Literature Survey on Outlier Detection Techniques For Imperfect Data Labels

Priyanka W. Meshram[1], Prof. Sapna Khapre[2]

[1]Department of Computer Science and Engineering, G. H. Raisoni College of Engineering, Nagpur, Maharashtra, India

[2]Department of Computer Science and Engineering, G. H. Raisoni College of Engineering, Nagpur, Maharashtra, India

Abstract: ***A dataset may contain objects that do not comply with the general behaviour or model of data .These data objects are outlier. Outlier detection has attracted increasing attention in machine learning, data mining and and statistics literature. A well-known definition of "outlier" is given as "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 Common general techniques for data classification include both unsupervised and supervised pattern classification methods. Some common approaches use clustering instead of simple feature selection, linear discriminant methods,neural networks and support vector machines Feature selection forms an important subset within the much larger area of data classification. Correctly identifying the relevant features in a data is of vital importance to the task of text classification. Our objective would be to actively select instances with higher probabilities to be informative in determining feature relevance so as to improve the performance of feature selection without increasing the number of sampled instances. Active sampling used in active feature selection chooses instances in two steps: first, it partitions the data according to some homogeneity criterion; and second, it randomly selects instances from these partitions.***

Keywords: Outlier Detection, Data of uncertainty, Feature Selection

1. Introduction

As the dataset contain the objects ,the task of outlier detection is to identify data object that are marketdly different from or imperfect labels, introduce likelihood values for each input data which denote the degree of membership of an example toward the normal and abnormal classes respectively. 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-based, clustering-based, density-based and model-based approaches, all of them with long history. This paper presents a novel outlier detection approach to address data with imperfect labels and incorporate limited abnormal examples into learning. Our main objectives are

- Make a research work on the existing algorithm for outlier detection with imperfect data labels
- Designing a unique and effective algorithm
- Testing the algorithm on different data labels
- The proposed scheme overcomes the drawbacks of existing scheme such as inefficiency and inaccuracy. It provides less search time and high retrieval accuracy.

2. Related Work

A. Efficient approach for outlier detection with imperfect data labels

The detail review of efficient approach for outlier detection has been given by Bo Liu, Yanshan Xiao, Philip S. Yu, Zhifeng Hao, and Longbing Cao. There are different approaches for outlier detection:

1) Support vector data iteration

The support vector data description (SVDD) has been proposed for one-class classification learning. Given a setof target data {**x***i*}, *i* = 1, . , *l*, where **x***i* ∈ *Rm*, the basic idea of SVDD is to find a minimum hyper-sphere that contains most of target data in the feature space, as illustrated in fig.1

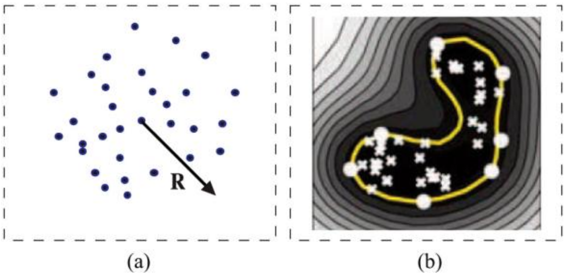

Figure 1: (a) Illustration of SVDD hyper-sphere in feature space.(b) Illustration of SVDD decision boundary in input space.

$$\text{Min } F(R,o,\xi i) = R^2 + C\sum_{i=1}^{l} \xi i\ ,$$

$$\text{s.t. } \| \phi(xi) - o \|^2 <= R^2 + \xi i,$$

$$\xi i \geq 0, (1)$$

where *φ(.)* is a mapping function which maps the input data from input space into a feature space, and *φ(*x*i)* is the image of **x***i* in the feature space, *ξi* are slack variables to allow some data points to lie outside the sphere, and *C* > 0 controls the tradeoff between the volume of the sphere and the number of errors.$\sum_{i=1}^{l} \xi i$ is the penalty for misclassified Samples.

By introducing Lagrange multipliers *αi*, the optimization problem (1) is transformed into:

$$\max \sum_{i=1}^{l} \alpha i K(xi, xi) - \sum_{i=1}^{l} \sum_{k=1}^{l} \alpha i \alpha\, k K(xi, xk)$$

s.t. 0<=αi<=C,

$$\textstyle\sum_i \alpha i = 1 \qquad (2)$$

in which kernel function *K(, .,)* is utilized to calculate the inner pairwise product of two vector *φ(***xi***)* and *φ(***xj***)*, that is *K(***xi**, **x***j) = φ(***xi***) ·φ(***x***j)*. The samples with *αi* > 0 are support vectors (SVs). For a test point **x**, it is classified as normal data when this distance is less than or equal to the radius *R*. Otherwise, it is flagged as an outlier

2) Kernel k-Means clustering-based method

We adopt the kernel *k*-means clustering algorithm to generate likelihood values for each input data. In kernel-based method, a nonlinear mapping function *φ(.)* maps the input samples into a feature space. Kernel *k*-means clustering minimizes the following objective function

$$J = \textstyle\sum_{i=1}^{k} \sum_{j=1}^{l+n} ||\emptyset(xj) - \emptyset(vi)||^{2}, \qquad (3)$$

where *k* is the number of clusters and **v***i* is the cluster center of the *ith* cluster. By solving this optimization problem, *k*-means clustering returns a set of local clusters, in which data samples belonging to a same cluster are more similar to each other. Intuitively, for a data sample, if most of data samples in the same cluster are normal, it would have a high probability of being normal, and if there is an outlying point that does not belong to any cluster, it would have a high probability of being an outlier. Therefore, we calculate the likelihood values for single likelihood model and bi-likelihood model as follows. For a given cluster *j*, assume there exist l^{p}_{j} normal examples and l^{n}_{j} negative examples.

3) Kernel LOF-based method

To cope with datasets with varying densities, we propose a local density-based method to compute likelihood values for each input data. Inspired by the LOF algorithm , the basic idea is to examine the relative distance of a point to its local neighbors in feature space. More specifically, we extend the original LOF into the kernel space by using kernel function and generate the likelihood values in the kernel space instead of the input space

B. Anomaly detection via online oversampling principal component analysis

The detail review of Anomaly detection via online oversampling principal component analysis has been given by Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang, Member

Anomaly (or outlier) detection aims to identify a small group of instances which deviate remarkably from the existing data. 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 .However, since only a limited amount of labeled data are available in the above real world applications, how to determine anomaly of unseen data (or events) draws attention from the researchers in data mining and machine learning communities

1) Anomaly detection via PCA

PCA is a well known unsupervised dimension reduction method, which determines the principal directions of the data distribution. To obtain these principal directions, one needs to construct the data covariance matrix and calculate its dominant eigenvectors. These eigen vectors will be the most informative among the vectors in the original data space, and are thus considered as the principal directions. Let A ¼ ½x>1 ; x>2 ; . . . ; x>n _ 2 IRn_p, where each row xi represents a data instance in a p dimensional space, and n is the number of the instances. Typically, PCA is formulated as the following optimization problem

$$\text{Max } \textstyle\sum_{i=1}^{n} U^{T}(xi\text{-}\mu)(xi\text{-}\mu)^{T}U, \ (1)$$

$U\varepsilon\ IR^{pxk}, ||U||=I$

where U is a matrix consisting of k dominant eigenvectors. From this formulation, one can see that the standard PCA can be viewed as a task of determining a subspace where the projected data has the largest variation. Alternatively, one can approach the PCA problem as minimizing the data reconstruction error, i.e.

$$\text{Min } J(U) = \textstyle\sum_{I=1}^{n} ||(xi - \mu) - UU^{T}(xi\text{-}\mu)||^{2}$$

$$U\varepsilon\ IR^{pxk}, ||U||=I \ (2)$$

where U>ðxi _ _Þ determines the optimal coefficients to weight each principal directions when reconstructing the approximated version of (xi _ _). Generally, the problem in either (1) or (2) can be solved by deriving an eigen value decomposition problem of the covariance data matrix, i.e.

$$\textstyle\sum_A U = U\Lambda, \qquad (3)$$

Where

$$\textstyle\sum_A = 1/n \sum_{i=1}^{n} (xi - \text{⍰})(xi - \text{⍰})^{T} \qquad (4)$$

is the covariance matrix, _ is the global mean. Each column of U represents an eigenvector of _A, and thecorresponding diagonal entry in _ is the associated eigenvalue. For the purpose of dimension reduction, the last few eigenvectors will be discarded due to their negligible contribution to the data distribution.While PCA requires the calculation of global mean and data covariance matrix, we found that both of them are sensitive to the presence of outliers. As shown , if there are outliers present in the data, dominant eigenvectors produced by PCA will be remarkably affected by them, and thus this will produce a significant variation of the resulting principal directions.

2) Oversampling PCA for anomaly detection

For practical anomaly detection problems, the size of the data set is typically large, and thus it might not be easy to observe the variation of principal directions caused by the presence of a single outlier. Furthermore, in the above PCA framework for anomaly detection, we need to perform n PCA analysis for a data set with n data instances in a p-dimensional space, which is not computationally feasible for large-scale and online problems. Our proposed oversampling

PCA (osPCA) together with an online updating strategy will address the above issues, as we now discuss

2.1 Oversampling Principal Components Analysis (osPCA)

As mentioned earlier, when the size of the data set is large, adding (or removing) a single outlier instance will not significantly affect the resulting principal direction of the data. Therefore, we advance the oversampling strategy and present an oversampling PCA (osPCA) algorithm for largescale anomaly detection problems. The proposed osPCA scheme will duplicate the target instance multiple times, and the idea is to amplify the effect of outlier rather than that of normal data. While it might not be sufficient to perform anomaly detection simply based on the most dominant eigenvector and ignore the remaining ones, our online osPCA method aims to efficiently determine the anomaly of each target instance without sacrificing computation and memory efficiency. More specifically, if the target instance is an outlier, this oversampling scheme allows us to overemphasize its effect on the most dominant eigenvector, and thus we can focus on extracting and approximating the dominant principal direction in an online fashion, instead of calculating multiple eigen vectors carefully.

3) Online anomaly detection for practical scenario

For online anomaly detection applications such as spam mail filtering, one typically designs an initial classifier using the training normal data, and this classifier is updated by the newly received normal or outlier data accordingly However, in practical scenarios, even the training normal data collected in advance can be contaminated by noise or incorrect data labeling. The flowchart of our online detection procedure is shown in Fig. 2. As can be seen in Fig. 2, there are two phases required in this framework: Data cleaning and online detection. In the data cleaning phase, our goal is to filter out the most deviated data using our osPCA before performing online anomaly detection. This data cleaning phase is done offline, and the percentage of the training normal data to be disregarded can be determined by the user. In our implementation, we choose to disregard 5 percent of the training normal data after this data cleaning process, and we use the smallest score of outlierness (i.e., st) of the remaining training data instances as the threshold for outlier detection. More specifically, in the second phase of online detection, we use this threshold to determine the anomaly of each received data point. If st of a newly received data instance is above the threshold, it will be identified as an outlier; otherwise, it will be considered as a normal data point, and we will update our osPCA model accordingly. In the online detection phase, we use the dominant principal direction of the filtered training normal data.

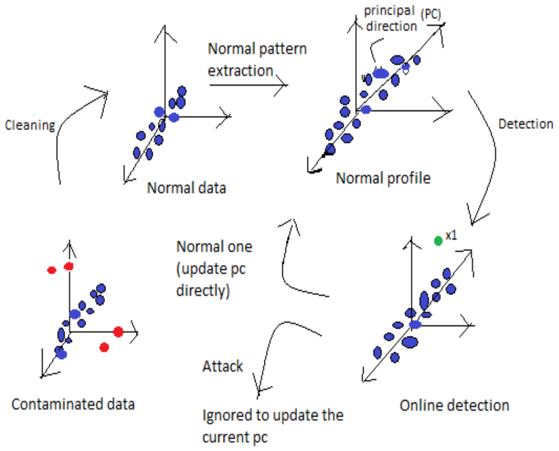

Figure 2: The framework of our online anomaly detection

C. A survey of uncertain data algorithms and applications

The detail description of a survey of uncertain dataalgorithms and application has been given by C. C. Aggarwal and P. S. Yu. A number of indirect data collection methodologies have led to the proliferation of uncertain data. Such databases are much more complex because of the additional challenges of representing the probabilistic information. In this paper, we provide a survey of uncertain data mining and management applications. It includes following steps:

- Modeling of uncertain data

A key issue is the process of modeling the uncertain data. Therefore, the underlying complexities can be captured while keeping the data useful for database management applications.

- Uncertain data management

In this case, one wishes to adapt traditional database management techniques for uncertain data. Examples of such techniques could be join processing, query processing, indexing, or database integration.

- Uncertain data mining

The results of data mining applications are affected by the underlying uncertainty in the data. Therefore, it is critical to design data mining techniques that can take such uncertainty into account during the computation.

1) Clustering uncertain data

The presence of uncertainty changes the nature of theunderlying clusters, since it affects the distance functioncomputations between different data points. A technique has been proposed . in order to find density-based clusters from uncertain data. The key idea in this approach is to compute uncertain distances effectively between objects which are probabilistically specified. The fuzzy distance is defined in terms of the distance distribution function. This distance distribution function encodes the probability that the distances between two uncertain objects lie within a certain user-defined range. In the deterministic version of the algorithm , data points are grouped into clusters when they are reachable from one another by a path which is such that every point on this path has a minimum threshold data density. To this effect, the algorithm uses the condition that the neighborhood of a data point should contain at least Min_Pts data points. The algorithm starts off at a given data point and checks if the neighborhood containsMin_Pts data points. If this is the case, the algorithm repeats the process for each point in this cluster and keeps adding points until no more points can be added. One can plot the density profile of a data set by plotting the number of data points in the neighborhood of various regions, and plotting a smoothed version of the curve

2) classification of uncertain data

A closely related problem is that of classification of uncertain data in which the aim is to classify a test instance into one particular label from a set of class labels. A method was proposed for support vector machine classification of uncertain data. This technique is based on a discriminative modeling approach which relies on a total least squares method.

3) Frequent Pattern Mining

The problem of frequent pattern mining has also been explored in the context of uncertain data. In this model, it is assumed that each item has an existential uncertainty in belonging to a transaction. This means that the probability of an item belonging to a particular transaction is modeled in this approach. In this case, an item set is defined to be frequent, if its expected support is at least equal to a userspecified threshold. In order to solve this version of the frequent pattern mining problem, the U-Apriori algorithm is proposed which essentially mimics the Apriori algorithm, except that it performs the counting by computing the expected support of the different item sets

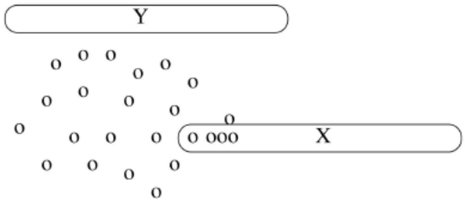

Figure 3: Effect of uncertainty on outlier detection.

4) Outlier detection with uncertain data

The problem of outlier detection has also been extended to the case of uncertain data. In the case of the outlier detection problem, differing levels of uncertainty across different dimensions may affect the determination of the outliers in the underlying data. For example, consider the case in Fig. 3.3, in which the contours of uncertainty for two data points X and Y are illustrated in the form of elliptical shapes. The data point X seems to be further away from the overall data distribution as compared to the data point Y . However, the contours of uncertainty are such that the concept of an outlier in terms of the probability that a given data point is drawn from a dense region of the overall data distribution.

D. Clustering-based outlier detection method

The detail description of clustering-based outlier detection method has been given by S. Y. Jiang and Q. B. An.

In this paper,the method consists of two stages, the first stage cluster dataset by one-pass clustering algorithm and second stage determine outlier cluster by outlier factor .the time complexity of CBOD is nearly linear with the size of dataset and the number of attributes, which results in good scalability and adapts to large dataset. The theoretic analysis and the experimental results show that the detection method is effective and practicable

1) One-Pass Clustering Algorithm

The goal of clustering is that the intra-cluster similarity is maximized while the inter-cluster similarity is minimized. Many efficient clustering algorithms have been proposed by the database research community. Clustering algorithm can be selected according to data, objective of clustering and application. In this paper, we use one-pass clustering algorithm divide dataset into hyper spheres with almost the same radius. The algorithm is described as follows.

Step 1: Initialize the set of clusters, S, as the empty set, read a new object *p*.
Step 2: Create a cluster with the object *p*.
Step 3: If no objects are left in the database, go to step 6, otherwise read a new object *p*, and find the cluster *C** in S that is closest to the object *p*. In otherwords, find a cluster *C** in *S*, such that for all *C* in S,*d*(*p*,*C**) ≤□*d*(*p*,*C*) .
Step 4: If *d*(*p*,*C**) □ □*r* , go to step 2.
Step 5: Merge object p into cluster *C** and modify the *CSI* of cluster *C** , go to step 3.
Step 6: Stop

2) Outlier Detection Method

On the basis of the outlier factor of cluster, we present a clustering-based outlier detection method (*CBOD*), which consists of two stages. Stage 1. Clustering: Cluster on data set *D* and produce clustering results C={C1,C2,.. *C*k}. Stage 2. Determining Outlier Clusters: Compute outlier factor *OF*(*Ci*)(1≤□*i* ≤□*k*) , sort clusters { , , , } 1 2 *k* *C* □ □*C* *C* □ *C* and make them satisfy: OF (C1)>=OF (C2)>=…>=OF (Ck) . Search the minimum*b* , which satisfies $\frac{\sum_{i=1}^{b}|Ci|}{|D|}$ >=1 ≥□ (0□ □ □ 1)finally, label clusters *C 1*,*C2* ,.. ,*C* b with 'outlier' (any object belonged to outlier class is regarded as outlier), while *Cb+1* ,*Cb+2* ,.. ,*C* □ □ with 'normal'. (Any object belonged to normal class is regarded as normal).

E. A survey of outlier detection methodologies

The detail review of a has been given survey of outlier detection methodologies has been given by V. J. Hodge and J. Austin, Artif. Intell. In this paper, introduce a survey of contemporary techniques for outlier detection. Identify their respective motivations and distinguish their advantages and disadvantages in a comparative review. In this paper chosen to call the technique outlier detection although we also use novelty detection.

1) Statistical models

Statistical approaches were the earliest algorithms used for outlier detection.Some of the earliest are applicable only for single dimensional data sets. Statistical models are generally suited to quantitative real valued data sets or at the very least quantitative ordinal data distributions where the ordinal data can be transformed to suitable numerical values for statistical (numerical)processing. This limits their applicability and increases the processing time if complex data transformations are necessary before processing.

1.1 Proximity-based

TechniquesProximity-based techniques are simple to implement and make no prior assumptions about the data distribution model. However, they suffer exponential computational growth as they are founded on the calculation of the distances between all records. The computational complexity is directly proportional to both the dimensionality of the data *m* and the number of records *n*. Hence, methods such as k-nearest neighbour with $O(n2m)$ runtime are not feasible for high dimensionality data sets unless the running time can be improved. There are various flavours of k-Nearest Neighbour (k-NN) algorithm for outlier detection but all calculate the nearest neighbours of a record using a suitable distance calculation metric such as Euclidean distance or Mahalanobis distance. Euclidean distance is given by equation .

1.2 Parametric Methods

Parametric methods allow the model to be evaluated very rapidly for new instances and are suitable for large data sets; the model grows only with model complexity not data size. However, they limit their applicability by enforcing a pre-selected distribution model to fit the data. If the user knows their data fits such a distribution model then these approaches are highly accurate but many data sets do not fit one particular model.

1.3. Non-Parametric Methods

Other techniques such as those based around convex hulls and regression and the PCA approaches assume the data follows a specific model. These all require *a priori* data knowledge. Such information is often not available or is expensive to compute. Many data sets simply do not follow one specific distribution model and are often randomly distributed. Hence, these approaches may be applicable for an outlier detector where all data is accumulated beforehand and may be pre-processed to determine parameter settings or for data where the distribution model is known. Non parametric approaches, in contrast are more flexible and autonomous

1.4. Semi-Parametric Methods

Semi-parametric methods apply local kernel models rather than a single global distribution model. They aim to combine the speed and complexity growth advantage of parametric methods with the model flexibility of non-parametric methods. Kernel-based methods estimate the density distribution of the input space and identify outliers as lying in regions of low density. Tarassenko & Robert and Bishop use Gaussian Mixture Models to learn a model of normal data by incrementally learning new exemplars. The GMM is represented by equation

$$p(t|x) = \sum_{j=1}^{M} \alpha j(x)\phi j(t|x)$$

3. Conclusion

The main objective is to demonstrate the data density by using Gaussian method.minimizing the noise ratio and searching time, by these we provide enhancement of the performance i.,e efficiency with the real time datasets also comparative study of results on the basis of graphs.In future work,hybrid algorithms applying for outlier detection.BR tree methods using for indexing providing microclassification of data and normalise the 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 Trans.Knowl. Data Eng., Vol. 26, No. 7, July 2014

[2] Y. J. Lee, Y. R. Yeh, and Y. C. F. Wang, “Anomaly detection via online over-sampling principal component analysis,” EEE Trans.Knowl. Data Eng.,vol. 25, no.7, pp. 1460–1470, May 2012.

[3] [3] C. C. Aggarwal and P. S. Yu, “A survey of uncertain data algorithms and applications,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 5, pp. 609–623, May 2009

[4] [4] S.Y.Jiang and Q. B. An,“Clustering-based outlier detection method,” in Proc.ICFSKD, Shandong, China, 2008, pp. 429–433.

[5] [5] V.J.Hodge and J. Austin, “A survey of outlier detection methodologies,”Artif Intell. Rev. vol . 22, no. 3, pp, 85–126, 2004.

[6] [6] D. M. J. Tax and R. P. W. Duin, “Support vector data description,”Mach. Learn., vol. 54, no. 1, pp. 45–66, 2004