IJARES/September 2014/Volume-2/Issue-9/1400-1404 **ISSN 2347- 9337**

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING AND SCIENCE

EXPOSURE TOWARDS ANOMALOUS NODES IN SEMANTIC GRAPHS

Sirisha Kalluri[1], R.Sathish Kumar[2]

[1]M.Tech Student, Dept of CSE, Malla Reddy Engineering College for Women, Hyderabad, T.S, India

[2]Assistant Professor, Dept of CSE, Malla Reddy Engineering College for Women, Hyderabad, T.S, India

ABSTRACT:

In real applications, a huge section or the total of data set is regularly presented in terms of categorical attributes. Existing methods of unsupervised outlier detection are effectual on data sets with numerical attributes. If the existing methods intended for outlier detection are classified consistent with availability of labels in training data sets, there are three broad groups such as supervised, semi-supervised, as well as unsupervised approaches. The unsupervised approach is additional widely used than other approaches since it does not require labelled information. Outlier detection, which is a dynamic research area, refers to difficulty of finding objects in a data set that do not conform to distinct notions of accepted behaviour. Outlier detection is a necessary measure in several realistic applications including intrusion detection, health system monitoring as well as criminal activity detection. Outlier detection can be put into practice as a pre-processing measure preceding to the application of superior data analysis method. We put forward two effective as well as efficient algorithms, named Information-Theory-Based Step-by-Step (ITB-SS) as well as Single-Pass (ITB-SP) techniques which require only number of outliers as an input parameter and totally distribute with parameters for characterizing outliers usually necessary by existing algorithms. These algorithms are building upon quite a lot of significant properties of the holoentropy that allocate equivalent importance to the entire attributes, while in actual applications, different attributes regularly put in differently to outline the overall structure of data set.

Keywords: Outlier detection, Unsupervised approach, Holoentropy, Intrusion detection.

1. INTRODUCTION:

Outlier detection is a necessary measure in several realistic applications including intrusion detection, health system monitoring as well as criminal activity detection in E-commerce, and can moreover be employed in scientific research in support of data analysis and knowledge discovery [1]. The supervised anomaly detection method gain knowledge of a classifier by means of labelled objects belonging to normal and anomaly classes, and allocate suitable labels to test objects. Outlier detection can be put into practice as a pre-processing measure preceding to the application of superior data analysis method. The supervised method has been considered expansively and numerous methods have been developed. Moreover, in a supervised method a training set have to be provided with labels in support of anomalies as well as labels of normal objects, in difference with training set with normal object labels alone necessary by semi-supervised approach. The semi-supervised anomaly detection method mainly learns a model representing regular behaviour from a specified training data set of normal objects, and subsequently calculates probability of a test object's being generated by learned representation. The unsupervised anomaly detection method notice anomalies in an unlabeled data set in the assumption that mainstream of objects in data set are standard. The unsupervised approach does not require any object label information therefore the three approaches have dissimilar limitations, and they fit different kinds of data sets with dissimilar quantity of label information [2][3]. To put into practice supervised as well as semi-supervised outlier detection methods, one have to first label training data. If one wants to make use of a supervised or semi-supervised method, an unsupervised means can be used as first move to discover a candidate set of outliers, which will help experts to construct the training data set. In real applications, a huge section or the total of data set is regularly presented in terms of categorical attributes. The difficulty of outlier detection in data set is additionally challenging as there is no inherent measurement of distance among the objects. Existing methods of unsupervised outlier detection are effectual on data sets with numerical attributes.

2. METHODOLOGY:

If the existing methods intended for outlier detection are classified consistent with

availability of labels in training data sets, there are three broad groups such as supervised, semi-supervised, as well as unsupervised approaches. In principle, models within supervised or semi-supervised approaches all necessitate to be skilled before use, whereas models adopting unsupervised method do not comprise the training phase [4][5]. Moreover, in a supervised method a training set have to be provided with labels in support of anomalies as well as labels of normal objects, in difference with training set with normal object labels alone necessary by semi-supervised approach. The unsupervised approach does not require any object label information therefore the three approaches have dissimilar limitations, and they fit different kinds of data sets with dissimilar quantity of label information. When faced with a huge data set with millions of high-dimensional objects and a small anomalous data rate, picking abnormal as well as normal objects to create a superior training data set is prolonged and labour-intensive. The unsupervised approach is additional widely used than other approaches since it does not require labelled information [6]. Outlier detection, which is a dynamic research area, refers to difficulty of finding objects in a data set that do not conform to distinct notions of accepted behaviour. We recommend two greedy algorithms to solve optimization problem in support of outlier detection. Outlier detection can be put into practice as a pre-processing measure preceding to the application of superior data analysis method. It can also be used as an effectual tool to find out interest patterns such as the expense behaviour of a to-be bankrupt credit cardholder.

3. AN OVERVIEW OF PROPOSED METHOD:

We put forward two effective as well as efficient algorithms, named Information-Theory-Based Step-by-Step (ITB-SS) as well as Single-Pass (ITB-SP) techniques which require only number of outliers as an input parameter and totally distribute with parameters for characterizing outliers usually necessary by existing algorithms. We recommend two greedy algorithms to solve optimization problem in support of outlier detection. Our algorithms are building upon quite a lot of significant properties of the holoentropy that allocate equivalent importance to the entire attributes, while in actual applications, different attributes regularly put in

differently to outline the overall structure of data set. In unsupervised outlier detection, mainstream of objects within a data set are supposed to be standard objects. We initiate three latest concepts such as the upper bound on outliers, anomaly candidate set and normal object set. These concepts are built on assumption that eliminating outliers will get better the purity of data set. In both algorithms, search is conducted merely within anomaly candidate set, though this does not create any dissimilarity for the algorithm ITB-SP as initialization of anomaly candidate set necessitate computation of outlier factors of the entire objects. ITB-SS does advantage, though, from reduced search space. In scheming two algorithms, we supposed that number of requested outlier's is constantly smaller than upper bound. For ITB-SS, attribute weights, initial outlier factors together with initialization of anomaly candidate set are worked out. To visualize data set, we illustrate a two-dimensional depiction in fig1, using principle of graph drawing. In this graph, vertices point to objects and the edges symbolize the similarity among objects, where the entire the similarities are 1.

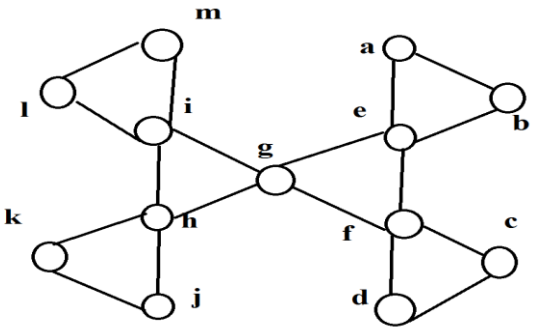

Fig1: An overview of synthetic data set.

4. CONCLUSION:

Outlier detection is a necessary measure in several realistic applications including intrusion detection, health system monitoring as well as criminal activity detection in E-commerce, and can moreover be employed in scientific research in support of data analysis and knowledge discovery. Outlier detection, which is a dynamic research area, refers to difficulty of finding objects in a data set that do not conform to distinct notions of accepted behaviour. It can also be used as an effectual tool to find out interest patterns such as the expense behaviour of a to-be bankrupt credit cardholder. When faced with a huge data set with millions of high-dimensional objects and a small anomalous data rate, picking abnormal as well as normal objects to create a superior training data set is prolonged and labour-intensive. The semi-supervised anomaly detection method mainly learns a

model representing regular behaviour from a specified training data set of normal objects, and subsequently calculates probability of a test object's being generated by learned representation. The unsupervised approach is additional widely used than other approaches since it does not require labelled information. To put into practice supervised as well as semi-supervised outlier detection methods, one have to first label training data. We put forward two effective as well as efficient algorithms, named Information-Theory-Based Step-by-Step (ITB-SS) as well as Single-Pass (ITB-SP) techniques which require only number of outliers as an input parameter and totally distribute with parameters for characterizing outliers usually necessary by existing algorithms. In unsupervised outlier detection, mainstream of objects within a data set are supposed to be standard objects. We initiate three latest concepts such as the upper bound on outliers, anomaly candidate set and normal object set which are built on assumption that eliminating outliers will get better the purity of data set.

REFERENCES

[1] Z. He, X. Xu, and S. Deng, "An Optimization Model for Outlier Detection in Categorical Data," Proc. Int'l Conf. Advances in Intelligent Computing (ICIC '05), 2005.

[2] S. Papadimitriou, H. Kitagawa, P.B. Gibbons, and C. Faloutsos, "Loci: Fast Outlier Detection Using Thelocal Correlation Integral," Proc. 19th Int'l Conf. Data Eng. (ICDE '03), 2003.

[3] J. Takeuchi and K. Yamanishi, "A Unifying Framework for Detecting Outliers and Change Points from Time Series," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 482-492, Apr. 2006.

[4] G.D. Battista, P. Eades, R. Tamassia, and I.G. Tollis, "Algorithms for Drawing Graphs: An Annotated Bibliography," Computational Geometry: Theory and Applications, vol. 4, pp. 235 282, 1994.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection for Discrete Sequences: A Survey," IEEE Trans. Knowledge and Data Eng., vol. 24, no. 5, pp. 823-839, May 2012.

[6] T. Leckie and A. Yasinsac, "Metadata for Anomaly-Based Security Protocol Attack Deduction," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1157-1168, Sept. 2004.