RESEARCH ARTICLE OPEN ACCESS

# Internet Traffic Classification Based on Constrained Based Clustering Technique and Genetic Algorithm

Divyashailley Jain[1], Kamlesh Chandravanshi[2]
1(Technocrats Institute of Technology, Bhopal)
2 (Computer Science Engineering, Technocrats Institute of Technology, Bhopal)

**Abstract:**

The clustering technique plays an important role in data mining process. For the mining of internet traffic data faced a lot of problem of noise and internet traffic number of iteration. The process of pattern generation used two type of technique such as supervised learning and unsupervised learning. In unsupervised learning clustering process are used. The varieties of clustering technique are used such as k-means, FCM and constraints clustering technique. The constraints clustering technique gives the two solution approach one is seed selection and another is mapping of seed in terms of constraint of center. In this dissertation modified the seed selection process using genetic algorithm technique. The genetic algorithm process select variable value one is seed value and another is constraint of center value. In constraints cluster technique used some value of center and generates new center value of new cluster for the better generation of cluster. For more improvement of constraints clustering technique used two level constraints clustering technique for better improvement of cluster technique. In this dissertation modified the constraints clustering technique for improvement. In the process of improvement used genetic algorithm technique. Genetic algorithm technique gives the better selection of seed for internet traffic database.

*Keywords* — **Internet Traffic, Clustering, Genetic Algorithm.**

## I. INTRODUCTION

Internet traffic classification is major area of research now days due to abnormal behaviour of traffic data. The internet traffic data is mixed categories data and the process of classification is very difficult due to unformatted and unstructured region. Various authors and researcher used various methods of classification and clustering to improve the categorization of internet traffic data[1]. In consequence of method used constrained based clustering and classification technique. In constrained based clustering and classification technique provide the different constraints function for the process of clustering. The constrained based clustering technique set the constrained function for different type of data for a separate constraints function [2]. The constrained based clustering technique is very complex process and defines separate constraints function for the processing of classification and cluster generation. For the selection of optimal constraint value used genetic algorithm. Genetic algorithm well knows optimization algorithm. The genetic algorithm optimized the minimal constraint function for the process of cluster generation. Network Traffic clustering has strained important consideration over the past few years. Classifying traffic flows by their generation applications plays very essential task in network security and management, such as, lawful interception and intrusion detection, Quality of Service control. Conventional traffic clustering methods [3] include the port-based prediction methods and payload-based deep inspection methods. In current network environment, the conventional methods suffer from a number of practical problems such as dynamic ports and encrypted applications. Recent research efforts have been absorbed on the application of machine learning techniques to traffic clustering constructed on flow statistical features. It can instinctively search for and describe practical structural patterns in a supplied traffic dataset, which is helpful to logically conduct traffic clustering. The flow statistical feature founded traffic clustering can be understood by using supervised clustering algorithms or unsupervised clustering algorithms. They strengthen that a similar level of clustering accuracy can be achieved when using several different algorithms with the similar set of features and training/testing data. Constraint based clustering algorithm is very efficient method for the internet traffic clustering. The diversity of internet traffic data required more constraint function for the purpose of selection of seed of cluster [4]. Also, the traffic data has big size and dimension so required reduction of data. The multiple features of internet traffic data need more constraint for the purpose of clustering. Now the process of constraint-based clustering technique is very complex [5]. Section-II gives the information of constraint-based clustering technique. In section-III discuss about genetic algorithm. In section IV discuss the proposed algorithm. In section V discuss experimental task and finally discuss conclusion and future work.

## II. CONSTRAINTS BASED CLUSTERING ALGORITHM

In the context of partitioning algorithms, instance-level constraints are a useful way to express a priori knowledge about which instances should or should not be grouped together. The must-link constraints define a transitive binary relation over the instances. Consequently, when making use of a set of constraints (of both kinds), we take a transitive closure over the constraints.1 the full set of derived constraints is then presented to the clustering algorithm. The major modification is that, when updating cluster assignments, we ensure that none of the specified constraints are violated. We attempt to assign each point di to its closest cluster cj . This will succeed unless a constraint would be violated. If there is another point d= that must be assigned to the same cluster as d, but that is already in some other cluster, or there is another point d1=d2 that cannot be grouped with d but is already in c, then d cannot be placed in c. We continue down the sorted list of clusters until we find one that can legally host d. Constraints are never broken; if a legal cluster cannot be found for d, the empty partition ({}) is returned[1].

```
Sbck(x1,….,xn,Ω,k)
Begin
Preparation rearrange x1,………,xn into equivalence set x1,……,xm according to the given set-based constraints Ω
Initialization set the means µ1,…..,µk random sample
Do
Assignment classify the samples in xs(s=1,……,m) to cluster l where
    $l = agmin \sum_{n=1}^{Ns} |x_n^s - \mu i|$
Update recomputed the means µ1,…..,µk and cluster assignments
End.
```

## III. GENETIC ALGORITHM

Genetic algorithms are search algorithms based on the mechanics of natural selection and natural genetics. They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some innovative flair of human search. These algorithms are started with a set of random solution called initial population. Each member of this population is called a chromosome [8]. Each chromosome of this problem which consists of the string genes. The number of genes and their values in each chromosome depends on the population specification. In the algorithm the number of genes of each chromosome is equal to the number of the pixel intensity value and the gene values demonstrate the denoising priority of the related filter to the process, where the higher priority means that noise must executed early. Set of chromosomes in each iteration of GA is called a generation, which are evaluated by their fitness functions. The new generation i.e., the offspring are created by applying some operators on the current generation. These are called crossover which selects two chromosomes of the current population, combines them and generates a new child (offspring), and mutation which changes randomly some gene values of chromosomes and creates a new offspring. Then, the best offspring's are selected by evolutionary select operator according to their fitness values[12]. The GA has four steps as shown below algorithms:

***Step 1:*** Read data (from traffic) and R values from matrix and get Np, Ng, Xr and Mr from the matrix
Np → (initial population size),
Ng → (the number of generations),
Xr → (crossover probability),
Mr→ (mutation probability)

***Step 2:*** Calculate the bottom-level and the top-level of each matrix in the data;
Generate initial population (pi);
Pcurrent ← Pi;
Non-optimal ← Decoding heuristic (Loptimal);
Best ← evaluate (optimal);

***Step 3:*** while stop criterion not satisfied, do begin
Poptimal ← {};
3 – 1: repeat for (Np/2) times
Father ← select (Poptimal, sum_of_fitness);
Mother ← select (Poptimal, sum_of_fitness);
Poptimal ← Poptimal Ů crossover (father, mother, child1, child2, Xr);
End repeat;
3 -2: for each chromosome € Poptimal do begin
Mutate (chromosomes, Mr);
End for
3-3:
Pnew ← Pnew Ů {four best chromosomes of Optimal}
Pnon-optimal ← Optimal;
Condition ← decoding heuristic (Optimal);
Best condition ← evaluate (condition);
End while
***Step 4:*** Repeat the best condition value

## IV. PROPOSED ALGORITHM

In this section discuss the improved set-based constraint clustering technique. The improved set-based constraint clustering technique using genetic algorithm. Initially genetic algorithm defines the constraints function for the process of different constraint according to the flow of traffic. Arrange the all flow level of traffic data and goes from the multiple constraints function. The multiple constraints satisfied the given threshold function for the selection of constraint function [8,9].

The process of seed selection used genetic algorithm. The process of genetic algorithm gives the better result in concern of constraints seed value. The seed selection process recalls with fitness function. The GA fitness function decides the selection process of seed parameter according to recall value. The fitness constraints parameter decides the selection criteria of constraints of cluster.

1: Define the dataset as (X1,………………………,XN)
2: Ω_list ← SBK-means (Ωi_list, )
3: Input Ω_list, the clustering number pn , population scale XN , probability auto P stop conditions cS ;
4: Code the data in real number and initialize population S(i),i = 0 at random;
5: Evaluate the fitness of all individual in the current instant D(s);
6: SBK clustering requires optimization of constraints, which way thrashing of data of waiting cluster. Hence the fitness function of algorithm is determined by f(x).

7: $G(s)=\frac{N(s)}{D(s)}=\frac{\sum_{i=0}^{n-1}A_i s^i}{\sum_{i=0}^{n}a_i s^i}$ Umpire the termination conditions. If the termination situation is satisfied, then turn to step 9, if not, turn to step 10;

$$p(U,Z,V,W)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in Gi}^{1} uiwtvjd(xij,zij)$$
$$+n\sum_{j=1}^{m} vjlog(vj)+\lambda\sum_{t=1}^{T} wtlog(wt)\ldots\ldots(1)$$

8: Subject to

$$\begin{cases}\sum_{l=1}^{k} ui.l=1, ui,l\in(0,1),1\le i\le n\\ \sum_{t=1}^{T} wt=1,0\le wt\le 1, \qquad \ldots..(2)\\ \sum_{j\in Gi} vj-1,0\le vj\le 1,1\le t\le T,\end{cases}$$

Where

9: U is a n×k portion matrix whose element ui,l are binary where ui,l=1 indicates that object I is allocated to cluster l;

10: Z={Z1,Z2,……….,Zk} is a set of k vectors representing the centers of the k clusters

11: W={W1,W2,………..Wt} are T constraints for T view

12: V={v1,v2,……………vm} are m constraints form variable

13: d(xij,zlj) is a distance or dissimilarity measure on the jth variable between the ith object and the center of the lth cluster. if the variable is numerical , then

14: d(xij ,zlj)=(xij-zlj)$^2$…………………………..(3)

15: if the variable is categorical, then

16: d(xij,zlj)={0 (xi.j=zl.j)……………………..(4)

17: 1 (xi.j=/zl.j)

18: The first term in (1) is the sum of the within cluster dispersions, the second and third terms are two negative constraints entropies. Two positive parameters are control the strength of cluster.

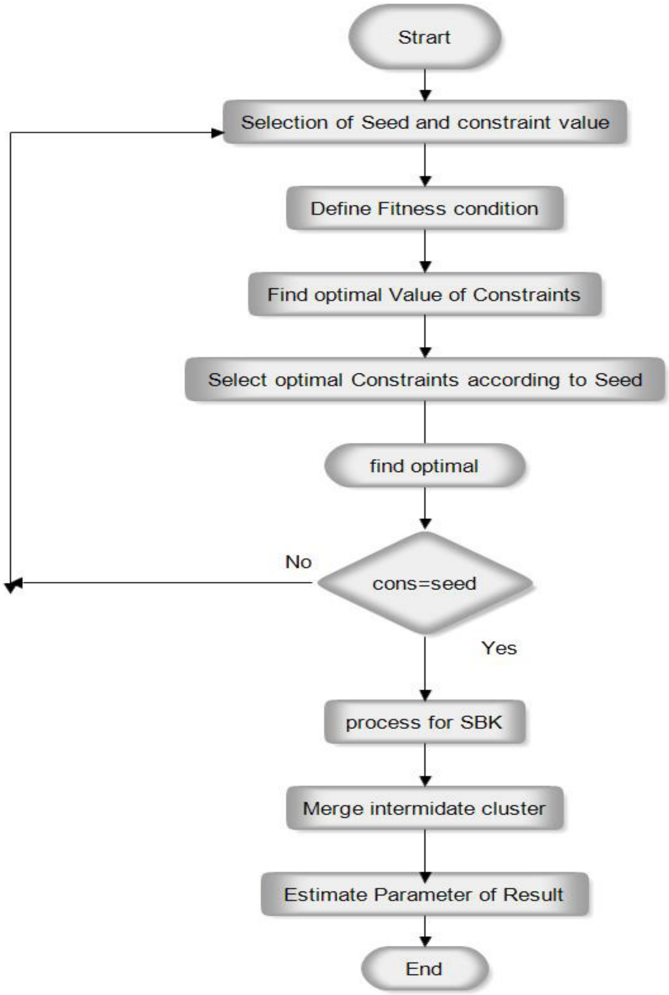

Strart
Selection of Seed and constraint value
Define Fitness condition
Find optimal Value of Constraints
Select optimal Constraints according to Seed
find optimal
No
cons=seed
Yes
process for SBK
Merge intermidate cluster
Estimate Parameter of Result
End

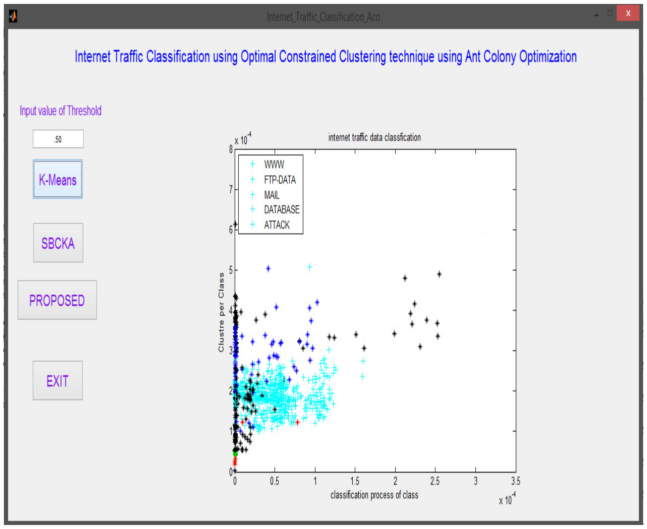
Figure 1: show that proposed model of constraint-based clustering technique for traffic classification.

## V. EXPERIMENTAL ANALYSIS

In this section we perform experimental process of internet traffic classification, the process of traffic classification done by two methods one is SBCKA and other one is proposed method with Genetic Algorithm for better classification. The proposed method implements in MATLAB 7.8.0 and tested with very reputed data set. In order to construct the sets of flows, the day trace was split into ten blocks of approximately 1680 seconds (28 minutes) each. In order to provide a wider sample of mixing across the day, the start of each sample was selected randomly (uniformly distributed over the whole day trace). It can be seen from Table 1 that there are a different number of flows in each data block, due to a variable density of traffic during each constant period. Since time statistics of flows are present in the analysis, we consider it to be important to keep a fixed time window when selecting flows. Each data set represents a period of time taken from within the day[9].

TABLE I
NUMBER GIVES THE DESCRIPTION OF DATASET

| Data-set | Start-time | End-time | Duration | Flows (Objects) |
|---|---|---|---|---|
| entry01 | 2003-Aug-20 00:34:21 | 2003-Aug-20 01:04:43 | 1821.8 | 24863 |
| entry02 | 2003-Aug-20 01:37:37 | 2003-Aug-20 02:05:54 | 1696.7 | 23801 |
| entry03 | 2003-Aug-20 02:45:19 | 2003-Aug-20 03:14:03 | 1724.1 | 22932 |
| entry04 | 2003-Aug-20 04:03:31 | 2003-Aug-20 04:33:15 | 1784.1 | 22285 |
| entry05 | 2003-Aug-20 04:39:10 | 2003-Aug-20 05:09:05 | 1794.9 | 21648 |
| entry06 | 2003-Aug-20 06:07:28 | 2003-Aug-20 06:35:06 | 1658.5 | 19384 |
| entry07 | 2003-Aug-20 09:42:17 | 2003-Aug-20 10:11:16 | 1739.2 | 55835 |
| entry08 | 2003-Aug-20 11:52:40 | 2003-Aug-20 12:20:26 | 1665.9 | 55494 |
| entry09 | 2003-Aug-20 13:45:37 | 2003-Aug-20 14:13:21 | 1664.5 | 66248 |
| entry10 | 2003-Aug-20 14:55:44 | 2003-Aug-20 15:22:37 | 1613.4 | 65036 |



Internet_Traffic_Classification_Aco

Internet Traffic Classification using Optimal Constrained Clustering technique using Ant Colony Optimization

Input value of Threshold: .50

Figure 2: show that our implementation simulation view in matlab.

In this section I show the selection of variable no. of attribute to process the given classification algorithms. The variable no. of attribute differs the classification rate and classification time. The evaluation parameter corresponding to attribute shown in given below table.

TABLE III
SHOWS THAT THE PERFORMANCE EVALUATION OF ACCURACY, F-MEASURE, NUMBER OF ITERATIONS AND ITERATION ERROR BY K-MEANS METHOD WITH INPUT VALUE IS 0.50

| **Method Name** | **Value** | **Data Set** | **Accuracy** | **F-Measure** | **Number of Iterations** | **Iteration error** |
|---|---|---|---|---|---|---|
| **K-Means** | 0.50 | Data Set I | 86.6000 | 81.6000 | 389.0000 | 24.9700 |
| | | Data Set II | 100.0900 | 95.0900 | 400.0000 | 25.5000 |
| | | Data Set III | 76.0900 | 71.0900 | 310.0000 | 21.2000 |
| | | Data Set IV | 49.9700 | 44.9700 | 200.0000 | 15.5100 |
| | | Data Set V | 102.0400 | 97.0400 | 400.0000 | 25.5000 |

TABLE IIII
SHOWS THAT THE PERFORMANCE EVALUATION OF ACCURACY, F-MEASURE, NUMBER OF ITERATIONS AND ITERATION ERROR BY SBCKA METHOD WITH INPUT VALUE IS 0.50

| **Method Name** | **Value** | **Data Set** | **Accuracy** | **F-Measure** | **Number of Iterations** | **Iteration error** |
|---|---|---|---|---|---|---|
| **SBCKA** | 0.50 | Data Set I | 76.1500 | 72.1500 | 379.0000 | 21.9700 |
| | | Data Set II | 100.1700 | 96.1700 | 390.0000 | 22.5000 |
| | | Data Set III | 76.2300 | 72.2300 | -7.0000 | 18.0000 |
| | | Data Set IV | 59.0200 | 55.0200 | 190.0000 | 12.5700 |
| | | Data Set V | 100.1000 | 96.1000 | 390.0000 | 22.5000 |

TABLE IVV
SHOWS THAT THE PERFORMANCE EVALUATION OF ACCURACY, F-MEASURE, NUMBER OF ITERATIONS AND ITERATION ERROR BY SBCKA METHOD WITH INPUT VALUE IS 0.50

| **Method Name** | **Value** | **Data Set** | **Accuracy** | **F-Measure** | **Number of Iterations** | **Iteration error** |
|---|---|---|---|---|---|---|
| **Proposed** | 0.50 | Data Set I | 86.8900 | 82.8900 | 369.0000 | 19.3700 |
| | | Data Set II | 92.9600 | 88.9600 | 380.0000 | 19.9000 |
| | | Data Set III | 86.1500 | 82.1500 | 290.0000 | 15.4000 |
| | | Data Set IV | 62.4500 | 58.4500 | 180.0000 | 9.9100 |
| | | Data Set V | 90.2200 | 86.2200 | 380.0000 | 19.9000 |

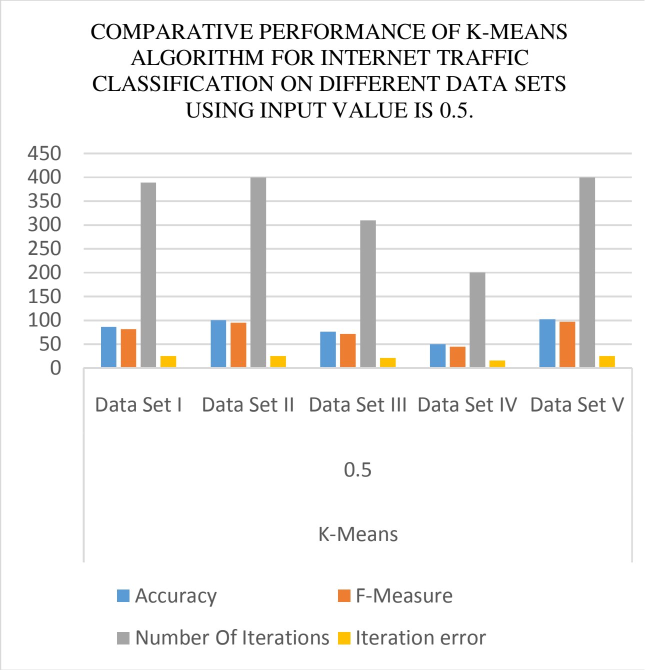


Figure 3: Shows that the performance evaluation of accuracy, F-measure, number of iterations and iteration error by K-Means method with input value is 0.50.

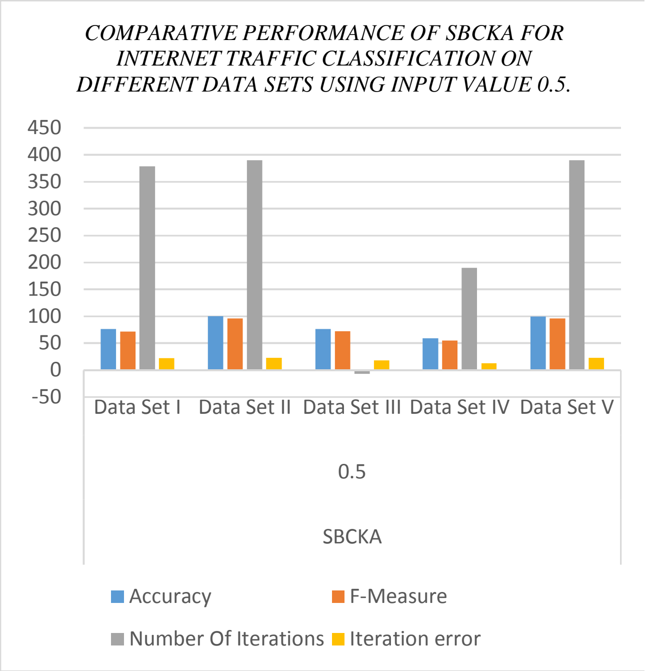


Figure 4: Shows that the performance evaluation of accuracy, F-measure, number of iterations and iteration error by SBCKA method with input value is 0.50.

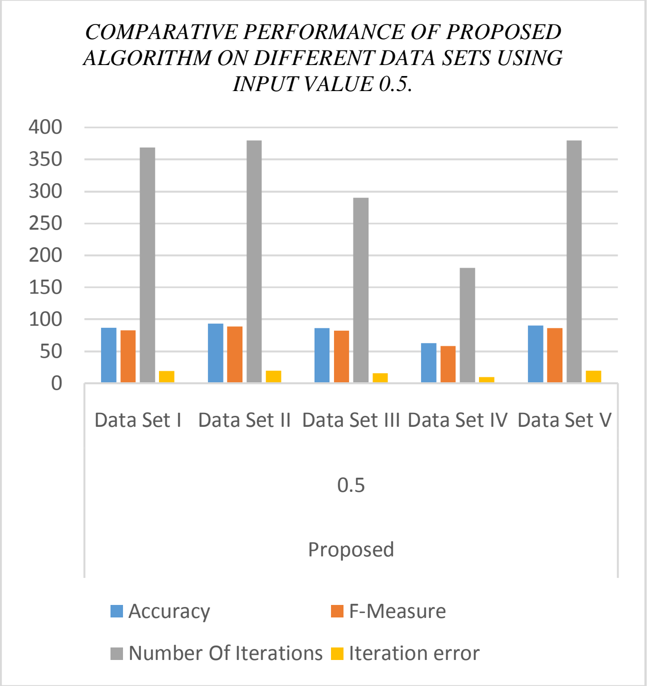


Figure 5: Shows that the performance evaluation of accuracy, F-measure, number of iterations and iteration error by proposed method with input value is 0.50.

## VI. CONCLUSION & FUTURE WORK

In this paper modified the constraints-based clustering technique using genetic algorithm. The genetic algorithm used for the selection of seed and constraints value. The optimal selection of seed and constraints value increases the accuracy of cluster technique. The cluster technique imposed the two processes for the selection of seed and constraints parameter. Proposed clustering algorithm for clustering of internet traffic data, proposed can compute constraints for views and individual variables simultaneously in the clustering process. With the two types of constraints, compact views and important variables can be identified and effect of low-quality views and noise variables can be reduced. Therefore, proposed can obtain better clustering results than individual variable constraint clustering algorithms for internet traffic data. We used four internet traffic data sets to investigate the properties of two types of constraints in Proposed. We discussed the difference of the constraints between Proposed and SBKC algorithms. The experiments also revealed the convergence property of the view constraints in Proposed. We compared Proposed with three clustering algorithms on internet traffic data sets and the results have shown that the proposed algorithm significantly outperformed the other two clustering algorithms in four evaluation indices. As such, it is a new variable constraint method for clustering of internet traffic data.

## REFERENCES

*[1] Yu Wang, Yang Xiang, Jun Zhang, Wanlei Zhou, Guiyi Wei, Laurence T. Yang "Internet Traffic Classification Using Constrained Clustering" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, 2013. Pp 1-11.*

*[2] Jun Zhang, Chao Chen, Yang Xiang, wanleizhou, Yong Xiang "Internet traffic classification by aggregating Correlated Naive Bayes Predictions" IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL-8, 2013. Pp 5-16.*

*[3] Yeon-sup Lim, Hyun-chul Kim, Jiwoong Jeong, Chong-kwon Kim "Internet Traffic Classification Demystified: On the Sources of the Discriminative Power" ACM, 2010. Pp 1-12.*

*[4] Yu Wang, Yang Xiang, Jun Zhang, Shunzheng Yu "A Novel Semi-Supervised Approach for Network Traffic Clustering" IEEE, 2011. Pp 169-174.*

*[5] Thuy T.T. Nguyen, Grenville Armitage "A Survey of Techniques for Internet Traffic Classification using Machine Learning" IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL-10. 2008. Pp 56-76.*

*[6] Alice Este, Francesco Gringoli, Luca Salgarelli "On the Stability of the Information Carried by Traffic Flow Features at the Packet Level" ACM, 2009. Pp 13-19.*

*[7] Jeffrey Erman, Martin Arlitt, Anirban Mahanti "Traffic Classification Using Clustering Algorithms" SIGCOMM'06 Workshops, 2006. Pp 281-286.*

*[8] Marcin Pietrzyk, Jean-Laurent, Guillaume Urvoy-Keller, Taoufik "Challenging Statistical Classification for Operational Usage: the ADSL Case" ACM, 2009. Pp 1-14.*

*[9] Jeffrey Erman, Anirban Mahanti, Martin Arlitt, Ira Cohen, Carey Williamson "Semi-Supervised Network Traffic Classification" ACM, 2010. Pp 1-2.*

*[10] Andrew W. Moore, Denis Zuev "Internet Traffic Classification Using Bayesian Analysis Techniques" ACM, 2005. Pp 50-61.*

*[11] Jeffrey Erman, Anirban Mahanti, Martin Arlitt "Internet Traffic Identification using Machine Learning" 2007. Pp 1-6.*

*[12] Rongjun Li, and Xianying Chang," A Modified Genetic Algorithm with Multiple Subpopulations and Dynamic Parameters Applied in CVAR model", IEEE Transactions on Intelligent Agents, Web Technologies and Internet Commerce, 2006.*