An Experimental Approach For Outlier Detection With Imperfect Data Labels

S.Dhiviya*
PG Scholar*
Department of CSE*
Kongu Engineering College
Perundurai, Erode-638052, Tamil Nadu, India
dhivselvaraj@gmail.com*

Dr.P.Jayanthi
Professor
Department of CSE
Kongu Engineering College
Perundurai, Erode-638052, Tamil Nadu, India
pjayanthikec@gmail.com

Abstract— **An anomaly or outlier is a deviation point which varies so much from other monitored data to indicate that it is different from others. Outliers are also mentioned as discordant, or anomalies. The chore of detecting outliers is to find data objects that are marked different from or incompatible with the original data. Data are imperfectly labeled due to data corruption or data of uncertainty. A data may be imperfectly labeled as outlier, although the data is not an outlier. These occur mainly due to negative examples or corrupted data. Typically, most outlier detection algorithms such as Index-based algorithm, Cell-based algorithm, Statistical based method use some distance measure of outlier or a statistical model. These methods cannot detect imperfectly labeled data and normal data which behave like an outlier. To find imperfectly labeled data, a membership value towards the normal and abnormal classes is proposed. The proposed approach works as follows: the Cuckoo k-means clustering method is used to detect the outliers and kernel LOF-based method will be used to reckon the membership values. The originated membership values and few negative examples are incorporated into Cuckoo-SVDD learning frame to develop a classifier for global outlier detection. The proposed method extensively deals with the imperfectly labeled data and attains the accuracy of detecting outliers.**

Keywords—Likelihood model, Support Vector Data Description, Outliers, Global Classifier

I. INTRODUCTION

Data mining also called as Knowledge Discovery in Databases (KDD), is the field of inventing useful patterns from huge amounts of data. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. The data mining tasks exhibit the general characteristics of data in the database and also test the current data in order to discover some patterns. Data mining systems should be able to learn patterns at various granularities that are at different levels of abstraction. A schema may contain data objects that are different from the model of the data. These data are called as outliers.

Outlier detection refers to the problem of finding data objects that are marked different from or incompatible with the original data. Applications such as health care, medical diagnosis, other kinds of detection include rare instances which are more interesting. Data are uncertain in nature for many real-life applications such as data points may correspond to objects, which are vaguely specified due to incompleteness and are therefore considered uncertain in representation.

The traditional outlier detection algorithms are classified into four categories: 1) Statistical-based method [9] uses a statistical model and apply deduction test to find the rare instance satisfies the model or not. 2) Density-based approach [8] takes that the original data occur nearer, while outliers occur distant from their neighbors, similar to Local Outlier Factor (LOF) which uses outlier score. 3) Clustering-based methods [10] mainly depend on the efficacy of clustering algorithms. 4) Model-based method [6] such as Support Vector Data Description (SVDD) learns the model from the labeled data and then classifies data based on the model.

Another important observation is that, the rare instances are not useful for developing a binary classifier, they can be used to form the decision boundary around the original data for outlier detection. The data are associated

with class label and membership values towards the positive and negative class labels. Some of the labeled abnormal examples are then incorporated into the dataset to build an accurate classifier.

The likelihood model is used to generate the pseudo training datasets by calculating the membership values based on their behavior. The originated membership values and few negative examples are then used to construct an outlier detection classifier from the learning framework. The integration of local and global outlier detection extensively handles the uncertain data and enhances the performance of outlier detection. The extensive experiment on datasets has been done using the proposed approach and this provides the effectiveness of detecting outliers.

The rest of the paper is organized as follows. Section II presents the previous work related to outlier detection problem. Section III discusses the proposed approach Cuckoo-SVDD in detail. Section IV reports the substantial experimental results on the real-world datasets. Section V provides the conclusion of the paper and possible direction for future work.

II. RELATED WORK

Tax et al [7] illustrated about the characterization of a data domain. The Support Vector Classifier is obtained as spherically shaped decision boundary by utilizing the kernel functions. The use of kernel function made the detection of outlier even capable with negative examples. Support Vector Data Description (SVDD) shows desirable results especially when anomaly information is used.

Abe et al [1] surveyed that most of the outlier detection methods are based on density estimation. Some of the demerits with these methods are flagging decisions and high computation. The demerits are rectified by a simple reduction of outlier detection to classification and a selective sampling mechanism based on the reduced classification problem. The method provides better results and also provides accuracy in detecting outliers.

Ghoting et al [8] presented Recursive Binning and Re-Projection (RBRP), a fast method based on distance or density based measures for large datasets. The outliers are detected based on the distance to their neighbors. RBRP is a two-phase process. In first phase it partitions the data using k-means and in second phase it finds the k-nearest neighbor. RBRP improves on the scaling behavior by employing an efficient preprocessing step.

Yu et al [2] formulated a methodology for handling uncertain data and data mining applications. The different methodology of collecting data has led to the escalation of uncertain data. The data points are specified as probabilistic function. The classification and clustering methods were also used for uncertain data. The methodology provides an efficient handling of uncertain data in various domains.

Chandola et al [3] illustrated a survey on anomaly detection. Anomalies are data points which vary from their normal behavior. Anomalies are classified into point anomalies and contextual anomalies. The existing techniques are grouped into different categories and analyzed. It provides the differentiation between natures of the data. The anomaly detection provides subsequent understanding and novel detection among the anomalous data.

Hido et al [9] proposed a new method to the problem of anomaly detection based on inlier, i.e., detecting outliers based on the training set of the data. The confidence score is used based on the density of training and test data. Among various density ratio estimation methods, Unconstrained Least-Square Importance Fitting (uLSIF) is utilized along with the parameters of regularization and the kernel width. This offers a computationally efficient and scalability to massive datasets.

Shi et al [10] presented that clusters and outliers are inseparable mainly for those datasets with noise. In Cluster-Outlier Iterative Detection (COID) algorithm, clusters are obtained and the intra-relationship and inter-relationship are defined. The alteration of clusters and outliers are performed iteratively. COID algorithm consistently outperforms natural clusters and outliers.

Bhaduri et al [5] proposed a new algorithm Vertically Distributed Core Vector Machine (VDCVM) is utilized on vertically distributed data. It integrates the data directly with storage nodes using Rapid Basis Function (RBF) kernels. The central node communicates with the part of a node. This method achieves a

comparable accuracy. VDCVM has lower communication cost compared to one-class learning. The accuracy of VDCVM is estimated on synthetic and real world datasets of varying densities.

Banerjee et al [4] presented an analysis on the problem of detecting outliers in symbolic sequences. This provides the understanding of the problem of anomaly detection. The research includes three distinct categories. First is to find the anomalous sequence and then to find the anomalous subsequence. The next process is to find the pattern of a sequence. The symbolic sequence achieves comparable results in different domains.

Liu et al [6] dealt with an approach for handling uncertain data which occurs due to various errors or partial completeness. The key challenge is that how to reduce the impact of uncertain data on learned distinctive classifier. A new SVDD-based approach is used to detect outliers on uncertain data. This involves two processes. In the first process, a pseudo-training data is generated by assigning a confidence score. In second process, the generated similarity score is incorporated into SVDD to construct a global classifier. SVDD is perceptive to noise contained in the input data and outperforms in detecting outliers.

The above discussed techniques have various problems in detecting outliers in various domains. These problems can be resolved by using local reachability density and Soft-SVDD and Bi-Soft-SVDD. Cuckoo-SVDD obtains a global optimum solution. These SVDD methods consistently perform in detecting outliers and can detect outliers from imperfectly data labeled data. The accuracy of outlier detection method can be measured by incorporating some percentage of error labels into the normal data.

III. PROPOSED APPROACH

Outlier detection is used to detect the imperfectly labeled data present in the normal data. Preprocessing step involved is the subset selection of data from the dataset. The selected subset of data is computed using Cuckoo k-means is used to detect outliers and kernel LOF based method is used to reckon the likelihood values. The generated likelihood values are then incorporated into SVDD based learning framework which provides a global classifier.

The likelihood model provides CUCKOO-SVDD. The basic idea is to include normal data inside the cluster and exclude the abnormal data outside the cluster. The outliers are the abnormal examples which are far away from their centroids and which does not satisfy the local reachability density. SVDD provides the classified data and outliers of the dataset. The flow of outlier detection is shown in Fig. 1.

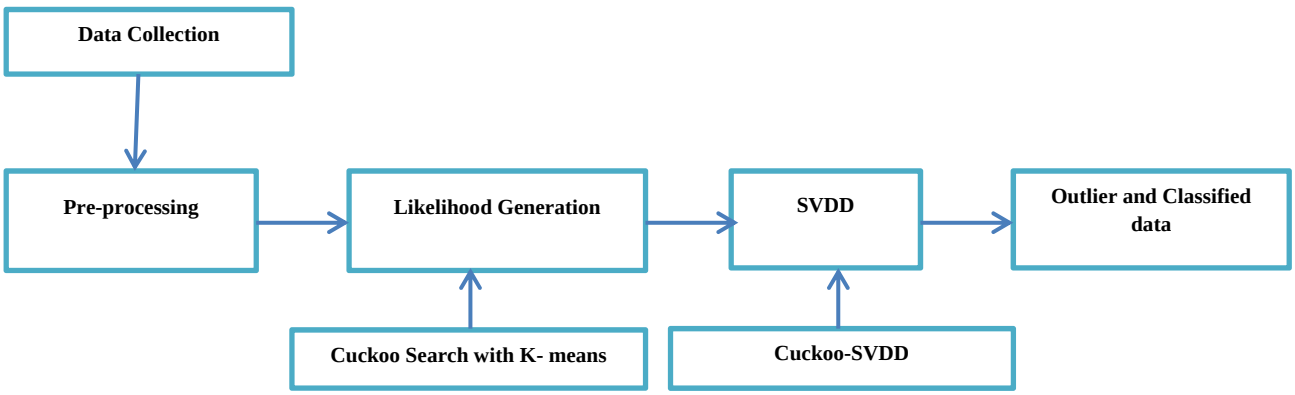


Figure 1. Flow of Outlier Detection

1.Input = 1,2,…n datapoints.

2.The kernel function used in all the methods is

$$k(v_l, v_n) = \exp(-\|v_l - v_n\|^2 / 2\sigma^2) + \frac{C\varepsilon}{2\sigma^2}$$

/* To form Clusters and detect outliers*/

3.Cuckoo search reckons the best cluster center position.

4.K-means Clustering –‘ n’ clusters

5.Repeat step 4 until convergence is reached.

/* Likelihood Generation using reachability and local reachability density*/

6.Kernel LOF- likelihood values of data points

7. Nearest local reachability density is calculated.

8.LOF provides the total number of positive and negative class

in the given data points.

/*Learning framework*/

9. Decision boundary is formed using the kernel function.

/* To enclose normal examples inside the sphere*/

10. The generated likelihood values are incorporated into Cuckoo-SVDD.

11.Output = global classifier for anomaly detection

Algorithm for Cuckoo-SVDD method

A. Cuckoo Search with K-means

Cuckoo Search algorithm is a population based stochastic global search metaheuristics algorithm. It is based on the general random walk system. The basic steps involved in cuckoo search include:

a) Each cuckoo lays one egg at a time, and dumps it in a randomly chosen nest.

b) The best nests with high quality of eggs (solutions) will carry over to the next generations.

c) The number of available host nests is fixed, and a host can discover an alien egg with a probability pa ∈ [0, 1]**.** In this case, the host bird can either throw the egg away or abandon the nest so as to build a completely new nest in a new location.

The ‘n’ nests are initialized and cuckoo is randomly selected using levy flight using the following equation

$$x_k(t+1) = x_k(t) + \mu * L \qquad (1)$$

where,

k = 1, 2….. n
n = no. of nests considered
μ = step-size
L = a value from Levy-distribution

The fitness function is calculated using (2) and replacing the nest with Cuckoo. A fraction of the nest will be replaced by new nests. The best nest is taken as an optimal solution and the cluster center will be the best nest position.

$$f(k) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (v_i - c_k)^2 \qquad (2)$$

where,

K = no. of clusters
v_i = patterns in the cluster
c_k = center of k^{th} cluster

K-means clustering partitions a subset of data into 'k' clusters in which cluster with the nearest mean contains most of the data. K-means partitions the data by using Euclidean distance between each point and centroid. The distance measure is calculated at different iterations until the convergence. The convergence here signifies that the data does not move to any other cluster. This provides the classified data and outliers. The outliers are detected using the classified data. The outliers are the data which departs from the centroids. K-means partition is done using the following function

$$J = \sum_{i=1}^{K} \sum_{j=1}^{L+N} \left\| \alpha(x_j) - \alpha(w_i) \right\|^2 \quad (3)$$

where,

K = no. of cluster
L = normal data
N = negative data
α(.) = a nonlinear mapping function like kernel function
w_i = cluster center

B. Likelihood Generation

The purpose of likelihood generation is to develop a pseudo training dataset model by computing membership values for each input data. Kernel LOF-based method is a density based approach. This determines the distance relative to its nearest neighbors in vector space using k-nearest neighbor. Kernel LOF method examines the distance reachable by object v_i with respect to object v_j in the vector space, using (4).

$$rd_k(v_i, v_j) = \max\{ \left\| \alpha(v_i) - \alpha(v_j) \right\| , max_{n' \epsilon N_k(v_i)} \{ \left\| \alpha(v_i) - \alpha(v_j) \right\| \quad (4)$$

where,

$N_k(v_i)$ = total number of k-nearest neighbors of point x_i
n' = nearest neighbor

Then the local reachability density is computed, using the mean reachable distance based on the closest neighbors of v_i. The local reachability distance is defined as

$$lrd_k(v_i) = \frac{1}{k} \sum_{v_j \in N_k(v_i)} rd_k(v_i, v_j) \quad (5)$$

where,

k = no. of clusters
$N_k(v_i)$ = total number of k-nearest neighbors of point v_i
rd_k = distance reachable by object v_i to object v_j

After local reachability density is computed, then neighborhood of lrd is calculated as

$$N_{lrd}(v_i) = \{ v_j \in D \left| \left\| \alpha(v_i) - \alpha(v_j) \right\|^2 \leq lrd_k(v_i) \right\} \quad (6)$$

where,

D = Data

The negative class is calculated from the l_t examples out of the nearest neighbor $|N_{lrd}(v_i)|$ belonging to the positive class l_t as

$$l_n = | N_{lrd}(v_i)| - l_t \quad (7)$$

The data are classified into positive class and negative class. The negative class is the outliers which does not satisfy the local reachability density. The single likelihood model denotes the membership value towards its own class label. The bi-likelihood model denotes the membership value towards positive and negative class labels. The likelihood model of a normal example v_t and abnormal example v_k for cuckoo k-means clustering is calculated as

$$m^t(v_t) = l_t / | N_{lrd}(v_i)|$$

$$m^n(v_k) = l_n / |N_{lrd}(v_i)| \quad (8)$$

where,

$|N_{lrd}(v_i)|$ =total number of closest neighbors in the lrd-neighborhood

l_t = positive class

l_n = negative class

Using the two likelihood models a pseudo training dataset model is generated by calculating the membership values based on the nature of the data in vector space. The generated likelihood values are then incorporated into Cuckoo-SVDD.

C. Cuckoo-SVDD

The generated likelihood values are incorporated into SVDD [11] to develop a global classifier for anomaly detection. The membership value used for Cuckoo-SVDD is based on its own class label. The cross validation technique is used along with the new kernel function. The SVDD is defined using the following objective function

$$\min F = R^2 + C_1 \sum m^t(v_i)\,\xi_i + C_2 \sum m^n(v_j)\,\xi_j$$

$$\text{s.t. } \|v_i - o\|^2 \le R^2 + \xi_i,\ v_i \in S_p$$

$$\|v_j - o\|^2 \ge R^2 - \xi_j,\ v_j \in S_n,\ \xi_i \ge 0, \xi_j \ge 0 \quad (9)$$

where,

R = radius of the sphere

C_1, C_2 = constants greater than zero

ξ_i , ξ_j = measure of error

S_p = positive examples

S_n = negative examples

$m^t(v_i)$, $m^n(v_j)$ = degree of membership

The Cuckoo k-means clustering along with SVDD is called as Cuckoo-SVDD and k-means clustering along with soft-SVDD is called as CS-SVDD and kernel-LOF along with soft-SVDD is called as LS-SVDD. The k-means clustering along with bi-soft-SVDD is called as CBS-SVDD and kernel LOF along with bi-soft-SVDD is called as LBS-SVDD.

IV. EXPERIMENTAL SETUP

The proposed model is implemented using MATLAB running on Intel core i3 processor with a 4 GB RAM capacity. The real life datasets were used for outlier detection which is taken from D.M.J. Tax. Outlier Detection Datasets [12]. The Segment dataset is collected from UCI Machine Repository [13]. The description of the dataset is given in Table I.

Table I. Dataset Description

Dataset	Description	# of dataset	Features
Abalone	Classes 1-8 vs rest	4177	10
Spambase	Others vs spam	4601	57
Thyroid	Class 2 vs rest 3	3428	21
Delft pump 5x3	Normal situations vs rest	1500	64
Diabetes	Present vs rest	768	8
Segment	Class 1 vs rest	2310	19
Waveform	Class 0 vs rest	900	21
Arrhythmia	Normal vs rest	420	278

A. Parameters for Evaluation

The outlier detection algorithms performance can be examined using detection rate. Detection rate estimates the total number of correctly identified anomalies.

In general, the AUC estimates the performance of outlier detection methods. The AUC values are explicitly computed to compare the performance of algorithms used. The new kernel function used in all the methods is

$$k(v_l, v_n) = \exp(-\|v_l - v_n\|^2 / 2\sigma^2) + \frac{C\varepsilon}{2\sigma^2} \quad (10)$$

The parameter σ used in the kernel function ranges from 2^{-4} to 2^3. The error rate ε is based on the SVDD and the constant C ranges from 1 to 3.

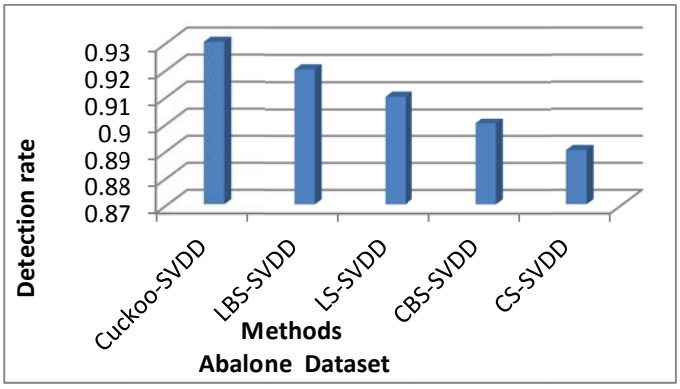
B. Results and Comparison

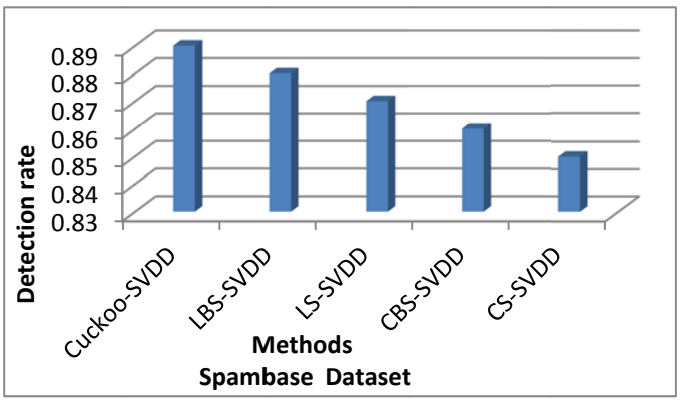
The k-means clustering along with soft-SVDD is called as CS-SVDD and kernel-LOF along with soft-SVDD is called as LS-SVDD. The k-means clustering along with bi-soft-SVDD is called as CBS-SVDD and kernel LOF along with bi-soft-SVDD is called as LBS-SVDD. These methods are compared along with Cuckoo-SVDD which provides a better performance in detecting outliers and provides global optimum solution.

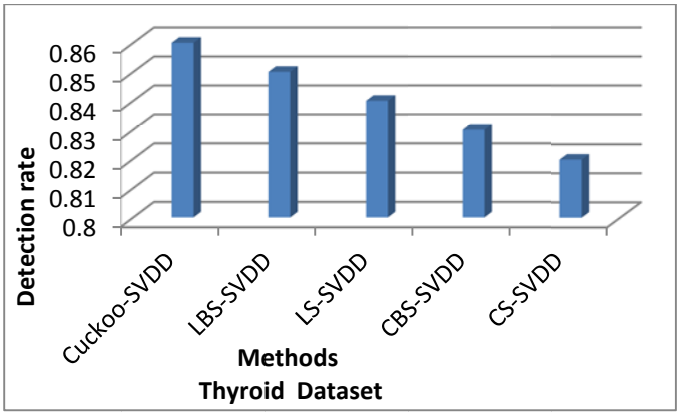
The AUC value of each dataset is higher in Cuckoo-SVDD in comparison with other methods. The CS-SVDD, LS-SVDD, CBS-SVDD, LBS-SVDD methods are based on their local data behavior whereas Cuckoo-SVDD is based on the global data behavior. The proposed method provides high accuracy than other methods.

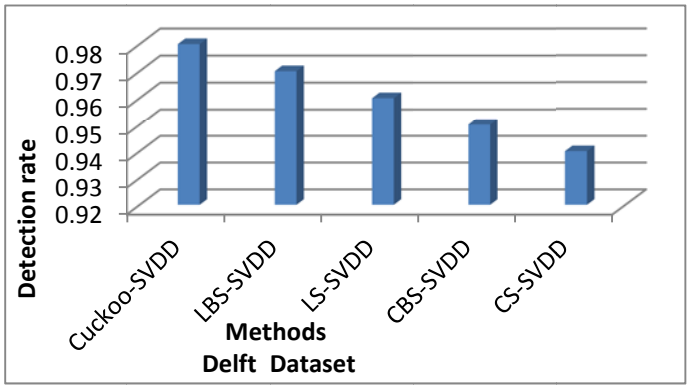
The original data labels are added with some percentage of noise and analyzed with the proposed method. This provides a consistent result when added with noise. The detection rate is higher and provides a consistent performance on detecting outliers even the data are misclassified as outliers.

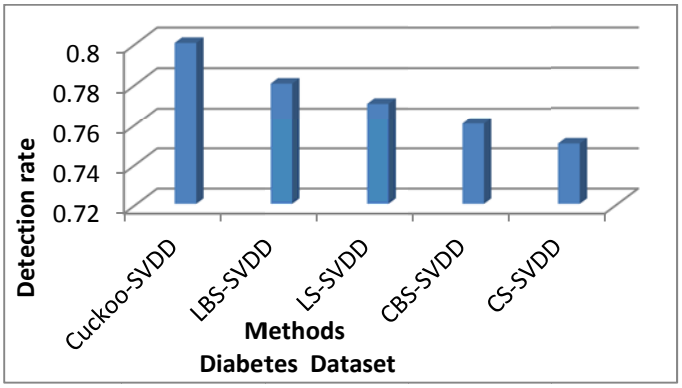
The performance comparison of outlier detection method on six datasets is shown in Fig 2.

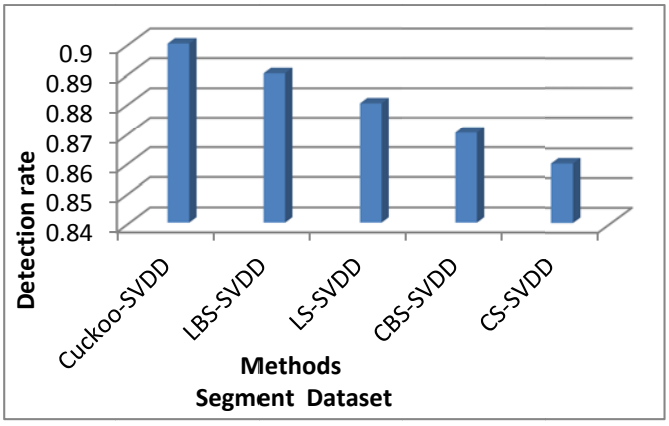





0.86
0.85
0.84
0.83
0.82
0.81
0.8
Detection rate
Cuckoo-SVDD
LBS-SVDD
LS-SVDD
CBS-SVDD
CS-SVDD
Methods
Thyroid Dataset


0.98
0.97
0.96
0.95
0.94
0.93
0.92
Detection rate
Cuckoo-SVDD
LBS-SVDD
LS-SVDD
CBS-SVDD
CS-SVDD
Methods
Delft Dataset


0.8
0.78
0.76
0.74
0.72
Detection rate
Cuckoo-SVDD
LBS-SVDD
LS-SVDD
CBS-SVDD
CS-SVDD
Methods
Diabetes Dataset


0.9
0.89
0.88
0.87
0.86
0.85
0.84
Detection rate
Cuckoo-SVDD
LBS-SVDD
LS-SVDD
CBS-SVDD
CS-SVDD
Methods
Segment Dataset

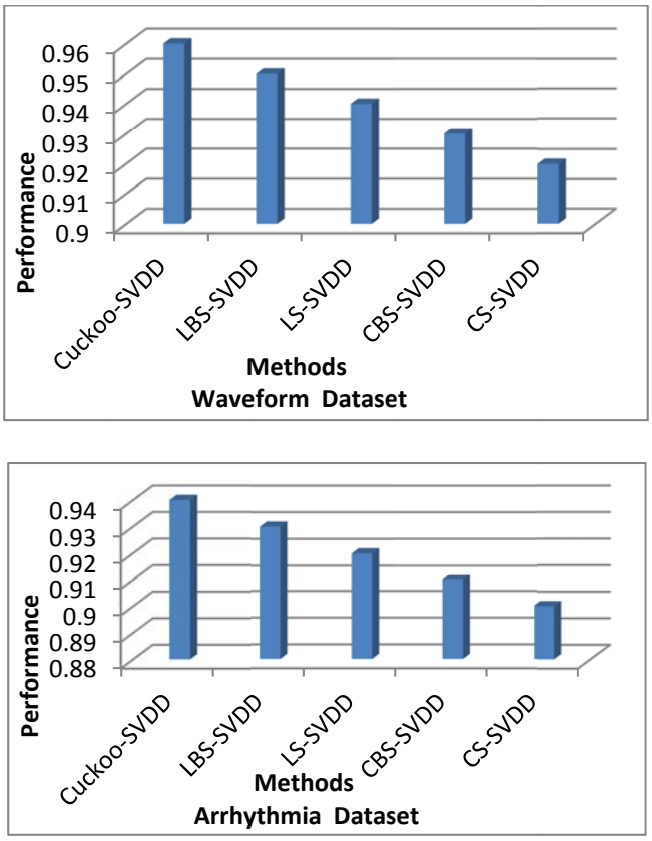


Figure 2. Performance comparison of outlier detection methods on datasets

V. CONCLUSION AND FUTURE WORK

The Cuckoo-SVDD handles imperfectly labeled data and efficiently detect outliers and provides global optimum solution based on the learned classifier. This is because the kernel function used for cross validation and kernel LOF-based method to calculate the membership values.This density based method consistently performs on the data with varying density. The proposed method will outperform well on various dataset and it provides high accuracy.

The work can be extended in several ways by enhancing a better method to generate the likelihood values. The bi-likelihood values can also be incorporated into Cuckoo-SVDD to further improve the performance of outlier detection associated with other classes.

REFERENCES

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

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 823–839, May 2012.

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

[4] C. C. Aggarwal and P. S. Yu,"A survey of uncertain data algo-rithms and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 5, pp. 609–623, May 2009.

[5] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Statistical outlier detection using direct density ratio estimation," *Knowl. Inform. Syst.*, vol. 26, no. 2, pp. 309–336, 2011.

[6] Y. Shi and L. Zhang,"COID: A cluster-outlier iterative detec-tion approach to multi-dimensional data analysis," *Knowl. Inform. Syst.*, vol. 28, no. 3, pp. 709–733, 2011.

[7] A. Ghoting, S. Parthasarathy, and M. E. Otey, "Fast mining of distance-based outliers in high-dimensionall datasets," *Data Min. Knowl. Discov.*, vol. 16, no. 3, pp. 349–364, 2008.

[8] K. Bhaduri, B. L. Matthews, and C. Giannella,"Algorithms for speeding up distance-based outlier detection," in *Proc. ACM SIGKDD Int. Conf. KDD*, New York, NY, USA, pp. 859–867, 2011.

[9] N. Abe, B. Zadrozny, and J. Langford,"Outlier detection by active learning," in *Proc. ACM SIGKDD Int. Conf. KDD*, New York, NY, USA, pp. 504–509, 2006.

[10] 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.

[11] Bo Liu, Longbing Cao, Philip S.Yu, Yanshan Xiao and Zhifeng Hao, "An Efficient Approach for Outlier Detection With Imperfect Data Labels", IEEE Transactions on Knowledge And Data Engineering, vol. 26, no. 7, 2014.

[12] http://homepage.tudelft.nl/n9d04/occ/index.html

[13] http://archive.lcs.uci.edu/ml/datasets.html