A Fast Greedy Algorithm for Outlier Mining

Zengyou He[1], Shengchun Deng[1], Xiaofei Xu[1], and Joshua Zhexue Huang[2]

[1] Department of Computer Science and Engineering, Harbin Institute of Technology, China
zengyouhe@yahoo.com, dsc@hit.edu.cn, xiaofei@hit.edu.cn
[2] E-Business Technology Institute, The University of Hong Kong, Hong Kong
jhuang@eti.hku.hk

Abstract. The task of outlier detection is to find small groups of data objects that are exceptional when compared with rest large amount of data. Recently, the problem of outlier detection in categorical data is defined as an optimization problem and a local-search heuristic based algorithm (LSA) is presented. However, as is the case with most iterative type algorithms, the LSA algorithm is still very time-consuming on very large datasets. In this paper, we present a very fast greedy algorithm for mining outliers under the same optimization model. Experimental results on real datasets and large synthetic datasets show that: (1) Our new algorithm has comparable performance with respect to those state-of-the-art outlier detection algorithms on identifying true outliers and (2) Our algorithm can be an order of magnitude faster than LSA algorithm.

1 Introduction

In contrast to traditional data mining task that aims to find the general pattern applicable to the majority of data, outlier detection targets the finding of the rare data whose behavior is very exceptional when compared with rest large amount of data. Studying the extraordinary behavior of outliers can uncover valuable knowledge hidden behind them and aid the decision makers to make profit or improve the service quality. Thus, mining for outliers is an important data mining research with numerous applications, including credit card fraud detection, discovery of criminal activities in electronic commerce, weather prediction, and marketing.

A well-quoted definition of outliers is firstly given by Hawkins [1]. This definition states: an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. With increasing awareness on outlier detection in data mining literature, more concrete meanings of outliers are defined for solving problems in specific domains [3-22].

However, conventional approaches do not handle categorical data in a satisfactory manner, and most existing techniques lack for a solid theoretical foundation or assume underlying distributions that are not well suited for exploratory data mining applications. To fulfill this void, the problem of outlier detection in categorical data is defined as an optimization problem as follows [22]: finding a subset of *k* objects such

W.K. Ng, M. Kitsuregawa, and J. Li (Eds.): PAKDD 2006, LNAI 3918, pp. 567 – 576, 2006.

that the expected entropy of the resultant dataset after the removal of this subset is minimized.

In the above optimization problem, an exhaustive search through all possible solutions with k outliers for the one with the minimum objective value is costly since for n objects and k outliers there are (n,k) possible solutions. To get a feel for the quality-time tradeoffs involved, a local search heuristic based algorithm (LSA) is presented in [22]. However, as is the case with most iterative type algorithms, the LSA algorithm is still very time-consuming on very large datasets.

In this paper, we present a very fast greedy algorithm for mining outliers under the same optimization model. Experimental results on real datasets and large synthetic datasets show that: (1) Our algorithm has comparable performance with respect to those state-of-the-art outlier detection algorithms on identifying true outliers and (2) Our algorithm can be an order of magnitude faster than LSA algorithm.

The organization of this paper is as follows. First, we present related work in Section 2. Problem formulation is provided in Section 3 and the greedy algorithm is introduced in Section 4. The empirical studies are provided in Section 5 and a section of concluding remarks follows.

2 Related Work

Statistical model-based methods, such as *distribution-based* methods [1,5] and *depth-based* methods [6], are rooted from the statistics community. In general, underlying distributions of data are assumed known a priori in these methods. However, such assumption is not appropriate in real data mining applications. *Distance based* methods [7-9] and *density based* methods [10,11] are recently proposed methods for mining outliers in large databases. However, they primarily focused on databases containing real-valued attributes. *Clustering-based* outlier detection techniques regarded *small* clusters as outliers [12, 14] or identified outliers by removing clusters from the original dataset [13]. *Sub-Space based* methods aim to find outliers effectively from high dimensional datasets [3,4]. *Support vector* based methods [15,16] and *neural network based* methods [17,18] are also widely used in outlier detection. *Outlier ensemble* based methods are investigated recently in [24,25].

The preceding methods may be considered as traditional in the sense that they define an outlier without regard to class membership. However, in the context of supervised learning (where data have class labels attached to them) it makes sense to define outliers by taking such information into account. The problem of class outlier detection is considered in [19-21].

3 Problem Formulation

Entropy is the measure of information and uncertainty of a random variable [2]. If X is a random variable, and S (X) the set of values that X can take, and p (x) the probability function of X, the entropy E (X) is defined as shown in Equation (1).