ANALYSIS OF CLASSIFICATION ALGORITHMS ON DIFFERENT DATASETS

S. Singaravelan, R. Arun, D. Arun Shunmugam, K. Ruba Soundar, R.Mayakrishnan, D. Murugan

(1) Department of Computer Science and Engineering, P.S.R Engineering College, Sivakasi, India
(2) Department of Computer Science and Engineering, P.S.R Engineering College, Sivakasi, India
(3) Department of Computer Science and Engineering, P.S.R Engineering College, Sivakasi, India
(4) Department of Computer Science and Engineering, P.S.R Engineering College, Sivakasi, India
(5) Department of Computer Science and Engineering, Manonmaniam Sundaranar, University, Tirunelveli, India

S. Singaravelan
Department of Computer Science and Engineering, P.S.R Engineering College, Sivakasi, India
singaravelan.msu@gmail.com

Article info
Paper category: Review paper
Received: 20.3.2018.
Accepted: 9.7.2018.
JEL classification: C38

Keywords:
Classification; Data Mining Techniques; Decision Tree; Sequential Minimal Optimization

ABSTRACT

Purpose. *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.*

Methodology. *It can be used for testing several datasets. The performance of J48 and Sequential Minimal Optimization has been analyzed to choose the better algorithm based on the conditions of the datasets. The datasets have been chosen from UCI Machine Learning Repository.*

Findings. *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.*

Originality. *This is the first implemented research work up to my knowledge, data sets classification problem handled in data mining using SMO with Weka interface.*

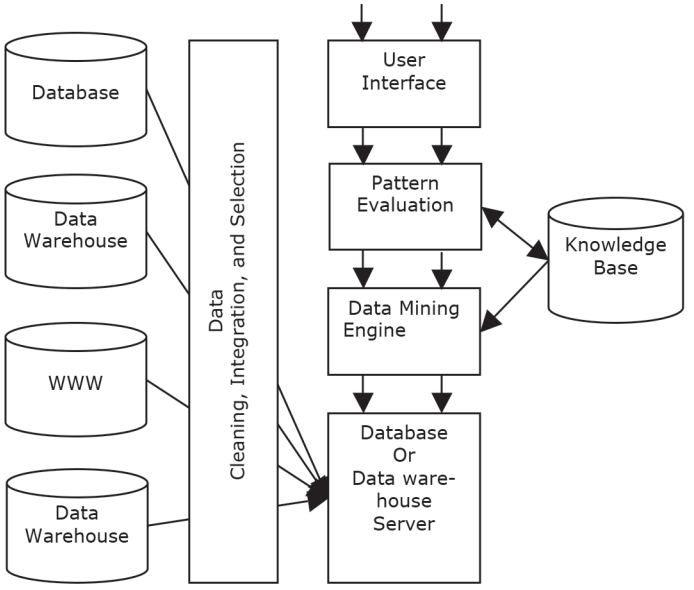
1. INTRODUCTION

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 analyzing the data set, it is used to guess with some certainty while only analyzing 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 (Haiyang, 2011).

The major components of the architecture for a typical data mining system are shown in Figure 1. (Han, 2006). 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 and proper way to exchange information with other systems which is adaptable to users with diverse requirements and change with time.

Figure 1.: Architecture of a Typical Data Mining System



Source: Authors.

2. 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 (Agrawal, 1993). 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 (Quinlan, 1986) and C4.5(Quinlan, 1993), have been devised for decision tree construction.

In (Bengio, 2000) 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 (Michie, 1994). In some cases, neural networks give a lower classification error rate than the decision trees but require longer learning time (Quinlan, 1994., Shavlik, 1991). 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 (Clark, 1989) or to simplify the set of rules corresponding to a decision tree (Quinlan, 1993). The alternating decision tree method (Freund, 1991) is a classification algorithm that tries to combine the interpretability of decision trees with the accuracy improvement obtained by boosting (Sharma, 2013).

Devendra Kumar Tiwari (2014), have comparatively tested four classification algorithms to find the optimum algorithm for classification. The Credit Card Approval dataset has been used for experimental purposes that contain 690 instances with 15 attributes and 1 class attribute to test and justify the differences among classification algorithms. Gupta (2016) explains the analysis of classification and clustering using some terms like Kappa Statistics, Mean Absolute Error, Confusion Matrix, Classification Accuracy correctly classified, incorrectly classified, root mean square error for different algorithms of classification and clustering. This paper considers the most extensively used tools, WEKA tool for this analysis purpose.

Kapur (2017) compared well performing classification algorithms such as Naïve Bayes, decision tree (J48), Random Forest, Naïve Bayes Multiple Nominal, K-star and IBk. Data that they have used is Student dataset and gauge students' potential based on various indicators like previous performances and in other cases their background to give a comparative account on what method is the best in achieving that end. They discussed about various statistical measure used to calculate the performance of each classifier. Neelamegam (2013) overview of several major kinds of classification method including decision tree, Bayesian networks, k-nearest neighbor classifier, Neural Network, Support vector machine are discussed.

Several major kinds of classification algorithms including C4.5, k-nearest neighbor classifier, Naive Bayes, SVM, and IB3 (Archana.S., 2013). This paper pro-

vides a general survey of different classification algorithms and their advantages and disadvantages.

In this paper, different kinds of classification techniques are discussed such as Association Rule Mining, Bayesian Classification, and Decision Tree Classification, nearest neighbor classifier, neural Networks and Support Vector Machine (Bharathi,2014).

3. METHODOLOGY

3.1. Datasets

Data pre-processing, classification, clustering, association, regression and feature selection these standard data mining tasks are supported by Weka. For classification purpose classify tab in Weka Explorer is used (Sudhir, 2013). Advantages of Weka tool:

- Available freely under the GNU General Public License.
- It is portable, as it is implemented in the Java programming language and thus runs on almost any platform.
- It is easy to use due to its graphical user interfaces.

There are four datasets we have used in our paper taken from UCI Machine Learning Repository (ml-repository). The details of each datasets are shown in Table 1.

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

Source: Authors.

In the diabetes dataset (Weka) several constraints were placed on the selection of instances from a larger database. 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

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

3.2. 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 (Witten, 2011). 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, 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).

3.3. 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 (Quinlan, 1993).

In Weka, the implementation of a 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 several classes as in Weka software they become difficult to comprehend and navigate (Werbos, 1990).

3.4. 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 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 linear and 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.

4. 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.

Table 2.: Confusion matrix of three classes of Iris

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

Source: Authors.

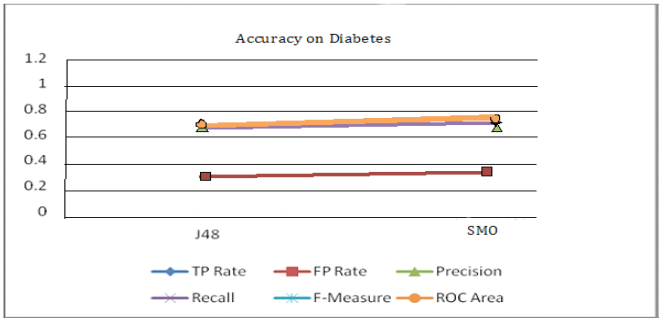
Accuracy = (TPA+TPB + TPC)/(Total number of classification)
i.e. Accuracy = (49+47+48)/150 = 96
The confusion matrix helps us to find the various evaluation measures like Accuracy, Recall, and Precision etc.

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

Source: Authors.

Figure 3.: Accuracy chart on Diabetes



Source: Authors.

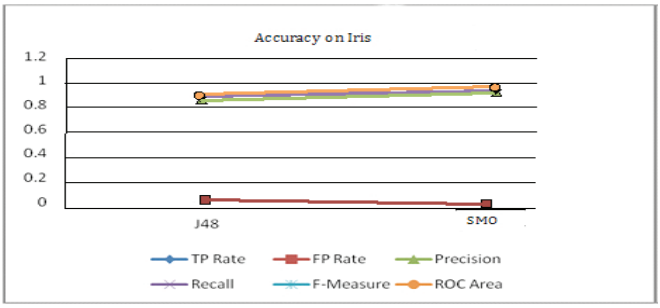
In diabetes dataset the accuracy parameters have shown in Table 3. and Figure 3. The above chart shows that it has almost equal accuracy measures except ROC Area measure in which SMO has higher accuracy on the diabetes dataset. So, SMO is better method for diabetes.

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

Source: Authors.

Figure 4.: Accuracy chart on Iris



Source: Authors.

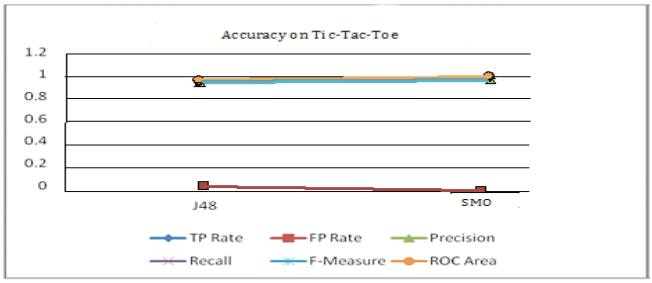
In iris dataset accuracy parameters have shown in Table 4. and Figure 4. Algorithm J48 having lower value than SMO. So, SMO is better method for iris dataset.

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

Source: Authors.

Figure 5.: Accuracy chart on Tic-Tac-Toe



Source: Authors.

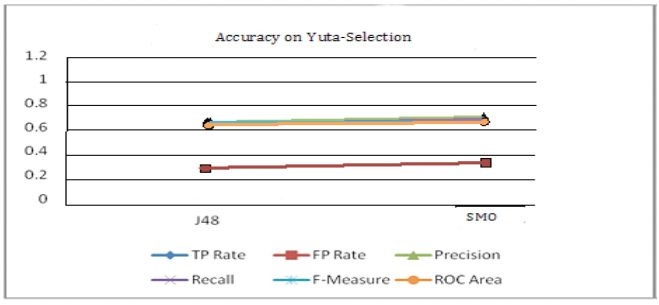
In tic-tac-toe dataset accuracy parameters have shown in Table 5. and Figure 5. The above chart shows that it has almost equal accuracy measures except ROC Area measure in which SMO has higher accuracy on the tic-tac-toe dataset. So, SMO is better method for tic-tac-toe dataset.

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

Source: Authors.

Figure 6.: Accuracy chart on Yuta-Selection



Source: Authors.

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

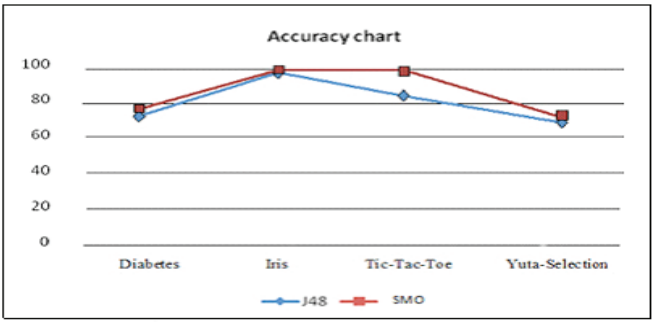
Table 8.: Accuracy measure of J48 and MLP

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

Source: Authors.

From the values of Table 8 and the chart shown in Figure 8., the accuracy measures are calculated on J48 and SMO algorithms.

Figure 8: Accuracy chart of J48 and MLP



Source: Authors.

The J48 and SMO classification algorithm applies on all the datasets for accuracy measure. From the above chart in Figure 8. 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.

5. 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 dataset. On diabetes, and tic-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.

6. FUTURE SCOPE

For the future work new algorithms from classification can be integrated and much more datasets should be taken or try to get the real dataset from the industry to have the actual impact of the performance of algorithms taken into consideration.

REFERENCES

Haiyang, H., A Short Introduction to Data Mining and Its Applications, 2011 International Conference on Management and Service Science, Wuhan (2011):1-4

Han, J., Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

Agrawal, R., Imielinski, T., Swami, A.N., Database Mining: A Performance Perspective, IEEE Trans. Knowledge and Data Engineering (1993):914-925

Quinlan, J.R., Induction of Decision Trees, Machine Learning (1963):81-106

Quinlan, J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

Bengio, Y., Buhmann, J. M., Embrechts, M., Zurada, J. M., Introduction to the special issue on neural networks for data mining and knowledge discovery, IEEE Trans. Neural Networks, (2000):545-549

Michie, D., Spiegelhalter, D.J., Taylor, C.C., Machine Learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence (1994): 263-266

Quinlan, J.R., Comparing Connectionist and Symbolic Learning Methods, In Proceedings of a workshop on Computational learning theory and natural learning systems (vol. 1): constraints and prospects: constraints and prospects, Stephen José Hanson, Ronald L. Rivest, and George A. Drastal (Eds.). MIT Press (1993): 445-456

Shavlik, J.W., Mooney, R.J., Towell, G.G., Symbolic and Neural Learning Algorithms: An Experimental Comparison, Machine Learning (1991):111-143

Clark, P., Niblett, T., The CN2 induction algorithm, Machine learning (1989):261-283

Freund, Y., Mason, L., The alternating decision tree algorithm, In Proceedings of the 16th International Conference on Machine Learning (1999):124-133

UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets.html

Weka: http://www.cs.waikato.ac.nz/~ml/weka/

Witten, I. H., Frank, E., Hall, M. A., Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann, 2011.

Werbos, P.J., Backpropagation Through Time: What It Does and How to Do It, in Proceedings of the IEEE (1990):1550-1560

Devendra Kumar Tiwary., A Comparative study of classification algorithms for credit card approval using weka, GIIRJ (2014):165-174

Shivangi Gupta., Comparative Analysis of classification Algorithms using WEKA tool, International Journal of Scientific & Engineering Research (2016):2014-2018

Bhrigu Kapur, Comparative Study on Marks Prediction using Data Mining and Classification Algorithms, IJARCS (2017):394-402

Trilok Chand Sharma, Manoj Jain, WEKA Approach for Comparative Study of Classification Algorithm, International Journal of Advanced Research in Computer and Communication Engineering (2013):1925-1931

Sudhir B., Jagtap., Kodge, B. G., Census Data Mining and Data Analysis using WEKA, International Conference in Emerging Trends in Science, Technology and Management, (2013):35-40

Archana, S., Elangovan, K., Survey of Classification Techniques in Data Mining, International Journal of Computer Science and Mobile Applications, (2014):65-71

Bharathi, A., Deepankumar, E., Survey on Classification Techniques in Data Mining, International Journal on Recent and Innovation Trends in Computing and Communication, (2014):1983-1986

Neelamegam, S., Ramaraj, E., Classification algorithm in Data mining: An Overview, International Journal of P2P Network Trends and Technology, (2013):1-5