Impact of Dimensionality Reduction On Outlier Detection Techniques

M. M. Tavakoli[1], Ashkan Sami[2]
[1]*MSc. CSE Department of Shiraz University, Kerman, Iran*
mtavakoli@cse.shirazu.ac.ir
[2]*Assistant Professor CSE Department of Shiraz University, Shiraz, Iran*
asami@ieee.org
Paper Reference Number: 07-02-9999
Name of the Presenter: Mohammad Mahdi Tavakoli

Abstract

In statistics, an outlier is an observation that is distant from the rest of the data. In other word mining or detecting outliers is referred to sequence of operations that lead to find exceptional objects that deviate from the rest of the data set. In this paper we use a correlation based feature selection algorithm, for dimensionality reduction as a preprocessing phase to outlier detection algorithms, which causes improvement on results of different types of outlier detection techniques.

Key words: Outlier Detection, Dimensionality Reduction, Unsupervised Learning

1. Introduction

According to Hawkins's (1980) definition, "*an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism*". From statistical intuition, normal data base objects generating mechanism are deviate from this and abnormal objects follow a generating mechanism.

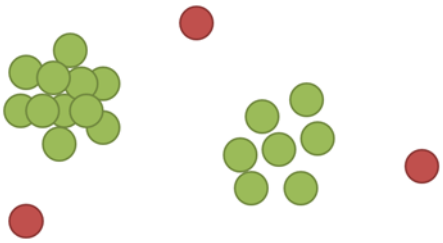


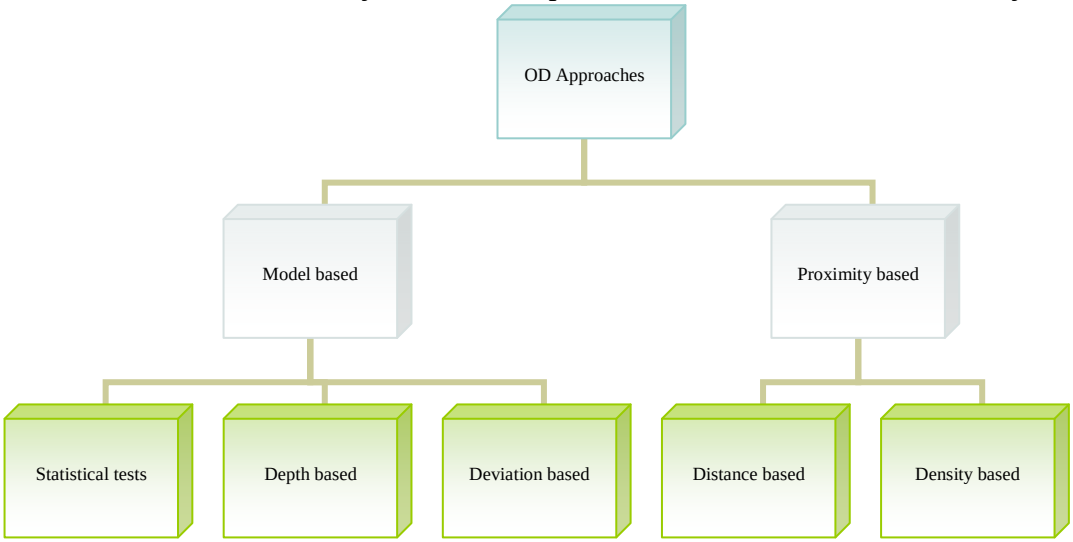
Fig. 1: intuition of outliers

The outlier detection technique as a subset of data mining techniques has many applications in actual world. For example according to survey of H. P. Kriegel et al. (2010), fraud detection, medicine, public health, sport statistics and detecting measurement errors, are some important applications of outlier detection.

The general scenarios for data mining techniques such as outlier detection are considered by V. J. Hodge and J. Austin (2004) and H. P. Kriegel et al. (2010):

- Supervised: in this case we assume that we have some pre labeled data with both normal and abnormal objects with this problem that these two classes of objects are highly imbalanced. There are many classification approaches for these types such as neural networks, decision trees, support vector machines and etc.
- Semi – supervised: models only normality or in few cases abnormality. It is considered semi – supervised as the normal class taught, but the algorithm learns to recognize abnormality.
- Unsupervised: in most applications there are no training data. The approach processes the data as a static distribution, pinpoints the most remote points, and flags them as potential outliers.

Based on another classification scheme by V. J. Hodge and J. Austin (2004) and H. P. Kriegel et al. (2010), the outlier detection techniques are classified as shown in Fig. 2:



In this paper we demonstrate the impact of dimensionality reduction on some unsupervised outlier detection techniques with different types of considered classification in fig. 2. Based on best of our knowledge, we did not find any similar prior work that address preprocessing for outlier detection techniques.

2. Literature Review

We investigate a spectrum of unsupervised outlier detection algorithms in order to demonstrate the influence of dimensionality reduction by feature subset selection on quality of results. One of the most popular OD algorithms is K – nearest neighbor algorithm that proposed by Sridhar Ramaswamy et al. (2000). this algorithm first partitions the input data set into separate subsets, and then prunes entire partitions as soon as it is determined that they cannot contain outliers. Markus M. Breunig et al. (2000)

Fig. 2: Classification of outlier detection techniques

employed a density based OD algorithm to find outliers. Recently, several works on outlier detection have been focused on objects that have significantly lower density than their local neighborhood. These objects are called essentially local outliers. As an objective measure, the degree of outlierness of an object p is defined to be the ratio of its density and the average density of its neighboring objects. Ranking Outliers Using Symmetric Neighborhood Relationship is another density based OD algorithm that is popular to INFLO and proposed by Wen Jin et al. (2006). To get a better estimation of the neighborhood's density distribution, they propose to take both the nearest neighbors (NNs) and reverse nearest neighbors (RNNs) which is investigated by F. Korn and S. Muthukrishnan (2000), into account. The RNNs of an object p are essentially objects that have p as one of their k nearest neighbors. For case of high dimensional data an angle based algorithm proposed by Hans-Peter Kriegel et al. (2008) named ABOD. The main contribution to detecting outliers in their approach is in considering the variances of the angles between the difference vectors of data objects. They claim that in high dimensional space, angles are less influenced than distance metrics such as Euclidean distance.

3. Data and Material

The data set used for evaluation is modified Breast Cancer Wisconsin (Diagnostic) Data Set (Wdbc) from UCI machine learning repository. The Wdbc data set is a medical data set and has 569 records and 32 attributes. The objects in Wdbc divided in two classes of benign and malignant that number of records for each of them is 357 and 212 respectively. We create a modified data set from Wdbc with 357 records of benign and first 10 malignant records, named Wdbc-v2. As this modified data set is highly imbalanced, we could use it for outlier detection, and compare the result of outlier detection algorithms before and after performing dimensionality reduction algorithm. This kind of data set generation for evaluating outlier detection is common in similar experiments such as K. Zhang et al. (2009).

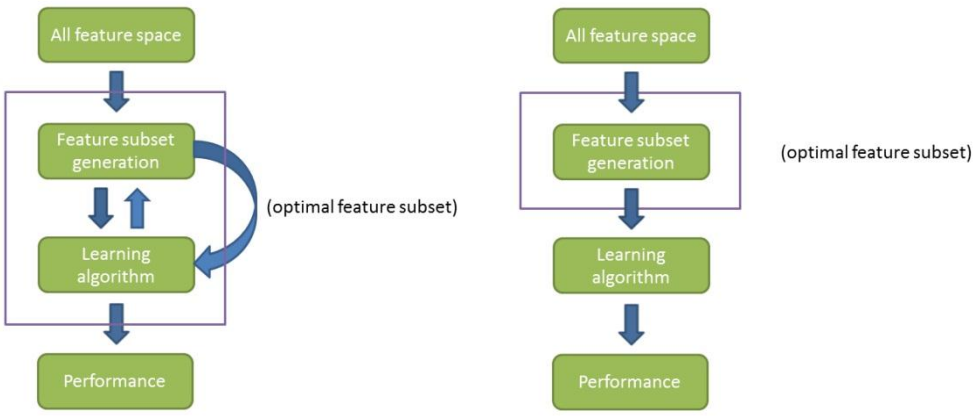
4. Research Methodology

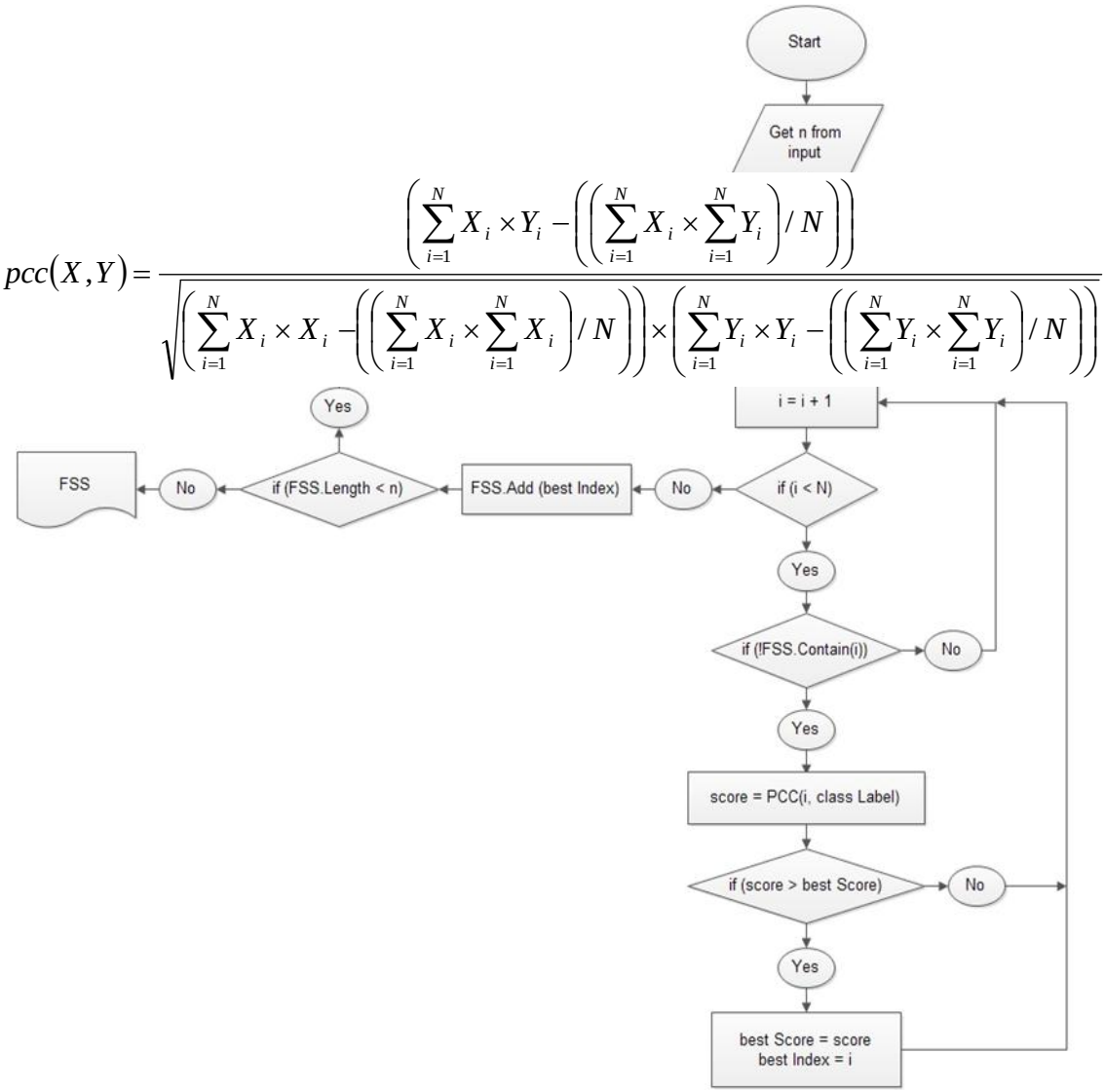
There are two main categories of feature subset selection techniques that could be used for dimensionality reduction (Fig. 3): filter and wrapper approaches that are reviewed by B. Kumari and T. Swarnkar (2011). Filter methods generally involve a non-iterative computation on the data set, which can execute much faster and is more general than a wrapper method, due to filters evaluate the fundamental properties of the data, rather than their interactions with a particular classifier, thus their results present more generality and lead to generation of more meaningful feature subsets. The drawback of filter methods is the tendency to select large subsets. This forces the user to select an arbitrary cutoff on the number of features to be selected. We want to use feature subset selection as a preprocessing session to the outlier detection algorithms. In case of high dimensional problems, the number of attributes causes complexity of learning, clustering and etc. to grow exponentially, this event is generally called “Curse of dimensionality” with Aggarwal and Yu (2005) literature. So in data sets with large attributes, dimensionality reduction could outperform the underlying problem.

Fig. 3: Wrapper approach in left and filter in right

4 – 1. Greedy forward selection

In order to select subset of features among feature space, we use a greedy forward selection algorithm with Pearson correlation coefficient as evaluation function that described in next section Eq. 1. This algorithm gets the number of features to be returned as input argument, and return the most important features as output. The steps of algorithm are shown in Fig. 4.





$$pcc(X,Y)=\frac{\left(\sum_{i=1}^{N}X_i\times Y_i-\left(\left(\sum_{i=1}^{N}X_i\times\sum_{i=1}^{N}Y_i\right)/N\right)\right)}{\sqrt{\left(\sum_{i=1}^{N}X_i\times X_i-\left(\left(\sum_{i=1}^{N}X_i\times\sum_{i=1}^{N}X_i\right)/N\right)\right)\times\left(\sum_{i=1}^{N}Y_i\times Y_i-\left(\left(\sum_{i=1}^{N}Y_i\times\sum_{i=1}^{N}Y_i\right)/N\right)\right)}}$$

Fig. 4: Steps for greedy forward selection algorithm

4 – 2. Evaluation metric for feature selection

In statistics, dependence refers to any statistical relationship between two random variables or two sets of data. We use Pearson correlation coefficient as degree of correlation between each feature and class label and as our evaluation metric for feature selection that giving a value between −1 and +1 inclusive.

Assume that X and Y are two objects with N features or attributes, the Pearson correlation coefficient between these objects calculated as below:

$$X=(X_1,X_2,X_3,...,X_N)$$

$$Y=(Y_1,Y_2,Y_3,...,Y_N)$$

$$pcc(X,Y)=\frac{\left(\sum_{i=1}^{N}X_i\times Y_i-\left(\left(\sum_{i=1}^{N}X_i\times\sum_{i=1}^{N}Y_i\right)/N\right)\right)}{\sqrt{\left(\sum_{i=1}^{N}X_i\times X_i-\left(\left(\sum_{i=1}^{N}X_i\times\sum_{i=1}^{N}X_i\right)/N\right)\right)\times\left(\sum_{i=1}^{N}Y_i\times Y_i-\left(\left(\sum_{i=1}^{N}Y_i\times\sum_{i=1}^{N}Y_i\right)/N\right)\right)}}$$

(1)

5. Results and Analysis

Based on best of our knowledge, we did not find any similar prior work that address preprocessing for outlier detection techniques. We applied our algorithm as a preprocessing phase to some unsupervised outlier detection algorithms that the output of them are the outlierness of each object in data set. We select features that have more than 20 percent correlation to class label. The features that selected by algorithm are twenty features by zero based indexes consist of: 2, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 22, 24, 25, 26, 27, 28, 29, 30, and 31.

We write our algorithm by Net Beans IDE 7.0. As we mention in section 1, we only considered unsupervised approaches that all of them compute the outlierness of each object and then sort the results, so the most probability outlier objects are in the beginning of output list. In order to compare results, we define some comparative metrics such as the position for first and last outliers that detected by algorithm, mean, median and standard deviation of outlier positions in output list. The obtained results for different values of nearest neighbors k = 5, k = 10, and k = 15 are shown in table 1, table 2 and table 3 respectively and the best results for each approach is shown in table 4. As you see in tables below, we compare results of four outlier detection techniques (ABOD, LOF, INFLO, and KNN) before and after performing dimensionality reduction by feature subset selection (ABOD – DM, LOF – DM, INFLO – DM, and KNN – DM) and then we highlight the cases that lead to obtain better results. The last column in tables below is the position of outliers detected by each approach.

Approach	First	Last	Mean	Median	STD	Outlier indexes
ABOD	2	94	23	7	31.5	[2, 3, 4, 5, 6, 8, 15, 28, 65, 94]
ABOD – DM	2	65	19.3	7	22.1	[2, 3, 4, 5, 6, 8, 15, 28, 57, 65]
LOF	2	57	35	41	22.2	[2, 10, 12, 22, 29, 53, 54, 55, 56, 57]
LOF – DM	2	65	39.2	48	23.9	[2, 6, 12, 29, 38, 58, 59, 60, 63, 65]
INFLO	2	141	93.5	103.5	49.3	[2, 37, 56, 79, 99, 108, 138, 139, 140, 141]
INFLO – DM	2	137	91.4	103.5	46.1	[2, 32, 53, 78, 101, 106, 134, 135, 136, 137]
KNN	1	29	12.4	9.5	10.3	[1, 2, 3, 4, 5, 14, 17, 21, 28, 29]
KNN – DM	1	29	12.4	9.5	10.3	[1, 2, 3, 4, 5, 14, 17, 21, 28, 29]

Table 1. Results for k = 5

Approach	First	Last	Mean	Median	STD	Outlier indexes
ABOD	1	44	13.8	7.5	14.4	[1, 2, 3, 4, 5, 10, 16, 22, 31, 44]
ABOD – DM	1	41	13.8	8	13.04	[1, 2, 3, 4, 5, 11, 17, 26, 28, 41]
LOF	1	18	8.5	8	6.1	[1, 2, 3, 4, 6, 10, 11, 14, 16, 18]
LOF – DM	1	17	8.5	8	5.7	[1, 2, 3, 4, 6, 10, 11, 15, 16, 17]
INFLO	1	147	33.2	19.5	41.5	[1, 10, 16, 17, 18, 21, 27, 35, 40, 147]
INFLO – DM	2	146	31.3	17	39.5	[2, 9, 13, 14, 15, 19, 25, 33, 37, 146]
KNN	1	38	13.1	9	11.8	[1, 2, 3, 4, 5, 13, 17, 21, 27, 38]
KNN – DM	1	35	12.8	9	11.2	[1, 2, 3, 4, 5, 13, 17, 21, 27, 35]

Table 2. Results for k = 10

Approach	First	Last	Mean	Median	STD	Outlier indexes
ABOD	1	38	13.5	8	13.5	[1, 2, 3, 4, 5, 11, 15, 22, 34, 38]
ABOD – DM	1	38	13.3	8	12.6	[1, 2, 3, 4, 5, 11, 15, 21, 33, 38]
LOF	1	21	9.3	7	7.4	[1, 2, 3, 4, 5, 9, 13, 17, 18, 21]
LOF – DM	1	21	9.4	7	7.05	[1, 2, 3, 4, 5, 9, 15, 16, 18, 21]
INFLO	1	27	12.3	11.5	8.8	[1, 4, 5, 6, 9, 14, 15, 17, 25, 27]
INFLO – DM	1	26	11.6	11	7.8	[1, 4, 5, 6, 10, 12, 13, 15, 24, 26]
KNN	1	42	14.4	9.5	13.5	[1, 2, 3, 4, 5, 14, 18, 22, 33, 42]
KNN – DM	1	40	14.2	9.5	13.1	[1, 2, 3, 4, 5, 14, 18, 22, 33, 40]

Table 3. Results for k = 15

In table 4 we demonstrate the best results obtained for each approach with equivalent values of k, for full feature space and reduced feature space.

k	Approach	First	Last	Mean	Median	STD	Outlier indexes
15	ABOD	1	38	13.5	8	13.5	[1, 2, 3, 4, 5, 11, 15, 22, 34, 38]
15	ABOD – DM	1	38	13.3	8	12.6	[1, 2, 3, 4, 5, 11, 15, 21, 33, 38]
10	LOF	1	18	8.5	8	6.1	[1, 2, 3, 4, 6, 10, 11, 14, 16, 18]
10	LOF – DM	1	17	8.5	8	5.7	[1, 2, 3, 4, 6, 10, 11, 15, 16, 17]
35	INFLO	1	23	8.7	6.5	7.4	[1, 2, 3, 4, 5, 8, 10, 12, 19, 23]
35	INFLO – DM	1	23	8.5	6	7.04	[1, 2, 3, 4, 5, 7, 9, 12, 19, 23]
5	KNN	1	29	12.4	9.5	10.3	[1, 2, 3, 4, 5, 14, 17, 21, 28, 29]
5	KNN – DM	1	29	12.4	9.5	10.3	[1, 2, 3, 4, 5, 14, 17, 21, 28, 29]

Table 4. Best results for each approach

6. Conclusion

Outlier detection is an attractive branch of data mining techniques. We consider data sets with highly imbalanced class labels as good benchmarks for outlier detection algorithms. In these data sets we assume positive class as outlier objects. In this paper we use a feature subset selection for dimensionality reduction, and as a preprocessing session to some unsupervised outlier detection techniques, which as you see in previous section, the result of these approaches after using dimensionality reduction in all of cases except two of them improved. According to section 4, in our approach we use some part of data with class labels to select most relevance features. The future work could be an unsupervised feature selection approach that does not need class labels for feature selection, and could be applied to unsupervised algorithms.

References

Hawkins, D. (1980). Identification of outliers, Chapman and Hall, London. 188p.

Hodge, V. J., & Austin, J. (2004). A survey of Outlier Detection Methodologies, EPSRC Grant No.GR/R55191/01, Kluwer Academic Publishers Printed in the Netherlands.

Kriegel, H. P., Kröger, P., & Zimek, A. (2010). Outlier detection techniques, 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets, In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers, In Proc. ACM, SIGMOD Conf. 2000, pages 93–104.

Jin, W., Tung, A. K. H., Han, J., & Wang, W. (2006). Ranking outliers using symmetric neighborhood relationship, In Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore.

Korn, F., & Muthukrishnan, S. (2000). Influence Sets Based on Reverse nearest Neighbor Queries, SIGMOD.

Kriegel, H. P., Schubert, M., & Zimek, A. (2008). Angle-Based Outlier Detection in High-dimensional Data, 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'08), Las Vegas, NV, pp. 444-452.

Zhang, K., Hutter, M., & Jin, H. (2009). A new local distance based outlier detection approach for scattered real world data, In Proc. PAKDD.

Kumari, B., & Swarnkar, T. (2011). Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 2 (3), 1048-1053

Aggarwal, C. C., & Yu, P. S. (2005). An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal 14: 211–221 / Digital Object Identifier (DOI) 10.1007/s00778-004-0125-5.