Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 7, July 2014, pg.47 – 59

SURVEY ARTICLE

SURVEY ON HETEROGENEOUS NETWORK TRAFFIC ANALYSIS WITH SUPERVISED AND UNSUPERVISED DATA MINING TECHNIQUES

D.Jayachitra[1], Dr. J. Jebamalar Tamilselvi[2]

[1]Research Scholar, Bharathiar University, Coimbatore and Assistant Professor, Department of Computer Science, Nehru Memorial College, Puthanampatti, Trichy District, Tamil Nadu, India

[2]Director, Department of MCA, Jaya Engineering College, Chennai, Tamil Nadu, India

[1] jayadchitra@gmail.com; [2] jebamalar@gmail.com

Abstract— Network Traffic Analysis (NTA) in heterogeneous networks is one of the emerging research areas receiving substantial attention from both the research community and traffic analyzers. Many tasks in NTA can be naturally cast in a supervised and unsupervised learning model. Many supervised classification models and unsupervised clustering learning models in data mining have been proposed for heterogeneous network. Due to the importance of network traffic analysis in data mining research with the rapid development of new models, To provide a comprehensive review on supervised classification and unsupervised clustering model on heterogeneous type of network in this paper and systematically give a summarization of the state-of-the-art techniques for network traffic analysis. It addresses the problem of network management such as traffic load, quality of service, and trend analysis. This survey covers real time supervised classification and unsupervised clustering algorithms and analyze techniques for heterogeneous networks. It provides taxonomy of the different supervised classification algorithms and unsupervised clustering algorithms and evaluates the various performance metrics that are significantly used for the purpose of comparison. A detailed review is provided covering fuzzy relational clustering algorithm, classification learning algorithms, global voting algorithm and hybrid algorithms. The survey evolve certain open issues, key research challenges for network traffic analysis using supervised classification and unsupervised clustering model in heterogeneous networks, and likely to provide productive research directions.

Key Words--- Supervised and Unsupervised Mining, Traffic Data Analysis, Heterogeneous Network

I. STATE OF ART

The growing population of the aged and the disable is leading to expansion of autonomous service systems. In data mining the data appear in limitless stream for classification of data stream.

The problem of data stream classification, where the data enter in an unreal unlimited stream and the probability to evaluate each record is briefed. The problem is solved with the existence of stream classification algorithm.

Sparse coding as demonstrated in [14] which fundamentally challenged to find an embedding for the data by assigning feature values based on subspace cluster membership. A direct application of sparse coding resulted in a collapse of knowledge relocate to sparse coding, by incorporating distribution distance approximate for the embedded data.

Bayesian learning and expectation-maximization (EM) techniques were developed under the proposed generative model as shown in [17] for recognizing new training data for learning new unseen sites. Previously unseen attributes combined with their semantic labels were also exposed through another EM- based on the generative model.

Segmentation algorithm is applied on this signal that automatically estimates the number of partitions and the partition borders as presented here [2] fails in holding each subtrajectory of the sampling set by different subtrajectories of the MOD (cluster), under the minimization of objective. Space-efficient algorithms maintain duplicate-insensitive order sketches so that rank-based queries are roughly processed with relative rank error that guarantees in the presence of data duplicates. Besides the space efficiency, the algorithm is time-efficient and highly accurate in [9]. Moreover, one scan algorithm is practical to the heavy hitter problem using distinct elements when compared to the existing fault-tolerant distributed communication techniques.

For the discovery of a significant arrangement of data generated by human behavior, a clustering technique capable of detecting outliers is often employed. To be specific, Possibilistic c-Means (PCM), Fuzzy Possibilistic c-Means (FPCM) and Possibilistic Fuzzy c-Means (PFCM) are robust against outliers. However, they suffer from the local optimum problem and need to find a suitable means of combining and adjusting several free parameters to achieve optimal performance.

Anomaly detection aims to recognize a minute group of instances which deviate remarkably from the accessible data. A well-known definition of outlier is that given an observation which deviates so much from other observations, as to arouse the uncertainties behavior generated by different mechanism, it gives the universal idea of an outlier and encourages many anomaly detection methods.

Detecting anomalous insiders in collaborative information systems as shown in [1] intend to analyze the impact of such information in the future. The goal of the current work was to determine the basic information in the access logs and Meta information for the subjects in anomaly detection. On line alert aggregation based on a active, probabilistic model in [18] essentially are regarded as a data stream version of a maximum likelihood approach for the estimation of the model parameters.

An online oversampling principal component analysis (OSPCA) illustrated in [8] aims in detecting the occurrence of outliers from a great amount of information via online update procedure. Fuzzy-state Q-learning (FSQL) process is incorporated, which is capable of learning human behavior patterns in a non-supervised manner and predicting subsequent human actions. In the latter case, interaction between certain users as shown in [7] often affects their choice of actions, and thus the situation of action learning is quite complicated.

Error terms augment the standard sum of squared error computational experiments as shown in [16] that the modified learning method helps to extract fewer rules without increasing individual rule complexity and without decreasing classification accuracy. Ontology-based fuzzy video semantic content model uses spatial/temporal relations in event and conception definitions supply a wide domain pertinent rule construction average.

Fuzzy video semantic content as shown in [10] allows the user to construct ontology for a given domain. In addition to domain ontology additional rule definitions are used to lower spatial relation computation cost and to identify some complex situations more effectively.

Fascinatingly, the idea of fuzzy partitioning based on relational data is not novel, and can be traced. The purity of a cluster is defined as the fraction of the cluster size that the largest class of objects assigned to that cluster in [3] fails to extend these ideas to the development of a hierarchical fuzzy relational clustering algorithm. To derive a novel method for measuring similarity between the data objects in sparse and high dimensional field as shown in [15], same principle can be made use of. But alternative forms are defined for the relative similarity and do not use average but have other methods to combine the relative similarities according to the different viewpoints.

More specially, show that hubness, i.e., the tendency of high-dimensional data to enclose points that frequently occur in k-nearest neighbor lists of other points in [20]. The hubness was successfully exploited in clustering. The cluster-adaptive distance bound based on separating hyper plane boundaries of Voronoi clusters in [5] enables well-organized spatial filtering, with a comparatively small preprocessing storage space overhead and is applicable to euclidean and Mahalanobis similarity measures. The large organizations will not retrieve and process petabytes of data, for various purposes such as data mining and decision support. Thus, there exist numerous applications that access large multimedia databases, which fail in efficient support.

Fast smallest amount spanning tree-inspired clustering algorithm uses an efficient implementation of the cut and the rotation property of the minimum spanning trees. The rich properties of the MST algorithms fails in adapting MST inspired clustering algorithm [12] to more general and larger data sets, primarily when the whole data set cannot fit into the main memory. MAximal Resemblance Data Labeling (MARDL) allocate each unlabeled data point into the matching suitable cluster based on the narrative categorical clustering representative. To detect the drifting concepts at different sliding windows, DCD contrast the cluster distributions in the middle of the last clustering consequence and the temporal current clustering result [11].

However, since only an inadequate amount of labeled data are available in the above real world applications, how to establish anomaly of unseen data (or events) draws attention from the researchers in data mining and machine learning communities. Deploying the semantics embedded in web services request and present semantic web services, but the sharing of knowledge is not addressed [4]. Multivariate Reconstructed Phase Space (MRPS) for recognizing multivariate temporal patterns as shown in [13] is characterized by identifying the anomalies or events in a dynamic data system.

Supervised model is based on training a data sample from data source with accurate clustering. Network traffic clustering is an important and challenging problem. The objective is to conclude the applications that produce a certain group of packets, such as video, peer-to-peer, gaming, email etc. However approach is no longer effective as there are now many different kinds of network applications, some of which deliberately change their behavior in order not to be detected.

Another complexity is that due to isolation requirements and computational problem, it is envisage that classification algorithms are allowed to use only partial information present in the network data and avoid deep packet inspection (DPI). Classification is one of the most frequently encountered decision making tasks. Extending pattern classification hypothesis and design methods to adversarial settings in [19] is extremely pertinent, which has not yet been pursuing in an efficient way.

Data stream classification techniques address the concept-evolution problem which is a major problem with data streams that must be dealt with. A more realistic solution to data stream classification introduces time constraints for postponed data labeling and creating classification decision in [6]. On the other hand, XM properly distinguish among concept drift and concept-evolution, stay away from false detection. Therefore, it fails in considering most of the novel classes as normal data, yielding very high false negative rate.

There are many industrial problems identified as classification problems. For the difficulty of solving such problem precisely lies in the accuracy and distribution of data properties and model ability.

Analyzing the flow sequences of packets during the communication between pairs of hosts aims in:

Clustering of network packets analyzes the traffic flow using supervised model

Considering seed points in such a way that they are distant enough to be perfectly classified into different categories and control the network traffic to perform non linear dimensionality reduction

Control and maintain the heterogeneous network traffic in an effective way by integrating the clustering of network packets and perfectly classified into different categories.

This paper is organized as follows: Section II discusses classification difficulty occurs on the network traffic, Section III shows the study and analysis of the existing clustering classification rule techniques in data mining, identifies the possible comparison between them and Section IV concludes the paper, key areas of research is given as making use of clustering and classification rule mining in network traffic analysis, control and maintenance.

II. SURVEY ON SUPERVISED AND UNSUPERVISED MINING TECHNIQUES

In machine learning, the difficulty of unsupervised learning on network traffic is that to discover hidden nodes in unlabeled data. Since the instances specified to the learner are unlabeled, there is error-free for estimating a possible solution. These instances distinguish unsupervised learning from supervised learning.

Unsupervised learning is associated to the problem of density judgment in network traffic analysis. However unsupervised learning also encompasses many other techniques that seek to recapitulate and explain key features of the data. Many of the existing work were based on data mining to preprocess information in unsupervised learning for effective development of complex models.

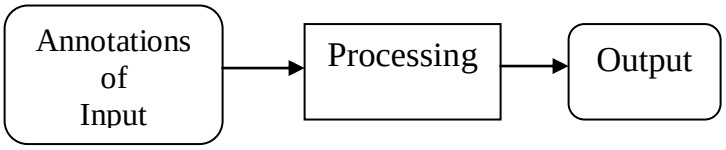

Fig 1 Supervised Model

Fig 2 Unsupervised Model

Figure 1 and 2 illustrates the difference in the fundamental structure of supervised and unsupervised learning. Supervised and unsupervised model are also combined together where both input annotations and latent variables are assumed to have caused the output annotations.

From the theoretical point of view, supervised and unsupervised learning change only in the causal structure of the network traffic model. In supervised learning, the model defines the result of annotations, called inputs, and has another set of annotations, called outputs. In other words, the inputs are assumed to be at the beginning step and outputs at the end of the fundamental chain. The models include intermediate variables between the inputs and outputs.

In unsupervised learning, all the annotations are assumed to be caused by hidden variables, that is, the annotations are assumed to be at the end of the causal chain. In supervised learning models, it repeatedly leaves the probability for inputs undefined. The model is not desirable as long as the inputs are obtainable, but if some of the input values are missing, it is not possible to infer anything about the outputs. If the inputs are also modeled, then the missing inputs reason causes no problem because the measured variables are unsupervised learning.

A. Community Anomaly Detection System

An unsupervised learning structure based on Community Anomaly Detection System (CADS) is used to notice the insider threats. The access logs of shared environments are based on the annotations of distinctive CIS users, who have a tendency to form neighborhood structures based on the subjects accessed. CADS consist of two working components namely relational pattern extraction and anomaly prediction.

Relational pattern extraction is a procedure to transform the access logs of a CIS into active neighborhood structures using amalgamation of graph-based modeling and dimensionality reduction techniques over the accessed subjects. Anomaly prediction leverages a statistical model to decide when the users have adequately diverged from communities.

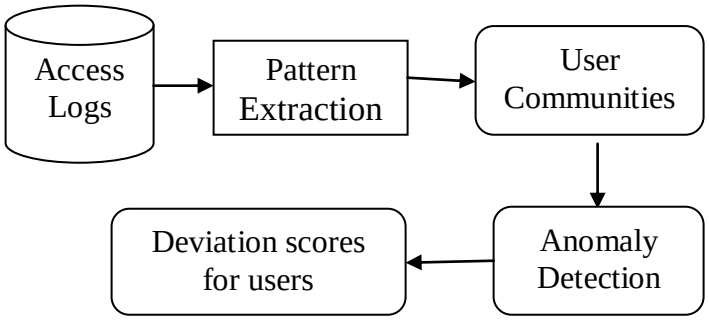

Fig 3 Architectural overview of the CADS framework

Fig 3 describes the architecture flow of the CADS framework with CIS access logs. CADS do not explicitly document the social structure of the organization. In recognition of this deficiency, CADS-PE leverages the associations between users and subjects to assume communities. To accomplish this derivation, the access transactions are translated into a tripartite graph of users, who are mapped to subjects, and then to semantic grouping. Users were found to deviate significantly from expected behavior and are considered to be anomalous.

To achieve this evaluation, the users are projected onto the spectrum of communities to compute the distance between every user and their neighbors in the network. The superior the distance between the user and their neighbors the better likelihood that the user is said to be anomalous. The algorithmic step for the network community profile minimization is shown

// Algorithm for network community profile minimization

Input: Distance Matrix
1: S: Initialize to all possible neighbors
2: For i= 1 to S do
3: N={}

4: For j= 1 to S do
5: N ←NUi
6: i-nearest neighbor network for user
7: End For
8: For j= 1 to S do
9: k ←i the conductance function
10: Reset Evaluation set
11: End For
12: End For

To detect anomalous insiders in a CIS, CADS, a community-based anomaly detection model utilizes a relational framework. To predict which users are anomalous, CADS analyze the deviation of users based on their nearest neighbor networks. CADS fail in extending into MetaCADS to incorpate the semantics of the subjects accessed by the users [1]. Unsupervised relational models in CADS display better performance at detecting anomalous users in collaborative domains than supervised models. Moreover, for better effectiveness, MetaCADS, the rate of intruding are generic and capable of inferring the mutual behavior of users in many settings.

B. Global Voting Algorithm Based on Representativeness

Global Voting Algorithm (GVA) is achieved based on local density and trajectory match information. The sequence of this descriptor over a trajectory gives the voting signal of the trajectory, where high values match to the majority of representative parts. Then, a novel segmentation algorithm is applied on this signal that estimates the number of partitions and the partition borders recognize homogenous division relating to their representativeness. As a final point, a sampling method over the ensuing segments gives up the majority representative subtrajectories in the Moving Object Database (MOD).

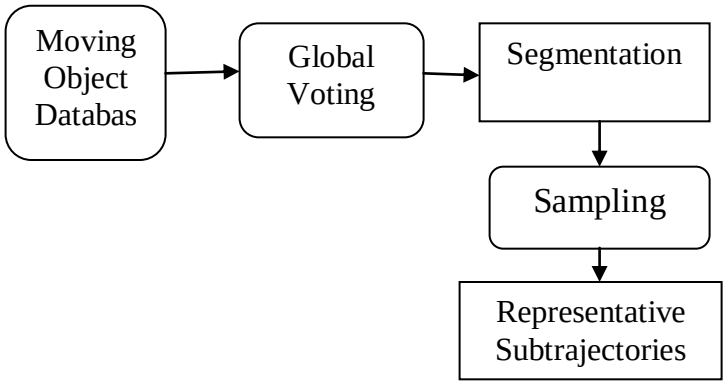

Fig.4 System Architecture of Global Voting Method

GVA is principally a stratified sampling technique, where strata are produced by the clustering algorithm, has the restriction that it is user supervised and it depends on the results of the clustering. The superiority of approach is compared to uniform random and stratified sampling techniques. Global Voting Algorithm is described below

Input: An Indexed database
Output: Voting vector V_k
1: For i=1 to L_k
2: $V_k(i) = 0$
3: Repeat

4: Normalized the trajectory voting vector V_k
5: Segmentation of L_k partitions
6: Subtrajectory Sampling Algorithm with normalized lifespan vector
7: Sampling Set Sorted $S_k(i)$
5: End For

The above steps are used for addressing the issue by segmentation and subtrajectory sampling based on global spatiotemporal similarity of trajectories. GVA extends the density biased sampling from point sets to trajectory segments providing a local trajectory descriptor per line segment that is related to line segment representativeness. Next, Trajectory Segmentation Algorithm (TSA) mechanically and efficiently estimates the number of subtrajectories and their borders, separating each trajectory of MOD into homogenous partitions concerning their representativeness.

To end with, Subtrajectory Sampling Algorithm (SSA) is applied over the resulting partitions providing the most representative subtrajectories of the MOD, also taking into account that high density regions of the MOD should not be oversampled. SSA is terminated by threshold, where the number of moving objects of the original MOD is represented.

C. Fuzzy Relational Clustering Algorithm at Sentence-Level Text

Fuzzy clustering algorithm operates on relational input information; (i.e.,) data in the form of pair wise comparison square matrix amongst data entity. The algorithm uses a graph demonstration of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an entity in the graph is interpreted as probability. Results of applying the algorithm to sentence clustering tasks show that the algorithm is competent of identifying overlapping clusters of semantically related sentences, and that it is consequently of potential use in a variety of text mining tasks.

// Fuzzy Relational Clustering Algorithm

Inputs: Pair wise similarity values between the sentence and cluster set
Output: Cluster membership values 'C'
//Initialization
1: For i =1to N
2. For m =1 to C
3. $P_i^m = rnd$
4. End for
//Expectation Step
5: For m = 1 to C
6: Weighted Similarity matrix for cluster m
7: End For
8: Calculate Page Rank scores for cluster m
9: Repeat until convergence
10: Assign Page Rank scores to likelihoods
11: Calculate new cluster membership values
// Maximization Step
12: For m = 1 to C
13: Update mixing coefficients
14: End For
15: End For

Fuzzy Relational Clustering algorithm uses the Page Rank score of an object within a cluster for determining the centrality to that cluster. The only parameters that need to be determined are the cluster membership values and mixing coefficients. The algorithm uses Expectation Maximization to optimize these parameters.

The cluster membership values are initialized arbitrarily, and normalized such that cluster membership for an object sums to unity over all clusters. E-step calculates the Page Rank value for each object in each cluster. Page Rank values for each cluster are calculated with the similarity matrix weights obtained by scaling the similarities by their cluster membership values. Once Page Rank scores have been determined, these are treated as likelihoods and used to calculate cluster membership values. Maximization step involves only the single step of updating the mixing coefficients based on membership values calculated in the Expectation Step.

D. Hybrid approach for context-aware service discovery

An integrated environment intended at providing user's context interest by deploying the semantics entrenched in web services. The main idea of the work is related to augment with qualitative representation of context underlying data by means of Fuzzy Logic in order to recognize the context and to consequently find the right set between the available ones. Semantic formalisms enable the context and services modeling in terms of domain ontology notion. Furthermore, the work defines hybrid architecture which achieves a synergy in the middle of the agent-based example and the fuzzy modeling.

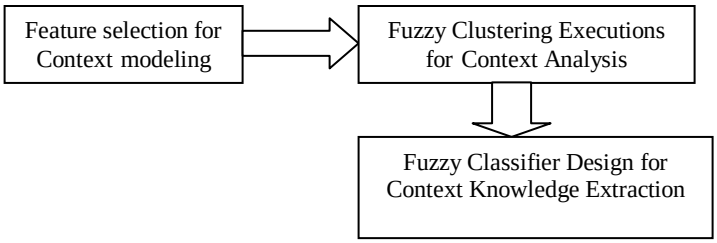

Fig 5 Diagrammatic Form of Context Training Phase

Context Training Phase use techniques of soft computing and semantic web in order to obtain and examine context information. Context Training Phase carries out mathematical models to process context data and trains itself according to the composed knowledge. The process of unsupervised fuzzy data analysis facilitates to augment context modeling with qualitative representation of underling data.

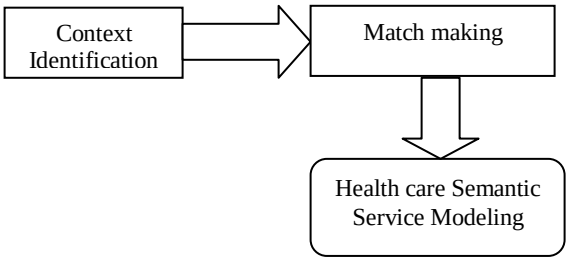

Fig 6 Diagrammatic Form of Context Aware Services Discovery Phase

Context Aware Services Discovery Phase retrieves semantic web services which suitably meet the user's context. A hybrid approach is described based on soft computing and using logic matching

evaluation in order to evaluate matchmaking in the middle of parameters and their values. Specifically, the location is stressed when no faithful match occurs between context and services. So, hybrid approach based on soft computing and merely logic matching assessment is defined.

III. COMPARISON OF CLUSTERING TECHNIQUE & SUGGESTIONS

In order to compare the execution time of the supervised and unsupervised framework, set of record classes are taken to perform the experiment. The initial metric is the execution time of different existing system, is defined as the time taken to perform the classification process on multi dimensional data. The second performance metric error rate is the number of bit errors occurred on supervised and unsupervised data stream.

The comparison takes place on existing Community Anomaly Detection System (CADS), Segmentation and Sampling of Moving Object Trajectories Based on Representativeness via Global Voting Algorithm (GVA), Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm (FRCA), Hybrid approach for context-aware service discovery in healthcare domain (Hybrid Approach). A survey and contribution on classification supervised and unsupervised techniques are available from recent research.

A. Classification Error rate

Existing Technique	Classification Error rate (%)
CADS	0.2647
GVA	0.4588
FRCA	0.6675
Hybrid Approach	0.8512

Table 3.1 Tabulation for Classification Error rate of different existing technique

The above table (Table 3.1) describes the error rate of the CADS classification, GVA and FRCA and hybrid approach. Error percentage of CADS [1] is lesser when compared to the all other existing technique. The raw data illustrating the effects of error rate on different techniques are shown in Fig. 3.1.

More classification unsupervised techniques developed feature-based and class-based measures searching but the classification accuracy are not improved in diagnosing the network traffic cause. Error rate of CADS is decreased using the partitions with dissimilar degrees of data stream class diversity. Comparatively, it is 8 – 12 % lesser in CADS when compared with FRCA.

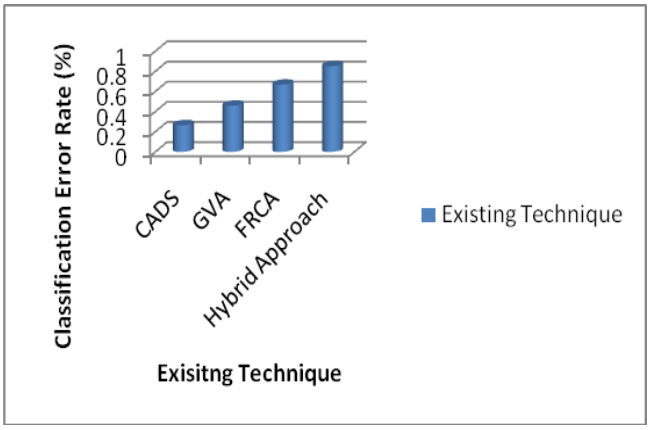

Fig 3.1 Classification Error rate of different technique

Fig 3.1 demonstrates the error rate of different techniques. The usage of clearance in the CADS, an unsupervised framework decreases the classification error rate when compared with the other existing algorithms. Classification rule improves the searching and classification accuracy reducing error rates on unsupervised learning. The cause for the network traffic is identified with the class labels in a classifying attributes.

Existing papers has reviewed the potential of the classification learning algorithms. An associative classifier does not follow the classification accuracy maximization paradigm i.e., it commonly portray the training data stream. Survey has reviewed the searching efficiently using classification algorithm for diagnosing the network traffic cause with improved classification accuracy.

Table 3.2: Execution Time

No. of Data stream classes	**Execution Time (ms)**			
	CADS	**FRCA**	**Hybrid approach**	**GVA**
5	891	757	702	656
10	881	751	794	658
15	883	774	717	674
20	895	793	716	683
25	884	709	724	682
30	883	795	726	691
35	874	776	738	696

Table 3.2 Tabulation for execution time on supervised and unsupervised framework

As the data stream classes on supervised and unsupervised framework increases, execution time is reduced in the segmentation and sampling of moving objects via GVA. The experiment shows that GVA primarily clusters whole trajectories and greatly brings down the time while performing the execution when compared with the CADS, FRCA and Hybrid approach. Graph shows that the clustering via GVA shows conspicuous advantage in controlling the network traffic.

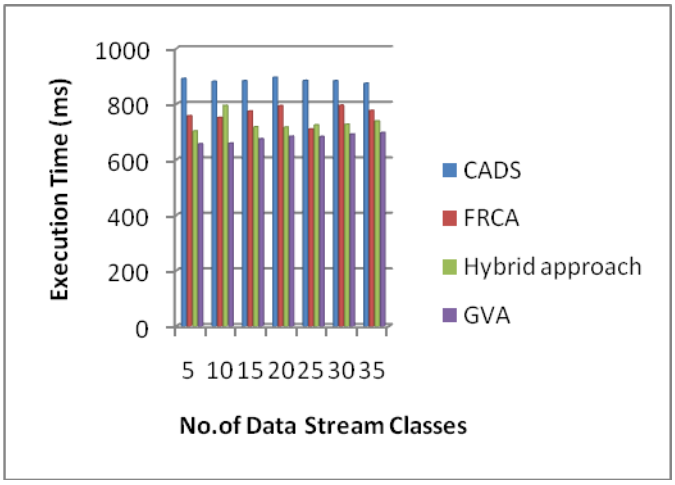

Fig. 3.2 Execution time on supervised and unsupervised framework

Fig 3.2 describes the execution time based on the record data stream classes. The execution time is measured in terms of milliseconds (ms). Clustering via GVA is 10 – 15 % lesser delay time taken when compared with the FRCA and 15 -20 % lesser running time taken when compared with the hybrid approach. Classification in CADS is approximately 25% lesser running time taken for when compared with the GVA.

The proposal cause a way to forward

- The clustering of network packets analyzing the sequence flow of the packets during the communication between pairs of hosts.
- Offers non linear dimensionality reduction to perfectly classify into different categories and control the network traffic.
- Reduced space dimension classifier with integration of supervised and unsupervised pattern to maintain the heterogeneous traffic.

Unsupervised and supervised learning models on heterogeneous network provides effective network traffic analysis, control and maintenance The cause for the heterogeneous traffic is identified with the class labels data stream in classifying attributes. The presence of the network traffic is identified with the unsupervised classification algorithm. Based on the training data stream, network traffic control rate is measured. Clustering accuracy is enhanced by enhancing the communication level. Unsupervised classification efficiency can be achieved easily using the principle component analysis form.

IV. **CONCLUSION**

Discussion about the existing Community Anomaly Detection System, Segmentation and Sampling of Moving Object Trajectories via Global Voting Algorithm, Novel Fuzzy Relational Clustering Algorithm, and Hybrid approach for context-aware service discovery in healthcare domain on execution time and classification error rate parameter dynamically adjusts based on the data stream.

Existing community anomaly detection system (CADS), an unsupervised learning framework based on the access logs of collaborative environments fails to develop the supervised clustering model for

network analysis. Existing Global Voting Algorithm based on local density and trajectory similarity information forms a local trajectory descriptor. GVA primarily clusters whole trajectories and is not customized to recognize the classified patterns of sub trajectories in an unsupervised way. In addition the efficiency is not considered so classification efficiency handling is focused. Hence the classification technique helps in satisfying the efficient way of controlling the network traffic.

Fuzzy clustering approaches based on prototypes of Gaussians are generally not applicable to integrate the supervised and unsupervised model for network control and maintenance. Extensive experiments evaluate the relative performance of the various algorithms and combinations of unsupervised and supervised model. The result shows that the classification technique outperforms consistently over a wide range of experimental parameters.

REFERENCES

[1] You Chen., Steve Nyemba., and Bradley Malin., *Detecting Anomalous Insiders in Collaborative Information Systems,* IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 9, NO. 3, MAY/JUNE 2012

[2] Costas Panagiotakis., Nikos Pelekis., Ioannis Kopanakis., Emmanuel Ramasso., and Yannis Theodoridis., *Segmentation and Sampling of Moving Object Trajectories Based on Representativeness*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 7, JULY 2012

[3] Andrew Skabar., and Khaled Abdalgader., *Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013

[4] G. Fenza., D. Furno., V. Loia ., *Hybrid approach for context-aware service discovery in healthcare domain,* Journal of Computer and System Sciences., Elsevier Journal., 2012

[5] Sharadh Ramaswamy., and Kenneth Rose., *Adaptive Cluster Distance Bounding for High-Dimensional Indexing,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[6] Mohammad M. Masud., Jing Gao., Latifur Khan., Jiawei Han., and Bhavani Thuraisingham., *Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[7] Sang Wan Lee., Yong Soo Kim., and Zeungnam Bien., *A Nonsupervised Learning Framework of Human Behavior Patterns Based on Sequential Actions,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 4, APRIL 2010

[8] Yuh-Jye Lee., Yi-Ren Yeh., and Yu-Chiang Frank Wang., *Anomaly Detection via Online Oversampling Principal Component Analysis,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013

[9] Ying Zhang., Xuemin Lin., Yidong Yuan., Masaru Kitsuregawa., Xiaofang Zhou., and Jeffrey Xu Yu., *Duplicate-Insensitive Order Statistics Computation over Data Streams,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 4, APRIL 2010

[10] Yakup Yildirim., Adnan Yazici., and Turgay Yilmaz., *Automatic Semantic Content Extraction in Videos Using a Fuzzy Ontology and Rule-Based Model,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013

[11] Hung-Leng Chen., Ming-Syan Chen., and Su-Chen Lin., *Catching the Trend: A Framework for Clustering Concept-Drifting Categorical Data,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009

[12] Xiaochun Wang., Xiali Wang., and D. Mitchell Wilkes., *A Divide-and-Conquer Approach for Minimum Spanning Tree-Based Clustering,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 7, JULY 2009

[13] Wenjing Zhang., and Xin Feng., *Event Characterization and Prediction Based on Temporal Patterns in Dynamic Data System,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2013

[14] Brian Quanz., Jun (Luke) Huan., and Meenakshi Mishra. Knowledge Transfer with Low-Quality Data: A Feature Extraction Issue, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 10, OCTOBER 2012

[15] Duc Thang Nguyen., Lihui Chen., and Chee Keong Chan., *Clustering with Multiview point-Based Similarity Measure,* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012

[16] Thuan Q. Huynh., and James A. Reggia, "Guiding Hidden Layer Representations for Improved Rule Extraction from Neural Networks," IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 22, NO. 2, FEBRUARY 2011.

[17] Tak-Lam Wong and Wai Lam., "Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 4, APRIL 2010.
[18] Alexander Hofmann., Bernhard Sick., "On-Line Intrusion Alert Aggregation with Generative Data Stream Modeling," IEEE TRANSACTIONS ON DEPEDABLE AND SECURE COMPUTING., 2009.
[19] Battista Biggio., Giorgio Fumera., Fabio Roli., "Security evaluation of pattern classifiers under attack," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING., 2013.
[20] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic., "The Role of Hubness in Clustering High-Dimensional Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, REVISED JANUARY 2013.

