World Engineering & Applied Sciences Journal 6 (2): 119-123, 2015
ISSN 2079-2204
© IDOSI Publications, 2015
DOI: 10.5829/idosi.weasj.2015.6.2.22162

Analysis of Classification Algorithms J48 and Smo on Different Datasets

[1]S. Singaravelan, [2]D. Murugan and [1]R. Mayakrishnan

[1]Department of Computer Science and Engineering,
P.S.R. Engineering College, Sivakasi,Tamilnadu, India
[2]Department of Computer Science and Engineering,
Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India

Abstract: Data mining is the forthcoming research area to solve different problems and classification is one of main problem in the field of data mining. In this paper, we use two classification algorithms J48 and Sequential Minimal Optimization alias SMO of the Weka interface. It can be used for testing several datasets. The performance of J48 and Sequential Minimal Optimization have been analysed so as to choose the better algorithm based on the conditions of the datasets. The datasets have been chosen from UCI Machine Learning Repository. Algorithm J48 is based on C4.5 decision based learning and algorithm Sequential Minimal Optimization uses the Support Vector Machine approach for classification of datasets. When comparing the performance of both algorithms we found Sequential Minimal Optimization is better algorithm in most of the cases.

Key words: Classification · Data Mining Techniques · Decision Tree · Sequential Minimal Optimization

INTRODUCTION

Data mining is the process to pull out patterns from large datasets by joining methods from statistics and artificial intelligence with database management. It is an upcoming field in today world in much discipline. It has been accepted as technology growth and the need for efficient data analysis is required. The plan of data mining is not to give tight rules by analysing the data set, it is used to guess with some certainty while only analysing a small set of the data.

In recent times, data mining has been obtained a great attention in the knowledge and information industry due to the vast availability of large amounts of data and the forthcoming need for converting such data into meaningful information and knowledge. The data mining technology is one comprehensive application of technology item relying on the database technology, statistical analysis, artificial intelligence and it has shown great commercial value and gradually to other profession penetration in the retail, insurance, telecommunication, power industries use [1].

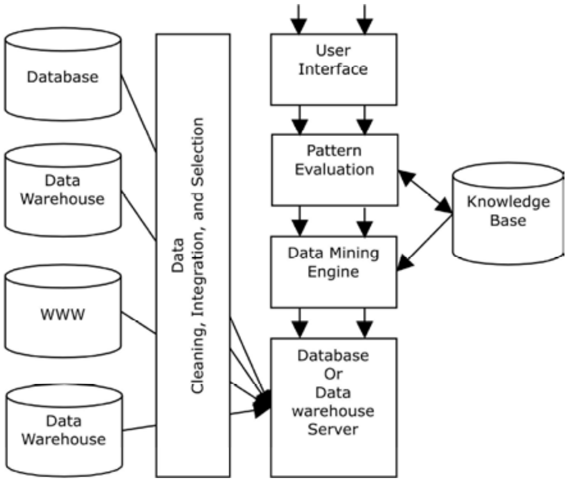

Fig. 1: Architecture of a Typical Data Mining System

The major components of the architecture for a typical data mining system are shown in Fig 1 [2]. Good system architecture will make possible the data mining system to make best use of the software environment. It achieves data mining tasks in an effective

Corresponding Author: S. Singaravelan, Department of Computer Science and Engineering,
P.S.R. Engineering College, Sivakasi,Tamilnadu, India.

and proper way to exchange information with other systems which is adaptable to users with diverse requirements and change with time.

Related Work: Recently studies have been done on various performance of decision tree and on back propagation.Classification is a classical problem in machine learning and data mining [3]. Decision trees are popular because they are practical and easy to understand. Rules can also be extracted from decision trees easily. Many algorithms, such as ID3 [4] and C4.5 [5], have been devised for decision tree construction.

In [6] neural networks are suitable in data-rich environments and are typically used for extracting embedded knowledge in the form of rules, quantitative evaluation of these rules, clustering, self-organization, classification and regression. They have an advantage, over other types of machine learning algorithms, for scaling.The use of neural networks in classification is not uncommon in machine learning community [7]. In some cases, neural networks give a lower classification error rate than the decision trees but require longer learning time [8, 9]. A decision tree can be converted to a set of rules, each one corresponding to a tree branch. Algorithms have been proposed to learn directly sets of rules [10] or to simplify the set of rules corresponding to a decision tree [5]. The alternating decision tree method [11] is a classification algorithm that tries to combine the interpretability of decision trees with the accuracy improvement obtained by boosting.

MATERIALS AND METHODS

Datasets: There are four datasets we have used in our paper taken from UCI Machine Learning Repository [12]. The details of each datasets are shown in Table 1.

In the diabetes dataset [12] several constraints were placed on the selection of instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

In the iris dataset contains 3 classes of 150 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

The tic-tac-toe dataset encodes the complete set of possible board configurations at the end of tic-tac-toe games, where "x" is assumed to have played first

Table 1: Details of 4 datasets

Datasets	Instances	Attributes	No. of Classes	Type
Diabetes	768	9	2	Numeric
Iris	150	5	3	Numeric
Tic-Tac-Toe	958	10	2	Nominal
Yuta-Selection	265	26	2	Numeric

The overview of all products by designer TakiroYuta. Refine your Designer takiroyuta selection and filter the overview by product group, manufacturer or theme.

Weka Interface: Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand [13]. The Weka suite contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

The original non-Java version of Weka was TCL/TK front-end software used to model algorithms implemented in other programming languages, plus data preprocessing utilities in C and a Make file-based system for running machine learning experiments.

This Java-based version (Weka 3.7.7) is used in many different application areas, in particular for educational purposes and research. There are various advantages of Weka:

- It is freely available under the GNU General Public License
- It is portable, since it is fully implemented in the Java programming language and thus runs on almost any architecture
- It is a huge collection of data preprocessing and modeling techniques
- It is easy to use due to its graphical user interface

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization and feature selection. All techniques of Weka's software are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported).

Classification Algorithm J48: J48 algorithm of Weka software is a popular machine learning algorithm based upon J.R. Quilan C4.5 algorithm. All data to be examined will be of the categorical type and therefore continuous data will not be examined at this stage. The algorithm will however leave room for adaption to include this capability. The algorithm will be tested against C4.5 for verification purposes [5].

In Weka, the implementation of a particular learning algorithm is encapsulated in a class and it may depend on other classes for some of its functionality. J48 class builds a C4.5 decision tree. Each time the Java virtual machine executes J48, it creates an instance of this class by allocating memory for building and storing a decision tree classifier. The algorithm, the classifier it builds and a procedure for outputting the classifier is all part of that instantiation of the J48 class.

Larger programs are usually split into more than one class. The J48 class does not actually contain any code for building a decision tree. It includes references to instances of other classes that do most of the work. When there are a number of classes as in Weka software they become difficult to comprehend and navigate [14].

Classification Function Sequential Minimal Optimization: Sequential Minimal Optimization (SMO) is used for training a support vector classifier using polynomial or RBF kernels. It replaces all missing the values and transforms nominal attributes into binary ones. A single hidden layer neural network uses exactly the same form of model as an SVM.

Training a Support Vector Machine (SVM) requires the solution of a very large quadratic programming (QP) optimization problem. SMO breaks this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop.

The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets. Because large matrix computation is avoided, SMO scales somewhere between linearand quadratic in the training set size for various test problems, while a standard projected conjugate gradient (PCG) chunking algorithm scales somewhere between linear and cubic in the training set size.

SMO's computation time is dominated by SVM evaluation, hence SMO is fastest for linear SVMs and sparse data sets. For the MNIST database, SMO is as fast as PCG chunking; while for the UCI Adult database and linear SVMs, SMO can be more than 1000 times faster than the PCG chunking algorithm [15].

RESULTS

For evaluating a classifier quality we can use confusion matrix. Consider the algorithm J48 running on iris dataset in WEKA, for this dataset we obtain three classes then we have 3x3 confusion matrix. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified. Let TPA be the number of true positives of class A, TPB be the number of true positives of class B and TPC be the number of true positives of class C. Then, TPA refers to the positive tuples that were correctly labeled by the classifier in first row-first column i.e. 49. Similarly, TPB refer to the positive tuples that were correctly labeled by the classifier in second row-second column i.e. 47. And, TPC refer to the positive tuples that were correctly labeled by the classifier in third row-third column i.e. 48 shown in Table 2.

The confusion matrix helps us to find the various evaluation measures like Accuracy, Recall and Precision etc.

In diabetes dataset the accuracy parameters have shown in Table 3 and Fig 2. The above chart shows that it have almost equal accuracy measures except ROC Area measure in which SMO has higher accuracy on the diabetes dataset. So, SMO is better method for diabetes[16].

In iris dataset accuracy parameters have shown in Table 4 and Fig 3. Algorithm J48 having lower value than SMO. So SMO is better method for iris dataset.

In tic-tac-toe dataset accuracy parameters have shown in Table 5 and Fig 4. The above chart shows that it have almost equal accuracy measures except ROC Area measure in which SMOhas higher accuracy on the tic-tac-toe dataset. So, SMO is better method for tic-tac-toe dataset.

In Yuta-Selection dataset accuracy parameters have shown in Table 6 and Fig 5. SMO has better accuracy measures except FP rate. So, SMO is better method for Yuta-Selection dataset.

Table 2: Confusion matrix of three classes of Iris

	Predicted class				
Actual Class		A	B	C	Total
	A	49	1	0	50
	B	0	47	3	50
	C	0	2	48	50
	Total				150

Accuracy = (TPA+TPB + TPC)/(Total number of classification)
i.e. Accuracy = (49+47+48)/150 = 96

Table 3: Accuracy on Diabetes

S.No	Parameter	J48	SMO
1	TP Rate	0.73	0.77
2	FP Rate	0.32	0.33
3	Precision	0.73	0.76
4	Recall	0.73	0.77
5	F-Measure	0.73	0.76
6	ROC Area	0.75	0.79

Table 4: Accuracy on Iris

S.No	Parameter	J48	SMO
1	TP Rate	0.98	0.99
2	FP Rate	0.01	0.00
3	Precision	0.98	0.99
4	Recall	0.98	0.99
5	F-Measure	0.98	0.99
6	ROC Area	0.98	0.99

Table 5: Accuracy on Tic-Tac-Toe

S.No	Parameter	J48	SMO
1	TP Rate	0.99	1
2	FP Rate	0.00	0
3	Precision	0.99	1
4	Recall	0.99	1
5	F-Measure	0.99	1
6	ROC Area	0.99	1

Table 6: Accuracy on Yuta-Selection

S.No	Parameter	J48	SMO
1	TP Rate	0.67	0.68
2	FP Rate	0.36	0.43
3	Precision	0.67	0.69
4	Recall	0.67	0.68
5	F-Measure	0.67	0.65
6	ROC Area	0.65	0.66

Table 7: Accuracy measure of J48 and MLP

S.No	Datasets	J48	SMO
1	Diabetes	73.828	77.343
2	Iris	96	96
3	Tic-Tac-Toe	84.551	98.329
4	Yuta-Selection	67.924	68.679

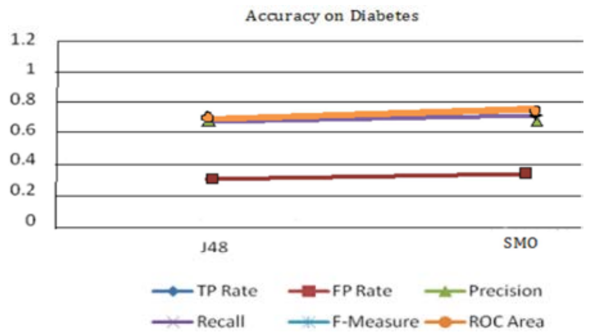

Fig. 2: Accuracy chart on Diabetes

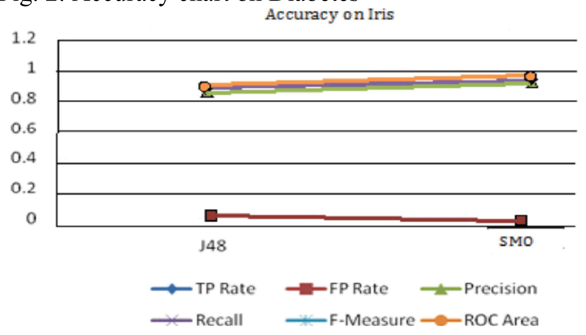

Fig. 3: Accuracy chart on Iris

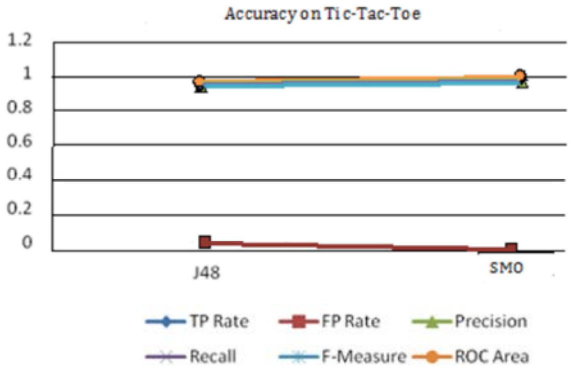

Fig. 4: Accuracy chart on Tic-Tac-Toe

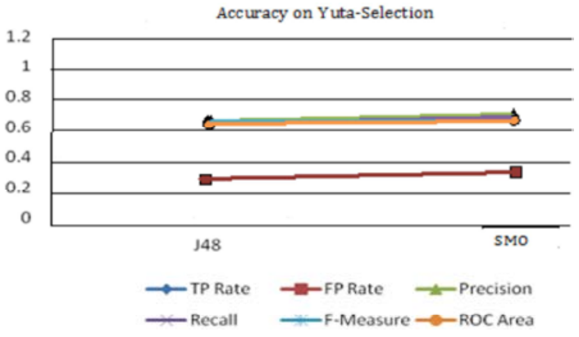

Fig. 5: Accuracy chart on Yuta-Selection

From the values of Table 7 and the chart shown in Fig 6, the accuracy measures are calculated on J48 and SMO algorithms.

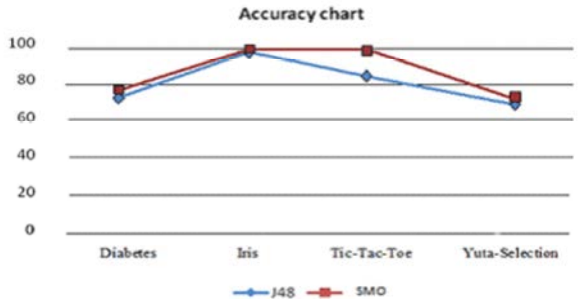

Fig. 6: Accuracy chart of J48 and MLP

The J48 and SMO classification algorithm applies on all the datasets for accuracy measure. From the above chart in Fig 6 it is clear that SMO gives better results for almost 3 datasets and approximate equal accuracy for iris dataset. Hence we can clearly say that SMO is better algorithm than J48 for the given 4 datasets.

CONCLUSION

In this paper, we evaluate the performance in terms of classification accuracy of J48 and Sequential Minimal Optimization algorithms using various accuracy measures like TP rate, FP rate, Precision, Recall, F-measure and ROC Area. Accuracy has been measured on each datasets. On diabetes andtic-tac-toe datasets Sequential Minimal Optimization is clearly better algorithm. On iris and yuta-selection datasets accuracy is almost equal and Sequential Minimal Optimization is slightly better algorithm. Thus we found that Sequential Minimal Optimization is better algorithm in most of the cases. Generally neural networks have not been suited for data mining but from the above results we conclude that algorithm based on neural network has better learning capability hence suited for classification problems if learned properly.

REFERENCES

1. Haiyang, Z., 2011. A Short Introduction to Data Mining and Its Applications, IEEE.
2. Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2nd.
3. Agrawal, R., T. Imielinski and A.N. Swami, 1993. Database Mining: A Performance Perspective, IEEE Trans. Knowledge and Data Engineering, 5(6): 914-925.
4. Quinlan, J.R., 1986. Induction of Decision Trees, Machine Learning, 1(1): 81-106.
5. Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.
6. Bengio, Y., J.M. Buhmann, M. Embrechts and J.M. Zurada, 2000. Introduction to the special issue on neural networks for data mining and knowledge discovery, IEEE Trans. Neural Networks, 11: 545-549.
7. Michie, D., D.J. Spiegelhalter and C.C. Taylor, 1994. Machine Learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence.
8. Quinlan, J.R., 1994. Comparing Connectionist and Symbolic Learning Methods,S.J. Hanson, G.A. Drastall and R.L. Rivest, eds. Computational Learning Theory and Natural Learning Systems, A Bradford Book, MIT Press, 1: 445-456.
9. Shavlik, J.W., R.J. Mooney and G.G. Towell, 1991. Symbolic and Neural Learning Algorithms: An Experimental Comparison, Machine Learning, 6(2): 111-143.
10. Clark, P. and T. Niblett, 1989. The CN2 induction algorithm. Machine Learning, 3(4): 261-283.
11. Freund, Y. and L. Mason, 1999. The alternating decision tree algorithm. In Proceedings of the 16th International Conference on Machine Learning, pp: 124-133.
12. UCI Machine Learning Repository: http://archive. ics. uci.edu/ml/datasets.html
13. Weka: http://www.cs.waikato.ac.nz/~ml/weka/
14. Witten, I.H., E. Frank and M.A. Hall, 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann.
15. Werbos, P.J., 1990. Backpropagation Through Time: What It Does and How to Do It, IEEE.
16. Lu, H., R. Setiono and H. Liu, 1996. Effective Data Mining Using Neural Networks, IEEE.