Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 329 (2016) 61–77

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

An Unsupervised Approach for Combining Scores of Outlier Detection Techniques, Based on Similarity Measures

José Ramón Pasillas-Díaz [1,2] Sylvie Ratté[3]

Department of Software and IT Engineeringg
École de Technologie Supérieure
Montreal, QC, Canada

Abstract

Outlier detection, the discovery of observations that deviates from normal behavior, has become crucial in many application domains. Numerous and diverse algorithms have been proposed to detect them. These algorithms identify outliers using precise definitions of the concept of outliers, thus their performance depends largely on the context of application. The construction of ensembles has been proposed as a solution to increase the individual capacity of each algorithm. However, the unsupervised scenario (absence of class labels) in the domains where outlier detection operates restricts the use of approaches relying on the existence of labels. In this paper, two novel unsupervised approaches using ensembles of heterogeneous types of detectors are proposed. Both approaches construct the ensemble using solely the results produced by each algorithm, identifying and giving more weight to the most suitable techniques depending on the particular dataset under examination. Through experimental evaluation in real world datasets, we demonstrate that our proposed algorithm provides a significant improvement over the base algorithms and even over existing approaches for ensemble outlier detection.

Keywords: outlier detection, ensembles

1 Introduction

Our capacity to collect and store data increases in an exponential manner but our capacity to analyze it has not followed the same trend. Despite the explosion of available data, the discovery of truly interesting patterns is a rare event. Outlier detection the discovery of observations that deviates from normal behavior has been widely studied in recent years [26,15,7], resulting in a set algorithms designed to detect these rare but potentially crucial events. In some specific contexts an outlier is a data point that can be considered either as an abnormality or noise,

[1] This work was supported by CONACYT Mexico, Scholarship 214609.

[2] Email: jose-ramon.pasillas-diaz.1@ens.etsmtl.ca

[3] Email: sylvie.ratte@etsmtl.ca

http://dx.doi.org/10.1016/j.entcs.2016.12.005

whereas anomaly refers to a special kind of outlier which is of interest to the analyst. However, the terms outlier and anomaly, in general, have been used interchangeably in the literature [7].

One of the core definitions of outliers was made in 1980 by Grubbs [12]: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. However, this definition lacks one important characteristic, this is, the case where the outlying points conglomerates to form their own group of outliers; Barnett and Lewis [2] improved the definition of outliers by considering as outlier not only a single and isolated point, but also a group of points deviating from the normal behavior.

The effect of undetected outliers in different application domains (i.e. medical, intrusion detection, fraud detection, geographical) could have deep and disastrous consequences. An example is the detection of breast cancer where an undetected positive case implies an untreated patient; another example is a failed attempt to detect strange behavior in the use of a stolen credit card resulting in a financial impact for the credit card holder. In both of these examples, the minority of the cases represents the class of interest.

The process of outlier detection represents a very specific classification scenario: first, the quantity of outliers is very small in proportion to the quantity of normal instances; and second, the use of labels (supervised approach) in outlier detection is limited due to the fact that, by definition, the outliers that we are trying to detect represent a new or unseen behavior. Despite the fact that some algorithms (techniques) can operate using only labels for the normal class [23] (semi supervised approach) and use this information to increase the detection rate, unsupervised approaches have the undeniable advantage of operating over unlabeled data. Furthermore, unlabeled data are usually easier to obtain and represents the more common scenario in outlier detection [10].

The use of an unsupervised outlier detection approach also has the benefit of avoiding the bias introduced by training an algorithm with anomalous observations, labeled wrongly as normal data, causing the misclassification of future similar observations.

Due to the large spectrum of domains where outlier detection can operate, there are a wide variety of outlier detection algorithms mainly based on: classification, clustering, nearest neighborhood and statistical approaches [7]. However, their use is application dependent; no single outlier detection algorithm is best suited for all the different data scenarios that we could encounter in real world datasets [19]. Some algorithms work better when the data tend to form clusters, whereas others are most suitable to use in the presence of neighborhoods in the data.

Despite the fact that by working on an unsupervised scenario it is not possible to know which algorithm is better for a specific dataset in advance, the performance of these algorithms can be improved.

Similar to ensemble classifier learning, where heterogeneous assumptions are used to produce a unified output [25,24], in ensemble outlier detection, diverse (heterogeneous) assumptions are also needed to produce a meaningful result, potentially