IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 6, Ser. 2 (Nov. – Dec. 2023), PP 42-44
www.iosrjournals.org

A Short Review Of One-Class Classification Algorithms

Sami M. Halawani and Mustasem Jarrah
IT Department, Faculty OfInformation Technology, King Abdulaziz University,
Jeddah, Saudi Arabia

Abstract:
One-class classification(OCC) is a special kind of classification problem in machine learning in which during training only the data points of one class is present. It has applications in many domains as it is difficult to collect the data of abnormal class. Many OCC algorithms have been proposed. In this paper, we will discuss many OCC algorithms. We will also discuss the performance measures used for OCC algorithms. Application domains of OCC algorithms will also be discussed.

Key Word*:OCC algorithms; Performance measures; Binary classification*

Date of Submission: 05-12-2023 Date of Acceptance: 15-12-2023

I. Introduction

Classification is an important task of machine learning in which the task is to predict the class of a testing point using the features of the testing point and trained models [7]. The training is done using the data points with class labels. Binary classification is a task in which the model is trained using data points belonging to two classes. However, there are many domains in which it is difficult to collect the data for abnormal class. Therefore, the training is done using data points belonging to one-class. This makes the classification task difficult. There are many domains which have this kind of data such as medical, attack detection in network, fraud detection on finance etc. Many OCC algorithms have been proposed [1, 2, 3]. In this paper, we will discuss some of the important OCC algorithms. We will also discuss the performance measures used for these OCC algorithms. Important application domains of these OCC algorithms will also be discussed.

The paper is organized in the following ways. Section 2 has a discussion about important OCC algorithms. Section 3 has related performance measures. Important application domains will be presented in Section 4. The paper will end with Section 5, in which the conclusion will be presented.

II. Important OCC algorithms

In this section, we will discuss some of the important OCC algorithms.

Autoencoders: Autoencoders are a special kind of neural networks which have two main components; encoders and decoders [1]. Encoders are used to create a low-dimensional representation of the original feature space. Decoders are used to convert back the low-dimensional into original feature space. The difference between the input and the output is considered as reconstruction error. For OCC problem, autoencoders are trained with normal data points. The motivation is that as the autoencoders are trained with normal data points the reconstruction error for a normal testing point will be small whereas the reconstruction error for aabnormal testing point will be large. In other works, reconstruction error is used to find whether a data point is normal or abnormal. Data points with large reconstruction errors will be considered as abnormal whereas data points with smaller points will be considered as normal data points. Autoencoders can be trained with a variety of data such as tabular, images etc. therefore they have applications in various domains to find the abnormal points. Autoencoders have various parameters such as number of hidden layers and the number of nods in eachhidden layer, therefore finding the correct parameters is difficult task for Autoencoders.

K-means clustering algorithm- K-means clustering algorithm is very popular clustering algorithm [7] because of its efficiency and accuracy. The algorithm has following steps

1- Initialize the data into desired number of clusters.
2- Find the center of each cluster.
3- Reassign the data points to a cluster for which it has minimum distance from the center
4- Iterate step 2 and step 3 until all the data points don’t change their clusters or any other predefined criterion is achieved.

For using K-means clustering algorithm as OCC algorithm, first we cluster the training data. To compute an outlier score for a testing data point, we first calculate the distances between the data point and the cluster centers. The minimum distance is considered as outlier score, in other words, the distance between the data point and its nearest cluster center will be considered as outlier score. The number of clusters and the distance measures re two important parameters of this algorithm.

Isolation Forests: Isolation forests is an ensemble of decision trees [5]. In other words, it consists of many decision trees. Decision trees are created using random splits. It was assumed that the path length (the distance between the root node and a leave) of an abnormal point will be less as compared to the path-length of a normal point. Therefore, a path-length of a data point can be used to find the outlier score of a testing data point. The data point with large path length will be considered as normal data points, whereas data points with smaller path lengths will be considered as abnormal points. Isolation forests is quite robust to the selection of parameters;therefore, it is very popular OCC algorithm. Isolation forests have shown excellent performance in many domains.

K-Nearest Neighbors: K-Nearest Neighbors is an OCC algorithm which is based on nearest neighbors [1]. It is based on the fact that to find out outlier of a data point, the nearest neighbors of the point are calculated and they are compared with the nearest neighbors of the nearest neighbors. If a point is an outlierthen the distance between the he point and its nearest neighbor will be large as compared to the distance between this nearest neighbor to its nearest neighbor.

Support Vector Machines: Support vector machines are an important OCC algorithm [4, 6]. This algorithm creates a decision boundary around the normal data points. To decide the class of a new data point, the position of a data point is found out with respect to the decision boundary. If the point is inside the decision boundary, it will be considered as normal points otherwise it will be considered as abnormal points.

III. Performance Measures

As the number of normal points is quite large as compared to abnormal data points, the accuracy is not a good performance measure. For example, if the training data has 99% normal data points and 1% abnormal data points, simply telling all the points as normal will give 99% accuracy. Various performance measures have been defined [7]. To describe them, we assume normal points as negative and abnormal points as positive. Confusion matrix is defined in the following way.

Table no 1 – Confusion Matrix. T represents True, F represents False.

Predicted

		Positive (P)	Negative(N)
Actual	Positive (P)	TP	FN
	Negative (N)	FP	TN

Three performance measures for a given threshold will be given below

Precision Precision = TP/(TP + FP)

Recall Recall = TP/P

F_measureF-measure = 2*Precision*Recall/(Precision + Recall)

Area under curve for receiver operating characteristic curve – This is a curve for true positive rate against the false positive rate (FPR) at each threshold setting. Area under the curve is calculated as a performance measure.

Area under curve for precision recall curve - This is a curve for precision against recall at each threshold setting. Area under the curve is calculated as a performance measure.

IV. Applications

OCC algorithms are used when there is highly class imbalanced data [1, 2, 3]. In other words, large majority class points and a small minority class points. There types of datasets are available in medical, health, finance, economics, Industry, network security etc. It shows that it has applications in wide range of areas.

V. Conclusion

In this paper, first we discussed an OCC problem. Then, we discussed many OCC algorithms. Various performance measures for OCC algorithms were also described. Applications areas for OCC algorithms were also presented.

In future, we will do experimental studies with these OCC algorithms to find out their applicability in various domains.

References

[1]. M. A. F. Pimentel, D. A. Clifton, L. Clifton, And L. Tarassenko, "Review: 378 A Review Of Novelty Detection," Signal Process., Vol. 99, Pp. 215–249, June 379 2014.

[2]. Program S. S. Khan And M. G. Madden, "One-Class Classification: Taxonomy Of Study And Review Of Techniques," Knowledge Eng. Review, Vol. 29, No. 3, Pp. 345–374, 2014.

[3]. M. J. Tax, One-Class Classification: Concept Learning In The Absence Of Counter-Examples. Phd Thesis, Technische Universiteit Delft, 2001.

[4]. Vapnik V (1998) Statistical Learning Theory. Wiley-Interscience, New York

[5]. F. T. Liu, K. M. Ting, And Z. Zhou, "Isolation Forest," In Proceedings Of The 2008 Eighth Ieee International Conference On Data Mining, Icdm '08, (Washington, Dc, Usa), Pp. 413–422, Ieee Computer Society, 2008.

[6]. Tax, D.M.J., Duin, R.P.W.: Support Vector Domain Description. Pattern Recognition Letters 20, 1191–1999.

[7]. Mla. Bishop, Christopher M. Pattern Recognition And Machine Learning. New York :Springer, 2006.