IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 9, October 2014.
www.ijiset.com
ISSN 2348 – 7968

Outlier Data Mining With Imperfect Data Labels

Mr.Dawange Y.P. , Prof. Durugkar S.R.

ME Computer Engg. SND COE & RC Yeola, (MH) India

HOD Department Of Computer Engg. SND COE & RC Yeola (MH) India

Abstract

To identify data objects that are markedly different from or inconsistent with the normal set of data is done by the outlier detection. Most existing solutions build a model using normal data and also identify outlier that do not fit represented model very proper. 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. It creates outlier detection very different than compared to that of traditional ones. To address data with imperfect labels and incorporate limited abnormal example into learning is done by a novel outlier detection approach. We have introduced likelihood values for each input data which denote the degree of membership of an example toward the normal and abnormal classes respectively, to deal with data with imperfect labels. There are two steps in our proposed approach work. First we generate a pseudo training dataset by computing likelihood values of each example based on its local behavior. To compute likelihood values, we present kernel k means clustering method and kernel LOF based method. The Second is we incorporate the generated likelihood values and limited abnormal examples into SVDD-based learning framework to build a more accurate classifier for global outlier detection. The performance of outlier detection method is based on the integrating local and global outlier detection, our proposed method explicitly handles data with imperfect data labels and enhances. Extensive experiments on real life datasets have demonstrated that our proposed approaches can achieve a better tradeoff between detection rate and false alarm rate as compared to state-of-the-art outlier detection approaches.

Keywords: Outlier detection, data of uncertainty, Imperfect Data Label**.**

1. Introduction

Outlier detection refers to the problem of detecting and analyzing patterns in data that does not conform to expected normal behavior. These anomalous patterns are often referred to as outliers means different form other, anomalies, discordant observations, exceptions, noise, errors, novelty, damage, faults, defects, aberrations, surprise, peculiarities or contaminants in different application domains. Outlier detection has been a widely researched data mining problem and finds immense use in a wide variety of application domains such as credit card, insurance, tax fraud detection, military surveillance for enemy activities, intrusion detection system for cyber security ,weather forecasting, fault detection in safety critical systems and many other areas. The importance of outlier detection is due to the fact that outliers in the data translate to significant information in a wide variety of application domains. For example, an anomalous pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized receiver. In public health data, outlier detection techniques are widely used to detect exceptional patterns in patient medical records which could be symptoms of a new disease means a outlier. Similarly, outliers in credit card transaction data could announce credit card theft or misuse. where the existence of an unusual region in a satellite image of enemy area could indicate enemy troop movement. Or exceptional readings from a space craft would signify a fault in some component of the craft. 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 outlier detection. We have identified the key aspects of any outlier detection technique, and used these as dimensions to classify current techniques into different categories. This survey intent at providing a better understanding of the distinct directions in which research has been done. Also it will help in determining the potential areas for future research.

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.

ISSN 2348 – 7968

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.

2. Literature Survey

In this 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) Distribution
2) Distance
3) density-based

first approaches assume that the data follows some standard or predetermined distributions, and this type of approach aims to find the outliers which deviate from such distributions. However, most distribution models are assumed univariate, and thus the lack of robustness for multidimensional data is a concern. Moreover, since these methods are typically implemented in the original data space directly, their solution models might suffer from the noise present in the database . Nevertheless, the assumption or the prior knowledge of the data distribution is not easily determined for practical problems

Second approach distance-based methods the distances between each data point of interest and its neighbors are calculated. If the result is above some predetermined threshold, the target instance will be considered as an outlier. While no prior knowledge on data distribution is needed, these approaches might encounter problems when the data distribution is complex (e.g., multi-clustered structure). In such cases, this type of approach will result in determining improper neighbors, and thus outliers cannot be correctly identified and third approach density-based methods are proposed One of the representatives of this type of approach is to use a density-based local outlier factor (LOF) to measure the outlines of each data instance Based on the local density of each data instance, the LOF determines the degree of outlines, which provides suspicious ranking scores for all samples. The most important property of the LOF is the ability to estimate local data structure via density estimation. This allows users to identify outliers which are sheltered under a global data structure. However, it is worth noting that the estimation of local data density for each instance is very computationally expensive, especially when the size of the data set is large.

3. Proposed System

The total contribution of our work can be summarized as follows

1) There are two likelihood models single likelihood and bi-likelihood model. In the first one, each input data is associated with one likelihood value which denotes the degree of membership towards its own class label. In the second bi-likelihood model, each of them has two likelihood values, which gives the degree of membership towards positive and negative class labels respectively. Based on the two likelihood models, we generate pseudo training datasets by computing likelihood values based on the local data behavior in the feature space. We put forward two methods based on the *k*-means clustering and local outlier factor (LOF) approaches respectively, to generate the likelihood values, which are called kernel *k*-means clustering-based method and kernel LOF-based method respectively. After that, we obtain two pseudo training sets for the two likelihood models respectively, in which each sample has likelihood values.

2) We construct two global classifier for outlier detection by generalizing the SVDD based learning process based on the two likelihood models, in this step. The soft SVDD is defined as the developed model derived from single likelihood model. Bi-soft SVDD is another classifier

ISSN 2348 – 7968

related with bi-likelihood. For both approaches, we incorporate the generated likelihood values of each sample and limited negative examples into the learning of support vector data description phase to build accurate outlier detection classifiers. Each sample makes different contribution to the learning of the outlier detection decision boundary based on their likelihood values. Our proposed approach explicitly handle the input data with imperfect labels and include a few labeled outliers into learning, by integrating local and global outlier detection.

3) To investigate the performance of our proposed approaches, we conduct extensive experiments on real life datasets. The results show that our proposed approaches can offer a better tradeoff between detection rate and false alarm rate and are less sensitive to noise in comparison of the state-of-the-art outlier detection algorithms.

3.1 Proposed System Block Digram

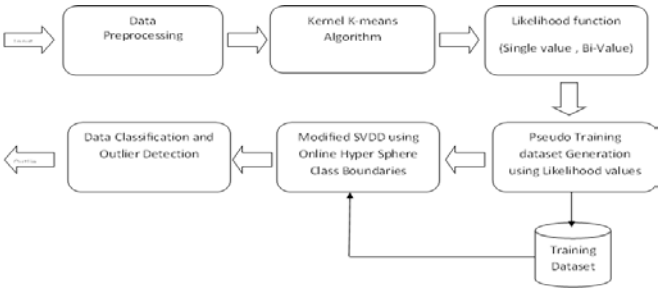

Fig. 1 Proposed System Block Digram.

3.2 Algorithm

Algorithmic steps for Kernel k-means clustering

Let X = {a_1, a_2, a_3, ..., a_n} be the set of data points and *'c'* be the number of clusters.

1) Randomly initialize 'c' cluster center.

2) Compute the distance of each data label point and the cluster center in the transformed space using:

$$\mathcal{D}(\{\pi_c\}_{c=1}^{k}) = \sum_{c=1}^{k}\sum_{a_i\in\pi_c} \|\phi(a_i) - m_c\|^2, \text{ where } m_c = \frac{\sum_{a_i\in\pi_c}\phi(a_i)}{|\pi_c|},$$

$$\phi(a_i)\cdot\phi(a_i) - \frac{2\sum_{a_j\in\pi_c}\phi(a_i)\cdot\phi(a_j)}{|\pi_c|} + \frac{\sum_{a_j,a_l\in\pi_c}\phi(a_j)\cdot\phi(a_l)}{|\pi_c|^2},$$

where,

c^{th} cluster is denoted by π_c.

'm_c' denotes the mean of the cluster π_c.

'$\Phi(a_i)$' denotes the data point a_i in transformed space.

$\Phi(a_i).\ \Phi(aj) = \exp^{-(\|ai - aj\|)*q}$ for gaussian kernel.

$= (c + a_i.a_j)^d$ for polynomial kernel.

3) Assign a data point to that cluster whose center distance is minimum.

4) Until data points are re-assigned repeat from the step 2).

3.3 Likelihood Values Generation

Single likelihood model:

In the model, we associate each input data with a likelihood value *(**x**i,m(**x**i))*, which indicates degree of membership of an example towards its own class label.

Bi-likelihood model:

In the model, each sample is associate with bi-likelihood values, denoted as (xi,mt(xi),mn(xi)), in which mt(xi) and mn(xi) indicate the degree of an input data xi belonging to the positive class and negative class respectively

To create a pseudo training dataset by computing likelihood values for each input data. For the single likelihood model, the generated pseudo training data consists of two parts for the *l* normal examples and *n* abnormal examples as follows.

*(**x**1,mt(**x**1)), . . . , (**x**l,mt(**x**l)), (**x**l+1,mn(**x**l+1)), . . . ,*

*(**x**l+n,mn(**x**l+n)),*

ISSN 2348 – 7968

in which *mt(***x***i)* and *mn(***x***i)* indicate the likelihood of example **x***i* belonging to the the normal class and the abnormal, respectively.

Similarly, the generated pseudo training data for bilikelihood model is:

*(***x**1,*mt(***x**1*),mn(***x**1*)), . . . , (***x***l,mt(***x***l),mn(***x***l)), (***x***l*+1,

*mt(***x***l*+1*),mn(***x***l*+1*)), . . . , (***x***l+n,mt(***x***l+n),mn(***x***l+n)),*

For each likelihood model, we propose two different schemes to compute likelihood values for each input data, which are inspired by the clustering-based and density based approaches to outlier detection. The basic idea of both schemes is to capture the local data uncertainty byexamining the relative distances of each input data to its local neighbors data in the feature space

4. Conclusions & Future Work

We proposed new model based approaches to outlier detection by introducing likelihood values to each input data into the SVDD training phase. In proposed system, we first capture the local uncertainty by computing likelihood values for each example based on its local data behavior in the feature space. 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 proposed. Extensive experiments on ten real life data sets have shown that our proposed 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] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM CSUR*, vol. 41, no. 3, Article 15, 2009.
[2] V. J. Hodge and J. Austin, “A survey of outlier detection methodologies,” *Artif. Intell. Rev.*, vol. 22, no. 3, pp. 85–126, 2004.
[3] D. M. Hawkins, *Identification of Outliers.* Chapman and Hall, Springer, 1980.
[4] 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.
[5] V. Barnett and T. Lewis, *Outliers in Statistical Data.* Chichester, U.K.: Wiley, 1994.
[6] 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.
[7] S. Y. Jiang and Q. B. An, “Clustering-based outlier detection method,” in *Proc. ICFSKD*, Shandong, China, 2008, pp. 429–433.
[8] 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.
[9] D. M. J. Tax and R. P. W. Duin, “Support vector data description,” *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, 2004.
[10] 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.

Mr. Dawange Yogesh Prakash Completed BE in Information Technology in 2012 and pursuing ME in Computer Engg. From SND COE & RC Yeola, (MH) India

Prof.Durugkar S.R.HOD Department Of Computer Engg.SND COE & RC Yeola (MH) India