ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

An Effective Progress for Extreme Deviation Spying with Flawed Data Labels

ABHINAYA.N.A

PG student, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science,

Thanjavur, TamilNadu, India

Abstract: The flaw detection will detect data objects that are inconsistent with normal dataset. In additional to normal data, there exist negative outliers in many applications and data will be imperfectly labeled. This paper represents an outlier detection approach to address data with imperfect data labels into learning. Our past approach works in two steps. In the first step, we develop a pseudo training dataset by computing possible values of each example based on its local action. We introduce kernel *k*-means clustering method and kernel Local Outlier Factor-based method to compute the likelihood values. In the next step, we introduce the obtained possible values and limited abnormal examples into SVDD-based learning to produce a more accurate classification for global outlier detection. The proposed system has three approaches. They are Naive Bayes approach, Logistic regression and . For classification of dataset we go for these two approaches which makes easy to find outliers.

Keywords: Flaw detection, abnormal data, local outlier factor

I. INTRODUCTION

FLAW detection has improved by increasing its performance in automatic learning, knowledge extracting and literature study. Outliers always points to the data sets which vary from the normal existing datasets. A popular definition of "flaw" is very great deviating data from normal datasets. Outlier detection is found in several applications such as duplicate detection of credit cards, life insurance, health care and intrusion detection in cyber security, military observation system[6].

Many outlier detection methods have proposed to find outliers from present normal data. In formal, the flaw detection is widely classified into distribution (statistical) based, clustering, density and method based manner[3]. In the model-based problem, they have used typically predictive model to characterize the normal data and then detect outliers as deviations. This category has support vector data description (SVDD) which demonstrates the capability of detecting flaws in several application domains. Even though there is much development 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, our work is to collect the data with imperfect labels due to noise or data of uncertainty. For examples, sensor networks will generate a large number of data related to sampling errors and instrument abnormality. This kind of imperfect data information may introduce labeling imperfections and errors into the training data, which at a greater distant limits the accuracy of upcoming outlier detection. Hence, it is necessary to implement outlier detection algorithms to handle imperfect data labels.

In addition to this another important work is that, negative examples or outliers, even though very few, also exist in many applications. For example, in the network domain, in addition to extensive data about the normal traffic conditions in the network, there also exist a small number of attacks that can be collected in improving the outlier detection. Though these outliers are not sufficient for building a binary classifier, they can be integrated into the training process to refine the decision boundary around the normal data for flaw detection.

In order to handle the outlier detection with imperfect labels, we propose a new approach to outlier detection by generalizing the support vector data description learning framework on imperfectly labeled training dataset. We must associate each example in the training dataset not only with a class label but also likelihood values which denotes the degree of membership towards the positive and negative classes. We then facilitate the few labeled negative examples and the generated likelihood values into the learning phase of SVDD to build a more accurate classifier[9].

1) We introduce two possible models, they are one and two possible model. In the first possible model, every input data is matched with one possible value which indicates the degree of relationship towards its own class label and in the bi possible model; each sample has the two probability values which indicates the degree of relationship to positive and negative class labels respectively. According to these two probability models, we generate trained datasets by computing probability values based on the local data behaviour in the feature space. We put forth two possibility models based on the kernel k means clustering classification and local outlier factor methods, to develop the possible values, which are called kernel k-means clustering-based method and kernel Local Outlier Factor based methods[10]. Then we obtain two trained datasets for the two possible models, in which each test data will have probability values.

2) In this second step, we construct two global classifiers for outlier detection by generalizing the SVDD

Copyright to IJARCCE
DOI 10.17148/IJARCCE.2015.4228

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

based learning process based on two likelihood models. The introduced model has obtained from the single possible model is called soft-SVDD. The classifier related with bi-likelihood model is called bisoft- SVDD. For both the 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 exact outlier detection classifiers[8]. In this process, each sample makes different contribution to the learning of the outlier detection decision boundary based on their likelihood values. By integrating local and global flaw detection, our proposed approaches explicitly handle the input data with abnormal and flawed labels and include a few labeled outliers into learning.

3) We conduct that the extensive experiments on real life datasets to investigate the performance of our proposed approaches. Thus result shows 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.

Compared to the previous work of outlier detection, such as Artificial Immune System (AIS), most of them did not explicitly cope with the problem of both outlier detection with very few labelled negative examples and outlier detection on the data with imperfect labels. Our proposed system first identifies the local data information by producing likelihood values for input examples, and then facilitates such information into support vector data description framework to build a more absolute flaw detection classifier.

II. RELATED WORK

In this section, we discuss previous approaches that are related to our study. Since we focus on outlier detection with imperfectly labeled outliers and data with imperfect labels, we concisely review past work on flaw detection. Another branch of related work on learning from imbalanced data. Finally, we will briefly review support vector data description.

1. Outlier Detection

In the previous, many outlier detection methods have been proposed. Typically, these existing approaches can be divided into four categories: distribution- based clustering-based, density-based and model-based approaches. The Statistical approaches will assume that the data follows some of the standard or predetermined distributions, and this type of approach targets to identify the outliers which deviate from such distributions. The methods in this category always assume a normal example that follow a certain of data distribution. Nevertheless, we can never have this kind of priori data distribution knowledge in practice, especially for high dimensional real data sets. For clustering-based approaches, they always conduct the clustering-based techniques on the samples of data which characterize the local data behavior[3]. In common, the sub-clusters contain significantly less data points than other clusters, are considered as outliers. For example, clustering techniques have been used to find flaws in the intrusion detection domain. In this work of, the clustering techniques iterative detect outliers to multidimensional data analysis. Since the clustering based approaches are unsupervised without requiring any labeled training data, the performance of unsupervised outlier detection is limited.

In addition to the density-based approaches has been proposed. One of the representatives of this type of approaches are the local outlier factor (LOF) and its variants. As per the local flaw density of each data instance, the LOF examines the percentage of outlier, which rates the value for all the samples[1]. The most important quality of the Local Outlier Factor has the ability to estimate local data structure through density estimation. The main advantage of this approach is that they do not need to make any assumption for the generative distribution of the data. However, these approaches exist with a high computational complexity in the testing phase, since they has to find the distance between each test instance and all the other instances to compute the nearest neighbours of data labels.

Apart from the above work, model-based outlier detection approaches have proposed. Among them, support vector data descriptions (SVDD) have demonstrated empirically to be capable of finding outliers in various domains. SVDD conducts a small sphere around the normal data and make use of the constructed sphere to detect an unknown sample as normal or outlier. The most attractive feature of SVDD is that it can transform the original data into a feature space through a kernel function and effectively detect the global outliers for high-dimensional data[5]. Though, its performance is sensitive to the noise involved in the input data.

Depending on the availability of a training dataset, outlier detection techniques mentioned above will operate in two different modes: supervised and automatic modes. Among four types of flaw detection methods, distribution based approaches and model based approaches will come into the category of supervised flaw detection, which checks for the availability of a training dataset that have been labeled instances for normal class. In addition to this, several techniques have been proposed that initiates artificial anomalies into a normal dataset to obtain a labeled training data set. Apart from this, the work will present a new method to identify outliers by utilizing the instability of the output of a classifier built on trained data.

In spite of much progress on flaw detection, most of the past work did not explicitly cope with the problem of outlier detection with very few labeled negative examples and data with flawed label as well. Our proposed system identifies local data information by generating the likelihood values of each input example towards the positive and negative classes. Such information is then

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

facilitated into the generalized support vector data description framework to speed up the global classifier for outlier detection.

The work in this paper has difference from our past work about flaw detection. Initially, the work will be, called uncertain SVDD (U-SVDD). This addresses the outlier detection only using the normal data without taking the outlier/negative examples into account. The second is, U-SVDD that calculates only the degree of membership of an example towards the normal example and takes the single membership into learning phase[7]. Based on this problem, we put forth the single likelihood model and bi-likelihood model to assign likelihood values to each examples based on their local behaviors. For single possibility model, the examples including positive and negative classes are assigned likelihood values indicating the degree of membership towards their own class labels. For bi-likelihood model, each example is not only with a class label but also bi-likelihood values which denote the degree of membership towards the positive and negative classes respectively. Based on two likelihood models, we put forth soft SVDD and bi-soft SVDD methods to facilitate the likelihood values together negative examples into SVDD-based learning phase[4]. Hence, the optimization model is said to be soft SVDD and the other model is called bi-soft-SVDD are completely different from the optimization problem.

2. Difference from Imbalanced Data Classification

The flaw detection problem that we consider in this paper is also closely related to the problem of imbalanced data classification, in which imperfect datas corresponding to the negative class are very small in proportion when compared to the normal data corresponding to the positive class.

Thus we briefly review the research on imperfect data as follows. In common, past work on imperfect data classification falls into two main categories. The first category tries to alter 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 ignores away part of the data in the majority class. The second category focuses on making a specific classifier learner cost sensitive, by setting the false positive and false negative costs very differently and facilitating the cost factors into the learning process. Representative methods include the cost-sensitive decision trees and the cost sensitive SVMs. In cost-sensitive SVMs, the cost factors of two classes are set as unique so that the cost factors can affect the decision boundary. When imperfect data are present, researchers have fought for the use of ranking-based metrics, such as the ROC curve and the area under ROC curve (AUC) instead of using exact.

The comparison between imperfect data classification and our flaw detection problem is that: in imperfect data classification, the examples from one or more minority classes are always self-similar, potentially forming compact clusters, while in flaw identification, the outliers are typically spread around normal data so that the distribution of the negative class can never be well represented by the very few negative training examples[2]. To solve this problem, we can exploit cost-sensitive learning algorithms, but the false positive and false negative costs are commonly unknown to us in real life applications. Hence, we exploit one class classification method for flaw detection, which aims at building decision boundary around the normal data, and makes use of the few negative examples to refine the boundary to build a flaw identification classifier.

3. Support Vector Data Description

The most important feature of SVDD is that it can convert the input data into a feature space and can identify global outliers effectively. As mentioned in Figure.1(b), the hypersphere is the feature space that responds to a decent decision boundary in input space. However, the performance of SVDD is sensitive to the defects involved in the input data labels[5]. Our proposed method customizes the SVDD to facilitate the likelihood value towards the positive and negative classes, which simulates the effect of noise on flaw detection.

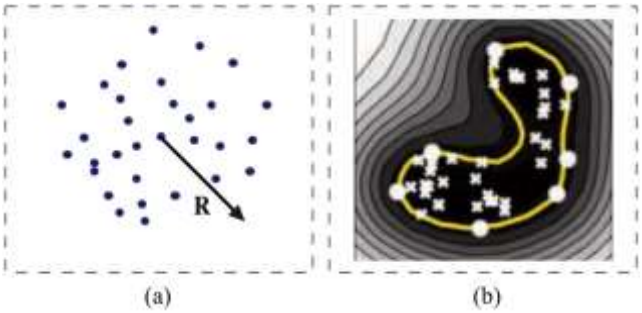

Figure 1: (a) Illustration of SVDD in feature space (b) Illustration of SVDD decision boundary in input space

III. OUR PROPOSED APPROACH

In this area, we provide a detailed description about our proposed approaches to flaw detection. Given a set of trained data S which consists of l normal examples and a small amount of n flaws (or abnormal) examples, the objective is to build a classifier using both normal and abnormal training data and the classifier is thereafter applied to separate unseen test data. However, finding the sample errors or device imperfections, a normal example might behave as a flaw, even though the sample itself might not be a flaw. Such mistakes might result in an imperfectly labeled training data, which makes the subsequent outlier identification become grossly inaccurate. To handle with this misery, we put forth two likelihood models as follows.

Single likelihood model: In this model, we do associate each input data with the likelihood value (xi, m(xi)), which

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

indicates the degree of membership of an example towards its own class label.

Bi-likelihood model: In this model, each sample is related with bi-likelihood values, explained as xi, mt(xi), mn(xi)), where mt(xi) and mn(xi) denoted the degree of an input data xi related to the positive class and negative class. The absolute difference of two models is that, single likelihood model consider only the degree of membership towards its own class label; while the other bi-likelihood model adds the degree of membership towards its own class and the opposite class. That likelihood values information is hence incorporated into the construction of a global classifier for outlier detection. Based on this proposed approaches work in two steps as follows:

• In the first step, for each likelihood model, we produce a pseudo training dataset by evaluating likelihood values for each input data based on local data behavior in the feature space.

• In the second step, we put forth the soft-SVDD and bi-soft-SVDD for single likelihood model and bilikelihood model respectively, by using both normal and abnormal samples as well as the produced likelihood values.

IV. ARCHITECTURE DIAGRAM

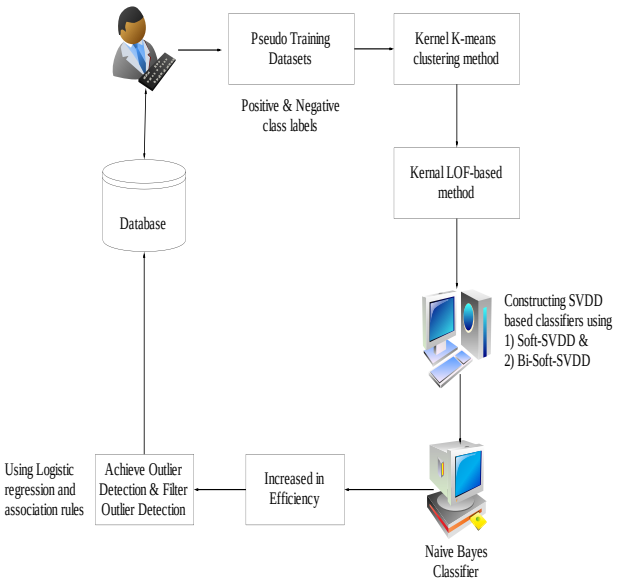

Figure 2: Architecture diagram

Architecture diagram consists of all the modules including data deployment, pseudo training datasets, kernel k-means clustering method, kernel Local Outlier Factor based method for constructing soft SVDD and bi soft SVDD, Naive bayes which consists of monotonic functions and checks for probability to enhance the efficiency of classification. Outlier is detected and it is highlighted.

Then it is further classified to detect the highest precision using Logistic Regression and association rules. Then it is updated by the user and stored those datas in the dataset.

V.DATA FLOW DIAGRAM

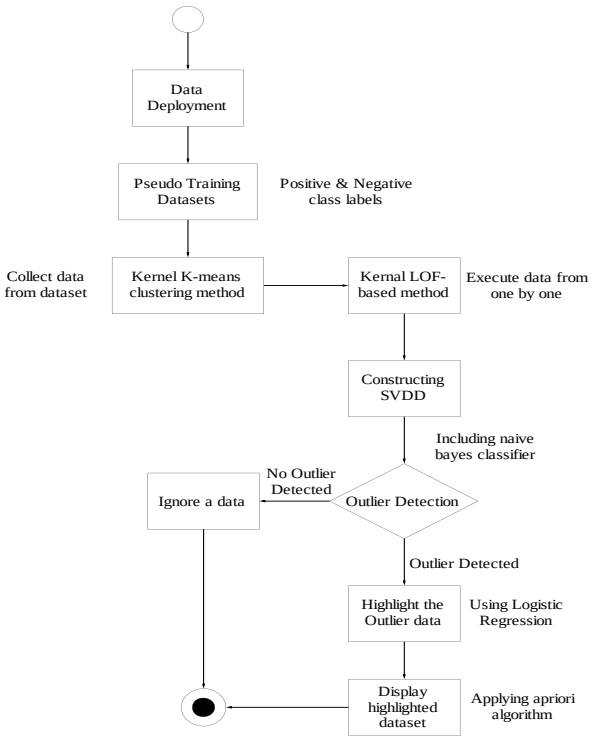

Figure 3: Data Flow representation`

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we propose new model-based approaches to flaw detection by developing the likelihood values to each input data into the SVDD training phase. Our system first identifies the local uncertainty by computing likelihood values for each sample based on its local data action in the feature space, and then constructs global classifiers for flaw detection by incorporating the negative examples and the likelihood values in the SVDD-based learning framework. We have proposed variants of approaches to mention the problem of data with flawed label in outlier identification. We have planned to elaborate our work in various directions. First, we would like to identify how to design better ways to produce likelihood values based on the data classification in a given application domain. Second, we will see for how to use an online process to learn the hyper-sphere boundary of soft-SVDD, linear classification in the streaming environments.

ACKNOWLEDGEMENT

My sincere thanks to my guide Mrs.R.Rebecca Asst.Professor, HOD, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur for her help and guidance.

REFERENCES

[1] 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.
[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM CSUR, vol. 41, no. 3, Article 15, 2009.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

[3] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Statistical outlier detection using direct density ratio estimation," no. 2, pp. 309–336, 2011.
[4] D. M. J. Tax and R. P. W. Duin, "Support vector data description," Mach. Learn., vol. 54, no. 1, pp. 45–66, 2004.
[5] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, "Svdd-based outlier detection on uncertain data," Knowl. Inform. Syst., vol. 34, no. 3, pp. 597–618, 2013.
[6] Bo Liu, Yanshan Xiao, Philip S. Yu, Zhifeng Hao, and Longbing Cao," An Efficient Approach for Outlier Detection with Imperfect Data Labels," VOL. 26, NO. 7, JULY 2014
[7] Y. Lin, Y. Lee, and G. Wahba, "Support vector machine for classification in nonstandard situations," Mach. Learn., vol. 46, no. 1–3, pp. 191–202, 2002.
[8] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in Proc. ECML, Pisa, Italy, pp. 39–50, 2004.
[9] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," Knowl. Inform. Syst., vol. 25, no. 1, pp. 1–20, 2010.
[10] S. Borah and M. Ghose, "Performance analysis of AIM-k-means & k-means in quality cluster generation," J. Comput., vol. 1, no. 1, pp. 175–178, Dec. 2009.

BIOGRAPHY

N.A.Abhinaya Completed B.E., (CSE) at Periyar Maniammai University, Thanjavur in 2013. Now pursuing M.E (CSE) at Parisutham Institute of Technology and Science, Thanjavur.

DOI 10.17148/IJARCCE.2015.4228