Voting-Based Ensemble of Unsupervised Outlier Detectors

Roy Thomas and J. E. Judith

Abstract Datasets may contain small sets of data objects whose characteristics are not in accordance with the mainstream characteristics of the data objects in a dataset. These data objects, which are not noise, may contain valuable information and are called outliers. Outlier detection is a topic of research in many fields like detecting malwares in cyber security, finding fake financial transactions, identifying defects in industrial products, detecting abnormality in health data, etc. Researchers have developed several application methods for detecting outliers and a few generic methods. These methods can be grouped into unsupervised methods, supervised methods and semi-supervised methods based on the readiness of class labels. We, in this paper, present the performance of three outlier detection algorithms using the realworld datasets. The algorithms used are one-class SVM, elliptic envelope and local outlier factor. In order to improve the performance, all these algorithms were selected and ensemble based on voting mechanism. The influence of dimensionality reduction on the proposed ensemble method has also been studied. Experiments using publicly available datasets show that the proposed technique outperforms individual outlier detectors.

Keywords Data mining · Dimensionality reduction · Ensemble · Outlier detection · Unsupervised

1 Introduction

Data mining is a field in computer science that is intended to find relevant and crucial information from large collections of datasets most of which are unstructured. Considering the volume, the data in the world is increasing tremendously at every moment and it becomes very difficult to extract the desired information from the datasets. Datasets are identified by the mainstream characteristics of the dataset

R. Thomas (✉) · J. E. Judith
Noorul Islam Centre for Higher Education, Kumaracoil, India
e-mail: roygptc@gmail.com

J. Jayakumari et al. (eds.), *Advances in Communication Systems and Networks*, Lecture Notes in Electrical Engineering 656,
https://doi.org/10.1007/978-981-15-3992-3_42

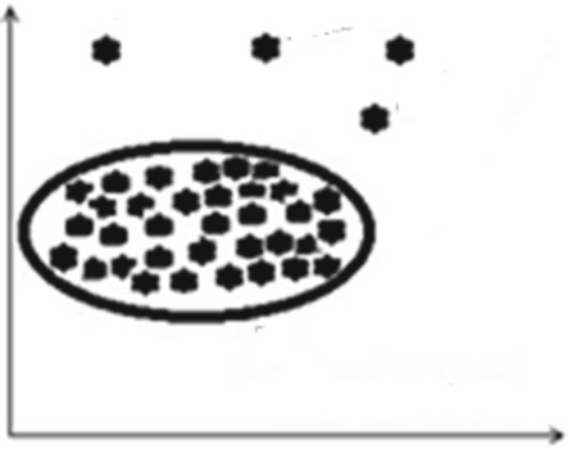

Fig. 1 Objects inside the ellipse are normal data and outside the ellipse are outliers

attributes, even though these datasets may contain data objects whose characteristics are very much different from the characteristics of the majority data objects in the dataset. These hidden data objects may contain valuable information and hence cannot be treated as noise. They are called outliers as their properties are not conformed to the balance data objects in the dataset. Outlier detection plays a vital role in many branches of study such as statistics and data mining. Detection of hidden and important information from large datasets has been a research field with diverse application areas for the past few decades.

Detection of hidden and important information has become an essential research field in many branches of statistics and computer science such as data mining, machine learning, information theory and spectral theory. Outlier detection is aimed at finding infrequent data objects containing valuable information and is not conformable with the majority of data objects in the dataset. Figure 1 shows the difference between normal data objects and outliers. Most of the normal objects are clustered together as they are having similar properties. An outlier is a data object that is distant from most of the remaining data objects in the dataset and is not clustered with the majority data objects. Outlier detection is defined as the discovery of data objects that are dissimilar, distant or uneven with respect to the majority of data objects in the dataset. Outlier detection provides critical information in a variety of domains such as military and aeronautical fields.

The applications of outlier detection are essential in many areas. Outlier detection methods can be applied to find fake financial transactions, unauthorized insurance claim, unauthorized computer network access, to find fake credit card transactions in banking, abnormality in medical and public health data, damages in industrial products, inadequacies in image processing, conflicts in Web applications, irregularities in robot behavior, variances in astronomical data, discrepancies in census data, etc.

Outliers can be categorized into point outliers, collective outliers and contextual outliers depending on the nature of outliers. An individual data object that differs considerably from other data objects in its set is called a point outlier or global

outlier. For example, the features of the credit card transactions like the type of items purchased, their quantity, amount spent, etc., by the fraud shall be very much different from the normal purchase features of the credit card transaction by the authenticated card owner. The credit card transactions that do not follow the regular pattern can be treated as an outlier. Collective outlier is a subgroup of data objects which together differs from the entire dataset. The distinct data objects in a collective outlier are not outliers individually. For example, heavy traffic for a limited time is a normal incidence in a metropolitan city. However, if it lasts for two or three days, it becomes a collective outlier. Contextual outlier is a data object that differs considerably from the dataset based on a selected situation and is a normal data object with reference to some other situation. In contextual outliers, the data objects are defined using two attributes—a context attribute and a behavior attribute. Rain in summer is a contextual outlier. Here, rain is the behavioral attribute and summer is the contextual attribute.

2 Related Work

Outlier detection methods have been a research subject in statistics and data mining from decades before, and now, it is extended to many fields such as deep learning and information theory. Different techniques have been developed to find outliers in various application domains. Hodge and Austin [4] observed outlier detection approaches as extracted mainly from three fields of computing—statistical, neural networks and machine learning. The classification outlier detection approaches done by Hodge and Austin into three categories—Type-I, Type-II and Type-III—are analogous to unsupervised, supervised and semi-supervised approaches, respectively. Patcha and Park [8] accumulated the methods into statistical-based, classification-based, clustering-based and nearest neighbor-based groups. Chandola et al. [1] added two more groups of outlier detection methods to these four groups called spectral methods and information theoretic methods. Han et al. presented methods for detecting collective outliers and contextual outliers in addition to global outliers. They also presented the methods for finding outliers in subspaces and high-dimensional data.

3 Outlier Detection Methods

The categorization of outlier detection methods can be done in different ways. One of these categorizations is based on the availability of class labels. Depending on the availability of class labels, outlier detection methods are grouped into supervised, semi-supervised and unsupervised methods [2].

3.1 Supervised Outlier Detection

Supervised approaches are suitable when the data objects are labeled and can be separated into normal objects and outliers based on the class label. This is a classification-based technique where the labels are used to make the model of the normal class or outlier class. Comparing to the existence of normal objects, the occurrences of outliers are very infrequent, and classification techniques which are able to handle highly imbalanced sets are needed to separate the data objects into normal objects, and outliers are needed in a supervised scenario.

3.2 Unsupervised Outlier Detection

Supervised outlier detection methods cannot be used when the datasets are not labeled. Unsupervised methods are used to find outliers in an unlabeled dataset by assigning each object an outlier score which indicates its degree of outlierness. These scores indicate how much different is a data object from other data objects in the dataset. Unsupervised methods do not use a training dataset, and clustering methods are normally used to find outliers, which are data objects that belong to neither of the clusters or to sparse clusters.

3.3 Semi-supervised Outlier Detection

Semi-supervised methods are used in situations when labels are available for only a small part of normal objects or outliers [5]. The labeled objects, either normal or outlier, are used in the training phase to obtain a model of the normal objects or outliers. This is similar to a binary classification problem in which only one set of objects are labeled. This label information is used to divide the data objects into two sets in which one set contains all the normal objects and the other set contains the outliers.

Researchers have developed a number of algorithms to find outliers using supervised, semi-supervised or unsupervised techniques. Some of them are given in the following sections.

3.4 Local Outlier Factor

Local outlier factor (LOF) algorithm is an efficient method to find outlier detection in moderately high-dimensional datasets [9]. The algorithm computes the outlier score called local outlier factor which reflects outlier degree of the object. It computes the

deviation in the local density of an object with respect to its neighbors. Objects whose local density is much lower than its neighbors are treated as outliers. Usually, the LOF score of an object is calculated by comparing its local density with the average local density of its k-nearest neighbors. The algorithm needs the value k for k-nearest neighbors and the threshold value for the outlier score as input parameters.

3.5 Elliptic Envelope

Outlier detection using elliptic envelope assumes that the normal data objects form a known distribution, like Gaussian distribution [6]. Data objects are classified into normal data objects and outliers based on this assumption in which the normal objects occur in high probability region of the distribution and outliers occur in low probability regions or do not follow this distribution. Elliptic envelope technique fits a robust covariance estimate to the data objects and fits an ellipse to the central data points. After finding the central data point, a distance measure is used to detect outlier degree of the data object. The method needs the percentage of outliers in the dataset as hyper parameter, which is not easy to estimate.

3.6 One-Class Support Vector Machine

One-class support vector machine (SVM) is a classification method when the dataset contains only one class [7]. In this, the support vector machine model is trained to gather the properties of normal data objects only. The goal is to predict whether a data object belong to this class or not. This method can be used to determine the presence of outliers in the dataset. A data object is considered as normal object or outlier depending on whether or not it belongs to the class.

4 Proposed Method

The proposed method is a voting-based ensemble of three existing methods. The technique involves the following steps.

Step 1. Obtain the predicted outlier values of the datasets using three different algorithms—one-class SVM, elliptic envelope and local outlier factor.
Step 2. Outlier is detected by aggregating the results of the individual predictors in the ensemble through voting mechanism.
Step 3. Detect the outlier after reducing the dimensionality of the datasets using principle component analysis and compare the results with the results before reducing the dimensionality.

Table 1 Characteristics of datasets used for experiment

Dataset	#instances	#attributes	Attribute type	#classes	#normal	#outlier
Iris	150	4	Real	3	50	5
Breast cancer	569	32	Real	2	300	30

5 Datasets

Datasets from the UCI machine learning repository are used for our experiments. These datasets are publicly available for experiments. The datasets used are 'iris' dataset and 'breast cancer' dataset. The reason for taking these two datasets is one dataset contains only four attributes, whereas the other contains 32 attributes. This helps to find out the influence of the dimensionality reduction in high-dimensional as well as low-dimensional datasets. The description of the datasets is given in Table 1.

6 Experiments

Experiments using publicly available datasets were conducted in an Intel core i3-based laptop using Python. Separate experiments were conducted for each outlier detection algorithm using different datasets. Only two classes of data objects were used for the experiment. Samples for experiments were taken from the dataset randomly in which about 90% of data belongs to the normal class and the remaining small set as outliers. The algorithms used for detecting outliers were one-class SVM, elliptic envelope and local outlier factor. The results obtained from the individual algorithms were compared with the actual dataset and evaluated the performance of each algorithm.

6.1 Performance of Individual Detectors

Figure 2 shows the scatter plot of the 'iris' data sample used for detecting outliers using different algorithms. The same sample is used for all t three algorithms and also for our proposed ensemble detector. The results obtained from the individual detectors using the 'iris' dataset are shown in Figs. 3, 4 and 5. The evaluation measures used for comparing the performance of individual algorithms are precision, recall and F1-score. The values of these measures obtained from the individual detectors using iris dataset and breast cancer dataset are shown in Tables 2 and 3, respectively.

All the predictors were able to detect the outliers in the iris dataset, but they wrongly classified some of the normal objects as outliers. Elliptic envelope predictor showed a better precision and F1-score. However, the algorithm took more time to complete. Result analysis of different algorithms shows that the performance of

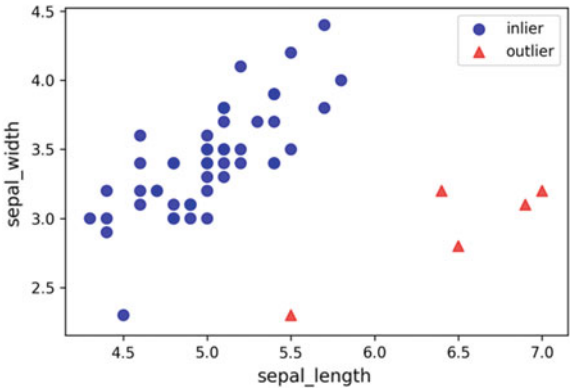

Fig. 2 Scatter plot of iris dataset

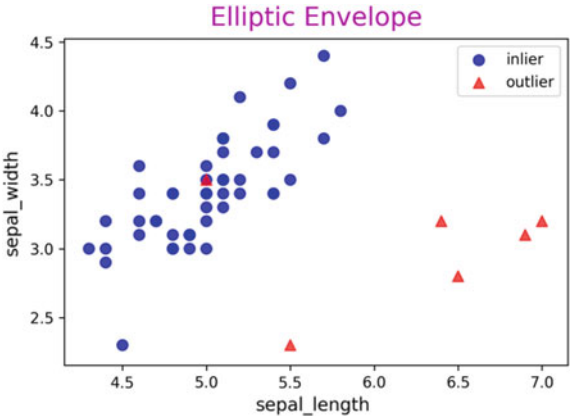

Fig. 3 Scatter plot of the results from elliptic envelope

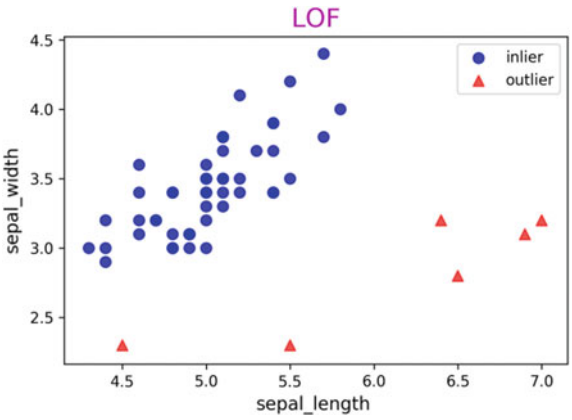

Fig. 4 Scatter plot of the results from LOF

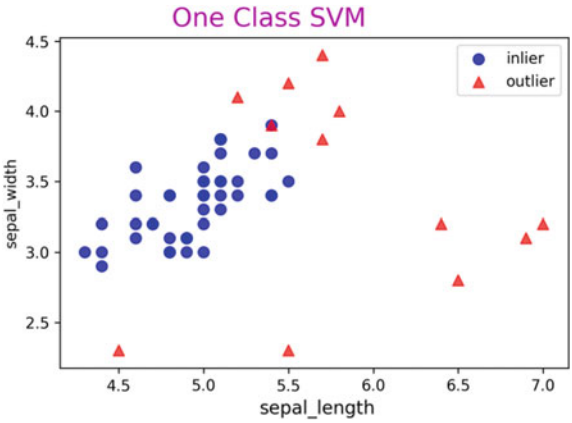

Fig. 5 Scatter plot of the results from one-class SVM

Table 2 Evaluation result of individual predictors using iris dataset

Predictor	Precision	Recall	F1-score	Duration
Elliptic envelope	0.833	1.000	0.909	0.065
One-class SVM	0.417	1.000	0.588	0.003
LOF	0.833	1.000	0.909	0.003

Table 3 Evaluation result of individual predictors using breast cancer dataset

Predictor	Precision	Recall	F1-score	Duration
Elliptic envelope	0.576	0.633	0.603	0.112
One-class SVM	0.485	0.533	0.508	0.005
LOF	0.485	0.537	0.507	0.006

an algorithm depends on the nature of the dataset. Hence, the performance can be improved by combining the results of various algorithms together. The following subsection shows the method used for combining the outputs of the predictors and also the results obtained from the proposed method.

6.2 Performance of Proposed Ensemble Detector

In the proposed ensemble detector, the outlier values are calculated by aggregating the results of the individual predictors in the ensemble through voting mechanism. Here, an object is considered as outlier if majority of the individual algorithms predicted that object as outlier; otherwise, it is considered as a normal object. The results obtained by the proposed ensemble detector are shown in Table 4. The precision

Table 4 Evaluation result of the proposed ensemble detector

Dataset	Precision	Recall	F1-score	Duration
Iris	1.000	1.000	1.000	0.071
Breast cancer	0.739	0.567	0.642	0.123

and F1-score of the proposed ensemble detector are higher than all other detectors. The proposed ensemble detector is a better method as it outperforms the individual detectors.

6.3 Effect of Dimensionality Reduction

Principal component analysis (PCA) is a popular method for reducing the dimensions of datasets [3]. Dimensionality reduction plays an important role in data mining as it finds the relationship among the attributes and helps in data visualization. Here, the performance of the outlier detection algorithms is evaluated after applying dimensionality reduction in the dataset. All three algorithms as well as the proposed ensemble method are used for analyzing the influence of dimensionality reduction in outlier detection techniques. Both the datasets, iris and breast cancer are used for this experiment. Experiments were performed by reducing to different dimensions, and the results obtained after reducing the dimensions are shown in Tables 5 and 6. The effect is more clear from the breast cancer dataset as the original dataset contained 32 attributes and reduced to much lower dimensions. Dimensionality reduction decreased the execution time of the outlier detection algorithms for both the datasets. The performance is improved for the breast cancer dataset, while there is little effect on the performance for iris dataset.

Table 5 Effect of dimensionality reduction on predictors (iris dataset)

Dimensionality	Predictor	Precision	Recall	F1-score	Duration
Dimension reduced from 4 to 3	Elliptic envelope	0.833	1.000	0.909	0.037
	One-class SVM	0.455	1.000	0.625	0.001
	LOF	0.833	1.000	0.909	0.001
	Proposed ensemble	0.833	1.000	0.909	0.039
Dimension reduced from 4 to 2	Elliptic envelope	0.833	1.000	0.909	0.036
	One-class SVM	0.455	1.000	0.625	0.001
	LOF	0.833	1.000	0.909	0.001
	Proposed ensemble	0.833	1.000	0.909	0.038

Table 6 Effect of dimensionality reduction on predictors (breast cancer dataset)

Dimensionality Reduction	Predictor	Precision	Recall	F1-score	Duration
Dimension reduced from 32 to 16	Elliptic envelope	0.636	0.700	0.667	0.074
	One-class SVM	0.471	0.533	0.500	0.003
	LOF	0.485	0.533	0.508	0.004
	Proposed ensemble	0.800	0.667	0.727	0.081
Dimension reduced from 32 to 10	Elliptic envelope	0.697	0.767	0.730	0.054
	One-class SVM	0.500	0.533	0.516	0.002
	LOF	0.485	0.533	0.508	0.003
	Proposed ensemble	0.815	0.733	0.772	0.059

7 Conclusion

In this paper, we studied three different techniques for detecting outliers and analyzed the performance of these detectors individually and also on our proposed ensemble technique using two publicly available datasets, namely iris and breast cancer. We also studied the influence of dimensionality reduction in detecting outliers using our proposed ensemble detector and also with the individual outlier detectors. We have found that our proposed ensemble-based detector outperforms individual outlier detectors, and we were able to achieve higher precision and F1-score than the scores of individual detectors. After applying dimensionality reduction, it is found that even though there is not much difference in the evaluation score for low-dimensional dataset, there is a reduction in execution time of the algorithms. The execution time as well as the performance of the outlier detection algorithms has improved much better for higher-dimensional data as a result of dimensionality reduction. The observations showed that dimensionality reduction has an important role in high-dimensional data, and it has little effect on low dimensional data.

References

1. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(3):1–58
2. Han J, Kamber M, Pei J (2012) Data mining: concepts and techniques. Morgan Kaufmann, Massachusetts, US
3. Hauberg S, Feragen A, Enficiaud R, Black MJ (2016) Scalable robust principal component analysis using Grassmann averages. IEEE Trans Pattern Anal Mach Intell 38(11):2298–2311. https://doi.org/10.1109/tpami.2015.2511743
4. HodgeVJ Austin J (2004) A survey of outlier detection methodologies. Artif Intel Rev 22(2):85–126
5. Ienco D, Pensa RG, Meo R (2017) A semi-supervised approach to the detection and characterization of outliers in categorical data. IEEE Trans Neural Networks Learn Syst 28(5):1017–1029. https://doi.org/10.1109/tnnls.2016.2526063

6. Kumar R, Kundu PP, Phoha VV (2018) Continuous authentication using one-class classifiers and their fusion. In: IEEE 4th international conference on identity, security, and behavior analysis (ISBA). https://doi.org/10.1109/isba.2018.8311467
7. Mũnoz-Marí J, Bovolo F, Gómez-Chova L, Bruzzone L, Camp-Valls G (2010) Semisupervised one-class support vector machines for classification of remote sensing data. IEEE Trans Geosci Remote Sens 48(8):3188–3197. https://doi.org/10.1109/tgrs.2010.2045764
8. Patcha A, Park J-M (2007) An overview of anomaly detection techniques: existing solutions and latest technological trends. In: Computer networks
9. Radovanovic M, Nanopoulos A, Ivanovic M (2015) Reverse nearest neighbors in unsupervised distance-based outlier detection. IEEE Trans Knowl Data Eng 27(5):1369–1382. https://doi.org/10.1109/tkde.2014.2365790