International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

SURVEY ON ANOMALOUS TOPIC DISCOVERY IN DISCRETE DATA

Brejit Lilly Abraham[1], Anjana. P .Nair[2]

[1]*M.Tech Computer Science & Engineering, Computer Science Department, Sree Buddha College of Engineering, Ayathil, Elavumthitta, Pathanamthitta, Kerala, India.*

[2]*Assistant Professor of Computer Science & Engineering, Computer Science Department, Sree Buddha College of Engineering, Ayathil, Elavumthitta, Pathanamthitta, Kerala, India.*

Abstract - *This survey provide a better understanding of the different directions in which research has been done on this topic, and how techniques developed in one area can be applied in domains for which they were not intended to begin with. Anomaly detection has been the topic of a number of surveys and review articles, as well as books. A substantial amount of research on outlier detection has been done in statistics and has been reviewed in several books, as well as other survey articles. The existing surveys discuss anomaly detection techniques that detect the simplest form of anomalies. This distinguishes simple anomalies from complex anomalies and provides a detailed discussion of the application domains where anomaly detection techniques have been used. For each domain it discusses the notion of an anomaly, the different aspects of the anomaly detection problem, and the challenges faced by the anomaly detection techniques. This also provides a list of techniques that have been applied in each application domain. The discussion of applications of anomaly detection reveals that for most application domains, the interesting anomalies are complex in nature to identifying the sources of the anomalies, while most of the algorithmic research has focussed on simple anomalies.*

Key Words*: *Data mining, Anomaly detection, Outlier Detection, and Topic Discovery.

1. INTRODUCTION

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, Statistics, and database systems. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and datamanagement aspects, datapreprocessing, model and inference considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

The Text Mining is the process of extracting unstructured information or extracts meaningful numeric indices from the text, and makes the information contained in the text accessible to the various data mining. This information can be used to extract summaries of the documents or to compute summaries of the documents based on the words included in them. Hence, users can analyze words, clusters of words in documents or could analyze documents and help in similarities between them or how it are related to other variables in the document. Text mining will "turn text into numbers" (meaningful indices), which can then be used in other analyses such as predictive data mining studies, the application of unsupervised learning methods (clustering), etc.

Anomaly detection (AD) is the problem of identifying items or patterns which do not conform to normal or expected behavior [1]. In some cases, no prior knowledge about normal behavior is available, and the goal is to detect anomalies (outliers) in a single data set consisting of normal and possibly abnormal instances, without any annotation of which samples are normal. More typically, there is a collection of normal data which sufficiently characterizes normal behavior. In the training phase, and use this data to build a (null) model. Then, in the detection phase, this model is used as a reference to help detect (possible) clusters of anomalous patterns in a different (test batch) data set.

Topic modeling is the method of expressing and explaining the content of a document to a leaner or to some publishing journals. This way of presenting ones idea or knowledge will be beneficial to the viewers and he/she will think, understand easily under what criteria or domain each document are. It mainly explains the core points of the topic. Generating the topic and cluster automatically will reduce time and effort to the viewer and include the main points of the paper like mining the main points from the paper. This survey is an attempt to provide a structured and broad overview of extensive research on anomaly detection techniques spanning multiple research areas and application domains.

1.1 PROBLEM STATEMENT

An algorithm for detecting patterns exhibited by anomalous cluster in high dimensional discrete data is shown in this paper with its various techniques, to detect group of anomalies; i.e., set of points which collectively exhibit abnormal behavior patterns with anomalous topic and its salient feature (words) subset under each topic models. In many applications, this can lead to a better understanding of

the nature of the atypical behavior and to identifying the sources of the anomalies.

Topic modeling is the method of expressing and explaining the content of a document to a leaner or to some publishing journals. This way of presenting ones idea or knowledge will be beneficial to the viewers and he/she will think, it will make others to understand easily under what criteria or which domain each document are. It mainly also explains the core points of the topic, by generating the topic and its salient feature.

This survey is an attempt to provide a structured and broad overview of extensive research on anomaly detection techniques spanning multiple research areas and application domains. Most of the existing surveys on anomaly detection either focus on a particular application domain or on a single research area. There are multiple-related works that group, anomaly detection into multiple categories and discuss techniques under each category. This survey builds upon this work by significantly expanding the discussion in several directions.

2. LITERATURE SURVEY

In this section, it reviews some previous works on group anomaly detection. These add two more categories of anomaly detection techniques, information theoretic and spectral techniques, to all the categories discussed on this survey. For each of the categories, it not only discusses the techniques, but also identifies unique assumptions regarding the nature of anomalies made by the techniques in that category.

2.1 OUTLIER DETECTION METHODOLOGIES

The most recent development in outlier detection technology is the data object that deviates significantly from the normal objects as if it were generated by a different mechanism, [2]. Outliers are different from the noise data that is Noise is random error or variance in a measured variable and Noise should be removed before outlier detection. These can be used for Applications such as, Credit card fraud detection, Telecom fraud, detection Customer segmentation, Medical analysis. It has three kinds of outliers:

- *Global outlier (or point anomaly)*: Object which significantly deviates from the rest of the data set Ex. Intrusion detection in computer networks and its issue is to find an appropriate measurement of deviation.
- *Contextual outlier (or conditional outlier)*: it deviates significantly based on a selected context and attributes of data objects should be divided into two groups, Contextual attributes: defines the context. Behavioral attributes are the characteristics of the object, used in outlier evaluation. It Can be viewed as a generalization of local outliers—whose density significantly deviates from its local area
- *Collective Outliers:* A subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers. Applications for collective outliers are intrusion detection. When a number of computers keep sending denial-of-service packages to each other, the Detection of collective outliers Consider not only behavior of individual objects, but also that of groups of objects, and the Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects.

Modeling normal objects and outliers properly are Hard to enumerate all possible normal behaviors in an application. The border between normal and outlier objects is often a gray area. Handling noise in outlier detection can distort the normal objects and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection.

In a database, outliers may indicate fraudulent cases or it may just denote an error by the entry clerk or a misinterpretation of a missing value code, either way detection of the anomaly is vital for data base consistency and integrity. A more exhaustive list of applications that utilize outlier detection is:

- *Fraud detection* - detecting fraudulent applications for credit cards, state benefits or detecting fraudulent usage of credit cards or mobile phones.
- Loan application processing - to detect fraudulent applications or potentially problematical customers.
- *Intrusion detection* - detecting un-authorized access in computer networks.
- *Activity monitoring* - detecting mobile phone fraud by monitoring phone activity or suspicious trades in the equity markets.
- *Network performance*- monitoring the performance of computer networks, for example to detect network bottlenecks.
- *Fault diagnosis* - monitoring processes to detect faults in motors, generators, pipelines or space instruments on space shuttles for example.
- *Structural defect detection*- monitoring manufacturing lines to detect faulty production runs for example cracked beams.
- *Satellite image analysis* - identifying novel features or misclassified features.
- *Detecting novelties in images* - for robot neotaxis or surveillance systems.

- *Motion segmentation* - detecting image features moving independently of the background.
- *Time-series monitoring* - monitoring safety critical applications such as drilling or high-speed milling.
- *Medical condition monitoring* - such as heart-rate monitors.
- *Pharmaceutical research* - identifying novel molecular structures. Detecting novelty in text - to detect the onset of news stories, for topic detection and tracking or for traders to pinpoint equity, commodities, FX trading stories, outperforming or underperforming commodities.
- *Detecting unexpected entries in databases* - for data mining to detect errors, frauds or valid but unexpected entries.
- Detecting mislabeled data in a training data set.

It uses Proximity-based techniques, which are simple to implement and make no prior assumptions about the data distribution model. An object is an outlier if the nearest neighbors of the object are far away, i.e., the proximity of the object is significantly deviates from the proximity of most of the other objects in the same data set. However, it suffers exponential computational growth as it is founded on the calculation of the distances between all records. Assumption of proximity-based approach: The proximity of an outlier deviates significantly from that of most of the others in the data set. The two types of proximity-based outlier detection methods:

- Distance-based outlier detection: An object o is an outlier if its neighborhood does not have enough other points.
- Density-based outlier detection: An object o is an outlier if its density is relatively much lower than that of its neighbors.

For each object o, examine the # of other objects in the *r*-neighborhood of o, where *r* is a user-specified distance threshold. An object o is an outlier if most (taking π as a fraction threshold) of the objects in D are far away from o, i.e., not in the r-neighborhood of o. An object o is a DB(r, π) outlier if it is calculate with the above equation , Equivalently, one can check the distance between *o* and its *k*-th nearest neighbor o_k, *o* is an outlier if dist(o,o_k)>r.

$$k = \lceil \pi \| D \| \rceil$$

Other Efficient computation is used for Nested loop algorithm, any object o_i, calculate its distance from other objects, and count the # of other objects in the r-neighborhood. If π·n other objects are within r distance, terminate the inner loop. Otherwise, o_i is a DB(r, π) outlier. The Efficiency of Distance-Based Outlier Detection is actually CPU time is not $O(n^2)$ but linear to the data set size since for most non-outlier objects, the inner loop terminates early.

Second it go for Clustering-Based Methods; Used for Normal data belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters. Since there are many clustering methods, there are many clustering-based outlier detection methods as well. Clustering is expensive for straightforward adaption of a clustering method for outlier detection can be costly and does not scale up well for large data sets.

2.2 ANOMALY DETECTION

The author [3] have tried to provide a broad sample of current techniques and has introduce a survey of contemporary techniques for outlier detection, Outlier detection has been used for centuries to detect and, where appropriate, remove anomalous observations from data. Outliers arise due to mechanical faults, changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations. Its detection can identify system faults and fraud before it escalates with potentially catastrophic consequences. It can identify errors and remove its contaminating effect on the data set and as such to purify the data for processing. It has also categories and analyzes broad range of outlier detection methodologies. And has point out how each handles outliers and make recommendations for when each methodology is appropriate for clustering, classification and/or recognition.

2.2.1 CHALLENGES

Due to these challenges, the anomaly detection problem, in most general form is not easy to solve. For each category, the author has identified the advantages and disadvantages of the techniques in that category. And also provide a discussion on the computational complexity of the techniques since it is an important issue in real application domains.

- Paragraph Availability of labeled data for training/validation of models used by anomaly detection techniques is usually a major issue.
- Often the data contains noise that tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.

2.3 FRAUD DETECTION

Fraud detection [4], refers to detection of criminal activities occurring in commercial organizations such as banks, credit card companies, insurance agencies, cell phone companies, stock market, and so on. The malicious users might be the actual customers of the organization or might be posing as customers (also known as *identity theft*). The fraud occurs when these users consume the resources provided by the organization in an unauthorized way. The

organizations are interested in immediate detection of such frauds to prevent economic losses.

2.3.1 CREDIT CARD FRAUD DETECTION

In this domain, anomaly detection techniques are applied to detect fraudulent credit card applications or fraudulent credit card usage (associated with credit card thefts). Detecting fraudulent credit card applications is similar to detecting insurance fraud. The data is typically comprised of records defined over several dimensions such as user ID, amount spent, time between consecutive card usage, and so forth. The credit companies have complete data available and also have labeled records. Moreover, the data falls into distinct profiles based on the credit card user. Hence profiling and clustering based techniques are typically used in this domain. The challenge associated with detecting unauthorized credit card usage is that it requires online detection of fraud as soon as the fraudulent transaction takes place. Anomaly detection techniques have been applied in two different ways to address this problem. The first one is known as *by-owner* in which each credit card user is profiled based on his/her credit card usage history. Any new transaction is compared to the user's profile and flagged as an anomaly if it does not match the profile. This approach is typically expensive since it requires querying a central data repository, every time a user makes a transaction. Another approach known as *by-operation* detects anomalies from among transactions taking place at a specific geographic location. Both *by-user* and *by-operation* techniques detect contextual anomalies. In the first case the context is a user, while in the second case the context is the geographic location.

2.4 STATISTICAL BASED SYSTEM FOR THE DETECTION OF FRAUD

The underlying principle of any statistical anomaly detection technique is: "An anomaly is an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed" [5], Statistical anomaly detection techniques are based on the following key assumption:

- *Assumption*: Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model.

Statistical techniques fit a statistical model to the given data and then apply a statistical inference test to determine if an unseen instance belongs to this model or not. Instances that have a low probability of being generated from the learned model, based on the applied test statistic, are declared as anomalies. Both parametric as well as nonparametric techniques have been applied to fit a statistical model. While parametric techniques assume the knowledge of the underlying distribution and estimate the parameters from the given data, nonparametric techniques do not generally assume knowledge of the underlying distribution. In the next two subsections it will discuss parametric and nonparametric anomaly detection techniques.

2.4.1 PARAMETRIC TECHNIQUES

As mentioned before, parametric techniques assume that the normal data is generated by a parametric distribution with parameters Θ and probability density function $f(\mathbf{x}, \Theta)$, where x is an observation. The anomaly score of a test instance (or observation) x is the inverse of the probability density function, $f(x, \Theta)$. The parameters Θ are estimated from the given data.

Alternatively, a statistical hypothesis test (also referred to as discordancy test in statistical outlier detection literature may be used. The null hypothesis ($H0$) for such tests is that the data instance x has been generated using the estimated distribution (with parameters Θ). If the statistical test rejects $H0$, x is declared to be anomaly. A statistical hypothesis test is associated with a test statistic, which can be used to obtain a probabilistic anomaly score for the data instance **x**. Based on the type of distribution assumed, parametric techniques can be further categorized as follows:

A. Gaussian Model-Based

Such techniques assume that the data is generated from Gaussian distribution. The parameters are estimated using Maximum Likelihood Estimates (MLE). The distance of a data instance to the estimated mean is the anomaly score for that instance. A threshold is applied to the anomaly scores to determine the anomalies. Different techniques in this category calculate the distance to the mean and the threshold in different ways.

B. Regression Model-Based

Anomaly detection using regression has been extensively investigated for time-series data

C. Mixture of Parametric Distributions-Based

Such techniques use a mixture of parametric statistical distributions to model the data.

2.4.2 NONPARAMETRIC TECHNIQUES

The anomaly detection techniques in this category use nonparametric statistical models, such that the model structure is not defined a priori, but is instead determined from given data. Such techniques typically make fewer assumptions regarding the data, such as smoothness of density, when compared to parametric techniques.

2.4.3 ADVANTAGES AND DISADVANTAGES OF STATISTICAL TECHNIQUES

The advantages of statistical techniques are:

i. If the assumptions regarding the underlying data distribution hold true, statistical techniques provide a statistically justifiable solution for anomaly detection.

ii. The anomaly score provided by a statistical technique is associated with a confidence interval, which can be used as additional information while making a decision regarding any test instance.

iii. If the distribution estimation step is robust to anomalies in data, statistical techniques can operate in a unsupervised setting without any need for labeled training data.

The disadvantages of statistical techniques are:

i. The key disadvantage of statistical techniques is that it relies on the assumption that the data is generated from a particular distribution. This assumption often does not hold true, especially for high dimensional real data sets.

ii. Even when the statistical assumption can be reasonably justified, there are several hypothesis test statistics that can be applied to detect anomalies; choosing the best statistic is often not a straightforward task. In particular, constructing hypothesis tests for complex distributions that are required to fit high dimensional data sets is nontrivial.

iii. Histogram-based techniques are relatively simple to implement, but a key shortcoming of such techniques for multivariate data is that it are not able to capture the interactions between different attributes. An anomaly might have attribute values that are individually very frequent, but whose combination is very rare, however an attribute-wise histogram-based technique would not be able to detect such anomalies.

2.5 PAYLOAD-BASED NETWORK AND BATCH OF NETWORK TRAFFIC FLOWS

The payload-based anomaly detector can be call as PAYL [6], for intrusion detection. PAYL models the normal application payload of network traffic in a fully automatic, unsupervised and very efficient fashion. It first computes during a training phase a profile byte frequency distribution and its standard deviation of the application payload flowing to a single host and port. Then use Mahalanobis distance during the detection phase to calculate the similarity of new data against the pre-computed profile. Mahalanobis distance is a standard distance metric to compare two statistical distributions. It is a very useful way to measure the similarity between the (unknown) new payload sample and the previously computed model. It then computes the distance between the byte distributions of the newly observed payload against the profile from the model computed for the corresponding length range.

Classifies intrusion detection systems into host-based and network based intrusion detection systems [7]:

A. Host-Based Intrusion Detection Systems:

Such systems (also referred to as system call intrusion detection systems) deal with operating system calls trace. The intrusions are in the form of anomalous subsequences (collective anomalies) of the traces. The anomalous subsequences translate to malicious programs, unauthorized behavior and policy violations. While all traces contain events belonging to the same alphabet, it is the co-occurrence of events that is the key factor in differentiating between normal and anomalous behavior.

B. Network Intrusion Detection Systems:

These systems deal with detecting intrusions in network data. The intrusions typically occur as anomalous patterns (point anomalies) though certain techniques model the data in a sequential fashion and detect anomalous subsequences (collective anomalies). The primary reason for these anomalies is due to the attacks launched by outside hackers who want to gain unauthorized access to the network for information theft or to disrupt the network. A typical setting is a large network of computers connected to the rest of the world via the Internet. A challenge faced by anomaly detection techniques in this domain is that the nature of anomalies keeps changing over time as the intruders adapt its network attacks to the existing intrusion detection solutions.

2.6 LATENT DIRICHLET ALLOCATION

Topic models such as Latent Dirichlet Allocation (LDA) [8] are widely used to model data having this kind of group structure. The original LDA model was proposed for text processing. It represents the distribution of points (words) in a group (document) as a mixture of K global topics $\beta_1, \ldots\ldots, \beta_k$ each of which is a distribution (i.e., $\beta_i \in S^f$ where S^f is the f-dimensional probability simplex). Let M (θ) be the multinomial distribution parameterized by $\theta \in S^k$ and Dir ($\propto$) be the Dirichlet distribution with parameter $\propto \in R^k t$. LDA generates the mth group by first drawing its topic distribution θm from the prior distribution Dir ($\propto$). Then for each point Xmn in the mth group it draws one of the K topics and then generates the point according to that topic. Although topic models are very useful in estimating the topics and topic distributions in groups, the existing methods are incapable of detecting group anomalies

comprehensively. In order to detect anomalies, the model should be flexible enough to enable complex normal behaviors. LDA is a matrix factorization technique, [9]. In vector space, any corpus (collection of documents) can be represented as a document-term matrix.

2.7 PARSIMONIOUS TOPIC MODELS

Extends LDA by proposing Parsimonious Topic Models (PTM) [10]. PTM controls the number of free parameters in the model by balancing model complexity and goodness of fit to the data set used for learning the model. Hypothesizes that, under each topic, only a modest number of words have topic-specific characteristics (salient words), which warrant its own probability parameters, while the rest of the words can be described by a universal shared model across all topics. PTM proposes that only sparse subsets of topics are present in each document, with the rest of the topics having zero proportions. Shows that PTM achieves better generalization accuracy than LDA evaluated on multiple text corpora. optimizes an objective function, a Bayesian Information Criterion (BIC), specifically derived for the PTM structure, to jointly learn the structure of the model and to estimate the model parameters (word probabilities and active topic proportions). Moreover, the PTM objective function is also optimized with respect to (thus estimating) the number of topics (model order) present in the corpus. PTM over LDA as the topic model for ATD algorithm is used for a number of reasons for its generalization.

2.8 BOOTSTRAP

The bootstrap is a general method for doing statistical analysis without making strong parametric assumptions. Efron's nonparametric bootstrap [11], re-samples the original data. It was originally designed to estimate bias and standard errors for statistical estimates much like the jackknife. The bootstrap is similar to earlier techniques which are also called re-sampling methods:

Jackknife, cross-validation, Delta method, Permutation methods, and Sub-sampling.

Technique was extended, modified and refined to handle a wide variety of problems including:

- Confidence intervals and hypothesis tests,
- Linear and nonlinear regression,
- Time series analysis and other problems.

2.9 GROUP AD- MGMM

Sample paragraph, consider the unsupervised detection problem, [12]. Which it do not know beforehand which data is normal and which is not. Gaussian Mixture Model-MGMM model that extends GLDA. Gaussian-LDA is effective for uni-modal group behaviors. With a set of typical topic mixtures/distributions. Proposes of MGMM is for group anomaly detection assumes each data point belongs to one group and that all points in a group are modeled by the group's Gaussian mixture model. Assumes each data point belongs to one group and that all points in a group are modeled by the group's Gaussian mixture model. Mixing proportions of the mixture model for each group, however, are not freely estimated, but rather, in a hierarchical way, are selected from a limited set of T possible mixing proportion "types" (genres). These types represent the normal behaviors. A test group is called anomalous if it has low likelihood under the normal types.

2.10 FLEXIBLE GENRE MODELS

FGM is designed to characterize data groups at both the point level and the group level so as to detect various types of group anomalies, [13]. Two types of group anomalies:

- A point-based group anomaly is a group of individually anomalous points.
- A distribution-based anomaly is a group where the points are relatively normal, but as a whole it are unusual.

Two key components are added:

- To model the behavior of topic distributions, use several "genres", each of which is a typical distribution of topic distributions.
- Use "topic generators" to generate adaptive topics for different groups.

Example: selected first 100 images from categories "mountain", "coast", and "inside city". These images are randomly divided: 80% are used for training and the rest for testing. Thus it creates anomalies by stitching random normal test images from different categories.

2.11 GLAD: SOCIAL MEDIA ANALYSIS

Ref. [14] addresses the first issue by presenting a method, specifically for network analysis, for jointly detecting groups of similar nodes and computing anomaly scores for the discovered groups. Nevertheless, it does not have an algorithmic procedure for discovering "hard" anomalous clusters one by one—some post-processing effort is required to hard-assign each data point to the cluster with highest membership degree. Moreover, [14] does not provide any statistical significance testing and relies on choosing an appropriate threshold for detecting anomalous clusters. And it mainly focused on a generative approach by proposing a hierarchical Bayes model: Group Latent Anomaly Detection (GLAD) model. GLAD takes both pair-wise and point-wise data as input automatically infers the groups and detects group anomalies simultaneously. To account for the dynamic

properties of the social media data, further generalize GLAD to its dynamic extension d-GLAD.

2.12 OCSVM FOR GROUP ANOMALY DETECTION

One-class support measure machines (OCSMMs) for group anomaly detection, [15]. Follows a discriminative approach to group anomaly detection and generalizes the idea of one-class support vector machines to a space of probability measures, proposing one-class support measure machines. A simple and efficient discriminative way of detecting group anomaly is illustrated in this work, M groups of data points are represented by a set of M probability distributions assumed to be i.i.d. realization of some unknown distribution. To handle aggregate behaviors of data points, groups are represented as probability distributions which account for higher-order information arising from those behaviors. Groups in this method are represented as probability distributions which are mapped into a reproducing kernel Hilbert space using kernel methods. Similar to MGMM, this method requires hard-clustering of the data prior to detecting any anomalous group. One-Class SVMs for Document Classification, [16].

Group anomaly detection may shed light in a wide range of applications. By working directly with the distributions, the higher-order information arising from the aggregate behaviors of the data points can be incorporated efficiently. The SVM appropriate for *one-class* classification in the context of information retrieval and probability measures.

3. CONCLUSIONS

These assumptions are critical for determining when the techniques in that category would be able to detect anomalies, and when it would fail. For each category, it provides a basic anomaly detection technique, and then shows how the different existing techniques in that category are variants of the basic technique. This template provides an easier and more succinct understanding of the techniques belonging to each category. Further, for each category it identifies the advantages and disadvantages of the techniques. It also provides a discussion of the computational complexity of the techniques since that is an important issue in real application domains. While some of the existing surveys mention the different applications of anomaly detection, it provides a detailed discussion of the application domains where anomaly detection techniques have been used. For each domain it discusses the notion of an anomaly, the different aspects of the anomaly detection problem, and the challenges faced by the anomaly detection techniques. It also provides a list of techniques that have been applied in each application domain.

REFERENCES

[1] R Hossein Soleimani and David J. Miller, Senior Member, "ATD: Anomalous Topic Discovery in High Dimensional Discrete Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 9, September 2016.

[2] R V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," Artif. Intell. Rev., vol. 22, no. 2, pp. 85–126, 2004.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surveys, vol. 41, pp. 1–58, 2009.

[4] A. Srivastava and A. Kundu, "Credit card fraud detection using hidden Markov model," IEEE Trans.DSC,vol. 5, Jan./Mar. 2008.

[5] J. Major and D. Riedinger, "EFD: A hybrid knowledge/ statistical based system for the detection of fraud," J. Risk Insurance, vol. 69, no. 3, pp. 309–324, 2002.

[6] K. Wang and S. Stolfo, "Anomalous payload-based network intrusion detection," in Proc. 7th Int. Symp. Recent Adv. Intrusion Detection, 2004, pp. 203–222.

[7] F. Kocak, D. Miller, and G. Kesidis, "Detecting anomalous latent classes in a batch of network traffic flows," in Proc. 48th Annu. Conf. Inform. Sci. Syst., 2014, pp. 1–6.

[8] D.M.Blei, A.Y.Ng, M.I.Jordan,"Latent dirichlet allocation," J. Mach. Learning Res., vol. 3, pp. 993–1022, 2003.

[9] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," ACM Commun., vol. 55, pp. 77–84, Nov. 2012.

[10] H. Soleimani and D. J. Miller, "Parsimonious topic models with salient word discovery," IEEE Trans. Knowl. Data Eng., vol. 27, no. 3, pp. 824–837, Mar. 2015.

[11] B. Efron, "Bootstrap methods: Another look at the jackknife," Ann. Statist., vol. 7, pp. 1–26, 1979.

[12] L. Xiong, S. P. Barnab_a, J. G. Schneider, A. Connolly, and V. Jake, "Hierarchical probabilistic models for group anomaly detection," in Proc. Int. Conf. Artif. Intell. Statist., 2011, pp. 789–797.

[13] L. Xiong, B. P_oczos, and J. Schneider, "Group anomaly detection using flexible genre models," in Proc. Adv. Neural Inform. Process.Syst., 2011, pp. 1071–1079.

[14] R. Yu, X. He, and Y. Liu, "GLAD: Group anomaly detection in social media analysis," in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2014, pp. 372–381.

[15] K. Muandet, B. Scholkopf, "One-class support measure machines for group anomaly detection," in Proc. 29th Conf. Uncertainty Artif. Intell, 2013, pp. 449–458.

[16] L.M. Manevitz and M. Yousef, "One-Class SVMs for document classification," J. Mach. Learning Res., vol. 2, pp. 139–154, 2001.

BIOGRAPHIES

Brejit Lilly Abraham received the Bachelor's Degree in Computer Science and Engineering from Karpagam University, Tamil nadu, India in 2017. She is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Kerala, India. His research area of interest includes the field of internet security, data mining and technologies in Department of Computer Science and Engineering.

Anjana.P.Nair received the bachelor's degree in LBS Institute of Technology for Women, Kerala, India. And master's degree in Computer Science and Engineering from Sree Buddha College of Engineering, Kerala, India in 2013. She is a lecturer in the Department of Computer Science and Engineering, Sree Buddha College of Engineering. Her main area of interest is Core Computers and has published more than 10 referred papers.