Volume 2, Issue 3, March 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey of Feature Selection Technique in Internet Traffic Data

Hardeep Singh
UIET, Panjab University, Chandigarh
Hardeep0808@gmail.com

Harish Kumar
UIET, Panjab University, Chandigarh

Abstract— ***Area of network traffic classification using application of machine learning has been increased enormously in recent years. Network traffic classificationis necessary today because of increase in no of users today in the internet and quality of service in the network. Network traffic classification algorithm works on various network traffic features. So in a huge amount of network traffic data not every feature is relevant. So a irrelevant feature increase the time of classification algorithm. So feature selection is needed to reduce the dimensionality of feature space and reduce the computational time of classifier. In this paper, different types of features selection method are used in traffic classification are presented.***

Keywords*— *Genetic Algorithm, Feature Subset Selection, Correlation Based Feature Selection,* *Classification Metrics, Feature Evaluation

I. INTRODUCTION

Network traffic measurement has recently gained more interest as an important network-engineering tool for networks of multiple sizes. Different applications such as worms and viruses or simply such as Web, malicious P2P affect the underlying network infrastructure. The traffic mix flowing through most backbones and links needs to be characterized in order to achieve a thorough understanding of its actual composition[1].

Feature selection has been a fertile field of research and development since the 1970s in machine learning and data mining, statistical pattern recognition as well as widely applied to solve the problems such as image retrieval ,text categorization, intrusion detection, customer relationship management in the network.

Feature selection is one of the important and frequently used techniques in data preprocessing for data mining. It reduces the number of features, or noisy data, and removes irrelevant, redundant and speed up a data mining algorithm, improving mining performance such as result comprehensibility and predictive accuracy in the network[1].

A number of researchers are looking closely at the application of Machine Learning to IP traffic classification. The application of Machine Learning techniques have a number of steps. First, *features* are defined then by these future unknown IP traffic is identified and differentiated. Second Machine Learning techniques help in classifying flows based on application protocol independent statistical features such as inter-arrival times and packet length. Each traffic flow is identified by the same set of features but with the different feature values[2].

Feature selection is a process that selects a subset of original features. Finding an optimal feature subset is usually intractable. Here the number of features increases proportional to the dimensionality of a domain in network traffic data .The optimality of a feature subset is decided by an evaluation criterion. A typical feature selection process consists of four basic steps that are shown in Figure 1. First steps is subset generation then second is subset evaluation there after stopping criterion and result validation. Subset generation is a search processs which produces candidate feature subsets for evaluation based on a certain method. Each candidate subset is evaluated and then compared with the previous best one according to a certain evaluation criterion used in the network . If the new subset is better then it replaces the previous best subset. The procedure of subset generation and evaluation is repeated until a some stopping criterion is met in the network. Then the selected best subset needs to be validated usually by prior knowledge. Feature selection can be found in many areas of data mining such as regression classification, association rules, and clustering[2]

II. TRAFFIC CLASSIFICATION METRICS

There are mainly three metrics used to evaluate the machine learning algorithms. If a classifier is trained to identify members of class *X* then:

1) Accuracy: the percentage of correctly classified instances over the total number of instances in the network.

2) Precision: the proportion of the instances which truly have class *X* among all those classified as class *X* in the traffic.

3) Recall: the proportion of class *X*'s instances which are correctly classified as belonging to class *X* in the traffic[3].

III. FEATURE SELECTION ALGORITHMS

Feature selection algorithms is broadly classified into filter method or wrapper method. Filter method algorithms make use of independent assessment that is based on some general characteristics of the data. They rely on a some metric to rate and select that best subset one before the learning starts. The

results obtained should not biased towards a particular machine learning algorithm. The wrapper model uses predetermined mining algorithms and uses its performance as the evaluation criterion. It searches for the features that better suit to the mining algorithm to improve mining performance, but it is more computationally expensive than filter method. Advantages of filter method are that they scale to very high-dimensional datasets easily and they are computationally fast and simple, also they are independent of classification method. So feature selection to be performed only once, and then the different classifiers can be evaluated. Advantages of wrapper method include interaction between feature subset search and model selection, and account totake feature dependencies. Drawback of these techniques are very computationally intensive if classifier has a high computational cost and they have a risk of overfitting . Table 1 provides a summary of the 249 per-flow features.

Table 1. Summary of some per-flow features [9]

1	Packet inter-arrival time (mean, variance, 1st and 3rd quartiles, median, minimum, maximum…)
2	Total packets (in each direction and total for flow)
3	Flow metrics (duration, packet-count, total bytes)
4	Size of TCP/IP control fields (mean, variance, 1st and 3rd quartiles, median, minimum, maximum…)
5	Total number of ACK packets seen carrying SACK information, minimum observed segment size
6	Effective bandwidth based upon entropy
7	Ranked list of top-ten Fourier-transform components of packet inter-arrival times

A. Clustering Ensembles Guided Feature Selection Algorithm (CEFS).

The main idea of CEFS[4]is to search for a subset of all features in the internet traffic such that the clustering algorithm trained on this feature subset can achieve the most similar clustering solution to the one obtained by a clustering ensembles method in the network.

CEFS has several advantages when compared with other existing unsupervised feature selection algorithms that are applied I network traffic: First, most existing unsupervised feature selection algorithms are dimensionality-biased in the network. For example, if scatter separability based feature selection algorithm is adopted,then high-dimensional feature subsets are selected more easily in the network. But CEFS does not bias to high-dimensional nor low-dimensional feature subsets. Because CEFS evaluates a candidate feature subset according to quality of the clustering solution obtained by the clustering algorithm trained on it. So it is independent of the dimension of the candidate feature subset. Second, CEFS leverages the consensus across multiple clustering solutions. in the internet traffic. So, it is able to obtain a more robust and stable feature subset when compared with other nonensembled unsupervised feature selection algorithms in the network.

B. Generic Algorithm based approach

This algorithm, which is an effective optimization method in wide search spaces, is preferred because it is the appropriate method for the solution of the problem in the network. To apply the genetic[5] algorithm, the problem should first be adapted to the genetic algorithm. In other words we can say that, the basic structures of the genetic algorithm, such as genes, chromosomes, and population, should be determined based on feature in the network. In this phase, coding, selection, crossover, mutation, and fitness functions should be chosen Individual's encoding :In the GA-based approach to feature selection, a candidate feature set in the network can be represented by a binary string called a chromosome. Chromosomes[8] comprising population are encoded in the form of binary vector in a manner to compose of genes as the number of feature in each feature space in the network. The ith bit in the chromosome represents the presence of the ith feature. Example is if chromosome corresponding to packet length have bit 1 then it means packet length is presents in the chromosome as a feature. Initialization of the population is commonly done by seeding the population with random values. If the value of the gene, which is coded in binary system, is ''1'', it means that the corresponding feature is selected, in the contrary, if the value of gene is ''0'', it means that the corresponding feature is not selected in the network.

Fitness function: Fitness function is used to decides which individuals are good to fit to optimum solution. Every individual has its own fitness value. A higher value of fitness means that the individual is more appropriate as a problem solution in the network; on the other hand, a lower value of fitness means that the individual is less appropriate as a problem solution in the network.

Selection**:** The objective of the selection process is to choose the individuals of the next generation according to the selected fitness function and selection method among the existing population selected. In the selection process, the transfer possibility of the fittest individual's chromosome to the next generation is higher than others. The decision of the individual's characteristic which will be transferred to the next generation is based on the values evaluated in internet traffic.

Crossover: In the pre-crossover phase, individuals are determined by using a mating process. Forming the new generation is called 'crossover'. The most widely used method is forming two new individuals from the two chromosomes in the network

Mutation: To increase the variety of the chromosomes which are applied on crossover, process mutation process can be applied. Mutation introduces local variations to the individuals for searching different solution spaces and keeps the diversity of the population in the network.

C. Correlation based feature selection

We measure the total amount of information enclosing in a feature is summation of inter-correlations to all of the rest of the features in the network, but CBFs only considers on a feature of rest ones at a time. Therefore, FCBF may be used in situation where the dependence between pair of features is weak but the total inter-correlated strength of one feature to the others is strong in the network. The result is that FCBF[7] possibly keeps a feature that its information can be found in the remaining selected subset of features in the network. We apply symmetric uncertainty measure to pairs of features in the network. If the measure of mutual information between a pair of features is low in the network then, it represents these two features are independent to each other. For each pair of features in the network, one feature only contains a little information about the other, i.e.so here knowing one feature cannot give any information about the other. Also the two features are highly inter-correlated with each other if they have a high mutual information measure in the network. It means that one feature contains lots of information about the other and so knowing one feature can provide necessary information about the other.Example In a network avg. packet length and packet size are correlated .So there is need to calculate mutual information between them. Under this circumstance, one of them can be considered as a redundant feature and can be discarded.

D. Methodology

First, the system captures all the packets passing through it, and aggregates them into traffic flow according to the 5- tuple (*i.e.*, source and destination address, source and
destination port and protocol). Second, the header information of each packet is collected and stored to the corresponding flow database. Thereafter flow statistics is computed in term of the each feature, to establish the statistics database. Third, we used feature selection to eliminate the redundant and irrelevant features, and yield the best feature subset.. Finally, the flow is classified using ML algorithm selected from the ML strategy database, and evaluated by classification accuracy and computational performance metrics. If satisfied, ML modeling is finished and ready for the traffic classification. If not, the process would be repeated until no further improvement is achieved or the evaluation for the specific application is satisfied in the Figure 1.

IV. CONCLUSION

Feature selection is an important academic issue in machine learning and data mining in the network. But the algorithms discussed in this paper mostly have quadratic or higher time complexity. More efficient evaluation criteria are needed for feature selection with large dimensionality. An efficient evaluation criteria more accurate reduce the feature space so that further classifier give good performance in the network.

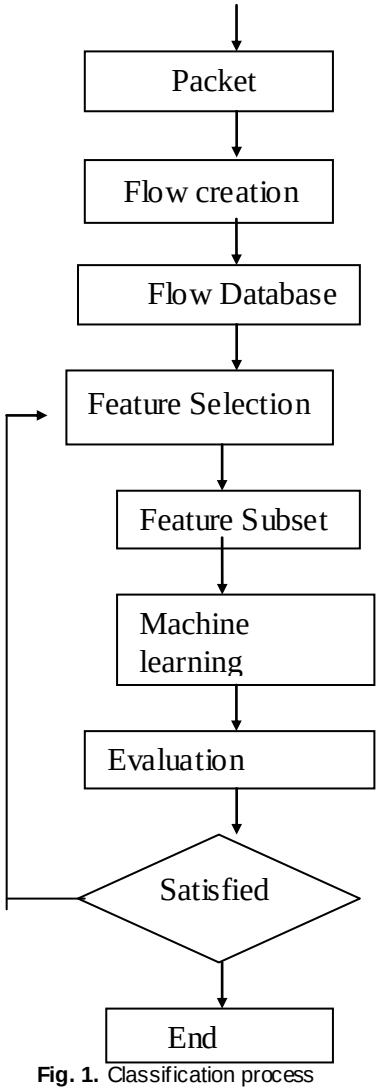

Fig. 1. Classification process

REFERENCES

[1] Huan Liu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions On Knowledge And Engineering, VOL. 17, No. 4, April 2005.

[2] Zhiping Chen, Kevin Lü, A preprocess algorithm of filtering irrelevant information based on the minimum class difference, Knowledge-Based Systems 19 (6) (2006) 422–429.

[3] M.E. ElAlami, "A filter model for feature subset selection based on genetic algorithm" Knowledge-Based Systems 19 (6) (2006) 422–429.

[4] Yi Honga, Sam Kwonga, Yuchou Changb, Qingsheng Renc, "Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm", Proceedings of the IEEE International Conference on Data Mining, 2002, pp. 115—122.

[5] I.-S. Oh, J.-S. Lee, B.-R. Moon, "Hybrid genetic algorithms for feature selection",IEEE Transaction on Pattern Analysis and Machine Intelligence 26 November 11, 2004.

[6] Harun Ug˘uz , "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm", Knowledge-Based Systems 22 (2009) 356–362.

[7] L. Yu, H. Liu, "Feature selection for high-dimensional data: A fast correlation based filter approach", in: Proceedings of the International Conference on Machine Learning, 2003

[8] J. Hurley E. Garcia-Palacios S. Sezer, "Classifying network protocols: a 'two-way' flow approach", Published in IET 8th December 2009

[9] Nguyen T T T, Armitage G. A survey of techniques for internet traffic classification using machine learning. IEEE Communications on Surveys and Tutorials.