# COMPARATIVE STUDY OF KSVDD AND FSVM FOR CLASSIFICATION OF MISLABELED DATA

**Rajani S Kadam[1], Prakash R. Devale[2]**

[1]*Research Scholar, Department of Information Technology, Bharati Vidyapeeth Deemed University College of Engineering, Maharashtra, India.*

[2]*Professor, Department of Information Technology, Bharati Vidyapeeth Deemed University College of Engineering, Maharashtra, India.*

## Abstract

*Outlier detection is the important concept in data mining. These outliers are the data that differ from the normal data. Noise in the application may cause the misclassification of data. Data are more likely to be mislabeled in presence of noise leading to performance degradation. The proposed work focuses on these issues. Data before classifying is given a value that represents its willingness towards the class. This data with likelihood value is then given to classifier to predict the data. SVDD algorithm is used for classification of data with likelihood values.*

***Keywords:*** *Confusion Matrix, FSVM, Outlier, Outlier Detection, SVDD*

----------***----------

## 1. INTRODUCTION

Outliers are those data whose behavior is very distinct from that of the other data[1],[6],[7]. Outlier detection is the process of finding these outliers. Outlier detection is applied in many areas like detection of faults in the system, intrusion detection, cyber attack, finding the abnormalities in medical fields and many more.

There are many outlier detection methods proposed [1],[2],[3],[10],[11]to detect outlier and the normal data. Support vector data description (SVDD)[5] is a variant of SVM. The algorithm is most widely used for outlier detection because of its performance with respect to other learning algorithm. The detection of these outliers plays a major role in many areas. The SVDD algorithm is effective but is sensitive to noise. Noise in the data leads to wrong prediction of data. The study deals with the specified problem. The data are subjected to likelihood value generation which defines the membership value of the data towards the class. These data with likelihood value are then classified using SVDD classifier. The SVDD with kernel k-means clustering is called KSVDD. The results are analyzed using the confusion matrix and algorithm performance is compared with the FSVM algorithm.

SVDD is a semi supervised learning method. It uses model based approach to classify the data. SVDD is a one class classification algorithm which uses only one class to train and build the model. The model built is then used to classify the test data. Usually positive class or normal class being more in number is used to train the model and classify the test data. Data that does not match the model are classified as negative or outliers. SVDD being a non-linear classifier tries to build a hyper-sphere around the normal data and classifying the data outside the boundary as outliers. The method builds the hyper sphere around the linear data to transform it into non-linear form by using the kernel function. This hyper sphere is given by following equation.

$$R^2+\frac{1}{\mu n}\sum_{i=1}^{n}\xi_i \quad \text{..... (1)}$$

*Such that* $||\phi(i)-b||^2 \leq R^2+\xi_i$

*Where R is the radius of the sphere*

*b is the center of the sphere*

$\frac{1}{\mu n}$ *is the constant (sometimes referred as C in the paper)*

$\xi_i$ *are the slack variables value* $\geq 0$

FSVM is nothing but Fuzzy SVM [13] used to classify the data by applying the fuzzy logic to each data.SVM provides a good classification for linear data. When data are overlapping kernel function needs to be applied to separate non linear data causing the slack variables $\xi_i$, that is the value of misclassification. By applying the fuzzy concept this misclassification can be corrected.

The optimization function is given by $\frac{1}{2}w.w + c\sum_{i=1}^{l} s_i\ \xi_i$ ..... (2)

*Here* $\xi_i$ *is misclassification*

$s_i$ is the membership degree which defines the point $x_i$ belongs to one class and ranges from 0 to 1($0< s_i <1$). $s_i$ applies the fuzzy concept to the algorithm. FSVM performs better when misclassification rate is more that is in the presence of noise.

## 2. THE PROPOSED WORK

In the work both the algorithms are implemented using matlab tool. The Fisher Iris data set is used to train and test the data using both algorithms. The work focuses on correctly predicting the data that is reducing the misclassification rate. This not only detects the outlier but also gives accurate result. To improve the accuracy and handle the misclassification data, likelihood value is generated for each data in the dataset [12]. Likelihood value is generated using kernel k-means clustering. The value for each data set decides the degree of membership towards positive or negative class. When the n numbers of small clusters are formed the data belonging to the cluster are similar to one-another and the data which does not belong to any class will be the obvious outlier. Hence the outliers are effectively detected, misclassification is reduced.

## 3. DATASET

Fisher Iris dataset is a floral data set. This was introduced by Ronald Fisher for his study of linear discriminant analysis. There are 150 samples in the dataset. Three classes of Iris flower are considered namely Iris-Setosa, Iris-Versicolor, Iris-Virginica. The characteristics of these flowers like sepal length, sepal width, petal length, petal width are measured in centimeters and are taken as parameters in the data set. Classification algorithms work on these parameters and classify the Iris sample to one of the three classes.

### 3.1. Flow of the Proposed Work

1. Iris dataset normalized
2. Dataset is divided into 70 and 30 ratio
3. 70% of train data is given to both the methods
4. KSVDD applies kernel K-means clustering to generate likelihood value
5. FSVM applies fuzzy logic $S_i$ to develop membership degree
6. Both built the model using training data
7. 30% of the data is used as test data
8. Output matrix is generated by the application

### 3.2. Evaluation Measures

To measure the performance of any classifier confusion matrix is used ,this gives the actual and predicted class values.

| | | Predicted Class | | |
|---|---|---|---|---|
| Actual Class | | Positive Class | Negative Class | Total |
| | Positive class | TP | FN | P |
| | Negative class | FP | TN | N |

**Figure. Confusion Matrix**

## 4. RESULT

The output of both the method in classifying Iris data is 3X3 matrix which gives the TPTNFPFN values as shown in the figure that follows. Here 1,2,3 represent the three types of iris in the dataset(1-setosa,2-versicolor,3-virginica).

| | | Predicted Class | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Actual class | 1 | TP | FP | |
| | 2 | FN | TN | |
| | 3 | | | |

**Figure :Output Matrix**

With this output matrix Detection rate, False alarm rate, Accuracy, Error rate are calculated

1. ***Detection rate***: gives information about number of correctly identified outliers

Detection rate =TP/ (TP+FN)

2. ***False alarm rate***: gives the number of outliers misclassified as normal data samples

False alarm rate=FP/ (FP+TN)

3. ***Accuracy***: Is also called as recognition rate and gives percentage of test set samples that are correctly classified by the classifier

Accuracy=(TP+TN)/(P+N)

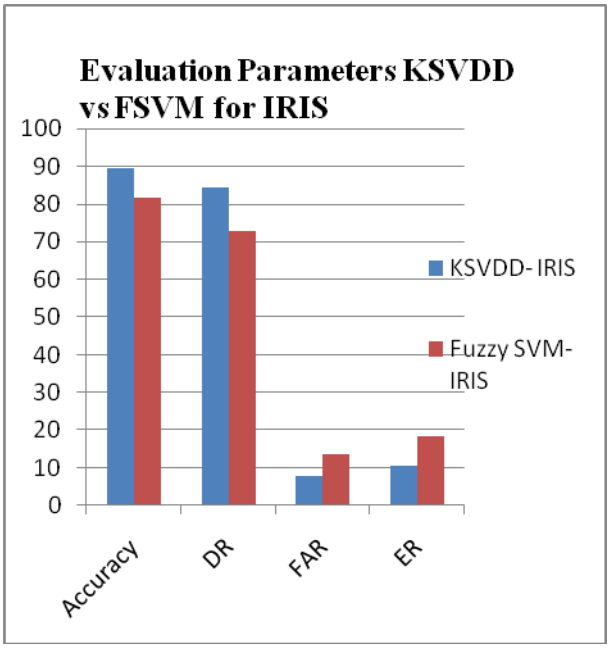
4. ***Error rate***: Is also called as misclassification rate and gives the percentage misclassification rate

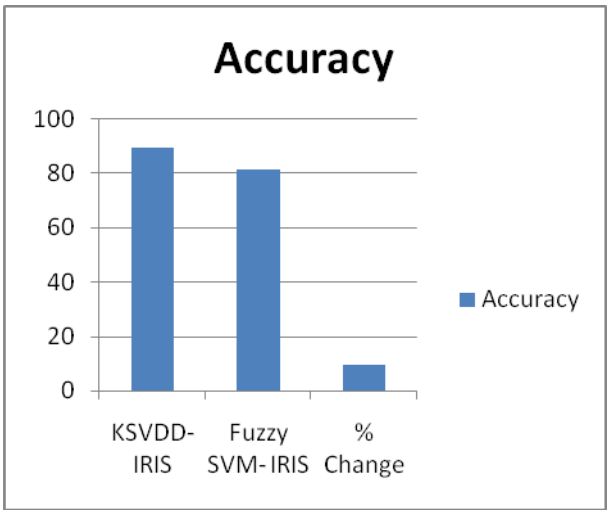
Error rate= (FP+FN)/(P+N)

About 10 outputs are taken. The average of these values are taken and are plotted in the graph as shown

| Confusion matrix for KSVDD- IRIS - Setosa | | | | | |
|---|---|---|---|---|---|
| **Iteration** | **TP** | **FP** | **FN** | **TN** | **Accuracy** |
| **1** | **22** | **4** | **5** | **44** | **88.00** |
| **2** | **25** | **3** | **2** | **45** | **93.33** |
| **3** | **23** | **3** | **3** | **46** | **92.00** |
| **4** | **20** | **2** | **6** | **47** | **89.33** |
| **5** | **22** | **4** | **2** | **47** | **92.00** |
| **6** | **22** | **4** | **1** | **48** | **93.33** |
| **7** | **23** | **4** | **4** | **44** | **89.33** |
| **8** | **23** | **4** | **5** | **43** | **88.00** |
| **9** | **19** | **7** | **6** | **43** | **82.67** |
| **10** | **21** | **4** | **2** | **48** | **92.00** |
| | **Avg. Accuracy for Setosa, Versicolor, Virginnica** | | | | **89.44** |

| Confusion matrix for FSVM- IRIS - Setosa | | | | | |
|---|---|---|---|---|---|
| **Iteration** | **TP** | **FP** | **FN** | **TN** | **Accuracy** |
| **1** | 18 | 8 | 7 | 42 | 80.00 |
| **2** | 19 | 8 | 5 | 43 | 82.67 |
| **3** | 18 | 8 | 9 | 40 | 77.33 |
| **4** | 17 | 5 | 8 | 45 | 82.67 |
| **5** | 18 | 7 | 7 | 43 | 81.33 |
| **6** | 17 | 7 | 7 | 44 | 81.33 |
| **7** | 18 | 6 | 7 | 44 | 82.67 |
| **8** | 17 | 7 | 7 | 44 | 81.33 |
| **9** | 18 | 6 | 5 | 46 | 85.33 |
| **10** | 17 | 6 | 6 | 46 | 84.00 |
| | **Avg. Accuracy for Setosa, Versicolor, Virginnica** | | | | **81.59** |




The result show that the accuracy of the KSVDD is more as compared to FSVM. KSVDD detects more outliers as compared to FSVM. The misclassification of KSVDD is less than FSVM. While classifying the error rate of SVDD is lass than the FSVM. Accracy rate between the KSVDD and FSVM show that the KSVDD is around 9% more efficient than FSVM.

## 5. CONCLUSION

Outlier detection is the important task in data mining. In the work of outlier detection, the difference between the KSVDD and FSVM resuts show that the KSVDD with the likelihood value for data is more effective in detecting the outliers.

## ACKNOWLEDGMENT

We would like to thank all the reviewers for their reviews, comments and suggestions

## REFERENCES


[1] D. M. Hawkins, Identification of Outliers. Chapman and Hall, Springer, 1980.

[2] M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2000

[3] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-Based Detection and Prediction of Outliers," IEEE Trans. Knowledge and Data Eng.,vol. 18, no. 2, pp. 145-160, 2006.

[4] 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, May 2012

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

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

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

[8] B.Liu, Y. Xiao, Philip S. Yu, Z.Hao, and L. Cao "An Efficient Approach for Outlier Detection with Imperfect Data Labels"IEEE Trans. Knowledge and data engineering, vol. 26, no. 7, July 2014

[9] 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.

[10] V. Barnett and T. Lewis, Outliers in Statistical Data. John Wiley&Sons, 1994.

[11] N.L.D. Khoa and S. Chawla, "Robust Outlier Detection Using Commute Time and Eigenspace Embedding," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2010.

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

[13] Pei-Yi Hao" Fuzzy one-class support vector machines", Elsevier, January 2008

[14] Yi-Hung Liu, Yan Chen Liu, Yen-Jen Chen," Fast Support Vector Data Descriptions for Novelty Detection", Neural Networks, IEEE Trans, Volume:21 , Issue: 8 , July 2010


## BIOGRAPHIES

**Ms.Rajani S. Kadam** is a student of M.Tech in Information Technology, Bharati Vidyapeeth Deemed University College of Engg, Pune-43.

**Prof. Prakash Devale** is Professor in Information Technology Dept., Bharati Vidyapeeth Deemed University College of Engineering, Pune-India. He is Pursuing Ph.D. from Bharati Vidyapeeth Deemed University in the area of Machine Translation .He is having 23yrs of experience in teaching. He is lifetime member of ISTE.