Proceedings of
National Conference on Innovative Trends in Computer Science Engineering (ITCSE-2015)
held at BRCMCET , Bahal on 4th April 2015

Discovering Outliers from Real-World Data using Decision Trees based Classifiers

Deepak Sinwar[#1], Dr. V. S. Dhaka[*2]
[#]*Research Scholar,* [*]*Professor*
[1,2]*Jaipur Nationl University, Jagatpura, Jaipur, Rajasthan (INDIA)*
[1]deepak.sinwar@gmail.com, [2]vijaypal.dhaka@gmail.com

Abstract— **Outliers (anomalies) are one of the important aspect now a day due to the fact that they may affect business decisions. Sometimes we ignore some kind of anomalies which was present in our data set. They may be some characteristics of different behaviour, may be missing/ misleading values or any other kind of characteristic. By we know that, no one decision maker is interested in making decisions on improper/ insufficient data. To cope with such anomalies and making optimal decisions, there is vital requirement of anomaly detection mechanism. Plenty of techniques exist for coping with different kind of anomalies. This paper has reviewed two decision tree based classification techniques viz. RandomForest and J48 (C4.5). Theoretical analysis and experimental results show that the performance of both RandomForest and RandomTree approaches is higher than J48 in terms of correct classification.**

I. Introduction

The aim of Data Mining as we know to provide some knowledge about our data, but if the data itself contains lots of misleading terms, then sometimes it is not possible to gain knowledge from that data. We may sometimes call such misleading values as Outliers. But this is not applicable to all kind of such values. For simplified definition of outliers we may refer the definition proposed by Hawkins [1980]: An Outlier is an observation that deviates so much from other observations as to arouse suspicion generated by a different mechanism. With the detection of outliers we can have significant knowledge about various aspects about the data. There are number of practical applications of outliers in different areas like credit/debit card fraud detection; but sometimes we ignore them by treating them as abnormal data items. Now days, it is the need of current scenario to detect and rank (if applicable) outliers for finding out their severity level/ influences/ benefits. Most researchers prune such anomalies in earlier stages called pre-processing stages of data, but one may want to handle these anomalies with very care due their sensitive nature. We may think about some real life applications where a little bit carelessness can cause production of wrong data, in such applications we have to be very careful about anomalies (outliers). In general the handling of anomalous or outlying observations in a data set is one of the most important tasks. Once outliers have been detected it may either retained or rejected. In order to successfully distinguish between noisy outlying data and noise free outliers, different kinds of information are normally needed. These should not only include various data characteristics and the context in which the outliers occur, but also includes relevant domain knowledge.

Motivation: Consider a supermarket scenario, containing a variety of items purchased by different kinds of customers. Some items are most preferred by most of the customers whereas some are neglected due to some reasons. This rejection will leads to the reduction in the sale frequencies of such items. Obviously no supermarket manager wants such rejection; so they may wish to take effective business actions against such items in order to increase these frequencies. They may adopt "selective profit maximization', which only deals with those items which are highly profitable but their sale frequencies are not so high; we may call these items as outliers.

Plenty of outlier detection algorithms exist in research literature. This paper will focus on various approaches for outlier detection in order to verify that which one find outs more outliers with less error and high efficiency. The rest of the paper is organized as follows. Section II will reviews some work related with outliers. Section III will discuss about various approaches included in this paper. Experimental results are discussed in Section IV. Finally Section V concludes the study with summary and future work.

II. Background

Most of the work in the area of outlier mining focused on statistical analysis of data. Let us review some of the work related with outliers detection included in [23]. Generally the problem of defining outliers is notrivial [8]. Regression based and Graphical based methods are most preferred in general, because human eyes can interpret them easily. Jingke Xi [9] classified outlier mining approaches in two classes: Classic Outlier approach and Spatial Outlier approach. The classic outlier approach deals with transactional data while spatial approach deals with spatial data. Same kind of classification has been given by J. Han and M. Kamber [8], they divided the computer based methods for outlier detection in four approaches: the *statistical distribution based approach*, the *distance-based approach*, the *density-based approach*, and the *deviation-based approach.*

As we know that the clustering is one of the famous techniques towards outliers' discovery. Jiang [16] generalize local outlier factor of object and propose a framework of clustering based outlier detection, which was effective enough. Another approach of this kind was also developed by Jiang [17], called a clustering-based outlier detection method (CBOD), which results in good scalability and adapts to large dataset. Zhang [21] also proposed a novel approach to detect outliers based on clustering, which combines probability with hierarchical agglomerative clustering. Same kind of approach based on distance to k-neighbours has been presented by Yu et al. [5]. They proposed two algorithms based on local sparsity and local isolation coefficient. They showed in their experiments that we can achieve better outlier mining results if their algorithms are utilized instead of the conventional

algorithms. Another outlier mining approach based on weighted attributes from data streams has been proposed by Yogita [22]. Such kind of outlier detection is a very challenging problem, because it is not possible to scan data streams multiple times. They assign weights to attributes depending upon their respective relevance. Sometimes weighted attributes are helpful in reducing or removing the effect of noisy attributes in mining tasks. Yousri et al. [13] proposed fuzzy outlier analysis approach which can combine any outlier analysis approach with any clustering approach. They introduced the concept of universal clusters and outlier clusters along with their memberships. Clustering based outlier detection approaches are common choices, same kind of approaches can be found in [6 & 3]. A Rough Set based approach to analyse outliers in high dimensional space has been proposed by Jin et al [20]. Their key concept behind the analysis is exceptional reduction algorithm (ERDA), which results in better understanding about the data. Comparison of various already developed outlier mining approaches has also been done by some researchers in [1 & 15]. Sometimes it becomes necessary to rank outliers according to their characteristics, such kind of ranking has been provided by Muller et al. [7]. This approach was very efficient in ranking outliers in high dimensional data. Zhou et al. [11] proposed a dissimilarity based approach to detect outliers called OMABD (Outlier Mining Algorithm Base on Dissimilarity). The key concept behind this approach is that they only check the objects in the dissimilarity matrix with the dissimilarity threshold. We can also categorize this concept in the class of clustering based outlier mining. There are numerous approaches of outlier mining; more detailed view about weighted frequent patterns based outlier mining, spatial outlier detection, entropy based, graph based and neural network based approaches can be found in [18, 14, 2, 4 & 19 respectively.

III. Techniques of Outlier Detection

Plenty of outlier mining approaches exist in research literature. The objective behind this work is to study various outlier mining methods in order to find out best amongst them. Sometimes it seems very difficult to categorise a particular data item that it is an outlier or not. Some items may be of outlying behaviour for a particular situation/ organization, but some may be not. These can be better visualized if they are more different than others, for example a 7 feet person can be identified clearly as an outlier from a group of people; but sometimes more difficult to identify whenever the difference criterion is so small. Han et al. [8] classified outlier detection methods into four categories as discussed in Section-II. Let us discuss them in brief:

1) Statistical Distribution-Based Outlier Detection: This method first assumes a distribution or probability model (i.e. Normal or Poisson) and then find outs outliers using discordancy test. This test tests two hypotheses: one is working and another is alternative. A working hypothesis is a statement that the entire data set of *n* objects comes from an initial distribution model, *F*, that is,

$$H : o_i \in F, \text{ where } i = 1,2,....,n.$$

The discordancy test will check whether the object o_i is large or small with respect to *F*. An alternative hypothesis, $\overline{H}$ states that o_i comes from another distribution model G. There are various kinds of alternative hypothesis also like, inherent, mixed, slippage etc.

2) Distance-Based Outlier Detection: We can think of distance based outliers as those objects which don't have 'neighbours'. Neighbours are the objects which are closer with respect to distance threshold. The distance based methods are most preferred outliers detection methods since human eyes can detect them easily.

3) Density-Based Outlier Detection: These are somewhat similar to distance based outliers. The only difference is that, density based methods detect outliers in local neighbourhood. We may think about local neighbourhood as the area nearest to the cluster boundary. Some clusters have the high density, while some are less dense. One object may be outlier in density based method if it is a neighbour of a cluster whose density is very high, but it may not be an outlier object for less dense cluster.

4) Deviation-Based Outlier Detection: These methods do not use the statistical tests or any distance based metrics to identify outliers, instead they identify outliers on the basis of characteristics of an outlier. The objects whose characteristics are different from the group are treated as outliers. Sequential exception technique is generally used to find out outliers from less dimensional data, whereas OLAP data cube computes regions of anomalies in multidimensional data.As discussed in section-II that there are numerous methods to detect outliers, let us discuss some of the famous decision tree methodologies which are considered as a part of this study:

A. J48 (C4.5 — an extension of ID3)Decision Tree:

Decision trees are one of the most preferred ways of doing classifications, because human eyes can easily analyse the graphical classifications done by decision trees. The idea behind decision trees is a simple tree as we are using in computer science. Root is classified in different classes (as per given classes), and children are further classified accordingly as per the behaviour of the data is concerned. There are various uses of Decision Trees in classification such as statistics, prediction/ forecasting, learning etc. Decision Trees are also called as classification trees or regression trees. J48 is basically an open source implementation class for generating pruned or unpruned C4.5 [Ross Quinlan 1993] decision tree.

B. Random Forest Decision Trees:

Another kind of decision tree that impact researchers is Random Forest Decision Tree. It is a special kind of tree which learns by operating a variety of decision trees and outputs as the mode of classes for classification and mean prediction for regression. We have used *RandomForest* [10] decision tree for validating this study, because of its nature to cope with over fitting by averaging multiple trees during training time.

IV. EXPERIMENTAL WORK

A comparative study between two decision tree based algorithms of classification viz. '*RandomForest'* (RF) and 'J48' for outlier mining has been conducted. The experiments are on two real life data sets obtained from UCI Machine Learning Repository. Both data sets contain only nominal values. The first data set (soybean) contains 683 instances and 35 attributes, whereas the second (contact-lenses) contains 24 instances and 4 attributes. We found some missing values in first data set, which have been eliminated by applying unsupervised attribute filter (*ReplaceMissingValues*[10]).

TABLE1. EXPERIMENTAL RESULTS OF BOTH APPROACHES ON TWO DATASETS

	First Data Set		Second Data Set	
	RF	J48	RF	J48
Correct Classification	99.70%	97.07%	100%	91.67%
Outliers	0.29%	2.93%	0%	8.33%
Kappa Statistic	0.99	0.96	1	0.84
MAE	0.006	0.005	0.075	0.08
Time Taken	0.03 s	0.02 s	0 s	0 s

All experiments were performed on Intel(R) Core(TM)2 Duo E7500 (each with 2.3 GHz clock) with 2 GB of main memory running on windows XP(32 bit) Service Pack 2. All algorithms were run on WEKA [10] version 3.7.9.

Description of the experimental results:

As shown in Table1, maximum outliers were discovered by *RandomForest* approach for both data sets. We can see from the table that RF approach has correctly classified 99% of the instances of first data set, whereas J48 has classified 97% instances correctly. Rest of the instances have been predicted in the category of outliers.

Fig.1. Outliers visualization of J48 Classifier from first data set

We may also view these unclassified instances (outliers) using pictorial representation as shown in Fig. 1 and Fig. 2. Outliers discovered by J48 decision tree classifier for first data set are shown in Fig. 1, whereas Fig. 2 shows outliers from second data set by the same approach. We can also visualize outliers discovered by *RandomForest* classifier using the same methodology. Table1 also shows three additional parameters along with correct classification and outliers viz. Kappa statistic, Mean Absolute Error (MEA) and Time take to build model.

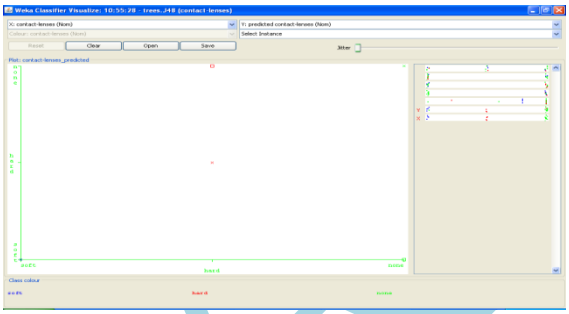

Fig.2. Outliers visualization of J48 Classifier from second data set

Kappa statistic is a measure of calculating observed accuracy with and expected accuracy. It means how closely the instances are classified correctly. *RandomForest* has highest kappa values for both data sets as compared to J48 Decision tree. Mean Absolute Error measure is a quantity to measure perditions. Both approaches have almost same absolute error rate ratio on both datasets. Time taken to build model is also used as a parameter for this comparative study to check which one takes less time in classification. Again both approaches have same time span for both data sets. Finally we may conclude that the performance of *RandomForest* approach is better than J48 decision tree in our experiments.

V. CONCLUSION

This paper has presented a study of two famous classification approaches viz. RandomForest and J48 based on decision trees. Theoretical analysis and experimental results on two real data sets shows that the RandomForest approach outperforms J48 in terms of number of outlier detection, whereas values of other parameters are almost same for both approaches for both data sets. This work may be extended for finding out outliers from other types of data i.e. data streams, spatial data etc. with some additional approaches.

VII. REFERENCES

[9] A. D. Bella, L. Fortuna, S. Grazianil, G. Napoli and M.G. Xibilia, "A Comparative Analysis of the Influence of Methods for Outliers Detection on the Performance of Data Driven Models", Instrumentation and Measurement Technology Conference - IMTC 2007, Warsaw, Poland, pp. 1-5.

[10] A. Daneshpazhouh and A. Sami, "Entropy-based outlier detection using semi-supervised approach with few positive examples", Pattern Recognition Letters (Elsevier), 49, 2014, pp. 77–84.

[11] A. Fawzy, H. M. O. Mokhtar and O. Hegazy, "Outliers detection and classification in wireless sensor networks", Egyptian Informatics Journal (Elsevier), 14, 2013, 157–164

[12] A. Rahmani, S. Afra, O. Zarour, O. Addam, N. Koochakzadeh, K. Kianmehr, R. Alhajj and J. Rokne, "Graph-based approach for outlier detection in sequential data and its application on stock market and weather data", Knowledge-Based Systems (Elsevier), 61, 2014, pp. 89–97.

[13] B. Yu, M. Song and L. Wang, "Local Isolation Coefficient-Based Outlier Mining Algorithm", International Conference on Information Technology and Computer Science" IEEE, 2009, pp. 448-51.
[14] C. Cassisi, A. Ferro, R. Giugno, G. Pigola and A. Pulvirenti, "Enhancing density- based clustering: Parameter reduction and outlier detection", Information Systems (Elsevier), 38, 2013, pp. 317–30.
[15] E. Muller, I. Assent, U. Steinhausen and T. Seidl, "OutRank: ranking outliers in high dimensional data", ICDE Workshop, IEEE 2008, pp. 600-603.
[16] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers – An Imprint of Elsevier, ISBN: 978-81-312-0535-8.
[17] J. Xi, "Outlier Detection Algorithms in Data Mining", Second International Symposium on Intelligent Information Technology Application", IEEE, 2008, pp. 94-97.
[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1.
[19] M. J. Zhou and X. J. Chen, "An Outlier Mining Algorithm Based on Dissimilarity", International Conference on Environmental Science and Engineering (ICESE 2011)", Procedia Environmental Sciences (Elsevier), 12, 2012, pp. 810-14.
[20] M. T. Hagan, H. B. Demuth and M. Beale, " Neural Network Design", PWS Publishing Company- a division of Thomson Learning, United States of America, ISBN: 7-111-10841-8
[21] N. A. Yousri, M. A. Ismail and M. S. Kamel, "Fuzzy Outlier Analysis A Combined Clustering - Outlier Detection Approach", IEEE, 2007, pp. 412-18.
[22] Q. Cai, H. He, and H. Mana, "Spatial outlier detection based on iterative self-organizing learning model", Neurocomputing (Elsevier), 117, 2013, pp. 161–72.
[23] S. H. Wu, D. Drmanac and Li-C. Wang, "A Study of Outlier Analysis Techniques for Delay Testing", INTERNATIONAL TEST CONFERENCE", IEEE, 2008, pp. 1-10.
[24] S. Jiang and A. Yang, "Framework of Clustering-Based Outlier Detection", Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, pp. 475-79.
[25] S. Jiang and Q. An, "Clustering-Based Outlier Detection Method", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, pp. 429-33.
[26] U. Yun, H. Shin, K. H. Ryu and E. Yoon, "An efficient mining algorithm for maximal weighted frequent patterns in transactional databases", Knowledge-Based Systems (Elsevier), 33, 2012, pp. 53–64.
[27] X. Zhang and Y. Zhang, "Outlier detection based on the neural network for tensor estimation", Biomedical Signal Processing and Control, 13, 2014, pp. 148–156.
[28] Y. Jin, Q. Zhu and Y. Xing, "An Exceptional Reduction Algorithm for Outliers Analyzing in High-Dimension Space", 6th World Congress on Intelligent Control and Automation, 2006, Dalian, China, pp. 5911-14.
[29] Y. Zhang, J. Liu and H. Li, "An Outlier Detection Algorithm based on Clustering Analysis", First International Conference on Pervasive Computing, Signal Processing and Applications, 2010, pp. 1126-28.
[30] Yogita and D. Toshniwal, "A Framework for Outlier Detection in Evolving Data Streams by Weighting Attributes in Clustering", 2nd International Conference on Communication, Computing & Security (ICCCS-2012), Procedia Technology (Elsevier), 6, 2012, pp. 214-22.
[31] D. Sinwar and V.S. Dhaka, "Outlier Detection from Multidimensional Space using Multilayer Perceptron, RBFNetwork and Pattern Clustering Techniques", IEEE Intl. conf ICACEA-2015

IJRRA