ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(25), DOI: 10.17485/ijst/2015/v8i25/80260, October 2015* ISSN (Online) : 0974-5645

Association-based Outlier Detection for Mixed Data

Young-Gi Kim and Keon Myung Lee*

Department of Computer Science, Chungbuk National University, Korea;
kmlee@cbnu.ac.krr

Abstract

Various methods have been developed to detect outliers which are significantly different from others. Most outlier detection methods assume the data lie in Euclidean space in which distances can be easily defined and computed. In reality, we meet many data with both numerical and categorical attributes together, so-called mixed-data, for which it is not easy to define widely-accepted distance metrics. This paper proposes an outlier detection method which can be applied to mixed data. The method focuses on the association among attribute values. It first selects the sets of potentially associated attributes, computes the degrees of outlierness for records with respect to the associated attributes, and then determines a collection of outliers using the degrees. In addition, this paper shows some experiment results of the proposed method and compares with some other methods.

Keywords: Data Analysis, Data Quality, Horizontal Consistency, Mixed Data, Outlier Detection

1. Introduction

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[1]. In such applications as fraud detection system and intrusion detection system, outliers are main targets for searching whereas outliers might be just erroneous data to be removed in some domains. Various outlier detection methods have been developed[2]. Most of them are assumed that data lie in Euclidean space in which distances can be easily defined and computed. In practice, a data object consists of both numerical and categorical attributes, in which it is hard to define widely-accepted distance. On the other hand, due to the curse of dimensionality[3], the distance is not a good indicator for the proximity of data objects in high dimensional space. In addition, high dimensional data might containnull values so that it is difficult to compute distances between data.

This paper is concerned with outlier detection for mixed data which contain both numerical and categorical attributes. For such data, the outlier detection methods for Euclidean space data cannot be applied. In mixed data, all attributes are not always related, but only some subsets of attributes are related or associated. Suppose that there are mixed data with attributes gender, disease, scholastic achievement score, Intelligence Quotient (IQ), and residence type. The attribute sets {gender, disease} and {scholastic achievement score, IQ} are associated ones, respectively. If a data has rare values for such attribute sets, the data might be an outlier. For example, if gender is male and his disease is breast cancer, such a data might be an outlier because breast cancer is generally a female disease regardless of what values other attributes have. The proposed method does not use a distance metric and pays attention to the rareness of values in the association attribute values. The method first uses a frequent pattern mining technique to determine the sets of associated attributes. Then, the degree of outlierness for each data is computed with respect to the associated attribute sets. To evaluate the outlierness, the so-called saliency degree is used. To get the final outlierness degree, the computed saliency degrees are aggregated.

The remainder of the paper is organized as follows: Section 2 briefly describes the outlier detection approaches. Section 3 presents the proposed outlier detection method

* *Author for correspondence*

in detail. Section 4 shows the experiment results compared with four existing methods and in final Section 5 draws conclusions.

2. Related Works

Outlier detection has been actively studied due to its practical applicability in various domains. The methodologies for outlier detection can be classified into statistical approaches, depth-based approaches, deviation-based approaches, distance-based approaches, cluster-based approaches, density-based approaches, high-dimensional approaches, and neural network-based approaches[2]. Most methods can be applicable only to numerical data for which distance functions can be easily defined. Few outlier detection methods have been developed for categorical and mixed data.

The statistical approaches build a statistical distributionmodel for the given data set and regard as outliers the data points with low probability with respect to the distribution. When a multivariate Gaussian is used to model the distribution, the Mahalanob is distance of a data point to the distribution is assumed to follow the a χ^2-distribution and the χ^2-test[3] can be applied to sort out potential outliers. Some models take non-parametric forms such as histogram-based methods[4].

The depth-based approaches organize nested convex hulls for the data set layer by layer[5]. The data points located at the border of the outer convex hulls become the candidates for outliers. It is usually not easy to construct convex hulls in higher dimensional spaces.

The deviation-based approaches determine outliers by searching for data points of which removal minimizes the variance of the data set[6]. The smoothing factor[6] is computed for each subset of data set, which indicates how much the variance of the remaining data set is reduced when the designated subset is removed. The potential outlier set is set to be the maximal subset with the largest smoothing factor. It is, however, expensive to compute directly the smoothing factors for every possible subset. Hence some heuristics have been applied to reduce the computational cost.

The Distance-based approaches make use of distances of data to their neighbors to judge outlierness. They assume that normal data points have a dense neighbor and outliers are far apart from their neighbors. The nested loop method[7] first computes all pair-wise distances and chooses the data points which are far away from the others. The *k*NN(*k*-th nearest neighbor) distance has sometimes been used, which is the distance to the *k*-th nearest neighbor. The data points with large *k*NN distances can be selected as outliers. Much effort have been paid to efficiently rank neighbors because *k*NN can be found in the ranked sequence of neighbors[7,8]. Some efficient indexing structures such as locality-sensitive hashing[9,10,19] have been developed to support distance-based outlier detection. Distance-based approaches cannot handle successfully the data set with different density distributions[11].

The clustering-based approaches construct clusters for the given data set and regard isolated small data sets as outliers. Distance-based clustering algorithms like *k*-means algorithm form spherical clusters whereas connectivity-based clustering algorithms like ROCK[6]and DBSCAN[6] form arbitrarily-shaped clusters in data space. FindCBLOF is a cluster-based outlier detection method that uses a measure called CBLOF to evaluate the degree of outlierness[27]. The cluster-based approaches usually ignore the density of data distribution.

The density-based outlier detection approaches take into account the density around a point and the density around its local neighbors. They are based on the assumption that the density around a normal data is similar to the density around its neighbors, and outliers have considerably different density compared to its neighbors[5] Local Outlier Factor (LOF)[11] is a ranking metric for outlier selection in which the reachability distance between two data points is defined as the maximum of the *k*NN distance of a data point and the distance between both points, local reachability distance of a data point is defined as the inverse of the average reachability distance of its *k* neighbors, and the LOF of a point is the average ratio of local reachability distance of its neighbors to that of the data point. The larger LOF, the more likely the data point is to be an outlier.

As a data set comes from higher dimension, it suffers from the curse of dimensionality[3], which means that the distribution of pair-wise data distances comes to get less contrasted with increasing dimensionality and thus the notion of neighborhood becomes meaningless[2]. Several outlier detection methods have been developed for high dimensional spaces[13-15]. Angles are more stable in high dimensional space than distance. The angle-based outlier detection method[13] computes the angles to other data, and regards data points as outliers when most other data are located in similar directions. Some method[14] partitions data space by an equi-depth grid and evaluates sparsity of each grid. The grid with lowest sparsity may contain outliers, for which outliers are searched.

Frequency-based methods use the frequency information of data values to determine outliers. Local Search Algorithm (LDA) is a method to find an outlier subset O of data set D which minimizes the entropy of $D–O$[27]. Find FPOF is a method that uses the frequent itemsets of association rule mining, in which a measure called FPOF (Frequent Pattern Outlier Factor) is defined as the ratio of the frequent item sets covered by a data record to the total number of frequent patterns[25].

Neural network-based approaches use neural network to model normal data and detect the outliers by measuring how much data deviates from the model. Replicator neural network is a typical model which can be applied in outlier detection[24]. It is an autoencoder model that is a multi-layer Perceptron for which input and output are the same in training data. If a new data is feed into the trained replicator network, it is expected to get similar output. Otherwise, the data is decided as an outlier.

3. The Proposed Outlier Detection Method

3.1 The Framework of The Proposed Method

Each data record contains multiple numerical and categorical attributes. The outlier detection methods for data having only numerical attributes assume that data lie in Euclidean space. To determine outliers, they make use of proximity, distance, or density information which is defined in Euclidean space. Such measurements in Euclidean space basically treatall attributes as making the same amount of contribution. In reality, it is not the case. Contributions of attributes usually are not the same and in addition all attributes are not always correlated together. There are some subsets of attributes which are associated together. When some data is deviated from the associated attribute subsets, the data could be regarded as an outlier. From these observations, we propose the outlier detection method which finds the outliers by examining the data with respect to the associated attribute subsets. In order to implement this idea, we need the techniques to identify associated attribute subsets, to evaluate the degree of outlierness of data with respect to an attribute set, and to decide whether a data is an outlier.

To identify the associated attribute subsets, the method uses a frequent itemset mining technique[6] for a sampled data set from the given data set. It selects the associated attribute sets by taking out the attribute names from the obtained frequent itemsets. In order to compute the degree of outlierness, a new measure calledsaliency degree is proposed. The degrees of saliency for a data record are aggregated to determine the overall degree of outlierness.

3.2 How to Determine the Associated Attribute Subsets

The proposed method examines the values of associated attributesto see whether it is deviated from the others. It is different from the existing methods which treat a data as a single vector in a space. The proposed method focuses on associated subspaces of the data space. In order to determine the associated attributes, it first discretizes the continuous attribute domains into intervals and treats them as categorical ones. Each interval is labeled with a pair of the attribute name and an interval label. This preprocessing operation transforms the data set into a data set with only categorical attributes. To find the frequent itemsets from the transformed data set, a frequent pattern mining algorithm such as Apriori or FP Growth is applied. For the frequent itemsets with more than 2 items, the method picks out the attribute names from frequent itemsets. The sets of selected attributes are the associated attributes for which we examine the data objects. The following procedure *Determine_Associated_Attributes* shows how to determine the associated attribute sets:

procedure *Determine_Associated_Attributes*
input : data set D, minimum support ms, minimum cover count mc
output : associated attribute sets AS
begin
$AS \leftarrow \varnothing$
for each numerical attribute A
Discretize the domain of attribute$A$ into intervals
for each value v of attribute A for each data d_i
Replace v of $d_i$with the pair (A, $v.label$) where $v.label$ is the assigned label for the interval for v
Apply a frequent pattern mining algorithm to the transformed data set D with the minimum support ms.
$FIS \leftarrow$ the mined frequent itemsets
$CS \leftarrow \varnothing$
foreach frequent item set IS of FIS
Pick out the attribute names IN from IS.
if $IN \in CS$
count(IN) ≥ count(IN) +1
else
$CS \leftarrow CS \cup \{IN\}$

count(IN) ← 1
for each $IN \in CS$
ifcount(IN) ≥mc
$AS \leftarrow AS \cup \{IN\}$
end.

3.3 How to Evaluate the Degree of Outlierness

To evaluate the outlierness of a combination of values for associated attributes, a measure called the saliency degree is proposed. The measure gives higher values for rare cases whereas it gives lowervalues for frequent cases. For the associated attributes, the rare value combinations are potentially outliers. For an associated attribute set AS_j and its value combination a_i, the saliency degree $sd(AS_j, a_i)$ is defined as follows:

$$sd(AS_j, a_j) = \begin{cases} \frac{1}{p_i}\left(-\log_2 p_i - \log_2 |v(AS_j)|\right) & \text{if } p_i \le \bar{p} \\ \text{otherwise} \end{cases} \quad (1)$$

where AS_j denotes an associate attribute set, $v(AS_j)$ is the set of distinct attribute value combinations for AS_j that occur in data set D, p_i is the proportion that attribute value combination a_i takes in D, and is $1/|v(AS)|$ which corresponds to the probability to see a value combination in the uniform distribution. In Equation (1), $-log$ is the amount of information according to the information theory and gives high value to rare cases. $-(1/)$ log is sensitive to smaller probability value. $-log_2|v(A)$ is the normalizing term which it is the maximum amount of information. If $p_i \le$, it is the case that the corresponding value combination has higher frequency than at the uniform distribution. Therefore, that case is treated as a normal case by giving 0 as the outlierness degree.

3.4 How to Determine the Overall Outlierness

To determine the overall outlierness degree $od(d_i)$ for a data record d_i, the saliency degrees for the associated attribute sets are aggregated by adding up them as follows:

$$od(d_i) = \sum_{ASj \in AS} sd(AS_j, d_i(AS_j)) \quad (2)$$

Where AS denotes the collection of the sets of associated attributes, AS_j is an associated attribute set, and $d_i(A\ S_j)$ is the attribute values of data d_i for attribute set AS_j. The higher the overall outlierness degree, the more likely the data is an outlier.

The outliers are reported by selecting top k data with respect to the overall outlierness degree, but there are no reasons for the data set to have k outliers. When the overall outlierness degree of a data record is much bigger than other data records, the data record is an outlier. In order to determine the top k data are outliers, the following strategy is used: First, the data are sorted in decreasing order of the overall outlierness degree and let the sorted data denoted by $DS = \{d(1), d(2), \ldots, d(m)\}$, where $od(d(i))) \ge od(d(j)$ if $i<j$. Then, select the top αk data from the sorted data, where α is a constant greater than 2. For the top k data to be outliers, their overall outlierness degree should be much greater than the remaining $(\alpha-1)$ k data in the selected data set. To compute the degree of contrastness between top k data and the remaining $(\alpha-1)k$ data, the overall outlierness degrees $nod(d_{(i)})$are normalized to be in the range [0,1] by dividing them by the largest degree as follows:

$$nod(d_{(i)}) = \frac{od(d_i)}{od(d_j)} \quad (3)$$

The degree $cd(k,\alpha)$ of contrastness is defined as follows:

$$cd(k, \alpha) = \sum_{i=1}^{\alpha k}\left(1 + \frac{i}{\alpha k}\right) * nod(d_{(i)}) * \frac{1}{\alpha k} \quad (4)$$

In Equation (4), $(1+i/\alpha k)$ is the weighting term to give penalty to uniform distributed data with respect to $nod(d_{(i)})$ and $cd(k,\alpha)$plays the role of computing the area under the curve of $nod(d_{(i)})$ weighted by $(1+i/\alpha k)$. The smaller the degree of contrastness, the larger possible for the top k data is.

4. Experiments

To show the performance of the proposed method, it was implemented and tested over two benchmark problems from the UCI Machine Learning Repository[28]: Wisconsin breast cancer data and Lymphography data. Wiscon breast cancer data set has 699 instances with 9 attributes, among which 458 instances are labeled as benign and 241 labeled as malignant. To compare with other existing experiment results, the test data set was sampled to have 444 benign instances and 39 malignant instances as in the experiment setting by He et al[27]. The lymphography data set consists of 148 instances with 18 categorical attributes and 4 classes (fibrosis, malign_lymph, metastasis, normal), in which the rare classes (4 fibrosis, 2 normal) were set as outliers.

The proposed method was implemented using R

language. To determine the associated attribute sets, the R-package 'arule' was used for frequent itemset mining with the minimum support 0.5. Figure 1 shows the computed overall outlierness degrees for the used Wisconsin breast cancer data which are sorted in decreasing order. In the figure, the higher the overall outlierness degree, the more likely the corresponding data is an outlier.

To show the effectiveness of the proposed method, it was compared against LSA algorithm[34], FindFPOF algorithm[25], Find CBLOF algorithm[24], and KNN algorithm[26]. Tables 1 and 2 show the experiment results for the sampled breast cancer data and the lymphography data. The evaluation results for the existing four algorithms are excerpted from He et al[27]. In the tables, the top ratio denotes the ratio of the number of data records specified as top k outliers to that of the data records in the dataset. The coverage indicates the ratio of the number of rare classes in top k members to that of the rare classes in the dataset. For the breast cancer data and the lymphography data, the proposed method showed the comparable performance to all the compared methods.

Table 2. Experiment results for lymphography data

Top Ratio (Number of Records)	Number of Rare Classes Included(Coverage)				
	The proposed method	LSA	Find FPOF	Find CBLOF	KNN
5%(7)	**6(100%)**	6(100%)	5(83%)	4(67%)	4(67%)
10%(15)	**6(100%)**	6(100%)	5(83%)	4(67%)	6(100%)
11%(16)	**6(100%)**	6(100%)	6(100%)	4(67%)	6(100%)
15%(22)	**6(100%)**	6(100%)	6(100%)	4(67%)	6(100%)
20%(30)	**6(100%)**	6(100%)	6(100%)	6(100%)	6(100%)

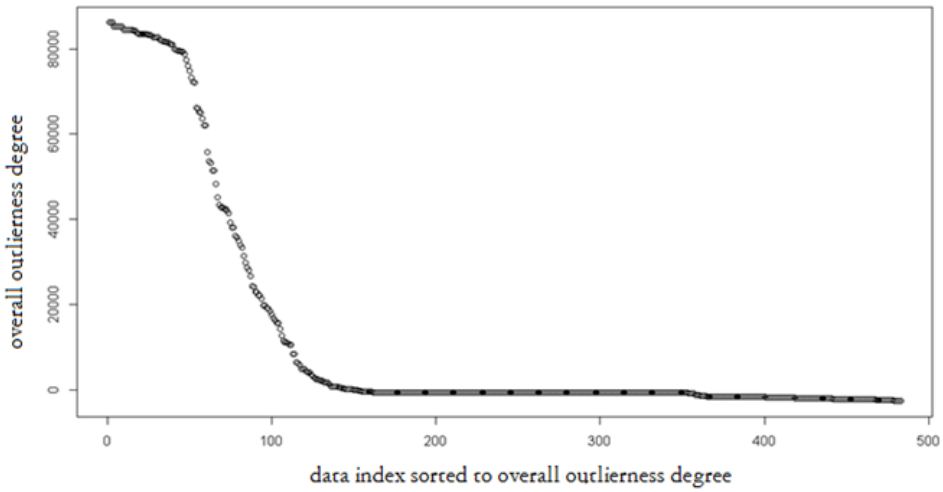

Figure 1. The sorted overall outlierness degrees for Winsconsin breast cancer data.

Table 1. Experiment results for Winsconsin breast cancer data

Top Ratio (Number of Records)	Number of Rare Classes Included(Coverage)				
	The proposed method	LSA	Find FPOF	FindCBLOF	KNN
1%(4)	**4(10.26%)**	4(10.26%)	3(7.69%)	4(10.26%)	4(10.26%)
2%(8)	**8(20.53%)**	8(20.53%)	7(17.95%)	7(17.95%)	8(20.53%)
4%(16)	14(35.90%)	**15(38.48%)**	14(35.90%)	14(35.90%)	16(41%)
6%(24)	**22(56.41%)**	22(56.41%)	21(53.85%)	21(53.85%)	20(51.28%)
8%(32)	28(71.79%)	**29(74.93%)**	28(71.79%)	27(69.23%)	27(69.23%)
10%(40)	**33(84.62%)**	33(84.62%)	31(79.49%)	32(82.05%)	32(82.02%)
12%(48)	**39(100%)**	38(97.44%)	35(89.74%)	35(89.74%)	37(94.87%)
14%(56)	**39(100%)**	39(100%)	39(100%)	38(97.44%)	39(100%)
16%(64)	**39(100%)**	39(100%)	39(100%)	39(100%)	39(100%)
18%(72)	**39(100%)**	39(100%)	39(100%)	39(100%)	39(100%)
20%(80)	**39(100%)**	39(100%)	39(100%)	39(100%)	39(100%)
25%(100)	**39(100%)**	39(100%)	39(100%)	39(100%)	39(100%)
28%(112)	**39(100%)**	39(100%)	39(100%)	39(100%)	39(100%)

5. Conclusion

We proposed a new outlier detection method which can be applied to data set with both numerical and categorical attributes. The method tries to find outliers by examining the associated attribute sets seperately, It is different from other existing approaches which consider all attribute values as vectors in a space. It has the following characteristics: First, it examines the associated subspaces of the data space and provides a strategy to search for such subspaces by employing frequent itemset mining. Second, it defines a new outlierness evaluation measure called the saliency degree. Third, it devises a mechanism to determine the set of outliers by using degree of contrastness. The proposed method was applied to two benchmark problems and compared with the exisitng four algorithms. It was observed that the method is comparable to the other methods.

6. References

1. Hawkins DM. Identification of Outliers. London: Chapman and Hall; 1980.
2. Aggarwal CC. Outlier Analysis. New York: Springer Science and Business Media; 2013.
3. Bishop CM. Pattern Recognitionand Machine Learning. New York: Springer. 2006; 4(4).
4. Kriegel HP, Kroger P, Zimek A. Outlier Detection Techniques. 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining; 2009.
5. Murphy KP. Machine Learning: A Probabilistic Perspective. MIT Press; 2012.
6. Arning A, Agrawal R, Raghavan P. A Linear Method for Deviation Detection in Large Databases Proc Int Conf on Knowledge Discovery and Data Mining. Portland, OR; 1996. p. 164–9.
7. Knorr EM. NgRT. Finding intensional knowledge of distance-based outliers. Proc of the 25th Int Conf on Very Large Data Bases. 1999; 99:211–22.
8. ChaudharyA, Szalay AS, Moore AW. Very Fast Outlier Detection in Large Multidimensional Data Sets. Proc. of ACM SIGMOD Workshop in Research Issues in Data Mining and Knowledge Discovery; 2002.
9. Wang Y, Parthasarathy S, Tatikonda S. Locality sensitive outlierdetection: A ranking driven approach. Proc Of ICDE; 2011. p. 410–21.
10. Lee KM. Locality-Sensitive Hashing Techniques for Nearest Neighbor Search. International Journal of Fuzzy Logic and Intelligent Systems. 2012; 12(4):300–7.
11. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. Proc ACM SIGMOD Int Conf on Management of Data. Dallas, TX. 2000; 29(2):93–104.
12. Ghoting A, Parthasarathy S, Otey ME. Fast mining of distance-based outliers in high-dimensional datasets. Data Mining and Knowledge Discovery. 2008; 16(3):349–64.
13. Kriegel HP, Schubert M, Zimek A. Angle-based outlier detection in high-dimensional data. Proc ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. Las Vegas, NV; 2008. p. 444–52.
14. Aggarwal CC, Yu PS. Outlier detection for high dimensional data. Proc ACM SIGMOD Int Conf on Management of Data. Dallas, TX. 2001; 30(2):37–46.
15. Kriegel HP, Kröger P, Schubert E, Zimek A. Outlier detection in axis-parallel subspaces of high dimensional data. Proc Pacific-Asia Conf on Knowledge Discovery and Data Mining. Bangkok, Thailand; 2009. p. 831–8.
16. Hodge VJ, Austin J. A survey of outlier detection methodlogies.Artificial Intelligence Review. 2004; 22(2):85–126.
17. Singh K, Upadhyaya S. Outlier detection: Applications and Techniques. International Journal of Computer Science Issues. 2012; 9(3):307–23.
18. Papadimitriou S, Kitagawa H, Gibbons PB. LOCI: Fast Outlier Detection Using the Local Correlation Integral Proc of the 19th Int Conf on Data Engineering; 2003. p. 315–26.
19. Lee KM, Lee KM. Similar pair identification using locality-sensitive hashing technique. Proc of 6th International Conference on Soft Computing and Intelligent Systems, and 13th International Symposium on Advanced Intelligence Systems, (SCIS/ISIS 2012); 2012. p. 2117–9.
20. Al-Zoubi MB, Ali AD, Yahya AA. Fuzzy clustering-based approach for outlier detection. Proc of the 9th WSEAS Int Conf on Applications of Computer Engineering; 2008. 192–7.
21. Yousri NA, Lsmail MA, Kamel MS. Fuzzy outlier analysis: A combined clustering – Outlier detection approach. Proc. of IEEE International Conference on Systems, Man and Cybernetics; 2007. p. 412–8.
22. Ostermark R. A fuzzy vector valued KNN-algorithm for automatic outlier detection. Applied Soft Computing. 2009; 9(4):1263–72.
23. He Z, Xu X, Deng S. Discovering Cluster Based Local Outliers. Pattern Recognition Letters. 2003; 24 (9-10):1651-60.
24. Harkins S, He H, Willams GJ, Baster RA. Outlier Detection Using Replicator Neural Networks. Proc of DaWaK02; 2002. p. 170-180.
25. He Z, Xu X, Huang JZ, Deng S. A Frequent Pattern Discovery Based Method for Outlier Detection. Proc (WAIM`04); 2004.
26. Ramaswamy S, Rastogi R, Kyuseok S. Efficient Algorithms for Mining Outliers from Large Data Sets. Proc Of SIGMOD; 2000. p. 93-104.
27. He Z, Xu X, Deng S. An Optimization Model for Outlier Detection in Categorical Data. Lecture Notes in Computer Science. 2005; 3644/2005:400-409.
28. Merz G, Murphy P. The UCI Reposity of Machine Learning. 1996. Available from: http://www.ics.uci.edu/mlearn/MLRepository.html.