Research Journal of Applied Sciences, Engineering and Technology 11(5): 473-479, 2015
DOI: 10.19026/rjaset.11.1850
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: March 14, 2015 Accepted: April 1, 2015 Published: October 15, 2015

Research Article

Traffic Analyzing and Controlling using Supervised Parametric Clustering in Heterogeneous Network

[1]D. Jayachitra and [2]J. Jebamalar Tamilselvi
[1]Department of Computer Science, Nehru Memorial College, Bharathiar University, Puthanampatti, Coimbatore, Tamil Nadu, India
[2]Department of MCA, Jaya Engineering College, Chennai, Tamil Nadu, India

Abstract: The aim of the study is to maximize class purity level of dynamic network structure using the Supervised Parametric Clustering (SPC) approach. An efficient means to ensure proper flow sequence on heterogeneous network is the effective analysis of network traffic and its smooth maintenance. Several works in the literature have focused on access logs for analyzing network traffic, but investigating supervised clustering model for network analysis has received relatively less attention. On the other hand, Peer-to-Peer Document Clustering maintain privacy inside the neighborhood data boundaries, but fails on extending dynamic structure that reduces the high class purity mapping. To maximize the class purity level of mapping, Supervised Parametric Clustering (SPC) approach is proposed in this study. SPC approach's initial work is to analyze the flow sequences of packets in the heterogeneous network. During analyzing the data traffic flow, to ensure smoothening of network data flow, a method called binning is designed that smoothen the network data flow information discussing with neighborhood value range. Secondly, traffic data feature selection uses the Candy Best First wrapper technique for balancing the heterogeneous network traffic. The Candy Best First wrapper technique extracts the content to the particular selected feature source and translated into the relational form. Then, SPC follows the separation of medoids value points to cluster the collaborative classes of traffic data points. Finally, mapping of the SPC resultant data clusters is carried out to effectively analyze and control the traffic flow in the heterogeneous network. Experiment is carried out on the factors such as performance speedup rate, false positive rate and class purity mapping rate.

Keywords: Best first wrapper technique, binning method, dynamic network structure, feature selection, fitness function, supervised parametric clustering

INTRODUCTION

A dramatic increase of network traffic leads to mixture of application problems running over internet. Network traffic occurred on the dynamic structure plays a crucial role on maintaining the data quality level. Another difficulty occurrence on the existing system is that the system is not effective on maintaining the privacy level on learning system.

Community Anomaly Detection System (CADS), is based on the access logs of mutual environments. The environment fails to expand supervised clustering model for network analysis. Locally Consistent Concept Factorization in extracts document concepts using clustering concept on semi-supervised model.

Semi-Supervised Semi-Riemannian Metric Map (S3RMM) follows geometric framework to maximize discrepancy of similarity measures. S3RMM consists of learning semi-Riemannian metrics but not integrated with active learning framework. A weighted accord function is established to guide active learning clustering rationale criteria with initial network prototype partitions.

A Peer-to-Peer document Clustering on the Hierarchically distributed system (HP2PC) permit privacy inside neighborhood data boundaries. Compression algorithm 2P2D is developed to attain the group movement patterns and reduce the amount of delivered data.

The K-means clustering and Self-Organizing Maps (SOMs) perform signal classification in the absence of training data. The data stream classification techniques address the concept-evolution problem by introducing time restraint for avoiding delaying data labeling and classification decision.

In this proposed study, main objective is to maximize class purity level of dynamic network structure using the Supervised Parametric Clustering (SPC) approach. The input data from the dataset are preprocessed by using the binning method where the data is partitioned in which the bin values are smoothened using the decision tree and remove the

Corresponding Author: D. Jayachitra, Department of Computer Science, Nehru Memorial College, Bharathiar University, Puthanampatti, Coimbatore, Tamil Nadu, India

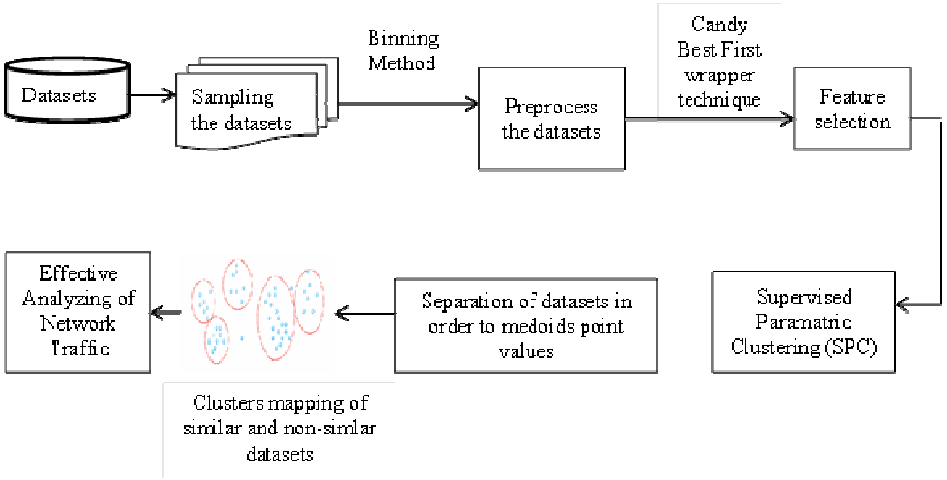

Fig. 1: Overall architecture diagram of SPC approach

noises. The preprocessed information along with the input data is given to the feature selection process where the Candy Best First wrapper algorithm extract the optimal features that makes the clustering process easier. Finally the Supervised Parametric Clustering (SPC) approach is applied to cluster the dataset, separate the datasets into similar and non-similar dataset which makes the dynamic network traffic analysis much better and increases class purity level on dynamic network structure.

LITERATURE REVIEW

An automated approach for activity tracking as illustrated in Rashidi *et al*. (2011) recognizes regular user behavior and logically measure the occurrence of an individual's schedule. The network routine fails to routinely select the number of objects based on the resident's lifestyle. Real world data from various fields achieves multi-relational and interconnected data. Yen *et al*. (2011) the two-step procedure is developed for analyzing the weighted graphs. But, the two step procedure is not applicable on real relational databases.

Guan *et al*. (2011), Seeds Affinity Propagation (SAP) based on a detailed similarity measurement and generic seeds construction strategy handles the cluster objects processing on varying time range. The concept-based mining model in Shehata *et al*. (2010) enhances the text clustering where the sentence, text and corpus level is increased. The direction is not linked to the web document clustering and not effect on performs the effective mapping concept. Personalized query clustering method as illustrated in Leung and Lee (2010) competent of deriving both of the user's positive and negative preferences. Lee and Lee (2010), a novel dissimilarity measure based on a dynamical system is associated with support estimating relational functions.

The aim of clustering is to locate intrinsic structures in data and systematize them into significant subgroups for further study and analysis. Multi-view point-Based Similarity Measure in Nguyen *et al*. (2012) measure similarity between data objects in sparse and high-dimensional domain mainly on the text documents The optic disc and optic cups (Sreelakshmi and Deepa, 2014) are segmented by utilizing the Principal Analysis Component (PCA) and Generalized Distance Function (GDF) to achieve the better resolution of grey scale images.

MATERIALS AND METHODS

The underlying principle of supervised clustering is used to analyze the dynamic network traffic. In the machine learning paradigm, SPC of a supervised learning algorithm partition the flows in the training data where it guides only by the similarity between the flows and predetermined labeling of the data flows. The result of the learning is a set of clusters, where the several systems are mapped with the different flow types in SPC method. The overall architecture Diagram of SPC approach on the dynamic network structure is described in Fig. 1.

As illustrated in Fig. 1, the general architecture of SPC method divides into different phases, namely preprocessing phase, feature selection phase, clustering phase and final stage is mapping phase. In preprocessing phase, traffic trace which composed is labeled to diverse application to improve the data class purity level. The class purity level is maintained using the binning method in SPC approach.

Subsequently, clear data points in SPC approach use the feature selection approach. The feature selection approach uses the Candy Best First wrapper technique to eliminate the redundant and irrelevant features to get optimal feature subset.

In the clustering phase, SPC employ machine learning approach called clustering to partition training data set that consists of scarce labeled flows combined

with abundant unlabeled flows. Separation in order of medoids value points partitions the training data set into different groups. Different groups are designed such that flows within a group are analogous to each other in SPC approach. The final stage in the SPC approach is the mapping phase, where the available labeled flows are mapped with the predefined labels to attain the effective network traffic result. The section given below describes the SPC approach step by step process.

Preprocessing: During preprocessing approach, binning method is employed to achieve higher class of purity level. The bin values are characterized based on the decision tree to perform smoothing process. The decision tree based smoothing process improves the accuracy of smoothing level by removing the noise on the data points. Decision tree uses the split a data points and log_2 is the logarithmic form of smoothing process to improve the speed rate. The separated impurity data points from the group are easily separated for speeding up the performance rate:

$$Binning\ based\ Splitting\ on\ different 'p' = \sum_{i \in p} \frac{|Splitting|}{|\ Splitting\ |}\ E\ (Splitting) \quad (1)$$

The binning method splits the pure and impure data points on 'P' different parameters. Splitting is carried out for the easy removal of outliers:

$$Smoothing\ factor = Binning\ based\ Splitting - Binning\ based\ Splitting\ on\ different\ 'p' \quad (2)$$

Feature selection: Feature selection is one of the most important and frequently used techniques in SPC method. Feature selection in SPC is carried out using the wrapper technique and each candidate feature subset is evaluated. Wrapper technique combines the preprocessing information in SPC approach to carry out the feature selection process.

Candy best first wrapper algorithm: Secondly, feature selection is carried out using the Candy Best First wrapper algorithm in SPC approach for balancing the network structure. The Candy Best First wrapper algorithm extracts the content to particular information source and translated into the relational form. The candy wrapper algorithm explores the graph by expanding the most promising nodes according to the specified rule. Efficient selections of the best features through the decision tree are normally used on producing the feasible feature selection result in SPC method. Candy best first wrapper technique is used for feature section on the combinatorial data points:

$$D(Splitting_i, Class) = \frac{1}{Splitting} \sum_{i \in splitting} Feature\ selected\ (i, classes) \quad (3)$$

The candy best first wrapper technique performs the effective feature selection to overcome error routines on different set of data points. The Candy Best First Wrapper algorithm is described as:

Begin
//Candy Best First Wrapper algorithm
Step 1 : Open with initial set of preprocessed data points
Step 2 : While data point on the Set 'S' is not empty
Step 3 : Do
Step 3.1 : Perform the Best First Selection of features
Step 3.2 : If 'i' is the feature selected data points, then return the information for clustering
Step 3.3 : Create 'i' successor of data points
Step 4 : Wrapper technique improves the feature selection processing on the 'i' successor of data points
Step 5 : Each such successor, add to the set and perform the evaluation
End

The above algorithmic step provides the step by step description of the feature selection processing on the SPC approach. The decision tree with different splitting purity classes identifies features on 'i' class samples. The wrapper technique develops feature selection processing on 'i' data points. The wrapper technique development improves the feature selection process for easier clustering process.

Supervised clustering process: The supervised parametric clustering with varying parameter uses fitness function ff (i) to perform clustering process. Here 'k' number of clusters is formed in SPC using different parametric input. The separation of the medoids point repeatedly and greedily adds the set of features to attain the higher fitness function. The supervised parametric clustering with mapping operation is shown through the diagrammatic form in Fig. 2.

The mapping operation is performed in SPC approach by exploring all possible replacements of single representatives. Initially, SPC based on separation of the medoids point value starts with set of feature selection representatives and best first approach inserts the new feature selected representatives into the set to improve the clustering efficiency. Secondly, SPC computes the fitness function to improve the clustering and mapping results together:

$$Fitness\ Function = \frac{ff(i)}{f(Higher\ class\ value\ of\ clustering)} \quad (4)$$

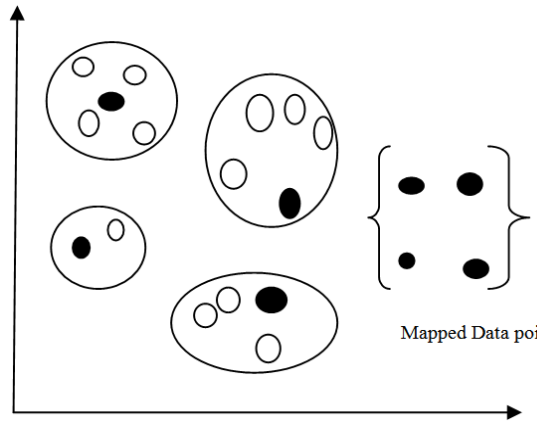

Fig. 2: Supervised parametric clustering with mapping operation

Finally, the fitness function is extended on improving the clustering efficiency and also effective analyzing of traffic on dynamic network structure through mapping operation. Separation in order of medoids value points in SPC approach clusters collaborative classes of data points.

RESULTS AND DISCUSSION

Supervised Parametric Clustering (SPC) approach on the dynamic network structure uses the KDD CUP Data Set from UCI repository. KDD cup dataset task is to differentiate various traffic data flow is analyzed using SPC mining of unstructured data portions of packets flow in the heterogeneous network. The two class classification problem with continuous input variables contains the 5 datasets of the traffic features computed using a two second time window. The data are split into training, justification and test set to implement the work using JAVA platform. Target values are provided only for the training and validation sets.

Supervised Parametric Clustering (SPC) approach are compared against the existing Community Anomaly Detection System (CADS), an unsupervised learning framework and Peer-to-Peer document Clustering on the hierarchically distributed system (HP2PC). Experiment is carried out on the factors such as performance speedup rate, false positive rate, class purity mapping rate, true positive rate, clustering efficiency and balancing time taken on dynamic network structure.

The performance speed up is characterized as the amount of effective clustering of the system based on the speed and time factor. The speed up process is measured in terms of speed percent:

$$SpeededUp\ rate = \frac{Balancing\ Time}{Level\ of\ Neighboring\ value} \quad (5)$$

The speed up process rate takes the balancing time of the dynamic structure for the easy identification of the process rate. The false positive rate usually refers to the probability of rejecting the faulty results from the sample classes. The false positive rate is measured in terms of percentage:

$$False\ Positive\ rate = \frac{False\ Probability\ ratio}{False\ positivie\ ratio + True\ Positive\ ratio} \quad (6)$$

The false and true positive rate is combined together and always value is equal to 1. The mapping is the rate at which the correct match result is carried out using the fitness function. The fitness function is computed in Eq. (5), where the high class function of clustering is carried out. True positive rate takes the entire accurate probability ratio on clustering the data points on the dynamic network structure. The true probability measure is formularized as:

$$True\ positive\ rate = \frac{Relevant\ Class\ sample\ \cap Retrieved\ Class\ sample}{Retrieved\ Class\ sample} \quad (7)$$

The above computed formula illustrates that the set of results relevant to the cluster divided by the retrieved results of class sample. Clustering efficiency is defined as the amount of effective clustering operation carried out using the separation in order of medoids value points. The clustering efficiency is measured in terms of efficiency percentage. The balancing time is computed as:

$$Balancing\ Time\ Bound = \frac{Start\ Time\ rate - End\ time\ rate}{Clustering\ Count} \quad (8)$$

The balancing time taken is defined as the amount of time taken to balance the overall process, measured in terms of seconds on the dynamic network structure. The below Table 1 and graph in SPC approach helps to easily analyze the result percentage rate using KDD CUP dataset.

The performance of the SPC approach is shown in Fig. 3 using the level of neighbor count level. The

Table 1: Tabulation for performance speedup rate

Level of neighbors count	Performance speed up rate (speed %)		
	CADS	HP2PC system	SPC approach
2	80	82	90
4	81	83	91
6	82	85	92
8	83	87	94
10	84	88	95
12	85	89	96
14	85	90	97

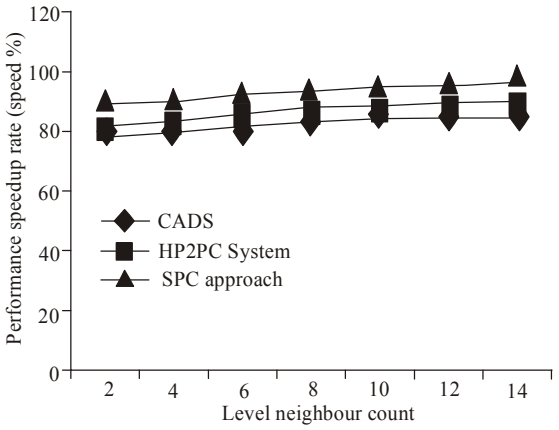

Fig. 3: Performance speedup rate measure

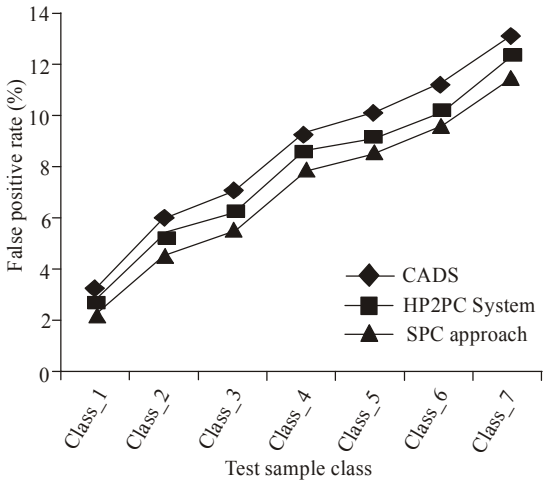

Fig. 4: Measure of false positive rate

Table 2: Tabulation of false positive rate

Test sample class	False positive rate (%)		
	CADS	HP2PC system	SPC approach
Class_1	3.18	2.672	2.21
Class_2	5.84	5.210	4.56
Class_3	6.92	6.120	5.51
Class_4	9.12	8.490	7.82
Class_5	9.99	9.030	8.45
Class_6	11.12	10.010	9.47
Class_7	12.97	12.210	11.38

Table 3: Class purity mapping rate tabulation

Data size for mapping (KB)	Class purity mapping rate (mapping %)		
	CADS	HP2PC system	SPC approach
150	80.5	82.5	90.5
300	81.6	83.6	91.2
450	83.5	84.9	93.5
600	84.9	85.2	94.1
750	85.3	87.6	95.6
900	87.6	89.1	97.0
1050	90.2	92.4	98.0

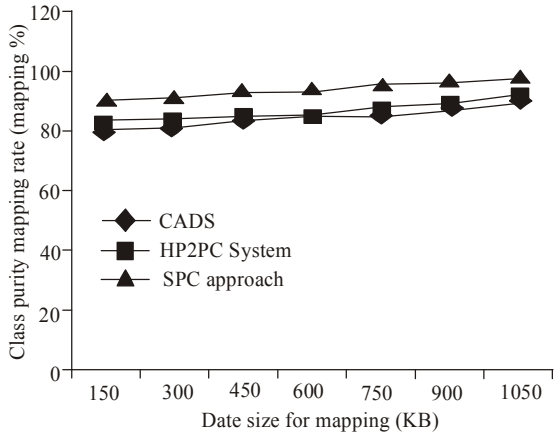

Fig. 5: Class purity mapping rate measure

performance speed up is computed on the CADS (Chen *et al*., 2012) HP2PC System (Hammouda and Kamel, 2009) and SPC Approach. The bin values usually characterize based on the decision tree and improves the performance speedup rate by 12-14% in SPC Approach when compared with existing CADS (Chen *et al*., 2012). The decision tree based smoothing process improves the performance rate without any noise level on the data points. Decision tree uses the split approach where the splitting is carried out using the class label of data points, thereby improves the performance rate by 7-9% when compared with existing HP2PC System (Hammouda and Kamel, 2009).

Table 2 describes the false positive rate based on the test sample classes. The sample class ranging from 1 to 7 is taken for the experimental work. The different set of the test sample produces the false positive ratio result.

Figure 4 illustrates the false positive ratio on CADS (Chen *et al*., 2012), HP2PC System (Hammouda and Kamel, 2009) and SPC Approach. The SPC Approach uses the Candy Best First wrapper technique to eliminate the redundant features. The removal of the redundant features reduces the false positive rate by 12-30% in SPC Approach when compared with CADS (Chen *et al*., 2012). For instance, Class_4 sample produce the 7.82% of false positive result in SPC Approach and 9.12% false positive result on CADS. The HP2PC System result is about 8.49% and also increased percent result when compared with proposed SPC Approach. The elimination is carried out to reduce the false rate, where the SPC Approach is 5-17% better in result when compared with HP2PC System (Hammouda and Kamel, 2009) (Table 3).

Figure 5 describes the class purity mapping based on the data size. The data size is measured in terms of Kilo Bytes (KB). The class purity mapping level is maintained using the binning method in SPC approach. The labeled flows are mapped with the predefined labels to attain the effective network traffic result with 8-12% improved result in SPC approach when compared with CADS (Chen *et al*., 2012). The class purity level on the dynamic network structure is

Table 4: Tabulation of true positive rate

Sample class user count	True positive rate (precision)		
	CADS	HP2PC system	SPC approach
20	83	88	91
40	84	90	93
60	85	91	95
80	86	91	96
100	87	92	96
120	87	93	97
140	88	94	98

Table 5: Tabulation for clustering efficiency

Data information size (KB)	Clustering efficiency (efficiency level value)		
	CADS	HP2PC system	SPC approach
1000	75	68	79
2000	77	69	80
3000	78	70	81
4000	80	71	82
5000	81	72	83
6000	82	74	86
7000	83	73	88

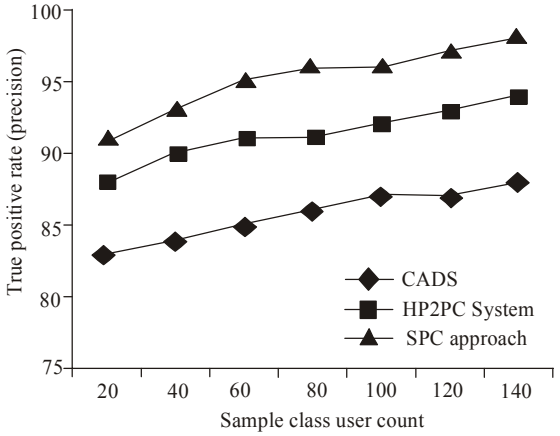

Fig. 6: Measure of true positive rate

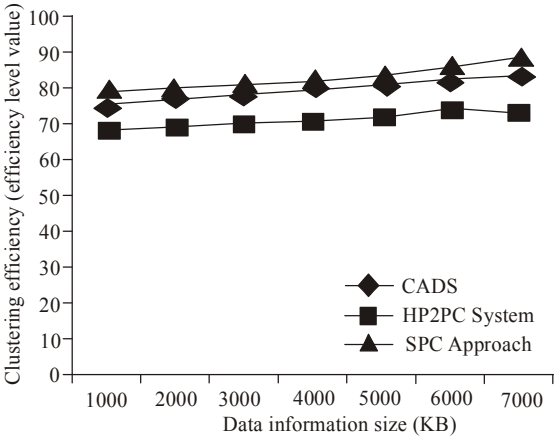

Fig. 7: Clustering efficiency measure

analyzed effectively, thereby improves the mapping rate by 6-10% when compared with existing HP2PC System (Hammouda and Kamel, 2009). The larger the data size range leads to the higher the mapping rate efficiency (Table 4).

Table 6: Tabulation of balancing time measure

Clustering count	Balancing time taken on dynamic structure (sec)		
	CADS	HP2PC system	SPC approach
5	12	702	685
10	1654	1677	1552
15	2965	3070	2822
20	3592	3632	3491
25	4628	4777	4519
30	5449	5477	5272
35	6648	6568	6337

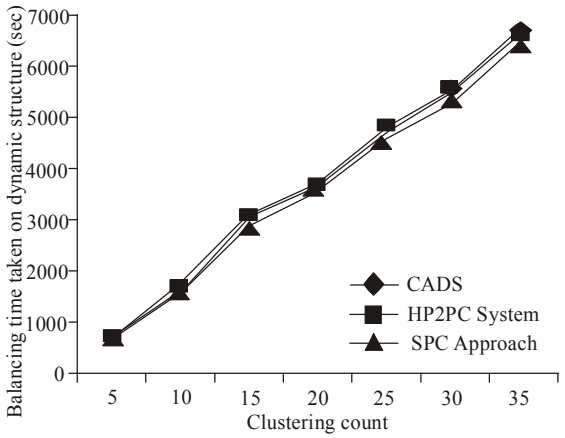

Fig. 8: Measure of balancing time taken on dynamic structure

The true positive rate compares the result of the proposed SPC approach with existing CADS (Chen *et al*., 2012), HP2PC System (Hammouda and Kamel, 2009) is shows in Fig. 6. True positive rate is measured based samples class user count. The true positive result percentage of SPC approach get is improved by using the best features through the decision tree. The candy wrapper technique uses the features of the decision tree with different splitting purity classes. The splitting classes help to improve the true positive rate by 9-11% in SPC approach when compared with CADS. The true positive rate is also improved by 3-5% when compared with existing HP2PC System (Hammouda and Kamel, 2009) (Table 5).

Clustering efficiency is measured based on the data information size and shown in Fig. 7. The data information size is taken about 1000 to 7000 KB for the experimental work. The mapping operation is performed in SPC approach by exploring all possible replacements of single representatives. The best first approach inserts the new feature selected representatives into the set to improve the clustering efficiency by 2-6% in SPC approach when compared with existing CADS (Chen *et al*., 2012). SPC based on separation of the medoids point value uses the representatives in SPC approach and then improves the clustering efficiency rate by 15-20% when compared with existing HP2PC System (Hammouda and Kamel, 2009) (Table 6).

Figure 8 depicts the balancing time taken on the dynamic structure. The dynamic structure measures the balancing time based on the clustering count. The clustering count ranges from 5, 10, 15, 20 up to 35, respectively. The scarce labeled flows combined with abundant unlabeled flows to balance the time factor and produce the 2-6% lesser time in SPC approach when compared with CADS. Separation in order of medoids value points in SPC approach separates the data objects for clustering. The effective clustering reduces the balancing time by 2-8% when compared with HP2PC System (Hammouda and Kamel, 2009).

As final point, Supervised Parametric Clustering (SPC) approach maximizes the class purity level with higher result percentage when compared with existing state of art methods. Candy Best First wrapper technique extracts the content to the particular information source using the relational form.

CONCLUSION

Supervised Parametric Clustering (SPC) approach is developed to increase class purity level on dynamic network structure. Initially, SPC approach achieved smoothening of information using preprocessing step. The preprocessing step is carried out using binning method. Binning method in SPC approach handles varying neighborhood value range among dynamic network structure information clustering. Secondly, smoothed information moves to next step of feature selection using Candy Best First wrapper algorithm. Candy Best First wrapper algorithm selects the features for easier clustering process with higher efficiency rate. Finally, the selected feature points achieve clustering process using the separation in order of medoids value points. The separation in order of medoids value points performs collaborative classes of data points using fitness function. In conclusion, mapping of the SPC result is carried out for effective analyzing of traffic on dynamic network structure with 9.462% lesser false positive rate and improves class purity mapping rate by 9.06%.

REFERENCES

Chen, Y., S. Nyemba and B. Malin, 2012. Detecting anomalous insiders in collaborative information systems. IEEE T. Depend. Secure, 9(3): 322-344.

Guan, R., X. Shi, M. Marchese, C. Yang and Y. Liang, 2011. Text clustering with seeds affinity propagation. IEEE T. Knowl. Data En., 23(4): 627-637.

Hammouda, K.M. and M.S. Kamel, 2009. Hierarchically distributed peer-to-peer document clustering and cluster summarization. IEEE T. Knowl. Data En., 21(5): 681-698.

Lee, D. and J. Lee, 2010. Dynamic dissimilarity measure for support-based clustering. IEEE T. Knowl. Data En., 22(6): 900-905.

Leung, K.W.T. and D.L. Lee, 2010. Deriving concept-based user profiles from search engine logs. IEEE T. Knowl. Data En., 22(7): 969-982.

Nguyen, D.T., L. Chen and C.K. Chan, 2012. Clustering with multiviewpoint-based similarity measure. IEEE T. Knowl. Data En., 24(6): 988-1001.

Rashidi, P., D.J. Cook, L.B. Holder and M. Schmitter-Edgecombe, 2011. Discovering activities to recognize and track in a smart environment. IEEE T. Knowl. Data Eng., 23(4): 527-539.

Shehata, S., F. Karray and M.S. Kamel, 2010. An efficient concept-based mining model for enhancing text clustering. IEEE T. Knowl. Data En., 22(10): 1360-1371.

Sreelakshmi, B.L. and M. Deepa, 2014. Optic disc and optic cup detection using superpixel classification based on PCA and mathematical morphology. Int. J. Invent. Comput. Sci. Eng., 1(2).

Yen, L., M. Saerens and F. Fouss, 2011. Link analysis extension of correspondence analysis for mining relational databases. IEEE T. Knowl. Data En., 23(4): 481-495.