An Efficient Approach for Mining of Outliers for Imperfectly Labeled Data

Dighavkar Aditi C. **Prof. N.M.Shahane**

Abstract **—Detection of outliers is a significant problem in the context of data analysis. Many of these existing techniques assume that the instance can be classified as either normal or abnormal class. But, many real-life applications contain data which are uncertain in nature because of many errors or partial completeness. This uncertainty in data gives rise to imperfect labelling of the data. Thus, important challenge lies in handling the imperfectly labelled data to nullify the effect of uncertain data on the classifier. This paper proposes a modified Support Vector Data Description based approach to detect outliers for uncertain data. The proposed system executes in two steps. The first step deals with generating the pseudo-training dataset by calculating the likelihood values. The second step deals with incorporating the generated confidence score into the support vector data description training phase to generate a global classifier. Here, the effect of the instances with the least confidence score on the construction of the decision boundary is reduced.**

Key Words **— Anomaly Detection; Likelihood Values; Support vector data description**

I. INTRODUCTION

Outlier detection refers to the problem of detecting and analyzing patterns in data that does not map to expected normal behavior. These patterns are often referred to as outliers, anomalies, surprise, exceptions, noise, defects, errors, damage, faults, aberrations, discordant observations, novelty, contaminants or peculiarities in different application domain [2]. Outlier detection has always been a widely researched problem. It finds immense use in a wide variety of application domains such as insurance, tax, military surveillance for enemy activities, credit card, fraud detection, fault detection in safety critical systems, intrusion detection for cyber security and many other areas [2]. Traditional outlier detection techniques typically assume that outliers are difficult or costly to obtain because of their rare occurrence. Hence, most of the previous techniques focus on modeling a representation of the normal data in order to identify outliers that do not fit the model. These outlier detection algorithms are mainly classified into four basic categories. Firstly, in Statistics-based algorithms [3], statistical techniques fit a statistical model to the given data and then apply a statistical inference test in order to determine if an incoming instance fits the model or not. Secondly, in Density based method [4] local outliers are identified by analyzing the distances to their nearest neighbors. Thirdly, Clustering based approaches [5] groups identical data instances into same clusters and considers clusters of smaller size as outliers. Fourthly, Model-based method [6] generates distinctive model from a set of training data instances to detect outliers which deviates from the model. In this category, Support Vector Data Descriptor (SVDD) [1, 7], determine a sphere around normal data with minimum volume. Another vital observation is that, data are uncertain in nature for many real-life applications. For example, the data points may correspond to objects, which are only vaguely specified due to data incompleteness, and are therefore considered uncertain in their representation; moreover, some new hardware technologies such as sensors usually collect large amounts of uncertain data due to sampling errors or instrument imperfections. Consequently, a labeled normal example corrupted by various errors or limitations of the underlying equipment always behaves like an outlier, even though the example itself may not be an outlier. This always makes the problem of outlier detection far more difficult from the perspective of data uncertainty. Therefore, it is worthwhile to develop techniques to refine the decision boundary of the distinctive classifier so as to improve the performance of outlier detection. The key challenge of handling of data with uncertainty in outlier detection is how to minimize the impact of this uncertain data on the learned distinctive classifier.

The remainder of paper is organized as follows: Section 2 discuss the related work done and its shortcoming. An overview of the proposed scheme is given in section 3. Section 4 discusses the expected results of the system. Finally, we conclude the paper.

II. LITERATURE SURVEY

A. Outlier Detection

Outlier detection deals with the problem of finding patterns in large data that do not conform to expected pattern or behavior. Lot of work has been done in this area which can be classified into the following four categories:

1) Statistics-based approaches: This tries to fit a statistical model to the given data and then it apply a statistical inference test in order to determine whether an unseen instance or record satisfies this model or not. The instances which have less probability of being generated from the learned model are declared as outliers. For example, we can assume the normal instance follow a certain data distribution like Gaussian distribution [8][9]. For this category, the main disadvantage lies in the assumption that the data is generated from a particular distribution.

2) Density-based approaches: Density-based approaches assume that normal data instances occur in dense neighborhoods, whereas outliers occur far from their closest neighbors [11]. One of the important methods is called LOF (local outlier factor)

[12][13] . LOF method assigns an outlier score or local outlier factor to any given data point, depending on its distances in the local neighborhood. But, in the case where data has normal instances with no enough close neighbors or if the data has outliers with enough close neighbors then the technique fails to label them correctly which further results in incorrect outliers.

3) Clustering-based approaches: Clustering-based methods [14, 15] is one of the highly researched method stream. After clustering, small clusters containing significantly fewer data points than other clusters are deemed as outliers. For clustering based techniques, the performance is highly dependent on the effectiveness of the clustering algorithm in the way of capturing the cluster structure of normal instances.

4) Model-based approaches: Model-based techniques are used to learn a model i.e. a classifier from a set of labeled data instances i.e. training and then to classify a test instance into one of the classes using the learnt model i.e. testing [16]. Model-based outlier detection approach operates in two-phase fashion. The training phase learns a classifier by using the available labeled training data. The testing phase classifies a test instance as normal or outlier using the classifier. In this approach, SVDD proposed by [17-21], has demonstrated experimentally to be capable of mining outliers in various domains. Model based techniques can find global outliers effectively for higher dimensional data without need to assume the prior distribution of data.

The main drawback of the previous work is that they assume that an input data sample either completely belong to normal class or outlier class. But, this is not appropriate for uncertain data. The problem of uncertainty in labels of the learning data can dramatically degrade the classification performance. It is because of the importance of labeled data in the learning procedure where a classifier tries to approximately fit to the training data and make decision about the class of a test sample based on the characteristics of the learned data. When the data labels have been imperfectly assigned to the learning samples, the optimal performance of the classifier on the test data is not expected even when it is strong enough to properly learn specifications of the training data. For instance, a labeled normal example which is corrupted by various errors or limitations of the underlying equipment always behaves like an outlier, even if the example itself may not be an outlier.

Hence, the key challenge of handling data with uncertainty in outlier detection is to reduce the impact of this uncertain data on the learned distinctive classifier. Most of the previous work is performed on the uncertain data; the decision boundary of these methods was impacted by the data containing uncertain information. Because of this consequently, performance is reduced.

B. Uncertain Data

Various techniques have been proposed in order to handle the uncertainty of data in query processing of uncertain data, indexing uncertain data, clustering uncertain data, classification of uncertain data, frequent pattern mining of uncertain data. Work in [22] considers uncertainty in data in the outlier detection problem where a probabilistic definition of outliers in conjunction with density calculation and sampling are used. Different from this work, to handle the problem of outlier detection in the presence of uncertain data, the proposed system proposes a model-based approach by introducing a likelihood value for each input data point into the SVDD training phase. The proposed approach operates in two important steps. In the first step, pseudo-training dataset is generated by assigning a confidence level to each input record, which indicates the likelihood of an input data record belonging to normal class. For this kernel-based clustering method is used to generate the confidence level for each input training record. In the second step, generated confidence score for each sample is incorporated into the SVDD training process. Due to introduction of confidence score into the training stage, each data record contributes differently to the generation of the decision boundary, which then is used for outlier detection.

III. PROPOSED WORK

A. *Proposed System:*

The proposed system is a systematic approach for outlier detection which consists of a classifier which is built by considering both normal and abnormal training data. The problem of uncertainty as discussed in previous section is solved by using likelihood values for each record in the dataset. For this, two likelihood models are generated i.e. single and bi likelihood model. For both likelihood models, pseudo training dataset is generated. These two pseudo training datasets are then used for training the support vector data description classifier separately. The contribution of the paper lies in selection of appropriate kernel parameters. Also the proposed system will deal with refinement of the support vector data descriptor classifier boundary. The proposed system architecture is as shown in the Fig.1

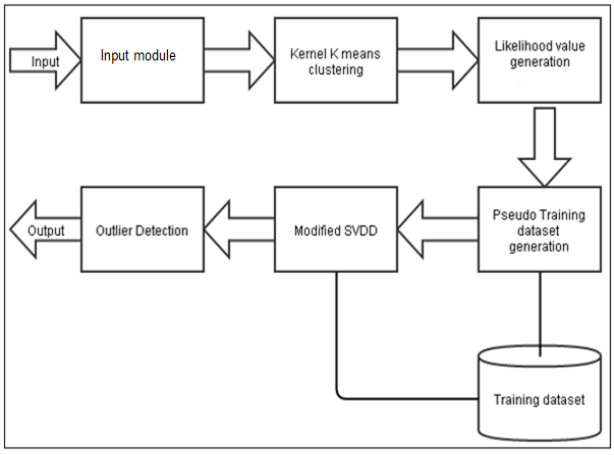

Fig. 1. Block diagram of proposed system

Following are the modules of the proposed system:

Input Module

Input module includes setting the environment for running the proposed system. This includes initialising the dataset for processing. The module sets the variables and parameters for the

system as per the dataset. Each instance in the d-dataset is represented as $X = x_1,x_2,x_3, .., x_d$, where $x_1,x_2,x_3, .., x_d$ are the features of the dataset.

Kernel k-means clustering

In kernel k-means clustering algorithm likelihood, a nonlinear mapping function Φ(.) is used to map the input sample into a feature space. For this it needs to minimize the objective function (1):

$$J= \sum_{i=1}^{k} \sum_{j=1}^{m} \| \Phi(x_j) - \Phi(v_i) \|^2 \quad (1)$$

k: Number of clusters
v_i: cluster center of i^{th} cluster
m=l+n
l: Total number of normal examples
n: Total number of abnormal examples

By solving this optimization problem, it returns set of local clusters. Thus, kernel k-means clustering algorithm takes input as the set of data points and the count of number of clusters. It randomly initializes 'k' cluster center and compute the distance of each data label point and the cluster center in the transformation space. It then assigns a data point to that cluster whose center distance is minimum. These steps are repeated till data points are reassigned. For a given cluster j, assume that there exist l^p_j normal examples and l^n_j negative examples.

Likelihood value generation function

Single likelihood model:

In this model, each input is associated with a likelihood value $(x_i,m(x_i))$, which represents degree of membership of an example towards its own class label.

Bi-likelihood model:

In the model, each input is associate with bi-likelihood values, denoted as $(x_i,m^t(x_i),m^n(x_i))$, in which $m^t(x_i)$ and $m^n(x_i)$ indicate the degree of an input data xi belonging to the positive class and negative class respectively.

To calculate the degree of membership value:

For the *single likelihood model,*

$m^t(x_t) = l^p_j /(l^p_j + l^n_j)$ where x_t belongs to the normal class

$m^n(x_k) = l^n_j/(l^p_j + l^n_j)$ where x_k belongs to the negative class

For the *bi-likelihood model,*

$m^t(x_t) = l^p_j /(l^p_j + l^n_j)$ where x_t belongs to the normal class

$m^n(x_t) = l^n_j/(l^p_j + l^n_j)$

Pseudo training dataset generation

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:

$(x_1,m_t(x_1)),\dots , (x_l,m_t(x_l)), (x_{l+1},m_n(x_{l+1})),\dots ,(x_{l+n},m_n(x_{l+n}))$

where,

$m_t(x_i)$: likelihood of example x_i belonging to the normal class

$m_n(x_i)$: likelihood of example x_i belonging to the abnormal class

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

$(x_1,m_t(x_1),m_n(x_1)),\dots,(x_l,m_t(x_l),m_n(x_l)),(x_{l+1},m_t(x_{l+1}),m_n(x_{l+1})), \dots , (x_{l+n},m_t(x_{l+n}),m_n(x_{l+n}))$

Modified SVDD

This module is a SVDD classifier which is trained by the pseudo training dataset. SVDD maps the input sample vectors to the kernel space and find a hyper sphere class boundary with the minimum volume as shown in Fig. 2. Modified SVDD is the proposed modified version of SVDD which improves the performance of SVDD by kernel parameter selection and refinement of hyper-sphere class boundary.

For single likelihood model,

Step 1. All positive examples are put into S_p set and all negative examples are put into S_n set. Thus, $m^t(x_i)$ is degree of membership associated with normal example and $m^n(x_j)$ is degree of membership associated with abnormal example.

Step 2. In order to obtain this hypersphere we need to minimize the objective function (2):

$$\text{Min } F = R^2 + C_1 \sum m^t(x_i)\, \xi_i + C_2\, m^n(x_j)\, \xi_j \quad (2)$$

Such that,

$\| x_i - o \|^2 \le R^2 + \xi_i$, x_i belong to S_p

$\| x_j - o \|^2 \ge R^2 - \xi_i$, x_j belong to S_n

$\xi_i \ge 0, \xi_j \ge 0$

Where,

R: Radius of the hypersphere

O: Center of the hypersphere

C_1, C_2: Parameters that control the tradeoff between the sphere volume and errors

ξ_i , ξ_j : Slack variables

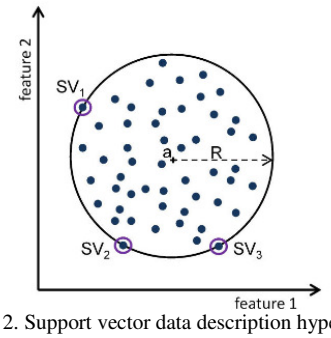

Fig. 2. Support vector data description hypersphere

Step 3. Use Lagranges method for solving above optimization problem, which includes maximizing the function (3):

$$\text{Max} \sum_{i=1}^{l+n} \alpha_i\, K(x_i , x_i) - \sum_{i=1}^{l+n} \sum_{j=1}^{l+n} \alpha_i \alpha_k\, K(x_i , x_j) \quad (3)$$

Such that, $0 \le \alpha_i \le C_i^m$, i = 1,2,… , l + n

$$\sum_{i=1}^{l+n} \alpha_i = 1$$

Step 4. After solving the above problem, we obtain Lagrange multiplier α_i which gives the centroid of the hypersphere by equation (4):

$$o = \sum_{i=1}^{l+n} \alpha_i \Phi(x_i) \quad (4)$$

Step 5. By assuming a point x_u on the surface of hypersphere, R can be calculated by equation (5):

$$R^2 = \|x_u - o\|^2$$
$$= K(x_u, x_u) + K(o,o) - 2K(x_u,o)$$
$$= K(x_u, x_u) + \sum_{i=1}^{l+n}\sum_{j=1}^{l+n} \alpha_i\alpha_k K(x_i, x_j) - 2\sum_{i=1}^{l+n} \alpha_i(x_i, x_u) \quad (5)$$

Step 6. To classify x, if the distance is less than or equal to R, then it is deemed as normal data as shown in equation (6)

$$\| x - o \|^2 \le R^2 \quad (6)$$

For bi likelihood model,

Step 1. All examples that totally belong to positive class are put into S_p set, i.e. whose $m^t(x_i) = 1$ and $m^n(x_i) = 0$. All examples that totally belong to negative class are put into S_n set i.e. $m^t(x_j) = 0$ and $m^n(x_j) = 1$. All the remaining examples which neither belong to positive nor negative class are put into S_b set i.e. $m^t(x_h) \neq 1$ and $m^n(x_h) \neq 0$. Thus, $m^t(x_i)$ is degree of membership associated with normal example and $m^n(x_j)$ is degree of membership associated with abnormal example.

Step 2. In order to obtain this hypersphere we need to minimize the objective function (7):

$$\text{Min } F= R^2 +C_1(\sum \xi_i + \sum m^t(x_h) \xi_h) +C_2 (\sum \xi_j + \sum m^n(x_k) \xi_k)$$

Such that, (7)

$$\| \Phi(x_i) - o \|^2 \le R^2 + \xi_i, \quad x_i \text{ belong to } S_p$$
$$\| \Phi(x_h) - o \|^2 \le R^2 + \xi_h, \quad x_h \text{ belong to } S_b$$
$$\| \Phi(x_k) - o \|^2 \ge R^2 - \xi_k, \quad x_k \text{ belong to } S_b$$
$$\| \Phi(x_j) - o \|^2 \ge R^2 - \xi_j, \quad x_j \text{ belong to } S_n$$
$$\xi_i \ge 0, \xi_h \ge 0, \xi_k \ge 0, \xi_j \ge 0$$

the meaning of the parameters is same as in single likelihood model.

Step 3. Use Lagranges method for solving above optimization problem, which includes maximizing the function (8):

$$\text{Max } \sum \alpha^t_i K(x_i, x_i) - \sum \alpha^n_j K(x_j, x_j) - \sum\sum \alpha^t_i\alpha^t_k K(x_i, x_k) + 2$$
$$\sum\sum \alpha^t_i\alpha^n_j K(x_i, x_j) - \sum\sum \alpha^n_j\alpha^n_v K(x_i, x_v) \quad (8)$$

Such that, $0 \le \alpha^t_i \le m_i^t(x_i)C_1$

$0 \le \alpha^n_i \le m_i^n(x_i)C_2$

$\sum \alpha^t_i - \sum \alpha^n_j = 1$

$x_i, x_k \in S_p \cup S_b$, $x_i, x_v \in S_b \cup S_n$,

Where $\alpha^t_i \ge 0$, $\alpha^n_i \ge 0$ are Lagrange multipliers.

Step 4. After solving the above problem, we obtain Lagrange multiplier α_i which gives the centroid of the hypersphere by equation (9):

$$o = \sum \alpha^t_i \Phi(x_i) - \sum \alpha^n_j \Phi(x_j) \quad (9)$$

Step 5. By assuming a point x_u on the surface of hypersphere, R can be calculated by equation (10):

$$R^2 = \|x_u - o\|^2 \quad (10)$$
$$= K(x_u, x_u) + \sum\sum \alpha^t_i\alpha^t_k K(x_i, x_k) + \sum\sum \alpha^n_j\alpha^n_v K(x_j, x_v)$$
$$-2\sum\sum \alpha^t_i\alpha^n_j K(x_i, x_j) -2\sum \alpha^t_i K(x_i, x_u) +2\sum\alpha^n_j K(x_j, x_u)$$

Step 6. To classify x, if the distance is less than or equal to R, then it is deemed as normal data as shown in equation (11)

$$\| x - o \|^2 = K(x, x) + \sum\sum \alpha^t_i\alpha^t_k K(x_i, x_k) +$$
$$\sum\sum \alpha^n_j\alpha^n_v K(x_j, x_v) -2\sum\sum \alpha^t_i\alpha^n_j K(x_i, x_j) -2\sum \alpha^t_i K(x_i, x)$$
$$+2\sum\alpha^n_j K(x_j, x) \le R^2 \quad (12)$$

In order to improve the performance of both SVDD classifiers, the proposed system will select the optimized parameters for the kernel function and will also refine the hypersphere boundary for effective and efficient results. SVDD parameters, C and σ massively influence the accuracy of classification. Here,

$$e^{(\|xi-xj\|^2/ \sigma^2)} \quad (13)$$

eq. (13) represents the kernel function which is to be optimized. For this, the best pair {Ci, σi} is determined. The ultimate task is to tune the parameters so that the accuracy on the test set and on unseen data is optimal. The optimization problem eq. (2) and (7) will be solved for a given set of parameters C and σ. Determination of parameters is not straightforward. The regularization parameter C, introduced in eq. (2) and (7), is lower-bound by 1/N, where N is the number of instances in the training data set. C = 1 corresponds to the hard-margin solution, where all instances are enclosed in the decision boundary. So the value range of C is $N \le C \le 1$. The second parameter to be optimized is the kernel width σ. For high values of σ the boundary will become spherical with the risk of under fitting, while for small values of σ a high fraction of instances are selected to be support vectors, hence the boundary is very flexible and is prone to over fitting.

Outlier detection

This module presents the detected outliers by the previous module and also gives the analysis about the outliers detected.

IV. EXPERIMENTAL SETUP

All Experimentation is performed using Pentium processor and 4 GB RAM. The operating system is windows 7(32 bit) with visual studio 10.

A. Dataset:

A large variety of datasets are available for experimentation such as Abalone, Spambase, Thyroid, Waveform, Satellite, Delft, Diabetes, Segment, Letter and Arrhythmia. For experimentation diabetes dataset is used. Diabetes files consist of glucose measurement pre and post of breakfast, lunch and supper. For normal data the pre-meal glucose measurement is 80 to 120 mg/dl and post-meal glucose measurement is 80-140 mg/dl. If it is less than 40 mg/dl then it is considered as hypoglycemic. And if it is more than 200 then it is considered as hyperglycemic. Diabetes dataset contains 768 patient records.

To perform outlier detection with very few abnormal data, we can randomly select 50% of positive data and a small number of abnormal data for training, such that 95 percent of the training data will belong to the positive class and only 5 percent belong to the negative class. All the remaining data will be used for testing. Experiments are designed to perform 10 fold validations and then further perform cross validation.

B.Performance Measure:

In order to evaluate the performance of the system two important performance parameters considered are False Positive Rate and Detection Rate.

This can be depicted as in Table I.

Table I. Confusion Matrix

		ACTUAL LABEL	
		Target Class	Negative Class
PREDICTED LABEL	Target Class	True positive	False negative
	Negative Class	False positive	True negative

False positive rate: The false positive rate is the ratio between the number of queries that are incorrectly misclassified as outliers and the total number of normal data.

Detection rate: The detection rate is the ratio between the number of correctly detected outliers and the total number of outliers. So the overall goal is to minimize the false positive rate and maximize the detection rate.

C.Results:

Initially kernel k means clustering is implemented for diabetes data set. Thus after clustering of the dataset records, k clusters are generated as output. These clusters are generated by considering two important features i.e. pre meal blood glucose measurement and post meal blood glucose measurement. For this, imperfectly labeled dataset is considered as input.

Table II is snapshot of the sample dataset with imperfect labels i.e. data with uncertainty. For example label for patient ID 5 is incorrectly termed as normal and label for patient ID 11 is incorrectly termed as anomaly.

Table II. Sample Imperfectly Labeled Dataset

Patient ID	Pre meal BG	Post meal BG	Imperfect label
1	85	120	Normal
2	90	110	Normal
3	40	70	Anomaly
4	120	220	Anomaly
5	40	80	Normal
6	120	220	Normal
7	40	50	Anomaly
8	10	50	Anomaly
9	90	100	Normal
10	120	140	Normal
11	100	135	Anomaly
12	115	130	Normal
13	166	200	Anomaly

After kernel k-means clustering, clusters are generated as shown in Table III, Table IV and Table V.

Table III. Cluster no. 1

ID	Pre meal BG	Post meal BG	Label	Likelihood value
3	40	70	A	3/4
7	40	50	A	3/4
5	40	80	N	1/4
8	10	50	A	3/4

Table IV. Cluster no. 2

ID	Pre meal BG	Post meal BG	Label	Likelihood value
1	85	120	N	5/6
2	90	110	N	5/6
9	90	100	N	5/6
10	120	140	N	5/6
11	100	135	A	1/6
12	115	130	N	5/6

Table V. Cluster no. 3

ID	Pre meal BG	Post meal BG	Label	Likelihood value
4	120	220	A	2/3
6	120	220	N	1/3
13	166	200	A	2/3

Table VI shows two pseudo training dataset generated for each SVDD classifiers:

Table VI. Pseudo Training Datasets

Single likelihood pseudo training datset			Bi- likelihood pseudo training datset		
Patient ID	$m^l(x_i)$	$m^n(x_i)$	Patient ID	$m^l(x_i)$	$m^n(x_i)$
1	0.833	-	1	0.833	0.167
2	0.833	-	2	0.833	0.167
3	-	0.75	3	0.25	0.75
4	-	0.667	4	0.333	0.667
5	0.25	-	5	0.25	0.75
6	0.33	-	6	0.33	0.667
7	-	0.75	7	0.25	0.75
8	-	0.75	8	0.25	0.75
9	0.833	-	9	0.833	0.167
10	0.833	-	10	0.833	0.167
11	-	0.167	11	0.833	0.167
12	0.833	-	12	0.833	0.167
13	-	0.667	13	0.333	0.667

After testing the dataset against the modified SVDD, we get the set of outliers in the dataset as shown in Table VII.

Table VII. Outliers Detected

Patient ID	Pre meal BG	Post meal BG
3	40	70
4	120	220
6	120	220
7	40	50
8	10	50

Thus, Fig. 3., depicts the increase in the performance of the proposed system with respect to the existing systems. The performance is calculated by considering the tradeoff between detection rate and false alarm rate.

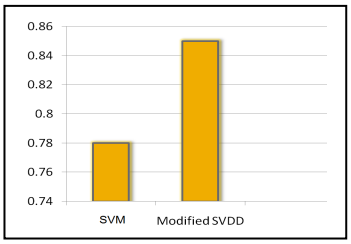

Fig. 3. Performance Increase in outlier detection

CONCLUSION

Outlier detection is a significant problem with direct application in a wide variety of domains. An important observation with outlier detection is that it is not a well-formulated problem. Several approaches have been proposed to target a particular application domain. A novel approach is introduced to handle outlier mining for unlabeled data. The proposed system is a new model-based approach for outlier detection which introduces likelihood values to each input data into the modified SVDD training phase. The proposed method first captures the local uncertainty by computing likelihood values for each example based on its local data behavior in the feature space, and then builds global classifiers for outlier detection by incorporating the negative examples and the likelihood values in the modified SVDD-based learning framework. Extensive experiments are designed for real life datasets with performance matrix to measure the performance.

ACKNOWLEDGMENT

I would like to express my special thanks to all people who have helped me to complete this work. I am very grateful to my guide, Prof. N M. Shahane, Associate Professor, Computer Engineering, for his guidance, encouragement and the interest shown in this project. He has continuously helped and encouraged me in my work.

REFERENCES

[1] Bo Liu et. al.,"An efficient approach for outlier detection with imperfect data labels", IEEE Transaction vol 26,no.7.,July 2014

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

[3] Eskin E, "Anomaly detection over noisy data using learned probability distributions.", In: Proceedings of the seventeenth international conference on machine learning, 2008, pp 255–262.

[4] Tan PN, Steinbach M, Kumar V, "Introduction to data mining. Addison-Wesley, Boston.", 2005

[5] Jiang SY, An QB, "Clustering-based outlier detection method.", In: Proceedings of the fifth IEEE international conference on fuzzy systems and knowledge discovery, 2008, 429 - 433

[6] Cheng L,WingHW, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.", In: Proceedings of the national academy of sciences, USA (98),2001, 31–36

[7] Tax DMJ et. al., "Support vector data description applied to machine vibration analysis.", In: Proceedings of fifth annual conference of the advanced school for computing and imaging (ASCI),1999 398 – 405

[8] Aggarwal C, Yu P, "Outlier detection for high dimensional data.", In: Proceedings of the ACM SIGMOD international conference on management of data. ACM Press, 2001, 37–46

[9] Chen D, Shao X, Hu B, Su Q, "Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra.", Anal Sci 21(2), 2005, 161–167

[10] Agarwal C, "An empirical bayes approach to detect anomalies in dynamic multidimen-sional arrays.", In: Proceedings of the 5th IEEE international conference on data mining, 2005, 26–33

[11] Hido S, et. al., "Statistical outlier detection using direct density ratio estimation.", Knowl Inf Syst 26(2), 2011, 309–336

[12] BreunigM, Kriegel H, Ng R, Sander J, "LOF: identifying density-based local outliers.", ACMSIGMOD, 2000, 93–104

[13] Yang X, Latecki LJ, Pokrajac D, "Outlier detection with globally optimal exemplar-based GMM.", SDM, 2009, 145-154

[14] Jain AK, Dubes RC, "Algorithms for clustering data.", Prentice-Hall, New Jersey, 1988

[15] Jiang SY, An QB, "Clustering-based outlier detection method.", In: Proceedings of the fifth IEEE international conference on fuzzy systems and knowledge discovery, 2008, 429-433

[16] Lee KY et. al.,"Density-induced support vector data description.", IEEE Trans Neural Netw, 2007,284–289

[17] Tax D, Duin R, "Support vector data description.", Mach Learn 54(1), 2004,45–66

[18] Lee KY et. al.,, "Density-induced support vector data description.", IEEE Trans Neural Netw 18(1), 2007,284–289

[19] Agarwal D, "Detecting anomalies in cross-classified streams: a bayesian approach.", Knowl Inf Syst 11(1), 2006,29–44

[20] Barbara D et. al., "Detecting novel network intrusions using bayes estimators.",first SIAM international conference on data mining,2001

[21] Matsubara Y et. al., "D-Search: an efficient and exact search algorithm for large distribution sets.", Knowl Inf Syst , 2011, 131–157

[22] Aggarwal C, Yu PS, "Outlier detection with uncertain data.", In: Proceedings of SDM, 2008 , 483–493

AUTHOR'S PROFILE

Aditi Dighavkar, received the B.E. degree in 2013 and is now pursuing M.E. in Computer Engineering in Computer Engineering from K.K.W.I.E.E.R., Nashik, Savitribai Phule Pune University, India.

Prof. N. M. Shahane, Associate Professor, Department of Computer Engineering, KKWIEER, Nashik. His research interests include Machine learning, Digital Signal Processing, Digital image processing, Probability & Statistics, Pattern Recognition, Data Mining.