Outlier Detection with Kernel Density Functions

Longin Jan Latecki[1], Aleksandar Lazarevic[2], and Dragoljub Pokrajac[3]

[1] CIS Dept. Temple University Philadelphia, PA 19122, USA
latecki@temple.edu
[2] United Technology Research Center 411 Silver Lane, MS 129-15 East Hartford, CT 06108, USA
aleks@cs.umn.edu
[3] CIS Dept. CREOSA and AMRC, Delaware State University Dover DE 19901, USA
dpokrajac@desu.edu

Abstract. Outlier detection has recently become an important problem in many industrial and financial applications. In this paper, a novel unsupervised algorithm for outlier detection with a solid statistical foundation is proposed. First we modify a nonparametric density estimate with a variable kernel to yield a robust local density estimation. Outliers are then detected by comparing the local density of each point to the local density of its neighbors. Our experiments performed on several simulated data sets have demonstrated that the proposed approach can outperform two widely used outlier detection algorithms (LOF and LOCI).

1 Introduction

Advances in data collection are producing data sets of massive size in commerce and a variety of scientific disciplines, thus creating extraordinary opportunities for monitoring, analyzing and predicting global economical, demographic, medical, political and other processes in the World. However, despite the enormous amount of data available, particular events of interests are still quite rare. These rare events, very often called outliers or anomalies, are defined as events that occur very infrequently (their frequency ranges from 5% to less than 0.01% depending on the application). Detection of outliers (rare events) has recently gained a lot of attention in many domains, ranging from video surveillance and intrusion detection to fraudulent transactions and direct marketing. For example, in video surveillance applications, video trajectories that represent suspicious and/or unlawful activities (e.g. identification of traffic violators on the road, detection of suspicious activities in the vicinity of objects) represent only a small portion of all video trajectories. Similarly, in the network intrusion detection domain, the number of cyber attacks on the network is typically a very small fraction of the total network traffic. Although outliers (rare events) are by definition infrequent, in each of these examples, their importance is quite high compared to other events, making their detection extremely important.

P. Perner (Ed.): MLDM 2007, LNAI 4571, pp. 61–75, 2007.

Data mining techniques that have been developed for this problem are based on both supervised and unsupervised learning. Supervised learning methods typically build a prediction model for rare events based on labeled data (the training set), and use it to classify each event [1,2]. The major drawbacks of supervised data mining techniques include: (1) necessity to have labeled data, which can be extremely time consuming for real life applications, and (2) inability to detect new types of rare events. On the other hand, unsupervised learning methods typically do not require labeled data and detect outliers (rare events) as data points that are very different from the normal (majority) data based on some pre-specified measure [3]. These methods are typically called outlier/anomaly detection techniques, and their success depends on the choice of similarity measures, feature selection and weighting, etc. Outlier/anomaly detection algorithms have the advantage that they can detect new types of rare events as deviations from normal behavior, but on the other hand suffer from a possible high rate of false positives, primarily because previously unseen (yet normal) data are also recognized as outliers/anomalies, and hence flagged as interesting.

Outlier detection techniques can be categorized into four groups: (1) statistical approaches; (2) distance based approaches; (3) profiling methods; and (4) model-based approaches. In statistical techniques [3,6,7], the data points are typically modeled using a stochastic distribution, and points are labeled as outliers depending on their relationship with the distributional model.

Distance based approaches [8,9,10] detect outliers by computing distances among points. Several recently proposed distance based outlier detection algorithms are founded on (1) computing the full dimensional distances among points using all the available features [10] or only feature projections [8]; and (2) on computing the densities of local neighborhoods [9,35]. Recently, LOF (Local Outlier Factor) [9] and LOCI (Local Correlation Integral) [35] algorithms have been successfully applied in many domains for outlier detection in a batch mode [4,5,35]. In addition, clustering-based techniques have also been used to detect outliers either as side products of the clustering algorithms (as points that do not belong to clusters) [11] or as clusters that are significantly smaller than others [12].

In profiling methods, profiles of normal behavior are built using different data mining techniques or heuristic-based approaches, and deviations from them are considered as outliers (e.g., network intrusions). Finally, model-based approaches usually first characterize the normal behavior using some predictive models (e.g. replicator neural networks [13] or unsupervised support vector machines [4,12]), and then detect outliers as deviations from the learned model.

In this paper, we propose an outlier detection approach that can be classified both into statistical and density based approaches, since it is based on local density estimation using kernel functions. Our experiments performed on several simulated data sets have demonstrated that the proposed approach outperforms two very popular density-based outlier detection algorithms, LOF [9] and LOCI [35].