International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 10, October 2014 **ISSN 2319 - 4847**

Outlier mining techniques for uncertain data

Ms. Aditi Dighavkar[1], Prof. N. M. Shahane[2]

[1]Student, ME Computer, KKWIEER, Savitribai Phule Pune University, India

[2]Associate Professor, Department of computer Engineering, KKWIEER, Savitribai Phule Pune University, India

ABSTRACT

Outlier detection has been a very significant concept in the realm of data analysis. Lately, many application domains have realized the direct relation between outliers in data and real world anomalies that are of immense interest to an analyst. Mining of outliers has been researched within vast application domains and knowledge disciplines. This paper provides a comprehensive overview of existing outlier mining techniques by classifying them along different dimensions. The motive of this survey is to identify the important dimensions which are associated with the problem of outlier detection, to provide taxonomy to categorize outlier detection techniques along the different dimensions. Also, a comprehensive overview of the recent outlier detection literature is presented using the classification framework. The classification of outlier detection techniques depending on the applied knowledge discipline can provide an idea of the research done by varied communities and also uncover the research avenues for the outlier detection problem.

Keywords: Outlier Detection, Anomaly Detection, Likelihood value, Data classification

1. INTRODUCTION

Outlier detection refers to the problem of detecting and analyzing patterns in data that does not map to expected normal behavior. These patterns are often referred to as outliers, anomalies, discordant observations, exceptions, noise, errors, novelty, damage, faults, defects, aberrations, contaminants, surprise or peculiarities in different application domains [1]. Outlier detection has been a widely researched problem. It finds immense use in a wide variety of application domains such as insurance, tax, credit card, fraud detection, military surveillance for enemy activities, fault detection in safety critical systems, intrusion detection for cyber security and many other areas [1]. Outlier detection is important due to the fact that outliers in the data translate to significant information in a wide variety of application domains. For example, in a computer network, an exceptional pattern could mean that a hacked computer is sending out sensitive data to an unauthorized receiver. Outliers in credit card transaction data could announce credit card theft or misuse. Similarly, in public health data, outlier detection techniques are used to detect exceptional patterns in patient medical records which could be symptoms of a new disease. Outliers can also translate to some critical entities such as in exceptional readings from a space craft which a fault in some component of the craft. In military surveillance, the existence of an unusual portion in a satellite image of enemy area could indicate enemy troop movement. Outlier detection has been found to be directly applicable in a large number of domains. This has resulted in a huge and highly distinct literature of outlier detection techniques. Many of these techniques have been developed to solve focused problems pertaining to a particular application domain, while others have been developed in a more generic fashion. This survey deals with providing a structured and comprehensive sketch of the research done in the field of anomaly detection. The key aspects of any outlier detection technique are identified, and are used as dimensions to classify current techniques into distinct categories. This survey intent at providing a good understanding of the distinct directions in which research has been done. Also it will help in determining the potential areas for future research. Outliers are patterns in data pool that do not conform to a well defined concept of normal behavior, or conform to a well specified notion of outlying style, even if it is typically easier to define the normal style. This survey discusses methods which find such outliers in data. Some of the major causes for outliers are as listed below

1. Malicious activity such as insurance, credit card or telecom fraud, a terrorist activity or a cyber intrusion
2. Instrumentation error such as defects in components of machines or wear and tear
3. Changes in the environment such as a climate change, mutation in genes, a new buying pattern among consumers
4. Human error such as an data reporting error or an automobile accident [1]

The “interestingness" or real life relevance of outliers is an important feature of anomaly detection and separates it from *noise accommodation* or *noise removal*, which deal with unwanted noise in the data. There is no real importance of noise in data by itself, but acts as a hindrance to data analysis. Before any data analysis to be performed on data, noise removal is required to remove the unwanted elements. Noise accommodation is nothing but immunizing statistical model estimation against outlying observations. Novelty *detection* aims mainly at detecting hidden patterns in the data. The difference between novel patterns and outliers is that the novel patterns are incorporated with the normal model after getting detected by the detection system.

2. ORGANIZATION

This survey is organized into three important sections which discuss the outlier detection technique. In Section 3 we identify the different aspects that establish an exact formulation of the problem. This section brings into focus the prosperity and complexity of the problem territory. In Section 4 we describe the distinct application domains where outlier detection has been applied. In Section 5, different outlier detection techniques are categorized depending on the knowledge discipline from which they are adopted. Section 6 deals with the proposed work to be undertaken.

3. DIFFERENT ASPECTS OF AN OUTLIER DETECTION PROBLEM

Input Data

Input data is a key component of any outlier detection technique where it has to detect the outliers. It can be:
Point data: data in which there is no assumption of structure among the data instances
Sequential data: data instances have a defined ordering where each data instance in entire data set occurs sequentially e.g. time-series data Spatial data: the data instances have a well defined spatial structure where the location of a data instance is significant well-defined with respect to others

Types of Supervision

Outlier detection is organized into three main classes based on the extent to which labels can be assigned to each group:

1. Supervised Outlier Detection

In supervised outlier detection mode, training dataset is available for both normal as well as outlier classes. This approach builds a predictive model for both classes and any new data instance is compared against these models. There are certain challenges in supervised outlier detection, such as outlier data instances are few as compared to normal data instances and even it is difficult to obtain the accurate class labels.

2. Semi-supervised Outlier Detection

In semi-supervised outlier detection mode, training dataset is available only for normal class. Hence it is widely used than supervised mode. The new target instance is compared against this normal class and the data instances which do not satisfies this class are considered as an outlier. This mode is not used commonly as it is difficult to cover each abnormal behavior to generate normal class.

3. Unsupervised Outlier Detection

In unsupervised outlier detection mode training data is not available. This technique makes an assumption that normal data instances are mere frequent than outliers. The data instances which are frequent or closely related are considered as normal instances and remaining are considered as outliers. These techniques are widely used as they do not require training data set.

Type of Outlier

Outliers can be classified into three categories depending on its composition and its relation to rest of the data.

1. Type I Outliers: An individual outlying instance in a given set of data instances is termed as a Type I outlier. It is the simplest type of outliers and is also the focus of many of existing anomaly detection methods. A data instance is an outlier because of its attribute values which are inconsistent with the values taken by normal instances.

2. Type II Outliers: These outliers are induced due to the existence of an individual data instance in the particular context in the given data. These outliers are also individual data instances like Type I outliers. The distinction is that a Type II outlier may not be an outlier in a distinct context. Hence Type II outliers are described with respect to a context. The concept of a context is induced by the structure in the data set and has to be specified as a part of the problem formulation. A context describes the neighborhood of a specific data instance.

3. Type III Outlier: These outliers occur due to a subset of data instances are outlying with respect to the whole data set. In a Type III, the individual data instances outlier are not outliers by themselves, but their existence together as a substructure is outlier. Only when the data has spatial or sequential nature, Type III outliers are worthwhile.

4. APPLICATIONS OF OUTLIER DETECTION

The ability to detect outliers is a major desirable feature in application domains for numerous reasons such as intrusion detection, network intrusion detection systems, fraud detection, credit card fraud detection, mobile phone fraud detection, insurance claim fraud detection, insider trading detection, medical and public health data, industrial damage detection, fault detection in mechanical units, novel topic detection in text, speech recognition, novelty detection in robot behavior, detecting faults in web applications, detecting outliers in biological data, detecting outliers in census data, detecting associations among criminal activities, detecting outliers in customer relationship management data, detecting outliers in astronomical data and detecting ecosystem disturbances, etc[1].

5. TECHNIQUES USED FOR OUTLIER DETECTION

Numerous outlier detection methods have been proposed until now. Generally, these approaches are classified into: statistical-based, clustering-based, density-based and model-based approaches [3]-[6].

Statistical Based

Statistical approaches make an assumption of some standards or predetermined distributions in order to find outliers which deviate from such distributions. The ways in this category invariably assume that the normal example follow an explicit of data distribution. Nonetheless, we cannot invariably have this type of priori data distribution information in practice, especially for top dimensional real datasets. The underlying principle of any statistical outlier detection technique is: "An outlier is an observation which is suspected of being partially or wholly irrelevant as it is not generated by the stochastic model assumed" [1]. In literature [1], the author applied parametric and non-parametric methods to fit a statistical model which was built for normal instances to test the unseen instances. Parametric techniques assume the knowledge of underlying distribution and estimate the parameters from the given data. For example, such method assumes that the data is generated from a Gaussian distribution. The parameters are estimated using Maximum Likelihood Estimates (MLE). A simple outlier mining approach, often used in process quality control domain, is to declare all data elements that are more than 3σ distance away from the distribution mean μ, where σ is standard deviation for the distribution. The $\mu \pm 3\sigma$ region contains 99.7% of the data instances [1, 2]. Of course, the normal instances are belonged to different distribution, because of the trait of the data, we should model diverse distribution. Non-parametric techniques do not generally assume knowledge of underlying distribution, such that the model structure is not defined a prior, but is instead determined from the given data. The statistical outlier detection approach depends on the nature of statistical model that is required to be fitted on the data. The main problem with these techniques is that in a number of situations, the user might not have much knowledge about the underlying data distribution [2].

Clustering Based

In clustering-based approaches, they regularly conduct clustering-based techniques on the samples of data to characterize the native data behaviour. The sub-clusters contain significantly less data points than remaining clusters, are termed as outliers. Cluster analysis [7] is a popular machine learning approach to group similar data instances into clusters. It is either used as a stand-alone tool to get an insight into the distribution of a data set, e.g. to focus further analysis and data processing, or as a preprocessing step for other algorithms operating on the detected clusters. The semi-supervised techniques mainly use normal data to generate clusters which represent the normal modes of behavior of the data [8]. Any new test instance is assigned to one of the clusters. If the new instance is not close to any of the learnt clusters, it is considered as an outlier. This approach is applied for novelty detection task in distinct domains such as novel topic detection in news data [9]. [10] incorporate the knowledge of labels to enhance their unsupervised clustering based outlier detection algorithm [11] by determining a measure called semantic outlier factor which is high if the class label of an object in a cluster is distinct from the majority of the class labels in that cluster. A semi-supervised approach is proposed by [12] where an instance can belong to one of the several class labels. The algorithm learns the parameters depending on a similarity measure for the clusters representing each class. A distance measure is used in order to classify a test point to a cluster and declaring it as an outlier if it is far from all clusters or if within a cluster it is far from all other points. Thus this approach finds global as well as local outliers with respect to the cluster. [13] Studied Self-Organizing Maps (SOM), K-means clustering, and Expectation Maximization (EM) to cluster training data and then used the clusters in order to classify test data. Unsupervised techniques in this category use some known clustering methods and then analyze each instance in the data with respect to the clusters. A bootstrapping technique [14] initially separates normal data from outliers using frequent item set mining. The data is categorized into segments based on time. For each and every segment, frequent item sets are generated. All item sets which exist in more than one segment are considered as normal. All data points corresponding to the frequent item set are considered as normal. Using the 'clean' data, clusters are obtained using COOLCAT clustering technique [15]. Most of the earlier clustering-based anomaly detection methods found outliers as the byproduct of a clustering. Hence any data point which does not comply with any cluster is called an outlier. As the main aim is to find clusters, these approaches are not optimized to find outliers. The advantage of the cluster based technique is that they do not have to be supervised. Moreover, clustering based techniques are capable of being used in an incremental mode i.e. after learning the clusters, new points can be inserted in to the system and tested for the outliers. Clustering based approaches are computationally expensive as they compromise huge computation of pair wise distances.

Density-Based

In addition, density-based approaches have been projected. One in all the representatives of this type of approaches are local outlier factor (LOF) and variants. Based on the local density of every data instance, the LOF determines the degree of outlierness, which provides suspicious ranking scores for all samples. The foremost necessary property of the LOF is the ability to estimate local organisation via density estimation. In density based method outlier are detected after clustering the data. The data objects that do not fit into the density of the cluster are considered as the outlier. Markus M. Breunig et al. has proposed a method in which, outlier is found on the basis of the local outlier factor that

how much the object is distinct from the other data objects with respect to the surrounding neighborhood [16]. Raghuvira Pratap et al. have used a technique based on density in which an efficient density based k-medoids clustering algorithm has been used in order to overcome the drawbacks of DBSCAN and k-medoids clustering algorithms [17][18]. The advantage of these approaches is that they do not need to make any assumption for the generative distribution of the data. But, these approaches consist a high process complexity within the testing part, as they have to determine the distance between every test instance and all the other instances to determine nearest neighbours.

Model-Based

Besides the above work, model-based approaches are projected. Among them, support vector data description (SVDD) has been incontestable empirically to be capable of detecting outliers in numerous domains. SVDD conducts a small sphere around the normal data and utilizes the constructed sphere to notice an unknown sample as normal or outlier. The foremost attractive feature of SVDD is that it can transform the original data into a feature space via kernel function and effectively notice global outliers for high-dimensional data. However; its performance is sensitive to the noise involved within the input data. Depending on the availability of a training dataset, the outlier detection techniques which are described above operate in two different modes: supervised and unsupervised modes. Among the four types of outlier detection approaches, distribution-based approaches and model based approaches fall under supervised outlier detection, which assumes that the training dataset has labeled instances for normal class (as well as anomaly class sometimes). In addition, some techniques [19]–[21] are proposed that inject artificial outliers into a normal dataset in order to get a labeled training data set. Also, the work of [22] presents a novel method to detect outliers by using the instability of the output of a classifier which is built on bootstrapped training data.

Distance Based

In order to overcome the disadvantage of statistical based, Knorr and Ng proposed the following distance-based definition for outliers that is both simple and intuitive as well as being computationally feasible for large sets of data points. This basic technique has been applied to detect land mines from satellite ground images and to detect shorted turns (anomalies) in the DC field windings of large synchronous turbine-generators [23, 24]. Given a distance measure on a feature space, there are many different definitions of distance-based outliers. Four popular definitions are as follows:

1. Outliers are the examples for which there are fewer than p other examples within distance d [24].
2. Outliers are the top n examples whose distance to the kth nearest neighbor is greatest [25].
3. Outliers are the top n examples whose average distance to the k nearest neighbors is greatest.
4. Outliers are the top n examples whose sum distance to the k nearest neighbors is greatest.

Neural Network Based

Neural network based outlier detection approaches works in two phases. In initial phase neural network is trained to build the normal classes. Secondly, each target instance is tested against those classes by providing input to neural network. If the neural network accepts the data instance then it is normal and if the neural network rejects data instance it is termed as an outlier. Distinct types of neural networks are derived from basic neural network. Replicator neural network has invented for one class outlier detection. A multilayer feed forward neural network is having same amount of input and output neurons. The training phase compresses the data and testing phase reconstructs it [1].

Rule Based

Rule based outlier detection approach learns the rule for normal behavior of the system. A test which is not covered by such rule is deemed as an outlier. Rule based approach uses two steps. In first step it learns rules from training data using rule learning algorithm. Each rule has associated confidence value that is equal to the ratio between number training instances correctly classified by the rule and total number of instances covered by the rule. The second step is to find the rule that best captures the test instance [1].

Information Theory Based

Information Theory based techniques analyze the information content of a data set using distinct information theoretic measures such as entropy, relative entropy etc. The general idea behind these approaches is that outlying instances affect the information content of the data set due to their surprising nature. These approaches typically run in an unsupervised mode. Lee and Xiang [2001] listed down different information theoretic measures which can be used to detect outliers in a sequence of operating system call. These are entropy, relative conditional entropy, conditional entropy, information gain, and information cost. The general approach is to measure the regularity of a data set with respect to each instance and classify the point as anomaly if it induces irregularity in the data. A new method has been adopted to determine surprising patterns in market basket data [26]. The authors encode a pattern by using a set of bits and then observe the change in the information content of the pattern over time. A sudden change in the encoding of a pattern shows a surprising change in the data. A variant of this method is also applied to detect anomalous substructures in graph data by [27]. In this approach, the data is not temporal but spatial and for each substructure an encoding of its surroundings is found. The substructures which need larger number of encoding bits are outliers, as they are distinct from their surroundings.

Spectral Decomposition Based

Spectral decomposition technique in general estimates the principle component vectors for a given data matrix. Thus in a way it tries to detect the normal modes of behavior in the data (using the principle components). Several techniques use Principal Component Analysis (PCA) for dimensionality reduction before actual outlier detection to detect a subset of features which capture the behavior of the data. Spectral approach can work in an unsupervised as well as semi-supervised setting. The simplest technique detected in PCA literature for outlier detection is based on the fact that the top few principal components capture the bulk of variability in a given data set. Thus one would expect that the smallest principal components result in constant values. Thus any data point that does not follow this structure for the smallest components is an outlier. [28] adopts this approach in order to find outliers in astronomy catalogs.

6. PROPOSED WORK

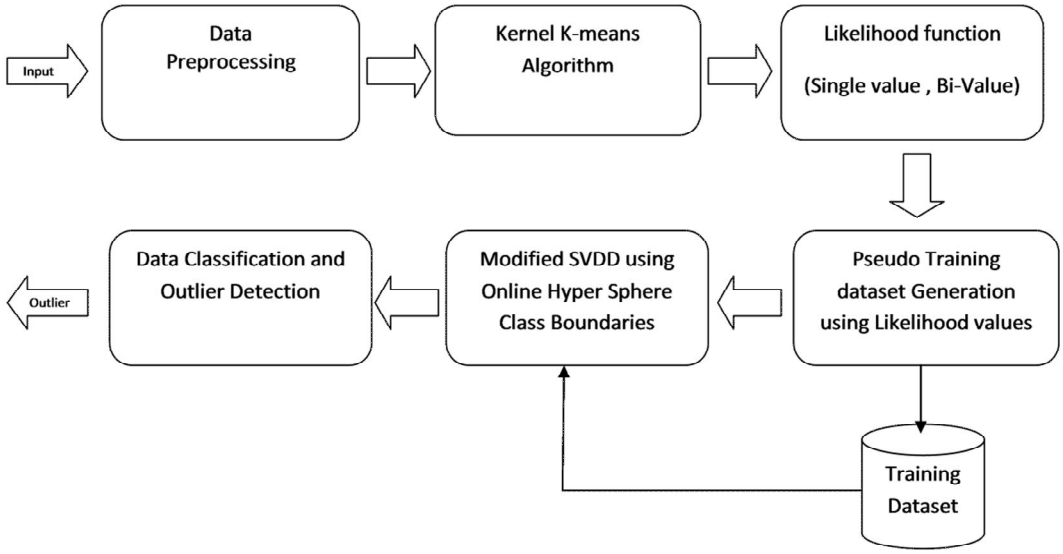

Figure1. Block diagram of proposed method

In recent years, many advanced technologies have been developed to store and record huge quantities of data continuously. This has given rise to the great need for uncertain data algorithms and applications. Many algorithms have been proposed for handling the uncertain data in query processing of uncertain data, indexing uncertain data, clustering uncertain data, classification of uncertain data, and frequent pattern mining of uncertain data. Meanwhile, considers uncertain data in the outlier detection problem where a probabilistic definition of outliers in conjunction with density estimation and sampling are used. Different from this work, proposed method is a model-based method, which does not need to pre-specify the density function of the dataset; therefore, proposed method can learn a distinctive classifier from the training set without assuming the distribution of the data. At the same time, proposed method models the uncertainty by assigning a confidence score to each sample and reduces the impact of the uncertain data on the construction of the classifier. In spite of much progress in outlier detection, much of the previous work did not explicitly cope with the problem of outlier detection having very few labeled negative examples and data having imperfect data label as well. In the figure 1, local data information is captured by generating the likelihood values of each input example towards the positive and negative classes. This information is then used in the generalized support vector data description framework in order to enhance a global classifier for the purpose of outlier detection.

7. DISCUSSIONS AND CONCLUSIONS

Outlier detection is an extremely vital problem with direct application in a wide variety of domains. An important observation with outlier detection is that it is not a well-formulated problem. We have discussed the distinct ways in which the problem has been formulated in literature. Every unique problem formulation has a different approach, resulting in a large literature on outlier detection techniques. Several approaches have been proposed to target a particular application domain. The survey can hopefully allow mapping of such existing approaches to other application domains.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM CSUR, vol. 41, no. 3, Article 15, 2009.
[2] Xue, A.: Study on Spatial Outlier Mining. Zhen Jiang, Jiang Su University (2008)
[3] V. Barnett and T. Lewis, Outliers in Statistical Data. Chichester, U.K.: Wiley, 1994.
[4] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in Proc. ACM SIGMOD Int. Conf. Manage. Data, New York, NY, USA, 2000, pp. 93–104.
[5] S. Y. Jiang and Q. B. An, "Clustering-based outlier detection method," in Proc. ICFSKD, Shandong, China, 2008, pp. 429–433.
[6] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection," in Proc. Natl. Acad. Sci. USA, 2001, pp. 31–36.
[7] Jain, A. K. and Dubes, R. C. 1988. Algorithms for Clustering Data. Prentice-Hall, Inc
[8] Marchette, D. 1999. A statistical method for proling network traffic. In Proceedings of 1st USENIX Workshop on Intrusion Detection and Network Monitoring. Santa Clara, CA, 119-128.
[9] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. 1998. Topic detection and tracking pilot study. In Proceedings of DARPA Broadcast News Transcription and Understanding Workshop. 194-218.
[10] He, Z., Deng, S., and Xu, X. 2002. Outlier detection integrating semantic knowledge. In Proceedings of the Third International Conference on Advances in Web-Age Information Management. Springer-Verlag, London, UK, 126-131.
[11] He, Z., Xu, X., and Deng, S. 2003. Discovering cluster-based local outliers. Pattern Recognition Letters 24, 9-10, 1641-1650.
[12] Vinueza, A. and Grudic, G. 2004. Unsupervised outlier detection and semi-supervised learning. Tech. Rep. CU-CS-976-04, Univ. of Colorado at Boulder. May.
[13] Smith, R., Bivens, A., Embrechts, M., Palagiri, C., and Szymanski, B. 2002. Clustering approaches for anomaly based intrusion detection. In Proceedings of Intelligent Engineering Systems through Artificial Neural Networks. ASME Press, 579-584.
[14] Barbara, D., Li, Y., Couto, J., Lin, J.-L., and Jajodia, S. 2003. Bootstrapping a data mining intrusion detection system. In Proceedings of the 2003 ACM symposium on Applied computing. ACM Press, 421-425.
[15] Barbara, D., Li, Y., and Couto, J. 2002. Coolcat: an entropy-based algorithm for categorical clustering. In Proceedings of the eleventh international conference on Information and knowledge management. ACM Press, 582-589.
[16] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander" LOF: Identifying Density-Based Local Outliers".
[17] Raghuvira Pratap, K Suvarna, J Rama Devi, Dr.K Nageswara Rao "Efficient Density based Improved K-Medoids " International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
[18] S.Vijayarani S.Nithya" An Efficient Clustering Algorithm for Outlier Detection" International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011.
[19]] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning," in Proc. ACM SIGKDD Int. Conf. KDD, New York, NY, USA, 2006, pp. 504–509.
[20] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," J. Mach. Learn. Res., vol. 6, pp. 211–232, Dec. 2005.
[21] J. Theller and D.M. Cai, "Resampling approach for anomaly detection in multispectral images," in Proc. SPIE, Orlando, FL, USA, 2003, pp. 230–240.
[22] D. Tax and R. Duin, "Outlier detection using classifier instability," in Proc. Adv. Pattern Recognit., London, U.K., 1998, pp. 593–601, LNCS.
[23] Yang, J., Zhong, N., Yao, Y.Y., et al.: Peculiarity analysis for classifications. In: Proceedings of the 2009 IEEE International Conference on Data Mining, pp. 607–616. IEEE Computer Society, Washington, DC, USA (2009)
[24] Knorr, E., Ng, R.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 12th International Conference on Very Large Data Bases, pp. 392–403 (1998)
[25] Ramaswamy, S., Rastogi, R., Kyuseok, S.: Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 427–438 (2000)
[26] Chakrabarti, S., Sarawagi, S., and Dom, B. 1998. Mining surprising patterns using temporal description length. In Proceedings of the 24rd International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 606-617.
[27] Noble, C. C. and Cook, D. J. 2003. Graph-based anomaly detection. In Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 631-636.
[28] Dutta, H., Giannella, C., Borne, K., and Kargupta, H. 2007. Distributed top-k outlier detection in astronomy catalogs using the demac system. In Proceedings of 7th SIAM International Conference on Data Mining.

AUTHORS

Aditi C. Dighavkar received the B.E. degree in Computer Engineering from K. K. Wagh Institute of Engineering Education and Research in 2009 and 2013, respectively. She is currently pursuing her Masters degree in Computer Engineering from K. K. Wagh Institute of Engineering Education and Research, Savitribai Phule Pune University Former UoP. This paper is published as a part of the research work done for the degree of Masters.

Prof. N. M. Shahane is an Associate Professor in Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Savitribai Phule Pune University. His current research interests include pattern recognition, digital signal processing, machine learning, data mining and mathematical modeling.