International Journal of Artificial Intelligence,
ISSN 0974-0635; Int. J. Artif. Intell.
Atumun (October) 2012 , Volume 9, Number A12

A Comparison Study between Data Mining Algorithms over Classification Techniques in Squid Dataset

Fartash. Haghanikhameneh[1], Payam. Hassany Shariat Panahy[2],

Nasim. Khanahmadliravi[3], Seyed Ahmad. Mousavi[4]

Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400, UPM, Serdang, Selangor

fartashh@gmail.com
payam_shp49@yahoo.com
nasim_khanahmad@yahoo.com
sam683@gmail.com

ABSTRACT

Classification is one of the most important supervised learning techniques in data mining. Classification algorithms can be extremely beneficial to interpret and demonstrate bandwidth usage pattern and predict the required bandwidth for different groups in distinct time interval, having the intention of improving efficiency. The dataset used in this study was collected over a year from a Squid proxy server's log file, on access.log file, from a computer institute. This study compares various classification algorithms to predict the bandwidth usage pattern in different time intervals among different groups of users in the network. Different classification algorithms including Decision Tree and Naïve Bayesian are compared using Orange, a data mining tool. The results of the experiment showed that the Decision Tree algorithm achieved 97% *accuracy and efficiency in predicting the required bandwidth inside the network.*

Keywords: Data Mining Algorithms; Classification Technique; Orange, Network Traffic, Bandwidth.

Computing Classification System: H.2.8

1. INTRODUCTION

Data mining is the process of discovering interesting knowledge from various perspectives and summarizing it in a useful form of information(Wahbeh et al., 2011, Shomona, 2011) .In fact, finding the hidden knowledge in the database is the major task of data mining (Santhi and Bhaskaran, 2010).Classification has been identified as a data mining technique for predicting classifier based on the proposed model(Han and Kamber, 2001). The proposed work will focus on various classification algorithms for extracting usage pattern among different groups of users, over different time intervals in the network, depending on the network traffic.
Squid proxy server is one of the most famous tools, which is used to manage the bandwidth inside the network (ElAarag and Romano, 2009); however Squid has lot of functionality such as traffic management, caching, and etc., this experiment focuses of its ability on managing the bandwidth (Wessels, 2004). Squid helps network administrators to define different groups based on their physical address, or IP address, and share the income bandwidth by allocating the fix amount of bandwidth to each group(Spare, 2001).

According to this problem this experiment focuses on finding effective and efficient algorithm to extract usage pattern for each group in different time for predicting the required bandwidth.

Although, it is not impossible to find out the amount of required bandwidth inside the network, detecting and calculating the required bandwidth for different groups in each time interval is complex and tedious job (Lee et al., 2006). Also, illustrating and determining the actual network traffic size is hyperbolic. Due to different usage pattern in various hours of a day, also depending on the day of the week, setting the fix amount of bandwidth for different groups of computers is not appropriate. Moreover, during periods of heavy traffic, assigning a fixed bandwidth for each coming connection is not reasonable while free bandwidth is available in the network (Wang and Li, 2008). Based on the given goal, this research mostly focuses on comparing different classification algorithms including Decision Tree and Naïve Bayesian for predicting efficient network traffic, using orange tool. To come up with this problem, firstly data was collected from the network for cleansing and preprocessing. Dataset used in this paper is Squid dataset consist of 9 attributes and 3000000 samples. At the time of data collection for analysis, the available Internet bandwidth on the institute was 2 Mbps, and it was shared by over 100 users in 4 different groups.

The main problem is comparative study of classification algorithms including Decision Tree and Naïve Bayesian using Squid dataset containing 9 attributes and 3000000 instances.

The organization of this paper is as follows. In Section 2, the classification techniques, Squid dataset collection, squid dataset preprocessing, building classifiers, measure for performance evaluation have been introduced. Section 3 discussed the results of the classification algorithms for predicting the bandwidth usage pattern in different time intervals among different groups of users over the network. Finally, section 4 includes conclusion and future works.

2. PRELIMINARY, DEFINITIONS AND CONCEPTS

Classification is a supervised learning or class prediction technique in data mining having the ability to give a specific rule to assign on new classes or samples (Ngai et al., 2009). Various classifications' algorithms such as support vector machines, k Nearest Neighbors, weighted voting and Artificial Neural Networks can be applied on data records belonging to a class for discovering set of models to predict the unknown class label. In fact, in classification data records are divided in to two sets randomly, called training set (dependent set) and test set (independent set). Data mining algorithm is applied on training set in order to use the predicting model on the testing set (Han and Kamber, 2001, Selvaraj and Natarajan, 2011).

In this study, selected attributes are Time Stamp, Response Time, IP Address, Transfer size, and object class label in network traffic. Every record in access.log file includes some attributes such as time stamp, response time, IP address, transfer size, method, requested address, requested IP address, content type, which are used to extract the bandwidth usage pattern. Classification technique was applied using $C4.5$ classifier and Naïve Bayesian classifier. At the first step of classification, the model is built on the training set with known class label and in the second step; the proposed model is applied by assigning class labels on the test set. Finally, based on applying different algorithms on dataset, accuracy of these algorithms were determined and compared. This paper compares the accuracy of two classification algorithms, namely Decision Tree and Naïve Bayesian through experimental study, for predicting bandwidth allocation in high usage times.

2.1. *Squid dataset collection*

Squid is a precious source of information with excellent performance, having the abilities to record access information, errors of system configuration and resource which use the memory. It has different log files such as explicitly activated during compile time and deactivated during run-time. All log file in squid have some common point such as time stamp, cache.log e.tc.,(Rousskov, 2012). The real dataset used in this study was collected over a year from a Squid proxy server's log file, on access.log file, from a computer institute which consists of 9 attributes and 3000000 instances to compare two classification algorithms, Naïve Bayesian and Decision Tree. A complete list of squid attributes is presented in Table1. For the rest of this paper we name the dataset as squid dataset.

Table1. Complete List of Attributes

Attributes	***Data Type***	***Description***
Time Stamp	Ratio	The time when the request is completed (socket closed). The format is "Unix time" with millisecond resolution(Lin and Choi, 2010).
Elapsed	Ratio	The elapsed time of the request, in milliseconds. This is the time between the accept () and close () of the client socket(Lin and Choi, 2010).
Client	Nominal	The “Client” field in the “access.log” file shows the IP address of the connecting client (Lin and Choi, 2010).
Action Code	Nominal	The HTTP reply code taken from the first line of the HTTP reply header(Lin and Choi, 2010).
Transfer Size	Ratio	For TCP requests, the amount of data written to the client. For UDP requests, the size of the request. (in bytes)(Lin and Choi, 2010).
Method	Nominal	The HTTP request method (GET, POST, etc), or ICP_QUERY for ICP requests(Lin and Choi, 2010).
URI	Nominal	The requested URI(Lin and Choi, 2010).
From	Nominal	Hostname of the machine where we got the object(Lin and Choi, 2010).
Content Type	Nominal	Content-type of the Object (from the HTTP reply header)(Lin and Choi, 2010).

2.2. *Squid dataset preprocessing*

Due to huge size of the data in the real world databases, they have missing, noisy and inconsistent data. Data preprocessing is essential for improving data quality (Han and Kamber, 2001). Data cleaning, integration, reduction and transformation are different techniques for data preprocessing (Han and Kamber, 2001, Santhi and Bhaskaran, 2010,). Before applying classification algorithms over the dataset, some preprocessing steps were performed.

At first step, data reduction is done by selecting the most informative attributes in a dataset, while attempting to lose no information required for the task at hand. So, the cleaning dataset consists of four attributes; Time Stamp, Response time, IP address and Transfer Size. At the second step, since the original time stamps and IP addresses are not able to be treated as computable parameters, they are converted to an appropriate format in order to use for some mathematical analysis. For instance, a time stamp turned into two parts which are normal date format (mm/dd/yyyy) and time (hour: minute: second). Also IP address has been converted to an integer number by using mathematical formula. Besides, according to the significant role of dates, date is divided into month; day and year as well as times are normalized by Min-Max normalization to make it much more meaningful for computation. Next, equal-width binning method is applied on three types of attributes; IP address into 4 groups to categorize the network zones, time into seven intervals and traffic into three classes. Due to enormous tuples in dataset which slows down the process of computation, data reduction is done for reducing the size without losing quality. An example is total “transfer size” and “total response time” for a particular group over a specific year, month, day and interval which are calculated by a summation of total transfers among those tuples in the same year, month, day and interval. At this level, each tuple is classified into high, normal or low classes based on division of total transfer time to total transfer size. The effect of reduction was to decrease the number of tuples from 3 million to 317284 .Consequently; 317284 tuples are used to classify the appropriate classes to start data mining. A list of attributes after preprocessing is presented in Table 2.

Table2. List of Attributes after Data Preprocessing

Attributes	*Data Type*	*Description*
Day	Nominal	Number of the week day (1-30)
Month	Nominal	Number of the months (1-12)
Year	Nominal	Number of the year
Group	Nominal	Four Groups (1, 2, 3, 4)
Total Response Time	Ratio	Amount of response
Total Transfer Size	Ratio	Amount of transfer
Class	Nominal	Normal , High, Low

2.3. Building classifiers

Bayesian classifiers is a supervised learning which can predict the probability that a given tuple belongs to a particular class (Batchu. et al., 2011). Naïve Bayesian classifiers have been proven as a powerful probabilistic model for solving classification problems (Geenen. et al., 2010).
For any given instance $X = \left(_{X1,X2,...Xn}\right)$, where $_{X1}$ is the value of attribute $_{X1}$, $P\langle C|X\rangle$ is calculated by Bayesian classifier for all possible class values C and predicts $C^* = \text{argma}_{xc} p\langle x|c\rangle$ as the class value for instance X. Hence, estimating a $P\langle X|C\rangle$ which is proportional to $P\langle X|C\rangle P(C)$ is the key step of a Bayesian classifier (Wong and Chang, 2010).

According to (Breiman et al., 1984, Han and Kamber, 2001, Xu et al., 2011) Decision Tree is a famous supervised learning and predictive model. Likes a tree structure, where each node denotes a test on an attribute value, leaves represent classes or class distribution that predict model for classification or regression of predictors. Branches represent conjunctions of features that lead to those classes. This structure has great potential to convert to classification rules. By applying algorithm to the entire dataset, the dataset will be treated as a single large set and then proceeds to recursively split the set. The algorithm applies the top down approach to construct the tree until some stopping criterion is met. This algorithm uses the gain in entropy to find out how to create the nods of the tree (Hamou. et al., 2011). Frequently partitioning the input space constructs the decision tree, so that the partition form a tree structure (Kováč, 2012).

Although Naïve Bayesian classifier provides short training time (computational time), fast evaluation and has proven to be suited for real world problems, solving complex classification problems is not possible with naïve Bayesian classifier. On the other hand, Decision tree can produce reasonable and interpretable classification trees which are used for making decision purposes. However, this algorithm is creating complex tree that cannot be generalized to all data as well.

The main question is how Decision Trees are used for classification? Given a tuple, X, for which the associated class labels, is unknown; the attribute values of the tuple are tested against the Decision Tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision Tree can easily be converted to classification rules (Batchu. et al., 2011).

2.4. Measure for performance evaluation

The data is evaluated based on three important performance measurements as follow:

I. Precision and Recall are two significance performance measure for evaluating classification algorithms (Cios et al., 1998) In this experiment Precision refers to proportion of data which is classified correctly using classification algorithm. Also, Recall refers to percentage of information which is relevant to the class and is correctly classified.

II. Accuracy is percentage of instances which is classified correctly by classifiers is called Accuracy (Han and Kamber, 2001).
III. F-Measure as (Kumar and Rathee, 2011) explained, performance metric is another F-Measure which combines Recall and Precision into a single measure.

3. EXPERIMENTAL RESULT, ANALYSIS AND PERFORMANCE EVALUATION

In this experiment the comparative study of Decision Three and Naïve Bayesian over Squid dataset is shown. During this experiment the preprocessed data set in CSV (comma separated format) file was applied to Orange tool as an input.

At the first step we divided the input dataset in 10 separate folds to apply Decision Tree and Naïve Bayesian operations to each fold separately. The results are shown in Table 3 and Table 4. As can be seen the average accuracy of these 10 folds for Decision Three and Naïve Bayesian are 97.38% and 43.81% respectively. Hence, Decision Three depicts more accuracy.

Table3. The Results of Decision Tree Algorithm for 10 Folds

FOLD	TRAINING DATASET COUNT	TESTING DATASET COUNT	CLASSIFICATION ACCURACY	F-MEASURE	Precision	RECALL
Fold1	285554	31729	0.9761	0.9753	0.9795	0.9712
Fold2	285554	31729	0.9736	0.9708	0.9737	0.9678
Fold3	285554	31729	0.9712	0.9675	0.9673	0.9677
Fold4	285555	31728	0.9714	0.9682	0.9696	0.9668
Fold5	285555	31728	0.9795	0.9751	0.9707	0.9795
Fold6	285555	31728	0.9739	0.9711	0.9690	0.9733
Fold7	285555	31728	0.9719	0.9695	0.9679	0.9710
Fold8	285555	31728	0.9748	0.9736	0.9726	0.9747
Fold9	285555	31728	0.9720	0.9683	0.9694	0.9672
Fold10	285555	31728	0.9740	0.9702	0.9690	0.9714
Average of classification Accuracy			0.97384			

Table4. The Results of Naïve Bayesian Algorithm for 10 Folds

FOLD	TRAINING DATASET COUNT	TESTING DATASET COUNT	CLASSIFICATION ACCURACY	F-MEASURE	Precision	RECALL
Fold1	285554	31729	0.4349	0.3041	0.4127	0.2407
Fold2	285554	31729	0.4354	0.3020	0.4215	0.2353
Fold3	285554	31729	0.4416	0.3092	0.4348	0.2400
Fold4	285555	31728	0.4423	0.3037	0.4161	0.2391
Fold5	285555	31728	0.4375	0.3084	0.4207	0.2434
Fold6	285555	31728	0.4461	0.3093	0.4271	0.2425
Fold7	285555	31728	0.4375	0.3133	0.4314	0.2460
Fold8	285555	31728	0.4319	0.3029	0.4098	0.2403
Fold9	285555	31728	0.4352	0.2977	0.4140	0.2324
Fold10	285555	31728	0.4387	0.4533	0.4584	0.4484
Average of classification Accuracy (Hand Folding)			0.43811			

Moreover, we built confusion matrixes for Decision Three and Naïve Bayesian (Table 5 and 6).

As Table 5 illustrates, there are 96096 items are classified into class Low using Decision Three algorithm,

- 93256 of these items are correctly classified into class Low,
- 2190 of these items are wrongly classified into class Normal,
- 650 of these items are wrongly classified into class High.

The same evaluation can be used for Normal and High traffic.
In addition to this, according to Table 6, there are 96096 items are classified into class Low using Naïve Bayesian algorithm,

- 23087 of these items are correctly classified into class Low,
- 48800 of these items are wrongly classified into class Normal,
- 24209 of these items are wrongly classified into class High.

The same evaluation can be used for Normal and High traffic.
Consequently, Decision Three Algorithm shows more accuracy in comparison with Naïve Bayesian algorithm over the Squid dataset using Orange tool. Naïve Bayesian algorithm operates on probability which can be High, Normal and Low during a day. The probability of traffic almost remains the same in a day for different groups. Hence, the accuracy of Naïve Bayesian algorithm is low in this case.

Table5. Confusion Matrix using Decision Tree

	L	*N*	*H*	
L	**93256**	2190	650	**96096**
N	1725	**117285**	626	**119636**
H	1031	2081	**98439**	**101551**
	96012	**121556**	**99715**	**317283**

Table6. Confusion Matrix Using Naïve Bayesian

	L	*N*	*H*	
L	**23087**	48800	24209	**96096**
N	19211	**71439**	28986	**119636**
H	12303	44321	**44927**	**101551**
	54601	**164560**	**98122**	**317283**

As the table clearly illustrates, the accuracy, precision and recall of Decision Tree are 97.38%,97.13%,97.04% respectively. Besides, the accuracy, precision and recall of Bayesian algorithm are 93.95%,42.28% and 24.02% respectively. It can be concluded Decision Tree shows better performance in comparison with Naïve Bayesian.

At the second step, Decision Three and Naïve Bayesian operations were applied to the dataset. The Orange tool automatically folded the dataset into 10 folds to run the process and get the results. The result is shown in Table 7.

Table7. The Results of Naïve Bayesian and Decision Tree Algorithms by Automatic Folding

Method	*Classification Accuracy*	*F Measure*	*Precision*	*Recall*
Decision Tree	0.9738	0.9709	0.9713	0.9704
Naïve Bayesian	0.4395	0.3064	0.4228	0.2402

4. CONCLUSION AND FUTURE WORK

This study has conducted a comparison between two classification algorithms namely; Decision Tree $(C4.5)$ and Naïve Bayesian on Squid, using Orange tool. The presented study illustrated that Decision Tree algorithms had 97% accuracy, 97% precision and 97% recall. On the other hand, Naïve Bayesian indicated 43% accuracy, 30% precision and 42% recall. By comparing the three evaluate parameters for the two algorithms it is concluded that Decision Tree has highest performance than Naïve Bayesian over the dataset using manual and automatic folding to adjust the required bandwidth inside the network. As future research, we are going to test other classification algorithms and to do comparison among them and applying these algorithms by other data mining tools for confirming the result.

5. REFERENCES

Batchu, V., Aravindhar, D.J., Thangakumar, J., Masillamani, M.R., n.d. 2011. A Classification based Dependent Approach for Suppressing Data. *International Journal of Computer Application (IJCA)* **1**:14-16.

Breiman, L., Friedman, J. H, Olshen, R.A, Stone, C.J. 1984. Classification and regression trees, *Belmont. CA, Wadsworth*.

Cios, K. J., Pedrycz, W., Świniarski, R., Swiniarski, R. 1998. Data mining methods for knowledge discovery. *USA, Springer*.

ElAarag, H., Romano, S. 2009. Improvement of the neural network proxy cache replacement strategy. *In Proceedings of the 2009 Spring Simulation Multiconference, Society for Computer Simulation International (SSM'09), San Diego, California*. **90**: 1-8.

Geenen, P.L., van der Gaag, L.C., Loeffen, W.L.A., Elbers, A.R.W. 2011. Constructing naive Bayesian classifiers for veterinary medicine: A case study in the clinical diagnosis of classical swine fever. *Research in Veterinary Science* **91**: 64–70.

Han, J., Kamber, M. 2001. Data mining: concepts and techniques, *San Francisco, CA. Morgan Kaufmann 5.*

Hamou, A., Simmons, A., Bauer, M., Lewden, B., Zhang, Y., Wahlund, L.O., Westman, E., Pritchard, M., Kloszewska, I., Mecozzi, P. 2011. Cluster Analysis of MR Imaging in Alzheimer's Disease using Decision Tree Refinement. *International Journal of Artificial Intelligence* **6**: 90–99.

Kováč, S. 2012. Suitability analysis of data mining tools and methods. Bachelor Thesis, Masaryk University, Brno, Czech Republic, 16-23.

Kumar, V., Rathee, N. 2011. Knowledge discovery from database using an integration of clustering and classification. *International Journal of Advanced Computer Science and Applications* **2:** 29-33.

Lee, J. F., Chen, M. C., Ko, M. T., Liao, W. 2006. Bandwidth allocation algorithms for weighted maximum rate constrained link sharing policy. *Information Processing Letters* **97:** 238-243.

Lin, H.C., Choi, M. S.,n.d. 2010. Mining web usage within a local area network. *IACSIT International Journal of Engineering and Technology*, **2:** 435-441.

Ngai, E. W. T., Xiu, L., Chau, D.C.K. 2009. Application of data mining techniques in customer relationship management: *A literature review and classification. Expert Systems with Applications* **36:** 2592-2602.

Rousskov, A. 2012. Squid-cache [Online]. Available: http://wiki.squid cache.org/SquidFaq/SquidLogs#Squid_Log_Files [Accessed 2 February 2012].

Santhi, P., Bhaskaran, V. M. 2010. Performance of clustering algorithms in healthcare database. *International Journal for Advances in Computer Science* **2:** 26-31.

Selvaraj, S., Natarajan, J. 2011. Microarray data analysis and mining tools. *Bioinformation* **6:**95-99.

Shomona, G.J. 2011. Discovery of knowledge patterns in clinical data through data mining algorithms*: Multi-class Categorization of Breast Tissue Data. International Journal* Computer Applications **32:** 46-53.

Spare, I. 2001. Deploying the squid proxy server on linux. *Linux Journal,* 2001**,** 5.

Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., Al-Shawakfa, E. M. 2011. A comparison study between data mining tools over some classification methods. *International Journal of Advanced Computer Science and Applications (IJACSA)* **2:**18-26.

Wang, D., Li, G. 2008. Efficient distributed bandwidth management for MPLS fast reroute. *Networking*, *IEEE/ACM Transactions on***16:** 486-495.

Wessels, D. 2004. Squid: the definitive guide*,* Sebastopol. *California, O'Reilly & Associates*, Inc.

Wong, T. T., Chang, L. H. 2010. Individual attribute prior setting methods for naive Bayesian classifiers. *Pattern Recognition* **44:** 1041-1047.

Xu, Y., Dong, Z. Y., Zhang, R., Wong, K.P. 2011. A decision tree-based on-line preventive control strategy for power system transient instability prevention. *International Journal of Systems Science***, 1:** 1-11.