ENACTMENT OF NOVELTY RECOGNITION TECHNIQUE FOR CLASSICAL DATASET

Sumeet V. Shingi[1], Yogita S. Pagar[2]
Computer Science Engineering Department, PES College of Engineering, India

Abstract
Data mining is an area which has been considered among several popular area of research and different thesis and their practical based results are being presented. Among various topics, finding an error from different types of datasets is one of the emerging section. In recent years, database querying methodologies are not sufficient to acquire required user information. Therefore, this adequacy provide a new platform for research scholars for developing new techniques o meet modern demands. Error recognition does not only depend upon the datasets present for practice but also different types of attributes defined for them and proper values been considered under them. Though the concept of outlier recognition seems to be very simplified to study in theoretical manner but when its practical approach is being considered on account , the actual behavior releases its significance. This provides an introduction of different types of evaluation techniques and provides results for the proposed algorithm. Further relatively variable attributes are mentioned for variable entities. Proposed weighted density based algorithm is very simple to understand and effective in nature so as to provide better results when it comes to the mode of its implementation for given datasets.

Keywords: Novelty recognition, Average density, Weighted density.

I. INTRODUCTION

Edward Amoroso defines the term outlier recognition as :
" The process of identifying and responding to malicious activity targeted at computing and network resources. "
In United States, there was a home or an enterprise which was broken into eleven minutes. During the few years back, based on information gathered through various resources, a computer was attacked more than once per second. There were different statistics been applied in order to study and develop the ability to identify and tracking of such differential attacks. The way for identification and several break downs is to be known as intrusion or error or outlier recognition.

An outlier recognition system is composed of hardware and software elements that are working together to find foreseen events that may indicate an attack will happen, is happening or has happened. Such systems provide an attempt to uncover behavior or configurations that indicate or could lead to malicious activity. Unfortunately, such terminologies which are defined in brief leads to believe that nearly anything, including papers describing system hardening, which further contribute to detect outliers.

There are different types of outlier recognition techniques which are capable of providing knowledge about attacking attempts by observing actions. They are used for finding out regular or irregular behavior of elements present in the dataset. In order to outcome mainstream

popularity as companies are moving most of their critical business interactions to the network-of-networks.

II. literature prospect

Outliers or errors are drastic values that fall a long way outside the other observations. For example in a normal distribution, outliers may be values on the trails of the distribution.

The process of detecting outliers can be identified by different norms in the data mining and machine learning such as outlier mining, outlier modelling, novelty recognition, anomaly recognition, etc.

A useful taxonomy of outlier recognition methods is as follows –

- Extreme value analysis
- Probabilistic and statiscal methods
- Linear models
- Proximity based models
- Information theoretic models
- High dimensional outlier recognition

Extreme value analysis

Determine the statistical tails of the underlying distribution of data
eg : statistical methods like the z-scorers on unvaried data

Probabilistic and statistical methods

Determine unlikely instances from a probabilistic model of the data
Eg. Gaussian mixture models optimized using expectation maximization

Linear models

Projection methods that model the data into lower dimensions using linear co-relations.
Eg. Principle component Analysis (PCA) and data with huge large residual errors may be outliers

Proximity based models

Data instances that are isolated from the mass of the data as determined by cluster, density or nearest neighbor analysis.

Information theoretic models

Outliers are detected as instances that increase the complexity i.e. minimum code length of the dataset.

High – dimensional outlier recognition

Methods that search subspaces for outliers give the breakdown of distance based measures in high dimensions.

III. PRE-REQIUESTIC

In actual world, a huge part or may be sometimes fully filled data packs are taken into consideration which may be represented in terms of classical properties.Transaction data, demographic data, several financial records in professional banks can be taken as examples.In specific way, data which is stored into the table form is determined as classical data where each row represents few facts of an object.

A classical data chart which has few assumptions can be described as follows –

U – universe (non-empty set of objects)

A – non-empty set of attributes

V – union of attribute domains

f : U*A-> V

Statement 1 : Let DT be the classical data table. A binary relation IND(P) which is called as indiscernibility relation or in distinguished relation.

Symbolically, it can be defined as follows –

$$IND(P)=\{(x,y) \in U*U|a \in P; f(x,a)=f(y,a)\}$$

Two objects, x & y, are indiscernible in the context of a set of attributes if they have the identical values for those properties.

IV. WEIGHTED DENSITY BASED OUTLIER RECOGNITION

Those values are irregular or infrequent in nature for given data table are considered as outliers. The higher the infrequency of the value of object for each property, the more likely the object is an outlier. Additionally, an " ideal " outlier in a classical data set is one whose each and every property value is extremely irregular or infrequent. This in frequentness can be easily calculated by computing the average density of the object based on equivalence class.

Statement 2 : The average density of an object in a given dataset is given as follows –

$$Adens(x) = \frac{\sum_{a \in A} ADens_a(x)}{\bigcup}$$

where,

$ADens_a(x)$ = density of object x in U with respect to attribute a

|U| = number of attributes

Liang et. al's concept of complementary entropy is used to measure information content and uncertainty for a categorical data table . The complement entropy not only can measure the uncertainty for a classical data table. The complement entropy not only can measure the uncertainty but also calculates fuzziness. Few applications, in which this concept is used can be observed through feature selection, clustering analysis, rule evaluation, uncertainty measurement, etc.

Statement 3 : The complement entropy for categorical data is given with respect to P,

$$E(P) = \sum_{i=1}^{m} \frac{|X_i|}{|U|} \frac{|X_i^c|}{|U|}$$

$$E(P) = \sum_{i=1}^{m} \frac{|X_i|}{|U|} \left(1 - \frac{|X_i|}{|U|}\right)$$

where,

X^c_i = complement set of X_i

|U| = number of objects

The higher the value of E, the more out-of-order the distribution.

The complementary entropy is maximum when there is uniform distribution, which means that it possess maximum uncertainty on the attribute a.

Statement 4 : A weighted density definition is given as,

$$WDens(x) = \sum_{a \varepsilon A} ADens_a(x).W(\{a\})$$

where,

W({a}) = weighting function with respect to attribute a $\underline{\in}$ A.

$$W(\{a\}) = \frac{1 - E(\{a\})}{\sum_{l \in A}(1 - E(l))}$$

----- (6)

The weighting function (W) is designed in order to measure the distribution of the value domain on each individual attribute by using information entropy.

Suppose there is a value domain of attribute 'a' whose distribution is uniform. The maximum uncertainty provides more outlier characteristics, then we should give more importance to the attribute a smaller weight in weighted density. The weight of every attribute has been normalized from zero to one and the sum of all weight is one. It is very easy to verify that the range of the weighted density value is [0,1] because the range of both the density of object and weighting function. The weighted density can be used as a good indicator to determine whether an object is an outlier or not. Further, the smaller the weighted density value of an object x, the more the likelihood of x to be an outlier.

Statement 5 : For any object $x \in U$, if $WDens(x) < \Theta$, then the object x is also called a weighted density-based outlier.

Here, Θ is the threshold value. The threshold value (Θ) or lower-limit value is playing a significant role in process of outlier recognition. However, it is quite difficult to define a uniform value that is applied on all datasets.

Following are some points which can be used to this parameters –

i) Detecting outliers in stable data sets requires a high threshold value
ii) Detecting outliers in unstable datasets requires a low threshold value

If one can able to gather previous knowledge of data from domain experts, thus it can help us to set proper parameter values.

Table 1 : Weighted Density based Outlier Recognition algorithm for classical dataset

Input: A categorical data table DT = (U, A, V, f) and threshold value Θ.
Output: A set O of weighted density-based outliers.

```
1  Let O = Θ
2  For every a [ A
3  {
4  Compute the partition U/IND({a}) according to definition 1;
5  Compute the complement entropy E({a}) according to definition 3;
6  }
7  For every x ∈ U
8  {
9  For every a ∈ A
10 {
11 Compute the average density ADens_a according to definition 2;
12 }
13 Compute the weighted densityWDens(x) according to definition 4;
14 If WDens(x) < Θ, then O = O U x:
15 }
16 Return O.
```

Table 2 : A Classical Dataset

U/A	A	b	c
X1	A	E	M
X2	A	D	N
X3	B	G	M
X4	C	D	N
X5	C	G	M
X6	C	F	N

Consider the data in Table 2. This is a categorical data table, where U = {x1, x2 ,..., x6} and A = {a, b, c}.

According to Definition 1, the partitions of U with respect to different attributes are given by –
U/IND({a}) = {{x1, x2}, {x3}, {x4, x5, x6}},
U/IND({b}) = {{x1}, {x2, x4}, {x3, x5}, {x6}}
U/IND({c}) = {{x1, x3, x5},{x2, x4, x6}}.

By Definition 3,
E({a}) = 2/6(1-2/6) + 1/6(1-1/6) + 3/6(1-3/6) = 11/18
E({b}) = 2*1/6(1-1/6) + 2*2/6(1-2/6) = 13/18
E({c}) = 3/6(1-3/6) + 3/6(1-3/6) = 9/18

One can easily observe that E({c}) < E({a}) < E({b}). In other words, E({b}) achieves its maximal value, which means that the attribute b contains maximal uncertainty and provides more outlier characteristics. Therefore, this attribute should contribute more in the process of outliers detecting.

The weights of every attribute are given by –
W(a)=7/21;
W(b) = 5/21
W(c) =9/21;

According to Definition 4, we can have the following weighted density values of objects in U:
WDens(x1) = (2/6)*(7/21) + (1/6)*(5/21) + (3/6)*(9/21) = 0.3651
WDens(x2) = (2/6)*(7/21) + (2/6)*(5/21) + (3/6)*(9/21) = 0.4048
WDens(x3) = (1/6)*(7/21) + (2/6)*(5/21) + (3/6)*(9/21) = 0.3492
WDens(x4) = (3/6)*(7/21) + (2/6)*(9/21) + (3/6)*(9/21) = 0.4603
WDens(x5) = (3/6)*(7/21) + (2/6)*(9/21) + (3/6)*(9/21) = 0.4603
WDens(x6) = (3/6)*(7/21) + (1/6)*(9/21) + (3/6)*(9/21) = 0.4206

If the outlier threshold value h is set 0.4 then, the objects x1 and x3 are considered as outliers.

V. EXPERIMENTAL OUTCOME

Here, we conduct effectiveness and efficiency tests to analyze the performance of the proposed algorithm. Most of the existing outlier recognition methods for categorical data need to have some user-defined parameters in advance. The parameter-laden results are heavily dependent on suitable parameter settings, which are very difficult to set without background knowledge about the data. For reasons of fairness, the proposed algorithm is compared with the the sequence-based outlier recognition algorithm (denoted as SEQ), which consist of only one parameter i.e., the number of outliers.

In Sect. 5.1, the real data sets downloaded from the UCI machine learning repository are described, and the results of effectiveness tests on these data sets are further reported. For the

efficiency test, we conduct evaluations on synthetic data sets to show how running time enhances with the number of objects, the number of attributes and the number of outliers in Sect. 5.2.

5.1 Effectiveness analysis

In order to test the effectiveness of an outlier recognition method, Aggarwal and Yu proposed a practicable way to test how well the method worked. That is, we can run the outlier recognition method on a given data set and test the percentage of points which belonged to one of the rare classes. Since the class labels of each object in the test data sets are known in advance, the points belonged to the rare classes are considered as outliers.

Following the experimental setup mentioned above, we use three real life data sets to demonstrate the performance of the proposed algorithm against the existing algorithms SEQ.

Lymphography This data set contains 148 objects and 18 categorical attributes and class attribute. These objects are partitioned into four classes, i.e.,‘‘normal find’’ (1.35 %), ‘‘metastases’’ (54.73 %), ‘‘malign lymph’’ (41.22 %) and ‘‘fibrosis’’ (2.7 %). These rare objects in classes 1 and 4 (‘‘normal find’’ and ‘‘fibrosis’’) are considered as the outliers.

The experimental results produced by the WDOD algorithm against the SEQ algorithm on this data set are summarized in Table 3. Here, the top ratio is ratio of the number of objects specified as top k outliers to that of the objects in the dataset. The coverage is ratio of the number of detected rare classes to that of the rare classes in the data set. For example, we let the WDOD algorithm find the top six outliers with the top ratio of 4 %. By examining these six objects, we found that 5 of them belonged to the rare classes. In contrast, when we ran the SEQ algorithm on this data set, we found that only 4 of top 6 outliers belonged to rare classes.

Table 3 Experimental results on lymphography data set

Top ratio (number of objects)	Number of rare classes included (coverage)	
	SEQ	WDOD
3% (4)	4 (67%)	4 (67%)
4% (6)	4 (67%)	5 (83%)
5% (8)	5 (83%)	5 (83%)
8% (12)	5 (83%)	6 (100%)
10% (15)	6 (100%)	6 (100%)

Wisconsin breast cancer : This data set was collected by Dr. William H. Wolberg at the University of Wisconsin Madison Hospitals. There are 699 records in this data set. Each record has nine attributes, which are graded on an interval scale from a normal state of 1-10, with 10 being the most abnormal state. In this database, 241 records are malignant and 458 records are benign. According to the experimental technique of Harkins et al, we randomly remove some of the records to form a very unbalanced distribution. The resultant data set had 39 (8%) malignant objects and 444 (92%) benign objects. That is to say, there are 39 outliers in this data set. we can see that for the wisconsin breast cancer data set, the performance of the WDOD algorithm is better than SEQ. The experimental results are summarized in Table 4.

Table 4 Experimental results on Wisconsin breast cancer data set

Top ratio (number of objects)	Number of rare classes included (coverage)	
	SEQ	WDOD
1% (4)	3 (7.7%)	3 (7.7%)
2% (8)	7 (17.8%)	7 (17.8%)
4% (16)	14 (35.9%)	14 (35.9%)
6% (24)	19 (48.7%)	18 (46.2%)
8% (32)	23 (59.1%)	24 (61.5%)
10% (40)	28 (71.8%)	30 (76.9%)
12% (48)	33 (84.6%)	34 (87.2%)
14% (56)	37 (94.9%)	39 (100%)
16% (64)	38 (97.45)	39 (100%)
18% (72)	39 (100%)	39 (100%)

4.2 Efficiency analysis

Experiment results on time consumption of the two outlier recognition algorithms with increasing number of objects, attributes and outliers. For the test of efficiency, we define a synthetic data generator to create a few categorical data sets with different number of data points and attributes. In all synthetic data sets, each attribute possesses five different values. The number of data points varies from 1,000 to 10,000, and the dimensionality is in the range of 10–50.

We conduct three types of test to see the change of each algorithm’s performance as parameters change, e.g., the size of the data set, the data set

dimensionality and the number of outliers. The scalability of the three algorithms with data size. This study fixes the dimensionality to 10, and the number of outliers to 20, and also varies the data size from 1,000 to 10,000. It can be seen that all the three outlier recognition methods are approximate linear with respect to the data size. However, the increasing rate of the execution time on the WDOD is much slower than SEQ algorithm. Therefore, the proposed WDOD algorithm can ensure efficient execution when the data size is large. The performance of the WDOD algorithm outperformed that of the SEQ algorithm

As we see from these experiment results with real and artificially generated data, the WDOD algorithm outperforms the SEQ algorithm. Moreover, the WDOD algorithm requires a short time with respect to the data size, data dimensionality, and the target number of outliers, in comparison to the other two algorithms. These experiment tests suggest that the WDOD algorithm is particularly appropriate for large data set with high dimensionality, and also suitable for data sets with high percentage of outliers.

VI. CONCLUSION

Outlier recognition is becoming critically important in many research communities and application domains. In this paper, a weighted density for measuring the uncertainty of every attributes and the density of each object was presented. Furthere, we gave an effective algorithm for outlier recognition in categorical data. Experiment results on real data sets and large synthetic data sets show that the proposed algorithm is superior to existing algorithms in detecting outliers for categorical data, especially in terms of efficiency. Extending our work for dynamic and mixed data is the focus of our future work.

REFERENCES

[1] UCI Machine Learning Repository 2012http://archive.ics.uci.edu/ml/datasets.html

[2] J. Han and M. Kamber, DATA MINING: Concepts and Techniques, J. Kacprzyk and L. C. Jain, Eds. Morgan Kaufmann, 2006, vol. 54, no. Second Edition.

[3] A simple and effective outlier recognition algorithm for categorical data by Xingwang Zhao, Jiye Liang , Fuyuan Cao in Springer-Verlag Berlin Heidelberg 2013

[4] Unsupervised Outlier Recognition in Streaming Data Using Weighted Clustering by Yogita Thakran, Durga Toshniwal in IEEE conference 2012.

[5] Chandola V, Banerjee A, Kumar V (2009) Anomaly recognition: A survey. ACM Comput Surv 41(3):Article 15.

[6] Z. He, X. Xu, and S. Deng, "Discovering cluster based local outliers," *Pattern Recognition Letters*, vol. 2003, pp. 9–10, 2003.

[7] Knorr E, Ng RT, "Algorithms for mining distance-based outliers in large datasets." *In: Proceedings of the 24th VLDB conference, New York*, pp 392–403

[8] M. O. Mansur, Mohd. Noor Md. Sap, "Outlier Recognition Technique in Data Mining: A Research Perspective" *Proceedings of the Postgraduate Annual Research Seminar*, 2005

[9] Karanjit Singh and Dr. Shuchita Upadhyaya, "Outlier Recognition: Applications And Techniques" *International Journal of Computer Science Issues*, Vol. 9, Issue 1, No 3, January 2012

[10] D.Hawkins: "Identification of outliers". Chapman and Hall,London. 1980.

[11] Amandeep Kaur, Ms. Kamaljit Kaur, "Different Outlier Recognition Algorithms in Data Mining",A Novel Approach to Face Recognition and Expression Analysis using Local Directional Number Pattern

[12] Karen Scarfone and Peter Mell, "Guide to Intrusion Recognition and Prevention Systems (IDPS)," Department of commerce, National Institute of Standards and Technology, Gaithersburg, 2007.

[13] Chris Petersen. (2012, February) LogRhythm website.[Online]. www.logrhythm.com

[14] V. Jyothsna, V.V Ramaprasad, and K Munivara Prasad," A Review of Anomaly based Intrusion," International Journal of Computer Applications, vol. 28, no. 7, pp.26-35, August 2011.

[15] International Journal of Computer Applications Technology and Research, Volume 2– Issue 2, 185 - 187, 2013

[16] P Garcia Teodora, J Diaz Verdejo, G Macia Farnandez and E Vazquez, "Anomaly-based network intrusion recognition: Techniques, Systems and Challenges," International Journal of Computers & Security, vol. 28, no. 1, pp. 18-28, February 2009.

[17] Rajdeep Borgohain, "FuGeIDS : Fuzzy Genetic paradigms in Intrusion Recognition Systems," International Journal of Advanced Networking and Applications, vol. 3, no. 6, pp. 1409-1415, 2012.

[18] Sang Jun Han and Sung Bae Cho, "Evolutionary Neural Networks for Anomaly," IEEE Transaction on Systems, Man, and Cybernetics, Part B: CYBERNETICS, vol. 36, no. 3, pp. 559-570, June 2006.

[19] Peng Ning and Sushil Jajodia. (2012, February) Intrusion Recognition Techniques.

[20] H. Hajji, "Statistical analysis of network traffic for adaptive faults recognition," IEEE Trans. Neural Networks, vol. 16, no. 5, pp. 1053–1063, Sep. 2005.

[21] K. Yamanish, J.-I. Takeuchi, G. Williams, and P. Milne, "On-line unsupervised outlier recognition using finite mixtures with discounting learning algorithms," Data Mining and Knowledge Discovery, vol. 8, no. 3, pp. 275–300, May 2004.

[22] M. S. Sadik and L. Gruenwald, DBOD-DS : Distance Based Outlier Recognition for Data Streams. Springer, 2011, vol. 6261, p. 1221

[23] Knorr, E., Ng, R., & Tucakov, V. "Distance-based outliers: Algorithms and applications." VLDB Journal: Very Large Databases, 8(3–4), 237–253, 2000

[24] Lin, T. Y. " Neighborhood systems-application to qualitative fuzzy and rough sets." In P. P. Wang (Ed.), Advances in machine intelligence and soft-computing (pp.132–155). Durham, North Carolina, USA: Department of Electrical Engineering, Duke University.

[25] Pawlak, Z. "Rough sets." International Journal of Computer and Information Sciences, 11, 341–356.

[26] Ramaswamy, S., Rastogi, R., & Kyuseok, S. "Efficient algorithms for mining outliers from large data sets.", In Proceedings of the ACM SIDMOD international conference on management of data, 2000

[27] Asmaa Shaker ashoor and Sharad Gore, "Intrusion Recognition System (IDS): Case Study," in IACSIT Press, Singapore, 2011, pp. 6-9.