Expert Systems with Applications 36 (2009) 4680–4687

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Some issues about outlier detection in rough set theory

Feng Jiang [a,*], Yuefei Sui [b], Cungen Cao [b]

[a] College of Information and Science Technology, Qingdao University of Science and Technology, Qingdao 266061, PR China
[b] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, PR China

ARTICLE INFO

Keywords:
Outlier detection
Rough sets
Distance metric
KDD

ABSTRACT

"One person's noise is another person's signal" (Knorr, E., Ng, R. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th VLDB conference, New York* (pp. 392–403)). In recent years, much attention has been given to the problem of outlier detection, whose aim is to detect outliers – objects which behave in an unexpected way or have abnormal properties. Detecting such outliers is important for many applications such as criminal activities in electronic commerce, computer intrusion attacks, terrorist threats, agricultural pest infestations, etc. And outlier detection is critically important in the information-based society. In this paper, we discuss some issues about outlier detection in rough set theory which emerged about 20 years ago, and is nowadays a rapidly developing branch of artificial intelligence and soft computing. First, we propose a novel definition of outliers in information systems of rough set theory – *sequence-based outliers*. An algorithm to find such outliers in rough set theory is also given. The effectiveness of sequence-based method for outlier detection is demonstrated on two publicly available databases. Second, we introduce traditional distance-based outlier detection to rough set theory and discuss the definitions of distance metrics for distance-based outlier detection in rough set theory.

1. Introduction

Knowledge discovery in databases (KDD), or data mining, is an important issue in the development of data- and knowledge-based systems. Usually, knowledge discovery tasks can be classified into four general categories: (a) dependency detection, (b) class identification, (c) class description, and (d) outlier/exception detection (Knorr & Ng, 1998). In contrast to most KDD tasks, such as clustering and classification, outlier detection aims to find small groups of data objects that are exceptional when compared with the remaining large amount of data, in terms of certain sets of properties. For many applications, such as fraud detection in E-commerce, it is more interesting to find the rare events than to find the common ones, from a knowledge discovery standpoint. Studying the extraordinary behaviors of outliers can help us uncover the valuable information hidden behind them. Recently, researchers have begun focusing on outlier detection, and attempted to apply algorithms for finding outliers to tasks such as fraud detection (Bolton & Hand, 2002), identification of computer network intrusions (Eskin, Arnold, Prerau, Portnoy, & Stolfo, 2002; Lane & Brodley, 1999), data cleaning (Rulequest Research), detection of employers with poor injury histories (Knorr, Ng, & Tucakov, 2000), and peculiarity-oriented mining (Zhong, Yao, Ohshima, & Ohsuga, 2001).

Outliers exist extensively in the real world, and are generated from different sources: a heavily tailed distribution or errors in inputting the data. While there is no single, generally accepted, formal definition of an outlier, Hawkins' definition captures the spirit: "an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980; Knorr & Ng, 1998). With increasing awareness on outlier detection in the literatures, more concrete meanings of outliers are defined for solving problems in specific domains. Nonetheless, most of these definitions follow the spirit of Hawkins' definition (Chiu & Fu, 2003).

Roughly speaking, the current approaches to outlier detection can be classified into the following five categories (Kovács, Vass, & Vidács, 2004):

(1) *Distribution-based approach* is the classical method in statistics. It is based on some standard distribution models (Normal, Poisson, etc.), and those objects which deviate from the model are recognized as outliers (Rousseeuw & Leroy, 1987). Its greatest disadvantage is that the distribution of the measurement data is unknown in practice. Often a large number of tests are required in order to decide which distribution model the measurement data follow (if there is any).

* Corresponding author.
E-mail addresses: jiangkong@163.net (F. Jiang), yfsui@ict.ac.cn (Y. Sui), cgcao@ict.ac.cn (C. Cao).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.019

(2) *Depth-based approach* is based on computational geometry and computes different layers of k–d convex hulls and flags objects in the outer layer as outliers (Johnson, Kwok, & Ng, 1998). However, it is a well-known fact that the algorithms employed suffer from the dimensionality curse and cannot cope with a large k.
(3) *Clustering approach* classifies the input data. It detects outliers as by-products (Jain, Murty, & Flynn, 1999). However, since the main objective is clustering, it is not optimized for outlier detection.
(4) *Distance-based approach* was originally proposed by Knorr and Ng (Knorr & Ng, 1998; Knorr et al., 2000). An object o in a data set T is a distance-based outlier if at least a fraction p of the objects in T are further than distance D from o.This outlier definition is based on a single, global criterion determined by the parameters p and D. Problems may occur if the parameters of the data are very different from each other in different regions of the data set.
(5) *Density-based approach* was originally proposed by Breunig, Kriegel, Ng, and Sander (2000). A Local Outlier Factor (LOF) is assigned to each sample based on its local neighborhood density. Samples with high LOF value are identified as outliers. The disadvantage of this solution is that it is very sensitive to parameters defining the neighborhood.

Rough set theory, introduced by Zdzislaw Pawlak in the early 1980s (Pawlak, 1982, 1991, Pawlak, Grzymala-Busse, Slowinski, & Ziarko, 1995), is for the study of intelligent systems characterized by insufficient and incomplete information. It is motivated by practical needs in classification and concept formation. The rough set philosophy is based on the assumption that with every object of the universe there is associated a certain amount of information (data, knowledge), expressed by means of some attributes. Objects having the same description are indiscernible. In recent years, there has been a fast growing interest in rough set theory. Successful applications of the rough set model in a variety of problems have demonstrated its importance and versatility.

To the best of our knowledge, there is no existing work about outlier detection in rough set community. The aim of this work is to combine the rough set theory and outlier detection to show how outlier detection can be done in rough set theory. We suggest two different ways to achieve this aim. First, we propose sequence-based outlier detection in information systems of rough set theory. Second, we introduce traditional distance-based outlier detection to rough set theory.

This paper is organized as follows. In the next section, we introduce some preliminaries in rough set theory and outlier detection. In Section 3, we give some definitions concerning sequence-based outliers in information systems of rough set theory. The basic idea is as follows: Given an information system $IS = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a set of attributes, V is the union of attribute domains, and f is a function such that for any $x \in U$ and $a \in A$, $f(x, a) \in V_a$. Since each attribute subset $B \subseteq A$ determines an indiscernibility (equivalence) relation $IND(B)$ on U, we can obtain the corresponding equivalence class of relation $IND(B)$ for every object $x \in U$. If we decrease attribute subset B gradually, then the granularity of partition $U/IND(B)$ will become coarser, and for every object $x \in U$ the corresponding equivalence class of x will become bigger. So when there is an object in U whose equivalence class always does not vary or only increases a little in comparison with those of other objects in U, then we may consider this object as a sequence-based outlier in U with respect to IS. An algorithm to find sequence-based outliers is also given. In Section 4, we apply traditional distance-based outlier detection to rough set theory. Since classical rough set theory is better suited to deal with *nominal* attributes, we propose the revised definitions of two traditional distance metrics for distance-based outlier detection in rough set theory – *overlap metric* and *value difference metric in rough set theory*, both of which are especially designed to deal with *nominal* attributes. Experimental results are given in Section 5, and Section 6 discusses the advantages of our sequence-based approach by comparing with other approaches to outlier detection. Section 7 concludes the paper.

2. Preliminaries

In rough set data model, information is stored in a table, where each row (tuple) represents facts about an object. All we know about an object from the table is the corresponding tuple in the table.

In rough set terminology, a data table is also called an information system. When the attributes are classified into decision attributes and condition attributes, a data table is also called a decision system. More formally, an information system is a quadruple $IS = (U, A, V, f)$, where

1. U is a non-empty finite set of objects.
2. A is a non-empty finite set of attributes.
3. V is the union of attribute domains, i.e., $V = \bigcup_{a \in A} V_a$, where V_a denotes the domain of attribute a.
4. $f : U \times A \to V$ is an information function such that for any $a \in A$ and $x \in U$, $f(x, a) \in V_a$.

We can split set A of attributes into two subsets: $C \subset A$ and $D = A - C$, conditional set of attributes and decision (or class) attribute(s), respectively. The condition attributes represent measured features of the objects, while the decision attributes are *a posteriori* outcome of classification.

Each subset $B \subseteq A$ of attributes determines a binary relation $IND(B)$, called indiscernibility relation, which is defined as follows:

$$IND(B) = \{(x, y) \in U \times U : \forall a \in B(f(x, a) = f(y, a))\} \quad (1)$$

It is obvious that $IND(B)$ is an equivalence relation on U and $IND(B) = \bigcap_{a \in B} IND(\{a\})$.

Given any $B \subseteq A$, relation $IND(B)$ induces a partition of U, which is denoted by $U/IND(B)$, where an element from $U/IND(B)$ is called an equivalence class. For every element x of U, let $[x]_B$ denote the equivalence class of relation $IND(B)$ that contains element x, called the equivalence class of x under relation $IND(B)$.

The distance-based approach is now widely used for outlier detection. An object o in a data set S is a distance-based (DB) outlier with parameters p and d, denoted by $DB(p, d)$, if at least a fraction p of the objects in S lie at a distance greater than d from o. The advantage of the distance-based approach is that no explicit distribution is needed to determine unusualness, and it can be applied to any feature or attribute space (the vector space spanned by some features is called a feature space) for which we can define a distance metric. It should be noted that metrics are measures possessing metric properties which express the degree or strength of a quality factor. And many measures that we often use are in fact not metrics. A distance metric is a distance function on a set of points, mapping pairs of points into the nonnegative real numbers. In general, any distance metric which obeys the following conditions can be used in similarity measures (Li, Chen, Li, Ma, & Vitnyi, 2003):

(1) $D(x, y) \geqslant 0$: Distances cannot be negative.
(2) $D(x, y) = 0$ if and only if $x = y$.
(3) $D(x, y) = D(y, x)$: Distance is symmetric.
(4) $D(x, y) + D(y, z) \geqslant D(x, z)$: Triangular inequality.