A Survey on Outlier Detection in a Sparse Coding Framework

[1]R. Swathi, [2]R. Suresh Kumar

[1]M.Phil Research Scholar, Sree Saraswathi Thiyagaraja College, Pollachi, India
[2]Associate Professor, Department of B.sc(Computer Science), Sree Saraswathi Thiyagaraja College, Pollachi , India

***Abstract*— Outlier detection is a significant problem that has been researched within various research areas and application domains. Many outlier detection methods have been particularly examined for certain application domains, as others are more standard. In this survey paper describes an outlier detection technique for high dimensional data sets accurately reduce the data from a root mapping at batch re-computation. For each outlier behavior activities is to identify the key factors, which are used by the methods to differentiate between normal and abnormal actions. This survey paper provides a best and brief understanding of the techniques belonging to each anomaly and sparse coding framework category. Further, for each clustering, to identify the improvements and drawbacks of the techniques in that category. It also provides a discussion on the computational complexity of the techniques since it is an important issue in real application domains hope that this survey will provide a good understanding of the many directions in which research has been done on this topic.**

***Index Terms*—Outlier, Sparse coding, nearest neighbor**

I. INTRODUCTION

Fraud detection must evolve continuously. Once criminals realize that a certain mode of fraudulent behavior can be detected, they will adapt their strategies and try others. Of course, new criminals are also attempting to commit fraud and many of these will not be aware of the fraud detection methods that have been successful in the past, and will adopt strategies that lead to identifiable frauds. This means that the earlier detection tools need to be applied as well as the latest developments. Statistical fraud detection methods may be 'supervised' or 'unsupervised'. In supervised methods, models are trained to discriminate between fraudulent and non-fraudulent behavior, so that new observations can be assigned to classes so as to optimize some measure of classification performance.

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains. Of these, anomalies and outliers are two terms used most commonly in the context of anomaly detection; sometimes inter- changeably. Anomaly detection finds extensive use in a wide variety of applications such as fraud detection for credit cards, insurance, or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities.

An outlier is an observation which appears to be inconsistent with the remainder of that set of data. In data mining, detection of outliers is an important research area. Most of the applications which apply outlier detection are high dimensional. Increase of dimensionality leads to sparsity of data. Sparse data is difficult to handle. The sparsity of high dimensional data signifies that every point is an almost equally good outlier [1]. Outlier (anomaly) detection refers to the process of finding patterns that do not conform to standard behavior. Outliers can be of three types such as point, contextual or collective outliers. Outlier detection techniques can be classified into three different categories such as supervised, semi supervised and unsupervised based on the existence of the labels for outliers. The unsupervised outlier detection is more applicable, where only a single data set without labels is given. The other techniques both require labeling data to produce the appropriate training set which is an expensive, time consuming and burdensome task [2]. In this paper, the proposed technique is applied to an unsupervised outlier detection. Hubness or k-hubness of an object x: Nk(x) is the number of times a point x is counted as one of the k nearest neighbors of any other point in a data set [3].

II. LITERATURE REVIEW

Richard J. Bolton and David J. Hand Et al[1] discussed the concerned with detecting behavioral fraud through the analysis of longitudinal data. These data usually consist of credit card transactions over time, but can include other variables, both static and longitudinal. Statistical methods for fraud detection are often classification (supervised) methods that discriminate between known fraudulent and non-fraudulent transactions; however, these methods rely on accurate identification of fraudulent transactions in historical databases – information that is often in short supply or non-existent. To particularly interested in unsupervised methods that do not use this information but instead detect changes in behavior or unusual transactions. To discuss two methods for unsupervised fraud detection in credit data in this paper and apply them to some real data sets.

AMANDA F. MEJIA et al [2] says Outlier detection for high-dimensional data is a popular topic in modern statistical research. However, one source of high-dimensional data that has received relatively little attention is functional magnetic resonance images (fMRI), which consists of hundreds of thousands of measurements sampled at hundreds of time points. At a time when the availability of fMRI data is rapidly growing— primarily through large, publicly available grassroots datasets consisting of resting-state fMRI data— automated quality control and outlier detection methods are greatly needed. To propose PCA leverage and demonstrate how it can be used to identify outlying time points in an fMRI scan. Furthermore, PCA leverage is a measure of the influence of each observation on the estimation of principal components, which forms the basis of independent component analysis (ICA) and seed connectivity, two of the most widely used methods for analyzing resting-state fMRI data.

VARUN CHANDOLA et al [3] addressed anomaly detection is an important problem that has been researched within diverse research areas and application domains. Many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic. This survey tries to provide a structured and comprehensive overviewoftheresearchonanomalydetection.Wehavegroupedexistingtechniquesintodifferentcategories based on the underlying approach adopted by each technique. For each category to have identified key as- sumptions, which are used by the techniques to differentiate between normal and anomalous behavior. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effectiveness of the technique in that domain. For each category, to provide a basic anomaly detection technique, and then show how the different existing techniques in that category are variants of the basic technique.

Mrs.R.Lakshmi Devi et al [4] address Identification of unsupervised outliers in a high dimensional data becomes an emerging technique in today's research in the area of data mining. Increase of dimensionality leads to various challenges. Hubness especially Antihubs (points that infrequently occur in k nearest neighbor lists) is the recently known concept for the increase of dimensionality pertaining to nearest neighbors. Outlier detection using AntiHub method is refined as Antihub2 to reevaluate the outlier scores of a point produced by the AntiHub method.

Charu C. Aggarwal et al [5] says outlier detection problem has important applications in the field of fraud detection, network robustness analysis, and intrusion detection. Most such applications are high dimensional domains in which the data can contain hundreds of dimensions. Many recent algorithms use concepts of proximity in order to find outliers based on their relation- ship to the rest of the data. However, in high dimensional space, the data is sparse and the notion of proximity fails to retain its meaningfulness. In fact, the sparsity of high dimensional data implies that every point is an almost equally good outlier from the perspective of proximity-based definitions. Consequently, for high dimensional data, the notion of finding meaningful outliers becomes substantially more complex and non-obvious. The authors discussed a new techniques for outlier detection which find the outliers by studying the behavior of projections from the data set.

Aleksandar Lazarevic et al [6] discussed outlier detection has recently become an important problem in many industrial and financial applications. In this paper, a novel feature bagging approach for detecting outliers in very large, high dimensional and noisy databases is proposed. It combines results from multiple outlier detection algorithms that are applied using different set of features. Every outlier detection algorithm uses a small subset of features that are randomly selected from the original feature set. As a result, each outlier detector identifies different outliers, and thus assigns to all data records outlier scores that correspond to their probability of being outliers. The outlier scores computed by the individual outlier detection algorithms are then combined in order to find the better quality outliers. Experiments performed on several synthetic and real life data sets show that the proposed methods for combining outputs from multiple outlier detection algorithms provide non-trivial improvements over the base algorithm.

Edwin M. Knox et al[6] address the deals with finding outliers (exceptions) in large, multidimensional datasets. The identification of outliers can lead to the discovery of truly unexpected knowledge in areas such as electronic commerce, credit card fraud, and even the analysis of performance statistics of professional athletes. Existing methods that we have seen for finding outliers in large datasets can only deal efficiently with two dimensions/attributes of a dataset. Here, we study the notion of DB-(Distance- Based) outliers. While we provide formal and empirical evidence showing the usefulness of DB-outliers, we focus on the development of algorithms for computing such outliers. First, we present two simple algorithms, both having a complexity of O(k N'), k being the dimensionality and N being the number of objects in the dataset. These algorithms readily support datasets with many more than two attributes. Second, we present an optimized cell-based algorithm that has a complexity that is linear wrt N, but exponential wrt k. Third, for datasets that are mainly disk-resident, we present another version of the cell-based algorithm that guarantees at most 3 passes over a dataset.

Sridhar Ramaswamy et al [7] discussed the addition to developing relatively straightforward solutions to finding such outliers based on the classical nested- loop join and index join algorithms, we develop a highly efficient partition-based algorithm for mining outliers. This algorithm first partitions the input data set into disjoint subsets, and then prunes entire partitions as soon as it is determined that they cannot contain outliers. This results in substantial savings in computation. We present the results of an extensive experimental study on real-life and synthetic data sets. The results from a real-life NBA database highlight and reveal several expected and unexpected aspects of the database. The results from a study on synthetic data sets demonstrate that the partition-based algorithm scales well with respect to both data set size and data set dimensionality.

Naoki Abe et al [8] says the existing approaches to outlier detection are based on density estimation methods. There are two notable issues with these methods: one is the lack of explanation for outlier flagging decisions, and the other is the relatively high

computational requirement. In this paper, authors presented a novel approach to outlier detection based on classification, in an attempt to address both of these issues. Our approach is based on two key ideas. First, we present a simple reduction of outlier detection to classification, via a procedure that involves applying classification to a labeled data set containing artificially generated examples that play the role of potential outliers. Once the task has been reduced to classification, we then invoke a selective sampling mechanism based on active learning to the reduced classification problem. We empirically evaluate the proposed approach using a number of data sets, and find that our method is superior to other methods based on the same reduction to classification, but using standard classification methods. It also show that it is competitive to the state-of-the-art outlier detection methods in the literature based on density estimation, while significantly improving the computational complexity and explanatory power.

III. Conclusion

In this survey paper have discussed some possible approaches to unsupervised credit card fraud detection through behavioral outlier detection techniques. The methods in this article describe early stages of research to produce some frameworks for unsupervised fraud detection and elementary examples are shown for illustrative purposes. We aim to proceed by incorporating other information, other than simply the amount spent, into the anomaly detection process and identifying the most useful and practical methods to develop for fraud detection.

References

[1] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," in Proc.Credit Scoring Credit Control VII, 2001, pp. 235–255.

[2] Z. He, X. Xu, J. Z. Huang, and S. Deng, "Mining class outliers: con- cepts, algorithms and applications in CRM," Expert Syst. Appl., vol. 27, no. 4, pp. 681–697, 2004.

[3] M. A. Rubin and A. M. Chinnaiyan, "Bioinformatics approach leads to the discovery of the TMPRSS2: ETS gene fusion in pros- tate cancer," Laboratory Investigation, vol. 86, no. 11, pp. 1099–1102, 2006.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surveys, vol. 41, no. 3, p. 15, 2009.

[5] E. M. Knox and R. T. Ng, "Algorithms for mining distance based outliers in large datasets," in Proc. Int. Conf. Very Large Data Bases, 1998, pp. 392–403.

[6] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 504–509.

[7] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2005, pp. 157–166.

[8] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," ACM SIGMOD Rec., vol. 30, no. 2, 2001, pp. 37–46.