ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2015

Review on Classification Mechanism for Outlier'S with Imperfect Data Labels for Streaming Data Environments

Sonali S.Patil [1], Hemant A.Tirmare [2]

P.G. Student, Dept. of Computer Science & Technology, DOT, Shivaji University Kolhapur, Maharashtra India[1]

Assistant Professor, Dept. of Computer Science &Technology, DOT, Shivaji University Kolhapur, Maharashtra, India [2]

ABSTRACT: Outlier detection is the set of data points that are considerably different than the remainder of the data. Consequently, for high dimensional data, the notion of finding meaningful outliers becomes substantially more complex and non-obvious. Earlier solutions typically build a model using the normal data and identify outliers that do not fit the represented model very well. 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. To overcome these difficulties proposed system focuses on outlier detection with limited labeled outliers and data with imperfect labels. Also capture local data information by generating the likelihood values of each input example towards the positive and negative classes respectively. Such information is then incorporated into the generalized support vector data description technique to enhance a global classifier for outlier detection. Here presented outlier detection approach to address data with imperfect labels and incorporate limited abnormal examples into learning. To deal with data imperfect labels, here introducing likelihood values for each input data which denote the degree of membership of the normal and abnormal classes respectively. Proposed approach works for online streaming data environment and different classification techniques such as K-means, Naives Bayes. Here considering combining approach to compute the likelihood values. By integrating local and global outlier detection, proposed method explicitly handles data with imperfect labels and enhances the performance of outlier detection.

KEYWORDS: Outlier detection, Classification techniques, likelihood values, imperfect labels.

I. INTRODUCTION

Outlier detection refers to the problem of detecting and analysing patterns in data that does not map to expected normal behaviour. 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. 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.

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. Identify data objects that are markedly different from or inconsistent with the normal set of data deal with outlier detection . Most existing solutions typically build a model using the normal data and identify outliers that do not suitable. 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 labelled. These make outlier detection far more difficult than the traditional ones. Imperfect labelling is the big issue in outlier detection approach. To deal with data with imperfect labels, here introduce likelihood values for each input data which denote the degree of membership of an example toward the normal and abnormal classes respectively.

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2015

Proposed 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 behaviour. Existing systems present kernel k-means clustering method and kernel LOF-based method to compute the likelihood values.
- In the second step, incorporating the generated likelihood values and limited abnormal examples into SVDD-based learning framework [7] to build a more accurate classifier for global outlier detection. By integrating local and global outlier detection, proposed method explicitly handles data with imperfect labels and enhances the performance of outlier detection.

Proposed approaches can achieve better tradeoffs between detection rate and false alarm rate as compared to state-of-the-art outlier detection approaches. Outlier detection is an important problem that has been studied within diverse research areas and application domains. Most existing methods are based on the assumption that an example can be exactly categorized as either a normal class or an outlier. However, in many real-life applications, data are uncertain in nature due to various errors or partial completeness. These data uncertainty make the detection of outliers far more difficult than it is from clearly separable data. The key challenge of handling uncertain data in outlier detection is how to reduce the impact of uncertain data on the learned distinctive classifier.

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.
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
- model-based approaches

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, 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. These 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. In addition, another important observation is that, negative examples or outliers, although very few, do exist in many applications. For example, in the network intrusion domain, in addition to extensive data about the normal traffic conditions in the network, there also exist a small number of cyber attacks that can be collected to facilitate outlier detection. Although these outliers are not sufficient for constructing a binary classifier, they can be incorporated into the training process to refine the decision boundary around the normal data for outlier detection.

In order to handle outlier detection with imperfect labels, here used a novel approach to outlier detection by generalizing the support vector data description learning framework on imperfectly labeled training dataset. Here 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. Then incorporate the few labeled negative examples and the generated likelihood values into the learning phase of SVDD to build a more accurate classifier.

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2015

II. RELATED WORK

In this section, presenting theory to discuss previous work related to this system. Since here focus on outlier detection with limited labeled outliers and data with imperfect labels, also briefly review previous work on outlier detection and discuss another branch of related work on learning from imbalanced data.

Outlier Detection

V. Chandola, A. Banerjee, and V. Kumar [2] , many outlier detection methods have been proposed . Typically, these existing approaches can be divided into four categories: distribution (statistical)- based clustering-based, density-based and model-based approaches.

E. Eskin , S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori[6][7],Statistical approaches assume that the data follows some standard or predetermined distributions, and this type of approach aims to find the outliers which deviate form such distributions. The methods in this category always assume the normal example follow a certain of data distribution. Nevertheless, we cannot always have this kind of priori data distribution knowledge in practice, especially for high dimensional real data sets.

S. Y. Jiang and Q. B, R. Smith, A. Bivens, M. Embrechts, C. Palagiri, and B, Y. Shi and L. Zhang [3][8][9], For clustering-based approache they always conduct clustering-based techniques on the samples of data to characterize the local data behaviour. In general, the sub-clusters contain significantly less data points than other clusters, are considered as outliers. For example, clustering techniques has been used to find anomaly in the intrusion detection domain . In the clustering techniques iterative detect outliers to multidimensional data analysis in subspace. Since clustering based approaches are unsupervised without requiring any labeled training data, the performance of unsupervised outlier detection is limited.

K. Bhaduri, B. L. Matthews, and C. Giannella[18] ,they proposed density-based approaches. One of the representatives of this type of approaches are local outlier factor (LOF) and variants. Based on the local density of each data instance, the LOF determines the degree of outlierness, 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. The advantage of these approaches is that they do not need to make any assumption for the generative distribution of the data. However, these approaches incur a high computational complexity in the testing phase, since they have to calculate the distance between each test instance and all the other instances to compute nearest neighbours.

E. M. Jordaan and G. F. Smits[20] have proposed model-based outlier detection approaches. Among them, support vector data description (SVDD) has been demonstrated empirically to be capable of detecting outliers in various domains. SVDD conducts a small sphere around the normal data and utilizes 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 via kernel function and effectively detect global outliers for high-dimensional data. However, its performance is sensitive to the noise involved in the input data.

B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng[10] The work in the paper has difference from previous work about outlier detection. First, the work in, called 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.

Mohiuddin Ahmed, Abdun Naser Mahmood[21] have proposed techniques consider either small clusters as outliers or provide a score for being outlier to each data object. These approaches have limitations due to high computational complexity and misidentification of normal data object as outliers. They provide a novel unsupervised approach to detect outliers using a modified k-means clustering algorithm. The detected outliers are removed from the dataset to improve clustering accuracy and validate approach by comparing against existing techniques and benchmark performance.

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2015

Varun Chandola, Arindam Banerjee[22], they provide a comprehensive and structured overview of the existing research for the problem of detecting anomalies in discrete/symbolic sequences. The objective is to provide a global understanding of the sequence anomaly detection problem and how existing techniques relate to each other. The key contribution of this survey is the classification of the existing research into three distinct categories, based on the problem formulation that they are trying to solve. These problem formulations are: 1) identifying anomalous sequences with respect to a database of normal sequences 2) identifying an anomalous subsequence within a long sequence 3) identifying a pattern in a sequence whose frequency of occurrence is anomalous. It show how each of these problem formulations is characteristically distinct from each other and discuss their relevance in various application domains. Also review techniques from many disparate and disconnected application domains that address each of these formulations. This approach shows how different techniques within a category are related or different from each other. the categorization reveals new variants and combinations that have not been investigated before for anomaly detection

Hasan Ferdowsi, Sarangapani Jagannathan, and Maciej Zawodniok[23], The main objective is to design an outlier scheme, which can detect, identify, and remove the outliers online, before an outlier triggers a false alarm during fault diagnosis. Therefore, the outlier detection must be performed online and prior to FD and root cause analysis. The OIR scheme uses a neural network (NN) to estimate the actual system states from measured system states involving outliers. With this method, the outlier detection is performed online at each time instant by finding the difference between the estimated and the measured states and comparing its median with its standard deviation over a moving time window.

Rikard Laxhammar and Göran Falkman[24], Detection of anomalous trajectories is an important problem in the surveillance domain. Various algorithms based on learning of normal trajectory patterns have been proposed for this problem. Yet, these algorithms typically suffer from one or more limitations: They are not designed for sequential analysis of incomplete trajectories or online learning based on an incrementally updated training set. Moreover, they typically involve tuning of many parameters, including ad-hoc anomaly thresholds, and may therefore suffer from overfitting and poorly-calibrated alarm rates. They propose and investigate the Sequential Hausdorff Nearest-Neighbour Conformal Anomaly Detector (SHNN-CAD) for online learning and sequential anomaly detection in trajectories. This is a parameter-light algorithm that offers a well-founded approach to the calibration of the anomaly threshold. The discords algorithm, originally proposed by Keogh et al., is another parameter-light anomaly detection algorithm that has previously been shown to have good classification performance on a wide range of time-series datasets, including trajectory data.

C. Elkan[11] have suggested The outlier detection problem that consider in this paper is also related to the problem of imbalanced data classification, in which outliers corresponding to the negative class are extremely small in proportion as compared to the normal data corresponding to the positive class.

C. Elkan and M. V. Joshi ,V. Kumar[11][12] they briefly review 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 .

G. Nakhaeizadeh, U. Knoll, B. Tausend, G. Fumera and F. Roli C. X. Ling, J. Huang, and H. Zhang,[14][15][17] the authors proposed representative methods include cost-sensitive decision trees and cost-sensitive SVMs. In cost-sensitive SVMs, the cost factors of two classes are set differently so that the cost factors can affect the decision boundary. When imbalanced data are present, researchers have argued for the use of ranking-based metrics, such as the ROC curve and the area under ROC curve (AUC) instead of using accuracy.

III.PROPOSED SYSTEM

In this section, here provide a detailed description about proposed approaches to outlier detection. Given a set of training data S which consists of l normal examples and a small amount of n outlier (or abnormal) examples, the

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2015

objective is to build a classifier using both normal and abnormal training data and the classifier is thereafter applied to classify unseen test data. However, subject to sampling errors or device imperfections, a normal example may behave like an outlier, even though the example itself may not be an outlier. Such error factors might result in an imperfectly labeled training data, which makes the subsequent outlier detection become grossly inaccurate. To deal with this problem, put forward two likelihood models as follows.

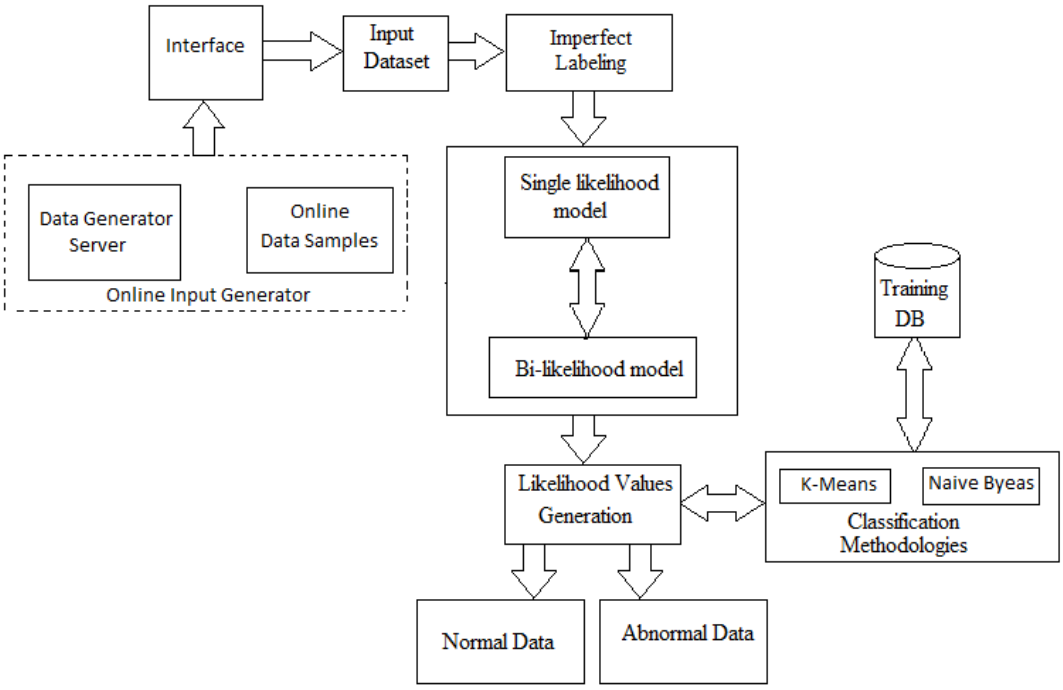

Fig: 1 Proposed System Architecture

The Modules of Proposed system is described as follows:

Online Input Generator:

Online input generator produces a data to further used to imperfect data labelling. noisy data can be added to analysis and data pre-processing.

Input dataset:

This is the input mechanism to the proposed system which consists of normal and a small abnormal, the objective is to build a classifier using both normal and abnormal training data

Imperfect Labelling:

In this, work involves a outlier detection approach to address data with imperfect labels. Here add noisy data or abnormal data in input dataset to classify is thereafter applied to classify unseen test data.

Single likelihood model: In the model, associate each input data with a likelihood value , 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, indicate the degree of an input data belonging to the positive class and negative class respectively.

Likelihood Value generator:

The main difference of above two models is that, single likelihood model only considers the degree of membership towards its own class label; while bi-likelihood model includes the degree of membership towards its own class and the opposite class. Such likelihood values information is thereafter 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, here generate a pseudo training dataset by computing likelihood values for each input data based on local data behaviour in the feature space.

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2015

Classification Methodologies:

Training data containing flow records of both normal and abnormal data transformed into feature datasets The datasets are divided into different clusters for normal and abnormal dataset using the K-means clustering algorithm and Naive byeas algorithm .The resulting cluster centroids are deployed for fast detection of abnormal data in new monitoring data based on simple distance calculations. Also Likelihood Value used to classification methodologies.

IV. CONCLUSION

Outlier detection has been a very significant concept in the data analysis. 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. Here Client and Server interface get the Input file, then use the Training dataset by adding abnormal data . This system will evaluate performance of Imperfect Labels against the noisy data .

REFERENCES

[1] Bo Liu,Yanshan Xiao,Philip S.Yu,Zhifeng Hao, and Longbing Cao, "An Efficient Approach for Outlier Detection with Imperfect Data Labels," IEEE Transaction on Knowledge and Data Engineering., vol. 26, no. 7, pp. 1602– 1616, 2014.
[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM CSUR, vol. 41, no. 3, Article 15, 2009.
[3] S. Y. Jiang and Q. B. An, "Clustering-based outlier detection method," in Proc. ICFSKD, Shandong, China, pp. 429–433, 2008.
[4] L. Chen and C. Wang, "Continuous subgraph pattern search over certain and uncertain graph streams," IEEE Transaction on Knowledge and Data Engineering, vol. 22, no. 8, pp. 1093–1109, 2010.
[5] Y. J. Lee, Y. R. Yeh, and Y. C. F. Wang, "Anomaly detection via online over-sampling principal component analysis," IEEE Transaction. On Knowledge and Data Engineering, vol. 25, no. 7, pp. 1460–1470, 2012.
[6] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," in Proc. ICML, San Francisco, CA, USA, pp. 255–262, 2000.
[7] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Statistical outlier detection using direct density ratio estimation," Knowledge and Information System, vol. 26, no. 2, pp.309–336, 2011.
[8] R. Smith, A. Bivens, M. Embrechts, C. Palagiri, and B. Szymanski,"Clustering approaches for anomaly based intrusion detection,"in Proc. Intell. Engineering System Artificial Neural Network ,pp. 579–584, 2002
[9] Y. Shi and L. Zhang, "COID: A cluster-outlier iterative detection approach to multi-dimensional data analysis," Knowledge and Information System, vol. 28, no. 3, pp. 709–733, 2011.
[10] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, "Svdd-based outlier detection onuncertain data," Knowledge and Information System, vol. 34, no. 3, pp. 597–618, 2013.
[11] C. Elkan, "The foundations of cost-sensitive learning," in Proc.IJCAI, San Francisco, CA, USA, pp. 973–978, 2001.
[12] M. V. Joshi and V. Kumar, "CREDOS: Classification using ripple down structure (a case for rare classes)," in Proc. SIAM Conf. Data Mining, 2004.
[13] P. Chan and S. Stolfo, "Toward scalable learning with non uniform class and cost distributions," in Proc. ACM SIGKDD International Conferences, KDD, pp. 164–168, 1998.
[14] G. Nakhaeizadeh, U. Knoll, and B. Tausend, "Cost-sensitive pruning of decision trees," in Proc. ECML, Catania, Italy, pp. 383–386, 1994.
[15] G. Fumera and F. Roli, "Cost-sensitive learning in support vector machines," in Proc Workshop Mach. Learn. Meth,2002.
[16] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," Knowledge and Information System, vol. 25, no. 1,pp. 1–20, 2010.
[17] C. X. Ling, J. Huang, and H. Zhang, "AUC: A statistically consistent and more discriminating measure than accuracy," in Proc. IJCAI, San Francisco, CA, USA, pp. 519–526, 2003.
[18] K. Bhaduri, B. L. Matthews, and C. Giannella, "Algorithms for speeding up distance-based outlier detection," in Proc. ACM SIGKDD International Conferences KDD, New York, NY, USA, pp. 859–867, 2011.
[19] Warusia Yassin, Nur Izura Udzir1, Zaiton Muda, and Md. Nasir Sulaiman," Anomaly- Based Intrusion Detection Through K-Means Clustering and Naives Bayes Classification," in Proc. ICOCI International Conferences, Sarawak, Malaysia, pp. 298-303, 2013
[20] E. M. Jordaan and G. F. Smits, "Robust outlier detection using SVM regression," in Proc. IJCNN, pp. 1098–1105, 2004
[21] Mohiuddin Ahmed, Abdun Naser, "A Novel Approach for Outlier Detection and Clustering Improvement" IEEE 8th Conference on Industrial Electronics and Applications (ICIEA),pp. 577-582, 2013.
[22] Varun Chandola, Arindam Banerjee, Vipin Kumar," Anomaly Detection for Discrete Sequences: A Survey", IEEE Transaction on Knowledge and Data Engineering, vol. 24, no. 5, pp. 823-839, 2012.
[23] Hasan Ferdowsi, Sarangapani Jagannathan, and Maciej Zawodniok," An Online Outlier Identification and Removal Scheme for Improving Fault Detection Performance", IEEE Transaction on Knowledge and Data Engineering, vol. 25, no. 5, pp. 908-919, 2014.
[24] Rikard Laxhammar and Göran Falkman," Online Learning and Sequential Anomaly Detection in Trajectories", IEEE Transaction on Pattern Analysis, Knowledge ,Data Engineering.,Vol.36,no.6,pp.1158-1173, 2014.