a new algorithm, PRISM which, although based on Quinlan's (1986) ID3 'Induction of decision trees' uses a different induction strategy to induce rules which are modular, thus avoiding many of the problems associated with decision trees.

Many algorithms have been derived from these approaches, like static versus Dynamic sampling for data mining by John and Langley (1996). See5.0 (Quinlan, 2011) for windows and its unix counterpart C4.5 (Quinlan, 1993) are sophisticated data mining tools for discovering patterns that delineate categories, assembling them into classifiers, and using them to make predictions. PART (Frank and Witten, 1998) shows how good rule sets can be learned one rule at a time without any need for global optimisation, prism (Cendrowska, 1987) has been developed to elaborate upon the 'separate and conquer' approach while IREP (Furnkranz and Widmer, 1994) that stands for incremental reduced-error pruning, splits the database into two parts: the growing set and the pruning set. Sometimes a small subset of rules is found by traditional classification techniques and detailed rules that may play an important role are missed (Pazzani et al., 1997).

3 Methodology

In this study, we have applied rule-based classification techniques to machine learning algorithms. Learning schemes in machine learning can be generally divided into two types: supervised learning, where the output has been labelled apriori or the learner has some previous knowledge about the data; and unsupervised learning where no previous information is available to the learner about the data or the output. Some of the commonly performed tasks of the learner are to classify, characterise and cluster the input data as required.

Classification is one of the most common tasks in machine learning, where, given two or more different sets of example data, the learner needs to construct a classifier to distinguish between the different classes. Classification can be looked upon as supervised learning.

Association rule enables us to establish association and relationships between large unclassified data items based on certain attributes and characteristics. Association rules define certain rules of associativity between data items and then use those rules to establish relationships.

3.1 Rule-based classification

Rule-based classification classifies records by using a collection of 'if…then…' rules. Rule: (Condition) → y where Condition is a conjunction of attributes and y is the class label e.g. (Blood Type = Warm) ^ (Lay Eggs = Yes) → Birds. A rule r covers an instance x if the attributes of the instance satisfy the condition of the rule. Classification rules can be built using one of the two methods – direct method and indirect method. Direct methods are those that extract rules directly from the data; e.g., RIPPER (Cohen, 1995). Indirect methods are those that extract rules from other classification models like decision trees, e.g. C4.5 rules (Quinlan, 1993). In the direct method, we first grow a single rule (Rule Growing), then remove instances from this rule (Instance Elimination). After this, we prune the rule according to the Stopping Criterion (Rule Pruning) and finally, add the rule to the current rule set.

Advantages of rule-based classification are, it is as highly expressive as decision trees, it is easy to interpret, it is easy to generate, and it can classify new instances rapidly. The rule-based classification techniques used in this study are described briefly below:

3.1.1 Decision table

Decision table (DT) (Kohavi, 1995; Holmes et al., 1999) is an accurate method for numeric prediction from decision trees and it is an ordered set of If-Then rules that have the potential to be more compact and therefore more understandable than the decision trees. A DT consists of a two-dimensional array of cells, where the columns contain the systems constraints and each row makes a classification according to each cell's value (case of condition). It builds a decision table majority classifier, evaluates feature subsets using best-first search, and can use cross-validation for evaluation. There is also an option to use the nearest neighbour method to determine the class for each instance that is not covered by a decision table entry, instead of the table's global majority, based on the same set of features.

3.1.2 JRIP

JRIP (Cohen, 1995) is a prepositional rule learner, and it implements the algorithm called RIPPER which stands for repeated incremental pruning to produce error reduction. Classes are examined in increasing size and initial set of rules for the class is generated using incremental reduced-error pruning or IREP. An extra stopping condition is introduced that depends on the description length of the examples and rule set. The description length DL is a complex formula that takes into account the number of bits required to send a set of examples with respect to a set of rules, the number of bits required to send a rule with k conditions, and the number of bits needed to send the integer k-times an arbitrary factor of 50% to compensate for possible redundancy in the attributes. Once a rule set has been produced for each class, each rule is reconsidered and two variants are produced, again using reduced-error pruning. But at this stage, instances covered by other rules for the class are removed from the pruning set and success rate on the remaining instances is used as the pruning criterion. If one of the variants produces a better description length, it replaces the rule.

3.1.3 NNGE

NNGE is a nearest-neighbour-like algorithm (Martin, 2002) for generating rules using non-nested generalised exemplars (which are hyper rectangles that can be viewed as if-then rules). Generalised exemplars are rectangular regions of instance space, called hyperrectangles because they are high-dimensional. When classifying new instances, it is necessary to modify the distance function to allow the distance to a hyper rectangle to be computed. When a new exemplar is classified correctly, it is generalised by simply merging it with the nearest exemplar of the same class. The nearest exemplar may be either a single instance or a hyperrectangle. In the former case, a new hyperrectangle is created, which covers the old and the new instance. In the latter case, the hyperrectangle is enlarged to encompass the new instance. Finally, if the prediction is incorrect and the reason is a hyperrectangle, then the boundaries of the hyperrectangle are altered so that it shrinks away from the new instance.

3.1.4 PART

PART stands for partial decision tree algorithm. The PART (Frank and Witten, 1998) technique avoids global optimisation step used in C4.5 rules (Quinlan, 1993) and RIPPER (Cohen, 1995). It uses a separate-and-conquer technique to build a partial C4.5 decision tree in each iteration and makes the ‘best’ leaf into a rule. In essence, to make a single rule, a pruned decision tree is built for the current set of instances; the leaf with the largest coverage is made into a rule and the tree is discarded.

3.1.5 RIDOR

RIDOR learns rules with exceptions by generating the default rule, using incremental reduced-error pruning to find exceptions, and iterating. Ripple down rule learner (Gaines and Compton, 1995) generates a default rule first and then the exceptions for the default rule with the least (weighted) error rate. It then generates the ‘best’ exception for each exception and iterates until pure. Thus, it performs a tree-like expansion of exceptions.

The exceptions are a set of rules that predict classes other than the default. IREP is used to generate the exceptions.

3.2 Association rule mining

Association rule mining, first proposed by Agrawal et al. (1993b), consists of "finding frequent patterns, associations, correlations or casual structure sets of items or objects in transaction database, relational database and other information repositories". The application of association rule ranges from business management, production control, and market analysis, to engineering design and science exploration. At present association rule mining is an important task of data mining and is used in market basket analysis that tries to find out the shopping behaviour of customers in the hope of finding patterns (Agrawal et al., 1993a). One of the most popular algorithms of finding association rules is Apriori (Agrawal and Srikant, 1994; Liu et al., 1998).

- Association rule (R): An implication expression of the form $X \rightarrow Y$ [s, c]. Where X and Y are itemsets. (X, Y subset of I) and $X \cap Y$ = empty.
- Support (s): Fraction of transactions that contain both X and Y. Probability that a transaction contains $X \cup Y$. i.e. $P(X \cup Y)$.
- Confidence (c): Measures how often items in Y appear in transactions that contain X. Conditional probability that a transaction having X also contains Y. i.e. $P(X \mid Y)$ = Support($X \cup Y$) / Support(X).

3.3 Measurement criteria

Confusion matrix

The general structure of n class confusion matrix is:

		Predicted class			
		A	*B*	*C*	*N*
Actual class	A	tpA	eAB	eAC	eAN
	B	eBA	tpB	eBC	eBN
	C	eCA	eCB	tpC	eCN
	N	eNA	eNB	eNC	tpN

$$\text{Classification accuracy} = (\text{tpA} + \text{tpB} + \text{tpC} + \ldots + \text{tpN}) / \text{total instances}$$

$$\text{Kappa statistics} = \frac{(\text{Observed agreement} - \text{Chance agreement})}{(1 - \text{Chance agreement})}$$

$$\text{Observed agreement} = (\text{tpA} + \text{tpB} + \text{tpC} + \ldots + \text{tpN}) / \text{total instances}$$

Let us define 2n variables A1, A2, B1, B2, C1, C2,…,N1, N2

$$\text{A1} = (\text{tpA} + \text{eAB} + \text{eAC} + \ldots + \text{eAN}) / \text{total instances}$$

$$\text{A2} = (\text{tpA} + \text{eBA} + \text{eCA} + \ldots + \text{eNA}) / \text{total instances}$$

$$B1 = (eBA + tpB + eBC + \ldots + eBN) / \text{total instances}$$

$$B2 = (eAB + tpB + eCB + \ldots + eNB) / \text{total instances}$$

$$C1 = (eCA + eCB + tpC + \ldots + eCN) / \text{total instances}$$

$$C2 = (eAC + eBC + tpC + \ldots + eNC) / \text{total instances}$$

$$N1 = (eNA + eNB + eNC + \ldots + eN(N-1) + tpN) / \text{total instances}$$

$$N2 = (eAN + eBN + eCN + \ldots + e(N-1)N + tpN) / \text{total instances}$$

$$\text{Chance agreement} = A1 \times A2 + B1 \times B2 + C1 \times C2 + \ldots + N1 \times N2$$

Actual target values: $a1\ a2 \ldots an$

Predicted target values: $p1\ p2 \ldots pn$

$$\text{Mean absolute error} = \frac{|p1 - a1| + \ldots + |pn - an|}{n}$$

$$\text{Root mean squared error} = \frac{(p1 - a1)2 + \ldots + (pn - an)2}{n}$$

$$\text{Relative absolute error} = \frac{|p1 - a1| + \ldots + |pn - an|}{|a - a1| + \ldots + |a - an|}$$

$$\text{Root relative squared error} = \frac{(p1 - a1)2 + \ldots + (pn - an)2}{(a - a1)2 + \ldots + (a - an)2}$$

4 Results and observations

For the purpose of this study, 11 medical datasets have been taken from the UCI Machine Learning Repository (Blake and Merz, 2000). All the datasets are in arff (attribute relation file) format. The datasets have both continuous and discrete attributes and also have missing values. For the classification, we first selected all attributes of the dataset, then chose cross validation of ten folds as the test method. This means, we performed the following steps:

1 Break the data into ten sets of size n/10.

2 Train on nine datasets and test on the remaining one dataset.

3 Repeat ten times and take a mean accuracy.

Next, we chose a particular classification technique (e.g. DT) and also specified the parameters, such as,

- search method = greedy stepwise
- seed value = 1 (seed is used for randomising the data)

- prune value = true, i.e. pruning is performed
- debug value = false i.e. debug information does not output to the console
- confidence factor = 0.25 (smaller values giver more pruning).

The analysis of these rule-based classification techniques have been done by using the same criteria on the number of rules generated, classification accuracy, kappa statistic, relative absolute error and root relative squared error. Analysis of association rules from the individual datasets has been done by using the same criteria on the number of best rules, different rules and their corresponding confidence values. In this paper, the WEKA version 3.5.7 (Witten and Frank, 2000, 2005) framework is used for performing the numerical calculations.

The dataset used in the experiments are described in Tables 1 and 2.

Table 1 Characteristics of datasets used

Dataset	*Instances*	*Classes*
Audiology	226	24
Breast cancer	286	2
Breast cancer-w	699	2
Colic	368	2
Diabetes	768	2
Heart-c	303	2
Heart statlog	270	2
Hepatitis	155	2
Hypothyroid	3,772	4
Lymph	148	4
Primary tumour	339	22

Table 2 More characteristics of datasets used

Dataset	*Continuous attributes*	*Discrete attributes*
Audiology	0	70
Breast cancer	0	10
Breast cancer-w	9	1
Colic	7	16
Diabetes	8	1
Heart-c	6	8
Heart statlog	13	1
Hepatitis	6	14
Hypothyroid	7	23
Lymph	3	16
Primary tumour	0	18

The numerical results for rule-based classification are given in Tables 3 through 8.

Table 3 Number of rules for different datasets

Dataset	*Rule-based classification*				
	Number of rules				
	DT	*JRIP*	*NNGE*	*PART*	*RIDOR*
Audiology	27	20	46	21	105
Breast cancer	25	3	105	20	3
Breast cancer-w	23	6	271	10	4
Colic	32	4	109	9	4
Diabetes	32	4	280	13	4
Heart-c	17	4	77	19	6
Heart statlog	16	5	102	24	6
Hepatitis	28	4	24	8	2
Hypothyroid	76	5	39	11	11
Lymph	18	6	33	13	10
Primary tumour	65	7	165	43	162

Table 4 Classification accuracy for different datasets

Dataset	*Rule-based classification*				
	Classification accuracy (%)				
	DT	*JRIP*	*NNGE*	*PART*	*RIDOR*
Audiology	69.47	73.00	71.24	78.32	73.00
Breast cancer	74.83	70.97	65.03	71.23	70.10
Breast cancer-w	95.42	95.42	96.00	93.84	95.85
Colic	83.69	84.24	80.43	84.78	83.69
Diabetes	71.09	76.04	74.00	75.26	75.00
Heart-c	76.24	81.52	80.86	79.86	79.54
Heart statlog	82.96	78.88	78.15	73.33	78.15
Hepatitis	81.93	78.06	84.51	84.51	78.71
Hypothyroid	99.33	99.33	98.70	99.41	99.44
Lymph	78.38	77.70	78.38	76.35	85.13
Primary tumour	39.82	39.23	40.70	40.70	37.19

Table 5 Kappa statistics for different datasets

Dataset	*Rule-based classification*				
	Kappa statistic				
	DT	*JRIP*	*NNGE*	*PART*	*RIDOR*
Audiology	0.63	0.68	0.66	0.74	0.68
Breast cancer	0.27	0.24	0.12	0.2	0.18
Breast cancer-w	0.89	0.89	0.91	0.86	0.90
Colic	0.64	0.65	0.56	0.66	0.63

Table 5 Kappa statistics for different datasets (continued)

Dataset	*Rule-based classification*				
	Kappa statistic				
	DT	*JRIP*	*NNGE*	*PART*	*RIDOR*
Diabetes	0.33	0.45	0.41	0.44	0.42
Heart-c	0.52	0.63	0.61	0.59	0.58
Heart statlog	0.65	0.56	0.55	0.46	0.55
Hepatitis	0.36	0.26	0.43	0.54	0.19
Hypothyroid	0.95	0.95	0.90	0.96	0.96
Lymph	0.58	0.57	0.58	0.55	0.71
Primary tumour	0.30	0.24	0.32	0.32	0.29

Table 6 Mean absolute error for different datasets

Dataset	*Rule-based classification*				
	Mean absolute error				
	DT	*JRIP*	*NNGE*	*PART*	*RIDOR*
Audiology	0.06	0.03	0.02	0.02	0.02
Breast cancer	0.37	0.38	0.35	0.36	0.29
Breast cancer-w	0.08	0.06	0.04	0.06	0.04
Colic	0.26	0.23	0.19	0.24	0.16
Diabetes	0.34	0.34	0.26	0.31	0.25
Heart-c	0.16	0.11	0.07	0.09	0.08
Heart statlog	0.27	0.29	0.21	0.27	0.22
Hepatitis	0.26	0.26	0.15	0.18	0.21
Hypothyroid	0.02	0	0	0	0
Lymph	0.20	0.14	0.10	0.13	0.07
Primary tumour	0.07	0.06	0.05	0.06	0.05

Table 7 Root mean squared error for different datasets

Dataset	*Rule-based classification*				
	Root mean squared error				
	DT	*JRIP*	*NNGE*	*PART*	*RIDOR*
Audiology	0.16	0.13	0.15	0.12	0.15
Breast cancer	0.44	0.45	0.60	0.47	0.54
Breast cancer-w	0.18	0.20	0.20	0.22	0.20
Colic	0.36	0.36	0.44	0.35	0.40
Diabetes	0.42	0.42	0.51	0.41	0.50
Heart-c	0.26	0.24	0.28	0.26	0.28
Heart statlog	0.37	0.41	0.46	0.49	0.46

Table 7 Root mean squared error for different datasets (continued)

Dataset	*Rule-based classification*				
	Root mean squared error				
	DT	*JRIP*	*NNGE*	*PART*	*RIDOR*
Hepatitis	0.36	0.41	0.39	0.36	0.46
Hypothyroid	0.07	0.05	0.08	0.05	0.05
Lymph	0.29	0.31	0.32	0.33	0.27
Primary tumour	0.19	0.19	0.23	0.19	0.23

Table 8 Relative absolute error for different datasets

Dataset	*Rule-based classification*				
	Relative absolute error (%)				
	DT	*JRIP*	*NNGE*	*PART*	*RIDOR*
Audiology	85.28	41.37	33.04	30.08	31.01
Breast cancer	89.45	90.78	83.56	87.22	69.36
Breast cancer-w	18.56	13.67	8.86	15.16	9.18
Colic	57.34	50.27	42.00	50.87	35.00
Diabetes	75.41	75.23	57.29	68.22	55.00
Heart-c	81.03	52.47	38.06	45.95	40.79
Heart statlog	55.23	58.60	44.24	55.97	44.24
Hepatitis	80.31	78.56	46.89	56.18	64.47
Hypothyroid	29.48	6.56	8.91	4.78	3.81
Lymph	76.08	52.74	40.31	48.69	27.71
Primary tumour	97.04	85.47	55.32	75.98	70.28

Table 9 Root relative squared error for different datasets

Dataset	*Rule-based classification*				
	Root relative squared error (%)				
	DT	*JRIP*	*NNGE*	*PART*	*RIDOR*
Audiology	85.74	70.44	81.70	64.76	79.15
Breast cancer	96.07	98.32	129.37	104.18	117.86
Breast cancer-w	39.28	42.54	42.11	47.00	42.85
Colic	75.61	76.41	91.67	72.47	83.64
Diabetes	89.17	88.93	107.06	87.04	104.90
Heart-c	84.50	76.71	87.83	84.00	90.81
Heart statlog	75.16	83.15	94.07	99.24	94.07
Hepatitis	89.46	101.79	97.17	88.94	113.94
Hypothyroid	40.27	29.10	42.32	26.42	27.70
Lymph	81.59	85.34	90.28	89.80	74.86
Primary tumour	96.47	95.69	115.41	97.04	118.81

From the numerical calculations performed, we observe the following:

a For the Audiology dataset, the PART algorithm gives a better result on classification accuracy and kappa statistic r than the other algorithms. This could be because the Audiology dataset has a large number of attributes (more than equal to 20).

b For the Breast cancer dataset, the DT algorithm gives a better result on classification accuracy and kappa statistic than the other algorithms.

c For the Breast cancer-w dataset, the NNGE and DT algorithm gives better result on classification accuracy, and kappa statistic than the other algorithms.

d For the Colic dataset, the PART algorithm gives a better result on classification accuracy, and kappa statistic than the other algorithms. It is noteworthy that the Colic dataset has a large number of attributes (more than or equal to 20).

e For the Diabetes dataset, the JRIP algorithm gives a better result on classification accuracy and kappa statistic than the other algorithms.

f For the Heart-c dataset, the JRIP algorithm gives a better result on classification accuracy and kappa statistic than the other algorithms.

g For the Heart statlog dataset, the DT algorithm gives a better result on classification accuracy and kappa statistic than the other algorithms. It is noteworthy that the Heart statlog dataset has all the continuous attributes except the class attribute.

h For the Hepatitis dataset, the PART algorithm gives better result on classification accuracy and kappa statistic than the other algorithms. Again it may be noted that the Hepatitis dataset has a large number of attributes (more than 20).

i For the Hypothyroid dataset, both the PART and RIDOR algorithms give better result on classification accuracy and kappa statistic than the other algorithms. The Hypothyroid dataset has a large number of attributes (more than 20).

j For the Lymph dataset, the RIDOR algorithm gives a better result on classification accuracy and kappa statistic than the other algorithms.

k For the Primary tumour dataset, both the NNGE and DT algorithm gives better result on classification accuracy and kappa statistic than the other algorithms. The Primary tumour dataset has only discrete attributes.

From the above observations, we can list some of our findings as follows:

1 The PART algorithm seems to be outperforming the other algorithms in classification accuracy and kappa statistics in the Audiology, Colic and Hepatitis datasets. The PART algorithm and the RIDOR algorithm are comparable in the Hypothyroid dataset.

2 The NNGE and DT algorithms seem to be outperforming the other algorithms in classification accuracy and kappa statistics in the Breast cancer-w and Primary tumour dataset.

3 The DT algorithm alone outperforms the other algorithms in the Breast cancer and Heart statlog dataset.

4 The JRIP algorithm works better in the Diabetes and Heart-c databases.

5 The RIDOR algorithm seems to outperform the other algorithms in the Lymph dataset.

One of the implications of our findings from the above observations is, if the total number of attributes (continuous and discrete) are large (say more than 20), the PART algorithm seems to be the algorithm of choice.

Another implication of our findings is that observations a, b and k made above agree with the observations made by Tan and Gilbert (2003), who stated that rule-based systems like DT and PART tend to perform better in discrete/categorical attributes. However, the observation k, which shows the results for Primary tumour dataset, only agrees partially with Tan and Gilbert's (2003) results as both NNGE and PART algorithms seem to outperform the other algorithms. It may be noted here that the Audiology, Breast Cancer and Primary tumour datasets have only discrete/categorical data.

5 Summary and conclusions

To summarise, we have studied five well-known rule-based classification techniques, namely DT, JRIP, NNGE, PART, and RIDOR in this paper and these classification techniques have been applied on 11 medical datasets. From the numerical results, (Table 3 to Table 9) we can make a comparative study of these algorithms and their applicability on medical databases.

From the numerical results, we can make the following empirical observations:

1 if the medical dataset has both continuous and discrete attributes, and the number of attributes is more than or equal to 20, then PART algorithm performs better than the others

2 if the medical dataset has only discrete attributes then the DT algorithm performs better than others

3 if the medical dataset has all continuous attributes except the class attribute, then the DT and JRIP algorithms perform better than others.

The values of mean absolute error, root mean squared error, relative absolute error and root relative squared error for these entire rule-based classification algorithms on the 11 medical datasets, make it possible to do a comparative study of these algorithms and see which one is better for these datasets.

From the numerical calculations shown in Table 6 to Table 9, we can make the following empirical observations:

1 RIDOR algorithm gives the lowest mean absolute errors for ten medical datasets out of 11 datasets

2 DT gives lowest root mean squared errors for ten medical datasets out of 11 datasets

3 RIDOR algorithm gives lowest relative absolute errors for ten medical datasets out of 11 datasets

4 DT and PART give the lowest root relative squared errors for eight medical datasets out of 11 datasets.

Therefore, in terms of mean absolute error, root mean squared error, relative absolute error and root relative squared error, the RIDOR algorithm seems to be the algorithm that performs best on most of the datasets chosen followed by the DT and the PART algorithms in terms of performance.

We now discuss some limitations of our study. The main aim of this research was to assist in the selection of an appropriate classification algorithm without the need for trial-and-error testing of the vast array of available algorithms. It may be pertinent to point out here that though better data often beats better algorithms, and designing good features goes a long way, if we have a huge dataset, our choice of classification algorithm might not really matter so much in terms of classification performance. So we may choose our algorithm based on speed or ease of use instead of classification performance.

We could have used the cross-validation approach, which is another popular way to evaluate a classifier. The hold-out, leave-one-out and rotation methods are different approaches to cross-validation. The main disadvantage of the cross-validation method is that all the samples are not used to construct the model when the samples are relatively small. Moreover, the hold-out method and leave-one-out method suffer from either large bias or variance (Jain et al., 2000).

Future work in this direction could include looking at other kinds of classification algorithms than rule-based classification techniques. One can also look at different kinds of datasets apart from the one used in this study, particularly datasets having varying combinations of discrete and continuous attributes and also datasets having a large number of overall records to see how a particular combination of discrete and continuous attributes and the large number of records in a dataset affect the classification capability of various algorithms.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments which helped us immensely to enrich the paper in its present form.

References

Agrawal, R. and Srikant, R. (1994) 'Fast algorithms for mining association rules in large databases', *20th International Conference on Very Large Data Bases*, pp.478–499.

Agrawal, R., Amielinski, T. and Swami, A. (1993a) 'Mining association rule between sets of items in large databases', *Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, DC, 26–28 May, pp.207–216.

Agrawal, R., Imielinski, T. and Swami, A. (1993b) 'Database mining: a performance perspective', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, No. 6, pp.914–925.

Asparoukhov, O.K. and Krzanowski, W.J. (2001) 'A comparison of discriminant procedures for binary variables', *Computational Statistics and Data Analysis*, Vol. 38, No. 2, pp.139–160.

Blake, C.L. and Merz, C.J. (2000) *UCI Repository of Machine Learning Databases, 1998* [online] http://www.ics.uci.edu/~mlearn/MLRepository.html (accessed 03-05-2014).

Cendrowska, J. (1987) 'PRISM: an algorithm for inducing modular rules', *International Journal of Man-Machine Studies*, Vol. 27, No. 4, pp.349–370.

Chiang, W.K., Zhang, D. and Zhou, L. (2006) 'Predicting and explaining patronage behavior toward web and traditional stores using neural networks: a comparative analysis with logistic regression', *Decision Support Systems*, Vol. 41, No. 2, pp.514–531.

Cohen, W.W. (1995) 'Fast effective rule induction', *Twelfth International Conference on Machine Learning*, pp.115–123.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996a) 'Knowledge discovery and data mining: towards a unifying framework', *Proc. 2nd Intl. Conf. on Knowledge Discovery and Data Mining (KDD'96)*, Portland, Oregon, pp.82–88.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (1996b) *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Cambridge, Massachusetts.

Finch, H. and Schneider, M.K. (2007) 'Classification accuracy of neural networks vs. discriminant analysis, logistic regression and classification and regression trees: three and five groups cases', *Methodology*, Vol. 3, No. 2, pp.47–57.

Frame, A.J., Undrill, P.E., Cree, M.Y., Olson, J.A. et al. (1998) 'A comparison of computer based classification methods applied to the detection of microaneurysms in ophthalmic fluorescein angiograms', *Comput. Biol. Med.*, May, Vol. 28, No. 3, pp.225–238.

Frank, E. and Witten, I.H. (1998) 'Generating accurate rule sets without global optimization', in Shavlik, J. (Ed.): *Proceedings of the Fifteenth International Conference on Machine Learning*, Morgan Kaufmann, Madison, Wisconsin, San Francisco, pp.144–151.

Frawley, W.J., Piatetsky-Shapiro, G. and Matheus, C.J. (1991) *Knowledge Discovery in Databases: An Overview*, pp.1–27, AAAI/MIT Press, Waltham, Massachusetts.

Furnkranz, J. (1996) *Separate-and-Conquer Rule Learning*, Technical Report TR-96-25, Austrian Research Institute for Artificial Intelligence, Vienna.

Furnkranz, J. and Widmer, G. (1994) 'Incremental reduced error pruning', *Machine Learning: Proceedings of the 11th Annual Conference*, Morgan Kaufmann, New Brunswick, New Jersey.

Gaines, B.R. and Compton, P. (1995) 'Induction of ripple-down rules applied to modeling large databases', *J. Intell. Inf. Syst.*, Vol. 5, No. 3, pp.211–228.

Hand, D.J. and Henley, W.E. (1997) 'Statistical classification methods in consumer credit scoring: a review', *Journal of the Royal Statistical Society*, Vol. 160, No. 3, pp.523–541.

Heikki, M. (1997) 'Methods and problems in data mining', in Afrati, F and Kolaitis, P. (Eds.): *Proceeding of International Conference on Database Theory*, Delphi, Greece, January, Springer Verlag.

Henley, W.E. and Hand, D.J. (1996) 'A k-nearest-neighbor classifier for assessing consumer credit risk', *The Statistician*, Vol. 45, No. 1, pp.77–95.

Herrmann, C.S., Halgamuge, S.K. and Glesner, M. (1995) 'Comparison of fuzzy rule based classification with neural network approaches for medical diagnosis', *European Congress on Fuzzy and Intelligent Technologies (EUFIT)*.

Holmes, G., Hall, M. and Frank, E. (1999) 'Generating rule sets from model trees', *Proc. 12th Australian Joint Conference on Artificial Intelligence*, Sydney, Australia, pp.1–12, Springer.

Jain, A.K., Duin, R.P.W. and Mao, J. (2000) 'Statistical pattern recognition: a review', *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 22, No. 1, pp.4–37.

John, G.H. and Langley, P. (1995) 'Estimating continuous distributions in Bayesian classifiers', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp.338–345, Morgan Kaufmann, San Mateo.

John, G.H. and Langley, P. (1996) 'Static vs. dynamic sampling for data mining', *Proceedings 2nd International Conference of Knowledge Discovery and Data Mining*, Portland, OR, AAAI Press, pp.367–370.

Kiang Y.M, (2003) 'A comparative assessment of classification methods', *Decision Support Systems*, July, Vol. 35, No. 4, pp.441–454.

Kim, Y-S. and Street, N.W. (2004) 'An intelligent system for customer targeting: a data mining approach', *Decision Support Systems*, Vol. 37, No. 2, pp.215–228.

Kohavi, R. (1995) 'The power of decision tables', *8th European Conference on Machine Learning*, pp.174–189.

Levin, N., Zahavi, J. and Olitsky, M. (1995) 'AMOS – a probability-driven, customer-oriented decision support system for target marketing of solo mailings', *European Journal of Operational Research*, Vol. 87, No. 3, pp.708–721.

Lin, M., Huang, S. and Chang, Y. (2004) 'Kernel-based discriminant technique for educational placement', *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 2, pp.219–240.

Liu, B., Hsu, W. and Ma, Y. (1998) 'Integrating classification and association rule mining', *Fourth International Conference on Knowledge Discovery and Data Mining*, pp.80–86.

Martin, B. (1995) *Instance-Based Learning: Nearest Neighbor with Generalization*, Hamilton, New Zealand.

Mazid, M.M., Ali, A.B.M.S. and Tickle, K.S. (2009) 'A comparison between rule based and association rule mining algorithms', *Third International Conference on Network Security 2009, NSS'09*.

Mitchell, T. (1997) *Machine Learning*, McGraw-Hill, New York, NY.

Morris, J.D. and Meshbane, A. (1995) 'Selecting predictor variables in two-group classification problems', *Educational and Psychological Measurement*, Vol. 55, No. 3, pp.438–441.

Nosofsky, R.M., Gluck, M.A. et al. (1994) 'Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961)', *Memory and Cognition*, Vol. 22, No. 3, pp.352–369.

Pazzani, M., Mani, S. and Shankle, W.R. (1997) 'Beyond concise and colorful: learning intelligible rules', *KDD-97*.

Quinlan, J.R. (1986) 'Induction of decision trees', *Machine Learning*, March, Vol. 1, No. 1, pp.81–106.

Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Francisco, CA.

Quinlan, J.R. (2011) *See5.0* [online] http://www.rulequest.com (accessed April 2011).

Schaffer, C. (1994) 'A conservation law for generalization performance', *Proceedings of the Eleventh International Conference on Machine Learning*, pp.259–265.

Shepard, R.N., Hovland, C.I. and Jenkins, H.M. (1961) 'Learning and memorization of classifications', *Psychological Monographs*, Vol. 75, No. 13, pp.1–42, Whole No. 517.

Tan, A.C. and Gilbert, D. (2003) 'An empirical comparison of supervised machine learning techniques in bioinformatics', *First Asia Pacific Bioinformatics Conference*, Adelaide, Australia.

Wilson, R.L. and Sharda, R. (1994) 'Bankruptcy prediction using neural networks', *Decision Support Systems*, Vol. 11, No. 5, pp.545–557.

Witten, I.H. and Frank, E. (2000) *Data Mining Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, Burlington, Massachusetts.

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

Wolpert, D.H. (2001) 'The supervised learning no-free-lunch theorems', *Proceedings of the Sixth On-line World Conference on Soft Computing in Industrial Applications*, pp.325–330.

Zhou, L., Burgoon, J.K., Twitchell, D.P. et al. (2004) 'A comparison of classification methods for predicting deception in computer-mediated communication', *Journal of Management Information Systems*, Spring, Vol. 20, No. 4, pp.139–165.