RESEARCH ARTICLE OPEN ACCESS

Internet Traffic Data Categorization Using Particle of Swarm Optimization Algorithm

Nikita Shrivastava [1], Prof. Amit Dubey [2]

M Tech Scholar [1], HOD [2],

Department of Computer Science and Engineering

OCT Bhopal, India

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 paper 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. For the performance evaluation of proposed algorithm used three real time dataset from UCI machine learning center. The proposed algorithm implemented in MATLAB software and measures some standard parameter for the validation of proposed methodology.

Keywords: - Clustering, Classification, FCM, UCI, GA.

I. INTRODUCTION

Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in the data. simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data [10]. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. "Knowledge mining," a shorter term, may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both "data" and "mining" became a popular choice [4]. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. The architecture of a typical data mining system has the following major components. Internet traffic classification is the process of identifying network applications and classifying the corresponding traffic, which is considered to be the most fundamental functionality in modern network management and security systems. It realizes a fine-grained visibility of types of traffic traversing the distributed network in a real-time basis, which enables the higher levels of controls such as Qos and per application security rule enforcement. Traffic classification, the branch of traffic measurement that studies mechanisms to associate traffic flows to the applications that generated them, in the last few years has focused on the statistical analysis of measurable features, such as packet size or flow duration. Many classification systems have been proposed, and their effectiveness has been proven on a series of different traffic traces, collected in various network locations and in various time periods [6]. Accurate classification of Internet traffic is important in many areas such as network design, network management, and network security. One key challenge in this area is to adapt to the dynamic nature of Internet traffic. Increasingly, new applications are being deployed on the Internet; some new

applications such as peer-to-peer (P2P) file sharing and online gaming are becoming popular. With the evolution of Internet traffic, both in terms of number and type of applications, however, traditional classification techniques such as those based on well-known port numbers or packet payload analysis are either no longer effective for all types of network traffic or are otherwise unable to deploy because of privacy or security concerns for the data. Traffic classification has gained substantial attention within the Internet research and operation community given recent events and struggles over the appropriate use and pricing of the Internet. Accurate and complete traffic classification is an essential task for understanding, operating, optimizing, and financing the Internet as well as planning improvements in future network architectures [13]. For example, traffic classification can be used to detect patterns indicative of denial of service attacks, worm propagation, intrusions, and spam spread. Section II discusses about classification overview and techniques with description, Section III discusses about the proposed methodology. Section IV discusses comparative result analysis. Finally, concluded in section V.

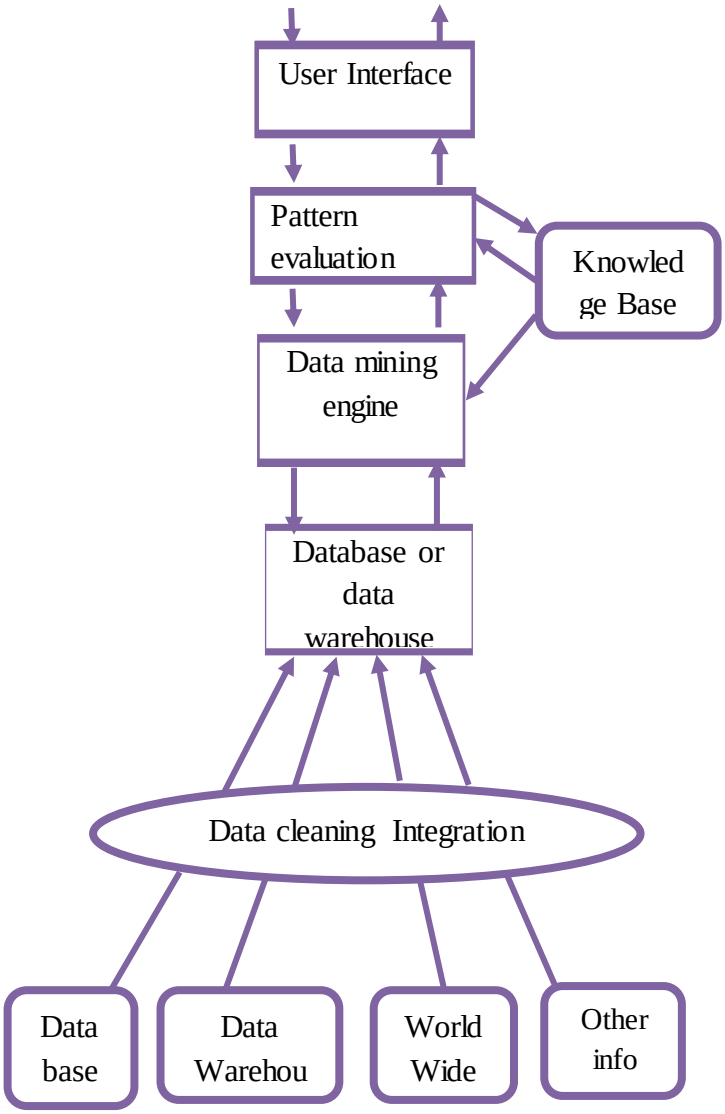

Figure 1: Basic architecture of data mining process.

II. CLASSIFICATION OVERVIEW

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome [9]. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. The following preprocessing steps may be applied to the data to help improve the accuracy, efficiency, and scalability of the classification or prediction process as shown in below figure.

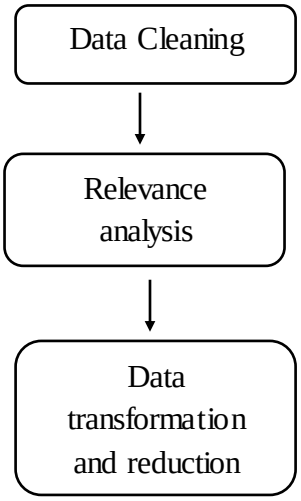

Figure 2: Steps to prepare data for classification.

DATA CLEANING

This refers to the preprocessing of data in order to remove or reduce noise (by applying smoothing techniques, for example) and the treatment of missing values (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics) [11]. Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help reduce confusion during learning.

RELEVANCE ANALYSIS

Many of the attributes in the data may be redundant. Correlation analysis can be used to identify whether any two given attributes are statistically related. For example, a strong correlation between attributes A1 and A2 would suggest that one of the two could be removed from further analysis. A database may also contain irrelevant attributes. Attribute subset selection can be used in these cases to find a reduced set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Hence, relevance analysis, in the form of correlation analysis and attribute subset selection, can be used to detect attributes that do not contribute to the classification or prediction task. Including such attributes may otherwise slow down, and possibly mislead, the learning step. Ideally, the time spent on relevance analysis, when added to the time spent on learning from the resulting "reduced" attribute (or feature) subset, should be less than the time that would have been spent on learning from the original set of attributes [8]. Hence, such analysis can help improve classification efficiency and scalability.

DATA TRANSFORMATION AND REDUCTION

The data may be transformed by normalization, particularly when neural networks or methods involving distance measurements are used in the learning step. Normalization involves scaling all values for a given attribute so that they fall within a small specified range, such as−1.0 to 1.0, or 0.0 to 1.0. In methods that use distance measurements, for example, this would prevent attributes with initially large ranges (like, say, income) from out weighting attributes with initially smaller ranges (such as binary attributes). The data can also be transformed by generalizing it to higher-level concepts. Concept hierarchies may be used for this purpose [3]. This is particularly useful for continuous valued attributes. For example, numeric values for the attribute in come can be generalized to discrete ranges, such as low, medium, and high. Similarly, categorical attributes, like street, can be generalized to higher-level concepts, like city. Because generalization compresses the original training data, fewer input/output operations may be involved during learning. Data can also be reduced by applying many other methods, ranging from wavelet transformation and principle components analysis to discretization techniques, such as binning, histogram analysis, and clustering.

STATISTICAL ALGORITHMS

The ID3 algorithm was originally developed by J. Ross Quinlan at the University of Sydney, and he first presented it in the 1975 book "Machine Learning" [1]. The ID3 algorithm induces classification models, or decision trees, from data. It is a supervised learning algorithm that is trained by examples for different classes. After being trained, the algorithm should be able to predict the class of a new item. ID3 identifies attributes that differentiate one class from another. All attributes must be known in advance, and must also be either continuous or selected from a set of known values. For instance, temperature (continuous), and country of citizenship (set of known values) are valid attributes. To determine which attributes are the most important, ID3 uses the statistical property of entropy. Entropy measures the amount of information in an attribute. This is how the decision tree, which will be used in testing future cases, is built. One of the limitations of ID3 is that it is very sensitive to attributes with a large number of values (e.g. social security numbers). The entropy of such attributes is very low, and they don't help you in performing any type of prediction. The C4.5 algorithm overcomes this problem by using another statistical property known as information gain. Information gain measures how well a given attribute separates the training sets into the output classes. Therefore, the C4.5 algorithm extends the ID3 algorithm through the use of information gain to reduce the problem of artificially low entropy values for attributes such as social security numbers.

GENETIC PROGRAMMING

Genetic programming (GP) has been vastly used in research in the past 10 years to solve data mining classification problems. The reason genetic programming is so widely used is the fact that prediction rules are very naturally represented in GP. Additionally, GP has proven to produce good results with global search problems like classification. The search space for classification can be described as having several 'peaks', this causes local search algorithms, such as simulated annealing, to perform badly. GP consists of stochastic search algorithms based on abstractions of the processes of Darwinian evolution. Each candidate solution is represented by an individual in GP. The solution is coded into chromosome like structures that can be mutated and/or combined with some other individual's chromosome. Each individual contains a fitness value, which measures the quality of the individual, in other words how close the candidate solution is from being optimal [12]. Based on the fitness value, individuals are selected to mate. This process creates a new individual by combining two or more chromosomes, this process is called crossover. They are combined with each other in the hope that these new individuals will evolve and become better than their parents. Additionally to matting,

chromosomes can be mutated at random. The running time of GPs is usually controlled by the user. There are many parameters used to determine when the algorithm should stop, and each data set can have very different settings. In all cases, the best individual is stored across generations and is returned when the algorithm stops. The most commonly used parameter is number of generations. Another stop parameters used is minimum expected hit ratio, in which case the algorithm will run until a candidate solution has a hit ratio greater than expected. This however can cause the algorithm to run forever. Combinations of stop conditions can also be used to ensure stoppage.

III. PROPOSED METHODOLOGY

In this section we define the proposed methods and architecture, the Proposed clustering constraints method based on Particle of swarm optimization where the compromise clustering is selecting optimal cluster criterion function using paticle of swarm optimization. This method uses a metric between clustering's based on the ant between partitions. It also uses class level method to solve the label correspondence problem. The search capabilities of paticle of swarm optimization s are used in these methods. It allows exploring partitions that are not easy to be found by other methods. However, a drawback of these algorithms is that a solution is better only in comparison to another; such an algorithm actually has no concept of an optimal solution or any way to test whether a solution is optimal or not. This selection function combines partitions obtained by using locally adaptive clustering SBK algorithms. When a SBK algorithm is applied to a set of objects X, it gives as an output a partition P = { C1,C2,….,Cq}, which can be also identified by two sets {c1,…..cq}and {w1,…..wq}, where ci and wi are the centroid and the confidence associated to the cluster Ci respectively. The SBK algorithms are designed to work with numerical data, i.e. this method assumes that the object representation in the dataset is made up of numerical features: X={x1,…,xn}, with xj € R, j =1, …..,n ; n. Also, ci € Rand wi €2 R, i = 1,….., k. The set of partitions P = {p1,p2,…….,pm} is generated by applying SBK algorithms m times with deferent parameters initialization. The process of paticle of swarm optimization is to use the constraints to choose, the selection of pheromone update (increment and decrement of constant deposit of interval value of phenomenon) parameters of Particle of swarm optimization to control the sensitive value. For a given cluster assembling the problem of optimal cluster selection can be stated as follows: given the variable set of cluster index, F ,of n data point, find optimal S ,which consists of m cluster$(m < n, ScF)$,such that the classification accuracy is maximized. The cluester index selection representation exploited by artificial ants includes the following:

1. n Data point that constitutes the cluster index set,$F = \{f_1 \ldots\ldots\ldots f_2\}$.
2. Distribution of the cluster index space (na ants).
3. τ_i , the intensity of pheromone trail associated with cluster index f_i ,which reflects the previous knowledge about the importance off_i .
4. For each ant j , a list that contains the selected optimal cluster subset, $R_j = \{R_1 \ldots\ldots\ldots R_m\}$.

SBK evaluation measure that is able to estimate the overall performance of subset as well as the local importance of cluster. A classification algorithm is used to estimate the performance of optimal cluster selection .On the other hand, the local importance of a given cluster measured using the correlation based evaluation function, which is a filter evaluation function. In the first iteration, each ant will randomly choose a cluster index of m cluster. Only the best k subsets, $k < na$, be used to update the pheromone trail and influence the optimal subset of the next move. In the second and following moves, each ant will start with $m - p$ cluster that are randomly chosen from the previously selected $k -$ best subsets, where p is a number that limit between 1 and $m - 1$. In this way, the constraints that constitute the best k subsets will have more chance to be present in the subsets of the next iteration.

Below are the steps of the algorithm:

1. Initialization:
   - Set $\tau_i = cc$ and $\Delta T_i = 0, (i = 1, \ldots., n)$,where cc is a constant and $\Delta\tau_i$ is the amount of change of pheromone trail quantity for variable cluster index f_i .
   - Assign the maximum number of moves.
   - Assign , where the $k - best$ subsets will influence the subsets of the next iteration.
   - Assign , where $m - p$ is the number of cluster indexes that each ant will start with in the second and following moves.
2. If the first iteration,
   - For$j = 1$ to na ,
     - Randomly assign a subset of m cluster to S_j.

- Goto step 4.

3. Select the remaining p cluster index for each ant:
   - For $mm = m - p + 1 \; to \; m$,
     - For $j = 1 \; to \; na$,
       - Given subset S_j, choose cluster index f_i that maximizes $USM_i^{S_j}$
       - $S_j = S_j \cup \{f_i\}$.
   - Merge the duplicated subsets, if any, with randomly chosen index.
4. estimated the selected index of each cluster ant using a chosen classification algorithm:
   - For $j = 1 \; to \; na$,
     - Estimate the Error (E_j) of the classification results obtained by classifying the optimal cluster of S_j.
   - Sort the subsets according to their E. Update the minimum E (if achieved by any ant in this iteration), and store the corresponding subset of cluster.
5. Using the constraints subsets of the best k ants, update the pheromone trail intensity:
   - For $j = 1 \; to \; k$,

$$\Delta\tau_i = \begin{cases} \dfrac{\max\limits_{g=1:k}(\varepsilon_g) - \varepsilon_j}{\max\limits_{h=1:k}\left(\max\limits_{g=1:k}(\varepsilon_g) - \varepsilon_h\right)} & if \; f_i \in S_j \\ 0 \quad otherwise \end{cases} \quad (3)$$

$$\tau_i = \rho . \tau_i + \Delta\tau_i \quad (4)$$

Where ρ is a constant such that $(1-\rho)$ represents the evaporation of pheromone trails.

6. If the number of moves is less than the maximum number of moves, or the desired E has not been achived, initialize the subsets for next iteration and goto step3:
   - For $j = 1 \; to \; na$,
     - From the selected cluster of the best k ants, randomly produce $m - p$ classifeir subset for ant j, to be used in the next iteration, and store it in S_j.

- Goto step 3

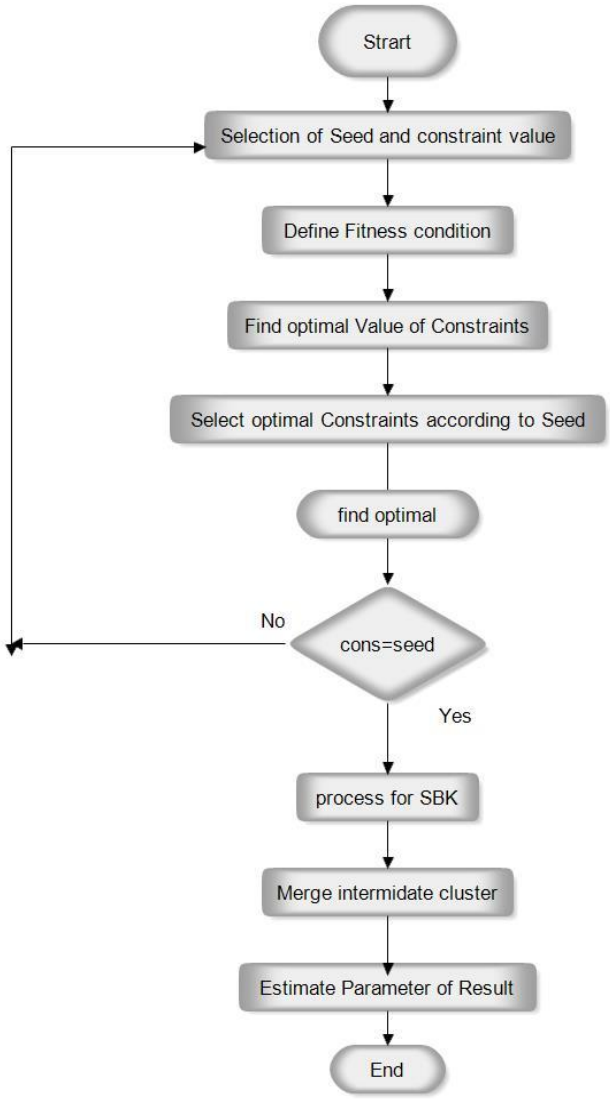

Figure 3: Proposed model for SBK.

IV. 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 particle of swarm optimization for better classification. The proposed method implements in MATLAB 7.14.0 and tested with very reputed data set.

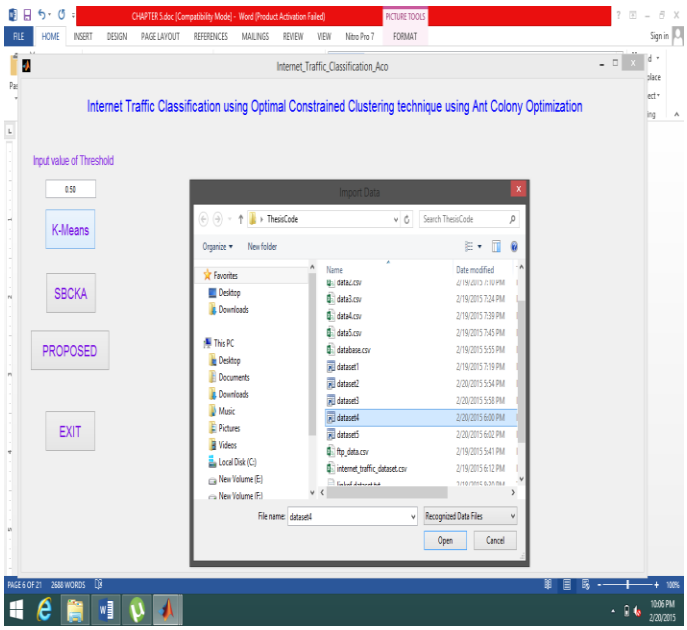

Figure 4: Shows that the windows for load and import the dataset.

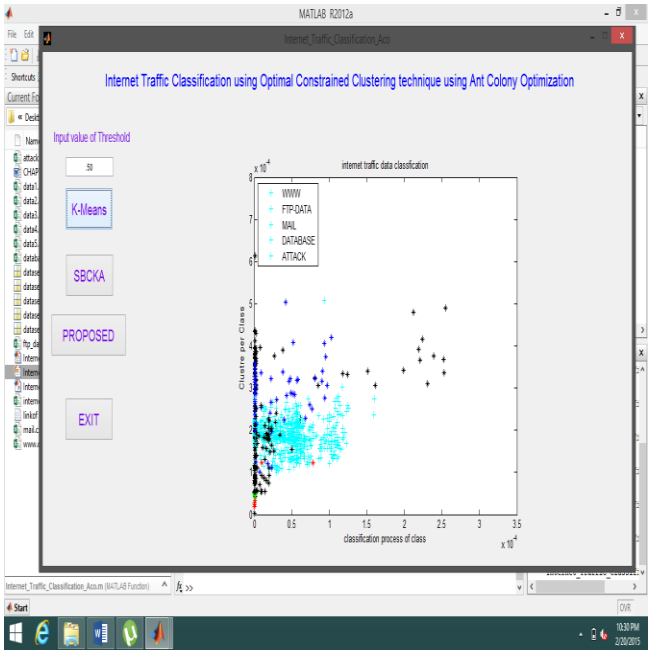

Figure 5: Shows that the window for classification of Internet traffic on data set 1 with threshold value is 0.50 and method is K-Means.

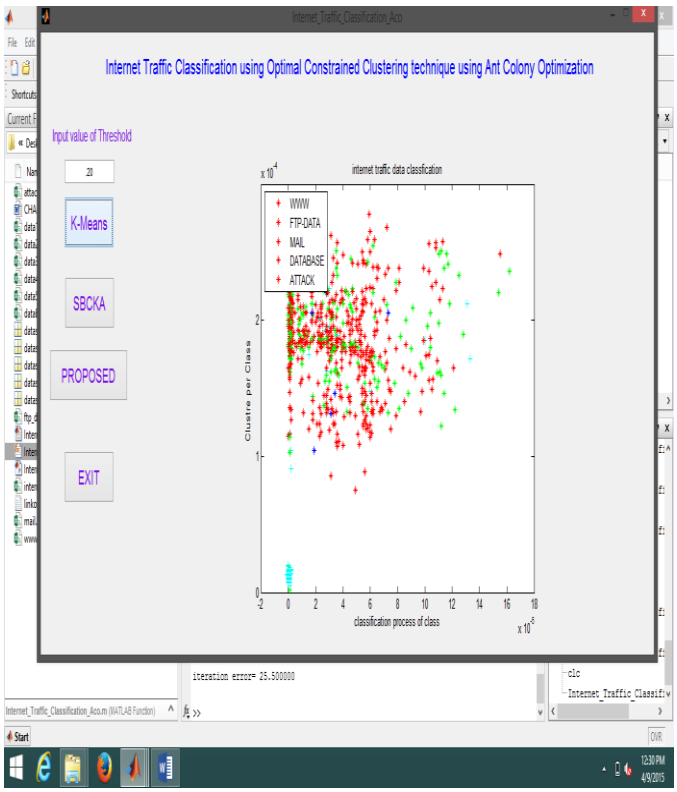

Figure 6: Shows that the window for classification of Internet traffic on data set 2 with threshold value is 0.20 and method is K-Means.

Method Name	Value	Data Set	Accuracy	F-Measure	Number Of Iterations	Iteration error
K-Means	0.50	Data Set I	86.6	81.60	389.0	24.9
		Data Set II	96.09	95.09	400.0	25.5
		Data Set III	76.09	71.09	310.0	21.2
		Data Set IV	49.97	44.9	200.00	15.5

Table 1: 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.

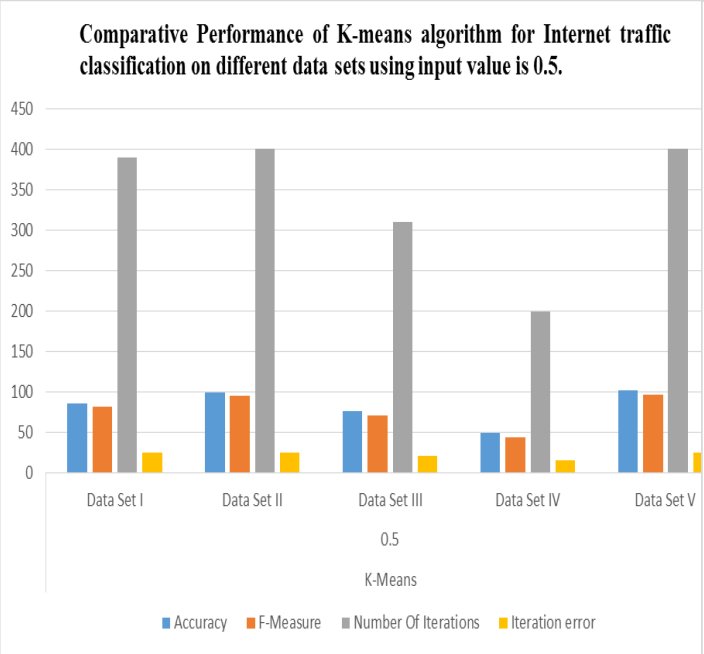

Figure 7: Shows that the performance evaluation of accuracy, F-measure, number of iterations and iteration error by K-Means method with input value is 0.50.

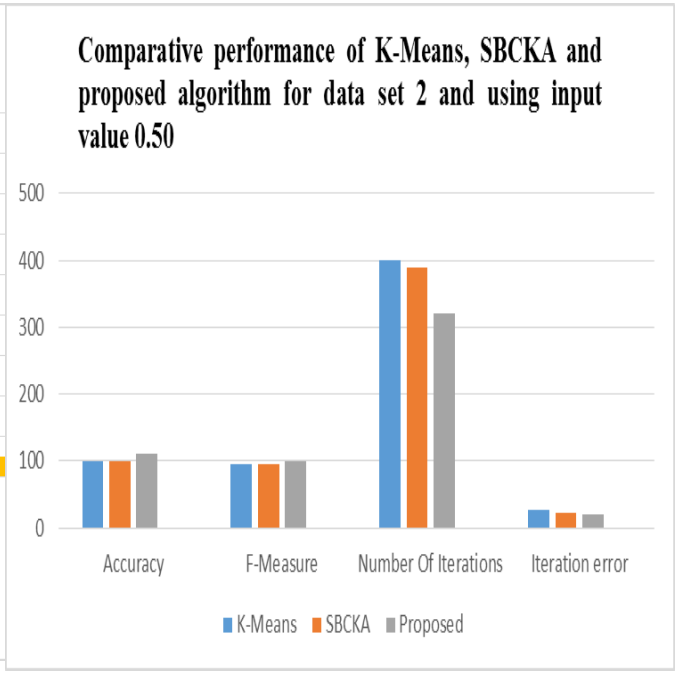

Figure 9: Shows that the performance comparison of accuracy, F-measure, number of iterations and iteration error by K-Means, SBCK and Proposed method.

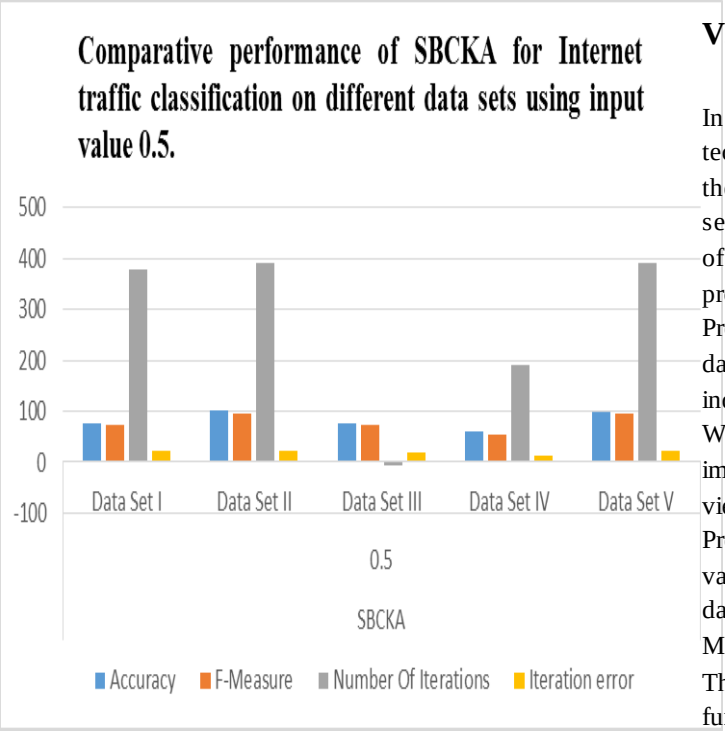

Figure 8: Shows that the performance evaluation of accuracy, F-measure, number of iterations and iteration error by SBCKA method with input value is 0.50.

V. CONCLUSION AND FUTURE WORK

In this dissertation modified the constraints based clustering technique using PSO algorithm. The PSO 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. For the evaluation of performance of algorithm used MATLAB software and three internet traffic data are used. The proposed algorithm work with PSO algorithm, so pos function of MATLAB is used. For the measuring the parameter used standard formula such as accuracy, precision, f-measure and recall. Our empirical result shows that our proposed algorithm shows better result in comparison of SBK algorithm. The exiting two algorithms not controlled the level

constraints of cluster and loss some data during the grouping of cluster. The proposed algorithm is very efficient for internet traffic data clustering technique.

The proposed algorithm is very efficient clustering technique for internet traffic data. The algorithm used PSO algorithm for controlling the constraints variable of cluster level generation during formation of cluster. The PSO algorithm takes more time for the selection of estimated value of constraints. The values of constraints influence the cluster quality during process of data. In future used optimization technique for self-selection of optimal cluster for 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, Wanlei Zhou,Yong Xiang"Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions" IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, 2012 Pp 1-11.

[3] Yeon-sup Lim,Hyun-chul Kim,Jiwoong Jeong,Chong-kwon Kim,Ted "Taekyoung" Kwon, Yanghee Choi.Internet Traffic Classification Demystified: On the Sources of the Discriminative Power. 2009Pp 1-12.

[4] Yu Wang, Yang Xiang, Jun Zhang, Shunzheng Yu "A Novel Semi-Supervised Approach for Network Traffic Clustering" 2011 Pp 1-7.

[5] Thuy T.T. Nguyen and Grenville Armitage," A Survey of Techniques for Internet TrafficClassification using Machine Learning" IEEE COMMUNICATIONS SURVEYS & TUTORIALS, 2008 Pp 1-21.

[6] Alice Este, Francesco Gringoli, Luca Salgarelli "On the Stability of the Information Carried by Traffic Flow Features at the Packet Level" 2009Pp 1-6.

[7] Jeffrey Erman, Martin Arlitt, AnirbanMahanti"Traffic Classification Using Clustering Algorithms" 2006 Pp 1-6.

[8] Marcin Pietrzyk and Jean-Laurent Costeux, Guillaume Urvoy-Keller and TaoufikEn-Najjary"Challenging Statistical Classification for Operational Usage: the ADSL Case" 2009 Pp 1-14

[9] Jeffrey Ermany, Anirban Mahantiy, Martin Arlittyz, Ira Cohenz, Carey Williamsony"Semi-Supervised Network Traffic Classification" 2007 Pp 1-2.

[10] Andrew W. Moore,Denis Zuevy "Internet Traffic Classification Using Bayesian Analysis Techniques" 2005 Pp 1-11,

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

[12] A. Este, F. Gringoli, and L. Salgarelli "Support Vector Machines for TCP traffic classification", Computer Networks, Vol. 53, No. 14, 2009, Pp 2476-2490.

[13] D. Schatzmann, W. Muhlbauer, T. Spyropoulos, X. Dimitropoulos "Digging into HTTPS: flow-based classification of webmail traffic" in Proceedings of the 10th annual conference on Internet measurement, 2010. Pp 322-327.

[14] Y. Wang and S. Yu "Supervised Learning Real-time Traffic Classifiers" Journal of Networks, 2009, Pp 622-629.

[15] T. T. Nguyen, G. Armitage, P. Branch, S. Zander "Timely and Continuous Machine-Learning-Based Classification for Interactive IP Traffic" IEEE/ACM Transactions on Networking, in press, 2012.

[16] S. Zander, G. Armitage, "Practical Machine Learning Based Multimedia Traffic Classification for Distributed QoS Management," in 36th Annual IEEE Conference on Local Computer Networks (LCN 2011), Bonn, Germany, 4-7 October 2011.

[17] M. Pietrzyk, J. Costeux, G. Urvoy-Keller, T. En-Najjary "Challenging statistical classification for operational usage: the ADSL case" in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference (IMC '09), 2009, Pp 122-135.

[18] A. Este, F. Gringoli, and L. Salgarelli "On the Stability of the Information Carried by Traffic Flow Features at the Packet Level" ACM SIGCOMM Computer Communication Review, Vol. 39, No. 3, Jul 2009, Pp 13-18

[19] Y. Lim, H. Kim, J. Jeong, C. Kim, T.Kwon, Y.Choi "Internet Traffic Classification Demystified: On the Sources of the Discriminative Power" in Proceeding of the ACM CoNEXT 2010.

[20] Y.Wang, Y. Xiang, and S.-Z. Yu "An automatic application signature construction system for unknown traffic" Concurrency Computat.: Pract. Exper., 2010. Pp 1927–1944.