Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

OUTLIER DETECTION APPROACH SURVEY FOR IMPERFECT DATA LABELS

Mr. Rohit U. Pawar[1] , Prof. Ujwala M. Patil[2]
[1]PG Student, Department Computer Engg., RCPIT,Shirpur, (MH) India
[2]Associate Professor, Department Of Computer Engg. , RCPIT,Shirpur, (MH) India

Abstract - The task of outlier detection is to identify data objects that are markedly different from or inconsistent with the normal set of data. However, in addition to normal data, there also exist limited negative examples or outliers in many applications, and data may be corrupted such that the outlier detection data is imperfectly labeled. These make outlier detection far more difficult than the traditional ones. Outlier detection approach to address data with imperfect labels and incorporate limited abnormal examples into learning. To deal with data with imperfect labels, use likelihood values for each input data which denote the degree of membership. Mainly approach works in two steps. In the first step, generate a pseudo training dataset by computing likelihood values of each example based on its local behavior. Kernel *k*-means clustering method and kernel LOF-based method to compute the likelihood values. In the second step, incorporate the generated likelihood values and limited abnormal examples into SVDD-based learning framework to build a more accurate classifier for global outlier detection. By integrating local and global outlier detection, the method explicitly handles data with imperfect labels and enhances the performance of outlier detection.
Keywords: Outlier detection, data of uncertainty, Imperfect Data Label**.**

I. INTRODUCTION

OUTLIER detection has attracted increasing attention in machine learning, data mining and statistics literature. Outliers always refer to the data objects that are markedly different from or inconsistent with the normal existing data [1], [2]. A well-known definition of "outlier" is given in [3]: "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism," which gives the general idea of an outlier and motivates many anomaly detection methods . Practically, outlier detection has been found in wide-ranging applications from fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, to military surveillance. Many outlier detection methods have been proposed to detect outliers from existing normal data. In general, the previous work on outlier detection can be broadly classified into distribution (statistical)-based, clustering-based, density-based and model-based approaches , all of them with long history. In the model-based approaches, they typically use a predictive model to characterize the normal data and then detect outliers as deviations from the model. In this category, the support vector data description (SVDD) has been demonstrated to be capable of detecting outliers in various application domains. In SVDD, a hyper-sphere is constructed to enclose most of the normal example with minimum sphere. The learned hyper-sphere is then utilized as a classifier to separate a test data into normal examples or outliers.

Though much progress has been done in support vector data description for outlier detection, most of the existing works on outlier detection always assume that input training data are perfectly labeled for

building the outlier detection model or classifier. However, we may collect the data with imperfect labels due to noise or data of uncertainty. For examples, sensor networks typically generate a large amount of data subject to sampling errors or instrument imperfections. Thus, a normal example may behave like an outlier, even though the example itself may not be an outlier. This kind of uncertain data information might introduce labeling imperfections or errors into the training data, which further limits the accuracy of subsequent outlier detection. Therefore, it is necessary to develop outlier detection algorithms to handle imperfectly labeled data.

Outliers, as defined earlier, are patterns in data that do not conform to a well defined concept of normal behavior of data, or conform to a well defined notion of outlying style, though it is typically easier to define the normal style. This survey discusses methods which find such outliers in data. Outliers exist in almost every real data set. Some of the prominent causes for outliers are listed below

1. Malicious activity such as insurance, credit card or telecom fraud, a terrorist activity or a cyber intrusion, weather forecasting.
2. Instrumentation error such as defects in components of machines or wear and tear.
3. Change 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

The “interestingness" or real life relevance of outliers is a important feature of outlier detection and distinguishes it from noise removal or noise accommodation, which deal with unwanted noise in the data. Noise in data does not have a real significance by itself, but acts as a hindrance to data analysis. Noise removal is driven by the need to remove the unwanted elements before any data analysis is performed on the data. Noise accommodation is nothing but immunizing statistical model estimation against outlying observations. Another related topic to outlier detection is novelty detection which aims at detecting unseen patterns in the data. The distinction between novel patterns and outliers is that the novel patterns are typically incorporated with the normal model after getting detected. It should be noted that the solutions for these related problems are often used for outlier detection and vice-versa.

II. APPLICATION AND SCOPE OF OUTLIER DETECTION

We shall highlight several applications of outlier detection. For each application we shall discuss following aspects:

- The notion of outlier.
- Nature of the data.
- Challenges associated with detecting outliers.
- Existing outlier detection techniques.

The Outlier Detection has scope in day todays useful things, examples are given below:

1) Fraud Detection:-

A general approach to outlier detection here would involve maintaining a usage profile for each customer and monitor the profiles to detect any deviations termed as activity monitoring. Some specific applications of fraud detection are discussed below.

Credit Card Fraud Detection: Outlier detection techniques are applied to detect:-

- Fraudulent Applications for Credit Card: This is similar to detecting insurance fraud.
- Fraudulent Usage of Credit Card: Associated with credit card thefts.

2) Mobile Phone Fraud Detection:-

In this activity monitoring problem the calling behavior of each account is scanned to issue an alarm when an account appears to have been misused. Calling activity is usually represented with call records. Each call record is a vector of continuous (e.g., Call-Duration) and discrete (e.g., Calling-City) features. However, there is no inherent primitive representation in this domain. Calls are aggregated by time, for example into call-hours or call-days or user
or area depending on the granularity desired. The outliers correspond to high volume of calls or calls made to unlikely destinations.

3) Medical and Public Health Outlier Detection:-

The data typically consists of patient records which may have several different types of features such as patient age, blood group, weight. The data might also have temporal as well as spatial aspect to it. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors. Most of the current outlier detection techniques in this domain aim at detecting anomalous records (point outliers). Typically the labeled data belongs to the healthy patient, hence most of the techniques adopt semi-supervised approach. Another form of data handled by outlier detection techniques in this domain is time series data, such as Electrocardiograms (ECG) and Electroencephalograms (EEG).

4) Image Processing :-

Outlier detection here aims to detect changes in an image over time (motion detection) or in regions which appear abnormal on the static image. This domain includes satellite imagery, digit recognition, spectroscopy, mammographic image, and video surveillance. The outliers are caused by motion or insertion of foreign object or instrumentation errors. The data has spatial as well as temporal characteristics. Each data point has a few continuous attributes such as color, lightness, texture, etc. The interesting outliers are either anomalous points or regions in the images (point and contextual outliers).

III. RELATED WORK

In last few year many research is going on the anomaly detection because this work is help to mine the important data from the large data ware house. This outlier detection algorithm is also help to identify credit card fraud and intrusion detection. Outlier detection aims to identify a small group of instances which deviate remarkably from the existing data. Outlier is define as “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” In the last few year many anomaly detection methods Practically, anomaly detection can be found in applications such as homeland security, credit card fraud detection, intrusion and insider threat detection in cyber-security, fault detection, or malignant diagnosis. The outlier detection algorithms are divide into the tree categories[1][2][3].

1. Statistics-based: This outlier detection techniques always fit a statistical model to the given data and then apply a statistical inference test to determine whether an unseen instance satisfies this model or not. In which Instances that have a low probability of being generated pattern from the learned model, based on the applied test statistic cases, are declared as outliers. For example, we can assume the normal examples follow a certain data distribution (such as Gaussian distribution),by estimating the parameter in the model, we can generate a Gaussian model to predict an unseen example into normal class or outliers. The statistics-based techniques always assume knowledge of the underlying distribution and

estimate the parameters from the given data such as Gaussian model based , in which the pre-specified data distribution is assumed to fit a Gaussian distribution; regression model based , where outlier detection using regression has been extensively investigated for time-series data; mixture of parametric distributions based , in which techniques use a mixture of parametric statistical distributions to model the data. For this category, the main disadvantage is that these techniques rely on the assumption that the data is generated from a particular distribution. However, this assumption often does not hold true in many applications, especially for high dimensional real data sets.

2. Density-based: This outlier detection technique always assumes that normal data instances occur in dense neighborhoods, while outliers occur far from their closest neighbors. One representative method is called LOF (local outlier factor) , which assigns an outlier score to any given data point, depending on its distances in the local neighborhood. In LOF there is a distance of a object or a parameter is be calculated and then as per LOF value the object or parameter is to be considered as a outlier. If LOF value is large than a other object value then this is a outlier. Recently, the work proposed by improves the accuracy of outlier detection by calculating an outlier score based on a Gaussian mixture model (GMM). However, if the data has normal instances that do not have enough close neighbors or if the data has outliers that have enough close neighbors, the technique fails to label them correctly, resulting in missed outliers.

3. Clustering-based: This outlier detection technique mainly relies on applying clustering techniques to characterize the local data behavior. As per by-product of clustering, small clusters that contain significantly fewer data points than other clusters are considered as outliers. The performance of clustering based techniques is highly dependent on the effectiveness of the clustering algorithm in capturing the cluster structure of normal instances. In that clustering algorithm like K-means, kernel K-means and other algorithm are used.

4. Model-based: This outlier detection techniques are used to learn a model from a set of labeled data instances and then to classify a test instance into one of the classes using the learnt model. Model-based outlier detection techniques operate in a similar two-phase fashion. The training phase learns a classifier using the available labeled training data. The testing phase classifies a test instance as normal or anomalous using the classifier. In this category, SVDD proposed by has been demonstrated empirically to be capable of detecting outliers in various domains. Model based approaches can detect global outliers effectively for high-dimensional data without need to assume the prior distribution of data.

Bo. Liu et.al. [1] presented kernel *k*-means clustering method and kernel LOF-based method to compute the likelihood values. In the second step, they incorporate the generated likelihood values and limited abnormal examples into SVDD-based learning framework to build a more accurate classifier for global outlier detection. By integrating local and global outlier detection, method explicitly handles data with imperfect labels and enhances the performance of outlier detection.

V. Chandola et. al. [5] defined three distinct formulations of the anomaly detection problem, and review techniques from many disparate and disconnected domains that address each of these formulations. Within each problem formulation, they group techniques into categories based on the nature of the underlying algorithm. For each category, author provide a basic anomaly detection technique, and show how the existing techniques are variants of the basic technique. This approach shows different

@IJMTER-2015, All rights Reserved

techniques within a category are related or different fromeach other. Our categorization reveals new variants and combinations that have not been investigated before for anomaly detection.

Shohei Hidoet. al. [12] this approach is expected to have better performance even in high-dimensional problems since methods for directly estimating the density ratio without going through density estimation are available. Among various density ratio estimation methods, they employ the method called unconstrained least-squares
importance fitting (uLSIF) since it is equipped with natural cross-validation procedures, allowing us to objectively optimize the value of tuning parameters such as the regularization parameter and the kernel width.

M. M. Breuniget. al. [7] innovated that for many scenarios, it is more meaningful to assign to each object a *degree* of being an outlier. This degree is called the *local outlier factor* (LOF) of an object. It is *local* in that the degree depends on how isolated the object is with respect to the surrounding neighborhood. Author provided a detailed formal analysis showing that LOF enjoys many desirable properties. Using real world datasets, they demonstrate that LOF can be used to find outliers which appear to be meaningful, but can otherwise not be identified with existing approaches.

D. M. J. Taxet. al. [11] presented the Support Vector Data Description (SVDD) which is inspired by the Support Vector Classifier. It obtains a spherically shaped boundary around a dataset and analogous to the Support Vector Classifier it can be made flexible by using other kernel functions. The method is made robust against outliers in the training set and is capable of tightening the description by using negative examples. They show characteristics of the Support Vector Data Descriptions using artificial and real data.

M. V. Joshi et. al. [13] were done the research on imbalanced data as follows. In general, previous work on imbalanced data classification falls into two main categories. The first category attempts to modify the class distribution of training data before applying any learning algorithms. This is usually done by over-sampling, which replicates the data in the minority class, or under-sampling, which throws away part of the data in the majority class. The second category focuses on making a particular classifier learner cost sensitive, by setting the false positive and false negative costs very differently and incorporating the cost factors into the learning process.

B. Liu et. al. [14] were defined uncertain-SVDD (U-SVDD) here, addresses the outlier detection only using normal data without taking the outlier/negative examples into account. Second, U-SVDD only calculates the degree of membership of an example towards the normal example and takes single membership into learning phase. However, the work by author addresses the problem of outlier detection with the few labeled negative examples, and takes data with imperfect labels into account.

Sr.	Author Name	Title	Year	Publication	Description
1.	Bo Liu, Yanshan Xiao, Philip S. Yu, ZhifengHao and Longbing Cao	An Efficient Approach for Outlier Detection with Imperfect Data Labels	2014	IEEE Transactions On Knowledge And Data Engineering	Kernel *k*-means clustering method and kernel LOF-based method to compute the likelihood values. In this system author used SVDD, by using it the define boundary to detect the outliers.

2.	Bo Liu, Yanshan Xiao, Longbing Cao, Zhifeng Hao, Feiqi Deng	SVDD-based outlier detection on uncertain data	2013	Knowledge Information System	Uncertain-SVDD only calculates the degree of membership of an example towards the normal example and takes single membership into learning phase.
3.	VarunChandola, Arindam Banerjee and Vipin Kumar	Anomaly Detection for Discrete Sequences: A Survey	2012	IEEE Transactions On Knowledge And Data Engineering	Author defined three distinct formulations of the anomaly detection problem, and review techniques from many disparate and disconnected domains that address each of these formulations. They group techniques into categories based on the nature of the underlying algorithm.
4.	ShoheiHido, YutaTsuboi, Hisashi Kashima, Masashi Sugiyama	Statistical outlier detection using direct density ratio estimation	2011	Knowl. Inform. Syst	To have better performance even in high-dimensional problems since methods for directly estimating the density ratio without going through density estimation are available.
5.	Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander	LOF: Identifying Density-Based Local Outliers	2005	Proc. ACM SIGMOD Int. Conf. Manage. Data, New York	It is more meaningful to assign to each object a degree of being an outlier. This degree is called the local outlier factor (LOF) of an object. It is local in that the degree depends on how isolated the object is with respect to the surrounding neighborhood.
6.	David M.J. Tax Robert P.W. Duin	Support vector data description	2004	in Proc. ASCI	It obtains a spherically shaped boundary around a dataset and analogous to the Support Vector Classifier it can be made flexible by using other kernel functions.
7.	Mahesh V. Joshi, Vipin Kumar	CREDOS: Classification using ripple down structure	2004	in Proc. SIAM Conf. Data Min	The first category attempts to modify the class distribution of training data

@IJMTER-2015, All rights Reserved

		(a case for rare classes)			before applying any learning algorithms, second category focuses on making a particular classifier learner cost sensitive, by setting the false positive and false negative

IV. CONCLUSIONS

We studied above all mechanisms, in it only some authors were worked on the criteria of outlier detection. In it we seen new model based approaches to outlier detection by introducing likelihood values to each input data into the SVDD training phase. We analyzed that, in the first step, generate a pseudo training dataset by computing likelihood values of each example based on its local behavior. Kernel *k*-means clustering method and kernel LOF-based method to compute the likelihood values, then builds global classifiers for outlier detection by incorporating the negative examples and the likelihood values in the SVDD-based learning framework. To address the problem of data with imperfect label in outlier detection, four variants of approaches to address the problem of data with imperfect label in outlier detection has been defined. Extensive experiments on ten real life data sets had shown that author approaches can achieve a better tradeoff between detection rate and false alarm rate for outlier detection in comparison to state-of-the-art outlier detection approaches.

REFERENCES

[1] Bo Liu, Yanshan Xiao, Philip S. Yu, ZhifengHao, and LongbingCao ” An Efficient Approach for Outlier Detection with Imperfect Data Labels ”IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 7, JULY 2014.
[2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM CSUR*, vol. 41, no. 3, Article 15, 2009.
[3] V. J. Hodge and J. Austin, “A survey of outlier detection methodologies,” *Artif.Intell.Rev.*, vol. 22, no. 3, pp. 85–126, 2004.
[4] D. M. Hawkins, *Identification of Outliers.* Chapman and Hall, Springer, 1980.
[5] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection for discrete sequences: A survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 823–839, May 2012.
[6] V. Barnett and T. Lewis, *Outliers in Statistical Data.*Chichester, U.K.: Wiley, 1994.
[7] 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.
[8] S. Y. Jiang and Q. B.An, “Clustering-based outlier detection method,” in *Proc. ICFSKD*, Shandong, China, 2008, pp. 429–433.
[9] 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.
[10] D. M. J. Tax and R. P. W. Duin, “Support vector data description,” *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, 2004.
[11] D. M. J. Tax, A. Ypma, and R. P. W. Duin, “Support vector data description applied to machine vibration analysis,” in *Proc. ASCI*, 1999, pp. 398–405.
[12] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, “Statistical outlier detection using direct density ratio estimation,”*Knowl. Inform. Syst.*, vol. 26, no. 2, pp. 309–336, 2011.
[13] M. V. Joshi and V. Kumar, “CREDOS: Classification using rippledown structure (a case for rare classes),” in *Proc. SIAM Conf. DataMin.*, 2004.
[14] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, “Svdd-based outlierdetection on uncertain data,” *Knowl. Inform. Syst.*, vol. 34, no. 3,pp. 597–618, 2013.