Expert Systems with Applications 37 (2010) 6338–6344

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An information entropy-based approach to outlier detection in rough sets

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

[a] College of Information Science and 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:
Information entropy
Outlier detection
Rough sets
Data mining

ABSTRACT

The *information entropy* in information theory, developed by Shannon, gives an effective measure of uncertainty for a given system. And it also seems a competing mechanism for the measurement of uncertainty in rough sets. Many researchers have applied the information entropy to rough sets, and proposed different information entropy models in rough sets. Especially, Düntsch et al. presented a well-justified information entropy model for the measurement of uncertainty in rough sets. In this paper, we shall demonstrate the application of this model for the study of a specific data mining problem – outlier detection. By virtue of Düntsch's information entropy model, we propose a novel definition of outliers – *IE* (*information entropy*)*-based outliers* in rough sets. An algorithm to find such outliers is also given. And the effectiveness of IE-based method for outlier detection is demonstrated on two publicly available data sets.

1. Introduction

Rough set theory introduced by Pawlak (1982, 1991), is as an extension of set theory for the study of intelligent systems characterized by insufficient and incomplete information. In recent years, rough set theory receives much attention in data mining since in rough sets, the new knowledge is formed based on the internal features of a given data. To obtain new knowledge from a given data, we need a good measure about the uncertainty of the given data and the uncertainty between the predicting knowledge and the given data. And the *information entropy* in information theory, developed by Shannon (1948), seems a competing mechanism for the measurement of uncertainty in rough sets. It is believed by Shannon that the physical entropy used in thermodynamics is more or less closely related to the concept of information as used in communication theory. Therefore, he defined the information entropy to give a measure of uncertainty.

Many researchers have applied the Shannon's information entropy to rough sets. For instance, Düntsch and Gediga first defined the information entropy and three kinds of conditional entropy in rough sets for predicting a decision attribute (Düntsch & Gediga, 1998). Beaubouef et al. proposed the information-theoretic measures of uncertainty for rough sets and rough relational databases (Beaubouef, Petry, & Arora, 1998). Sui et al. defined the information entropy and conditional entropy of similarity relations in rough relational databases (Sui, Xia, & Wang, 2003). And Wierman also presented the measures of uncertainty and granularity in rough set theory, along with an axiomatic derivation (Wierman, 1999). Furthermore, Liang et al. introduced some new definitions for *information entropy*, *rough entropy*, and *knowledge granulation* in rough sets and discussed the relationships among them (Liang & Shi, 2004).

By now, information entropy has been widely applied to rough sets (Düntsch & Gediga, 1998), rough relational databases (Beaubouef et al., 1998; Sui et al., 2003), and incomplete information systems (Liang & Shi, 2004). Many different information entropy models have been proposed, among which the model proposed by Düntsch is the most commonly used one (Düntsch & Gediga, 1998). In this paper, we shall mainly discuss the issues of how to apply Düntsch's model to the problem of outlier detection, whose aim is to detect outliers – objects who behave in an unexpected way or have abnormal properties. Detecting such outliers is important for many applications. And outlier detection is critically important in the information-based society. As an important task of data mining, outlier detection has gained much attention in recent years. Many 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), detection of employers with poor injury histories (Knorr, Ng, & Tucakov, 2000), and peculiarity-oriented mining (Zhong, Yao, & Ohshima, 2003).

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

* 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 © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.02.087

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).

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 model (Normal, Poisson, etc.) and those objects which deviate from the model are recognized as outliers (Barnett & Lewis, 1994). Its greatest disadvantage is that the distribution of the measurement data is unknown in practice.
(2) *Depth-based approach* is based on computational geometry and compute 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 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 (1998) and 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 their 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.

In this paper, we propose a new method for outlier detection, which exploits Düntsch's information entropy model in rough set theory. The main idea of our method can be described as follows. Given an information system $IS = (U, A, V, f)$, and a set of indiscernibility relations on U, since any indiscernibility relation $IND(B)$ induces a partition $U/IND(B)$ of U (i.e. a set of equivalence classes), we deem each equivalence class in the partition as a group of objects in U. That is, relation $IND(B)$ partition domain U into several groups. Then, by virtue of a given standard (i.e. the relative cardinality of every equivalence class in partition $U/IND(B)$), we can divide all the objects of U into two categories: objects belonging to the minority groups in U and objects belonging to the majority groups in U. Therefore, given any object $x \in U$, we first determine the category of x under each indiscernibility relation $IND(B)$. Then, based on the information entropy of relation $IND(B)$, we calculate the *relative entropy* of object x under $IND(B)$, which gives a measure of uncertainty for object x. Since uncertainty can be deemed as a kind of abnormal property, the aim of outlier detection is to find the small groups of objects in U that are exceptional. Given a set of indiscernibility relations on U, if object x always belongs to the minority groups in U and the relative entropies of x under these relations are always very high, then we may consider object x as a *IE* (*information entropy*)-based outlier in U wrt *IS*. In a word, an IE-based outlier in U wrt *IS* is such an element that always belongs to the minority groups in U, and the relative entropies of it are always high in view of the given knowledge.

The paper is organized as follows. In the next section, we introduce some preliminaries of rough set theory that are relevant to this paper. In Section 3, we give some definitions concerning IE-based outliers in information systems of rough set theory. An example and an algorithm to find IE-based outliers are also given. Experimental results are given in Section 4. Finally, Section 5 concludes the paper.

2. Preliminaries

In rough set terminology, a data table is also called an information system. If some of the attributes are interpreted as outcomes of classification, 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 which associates a unique value of each attribute with every object belonging to U, such that for any $a \in A$ and $x \in U$, $f(x, a) \in V_a$.

In a given information system $IS = (U, A, V, f)$, each subset of attributes $B \subseteq A$ determines a binary indiscernibility relation $IND(B)$ as follows:

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

In rough sets, indiscernibility relation $IND(B)$ is always deemed as the knowledge in *IS*. It is obvious that $IND(B)$ is an equivalence relation on U and $IND(B) = \bigcap_{a \in B} IND(\{a\})$. Given any $B \subseteq A$, equivalence 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* or *elementary set*. For every object $x \in U$, let $[x]_B$ denote the equivalence class of relation $IND(B)$ that contains element x, called the equivalence class of x under relation $IND(B)$.

Given any $B \subseteq A$ and $X \subseteq U$, the *B-lower and B-upper approximation* of set X is defined, respectively, as follows.

$$\underline{X}_B = \bigcup \{[x]_B \in U/IND(B) : [x]_B \subseteq X\}$$
$$\overline{X}_B = \bigcup \{[x]_B \in U/IND(B) : [x]_B \cap X \neq \emptyset\}$$

The set $BN_B(X) = \overline{X}_B - \underline{X}_B$ is called the B-boundary region of X. The set $NEG_B(X) = U - \overline{X}_B$ is called the B-negative region of X.

3. Information entropy-based outliers

Rough set theory has been found to have many interesting applications. The rough set approach seems to be of fundamental importance to AI and cognitive sciences, especially in the areas of machine learning and data mining (Lin & Gereone, 1996; Pawlak, Grzymala-Busse, Slowinski, & Ziarko, 1995; Skowron & Rauszer, 1992; Yao, Zhao, & Maguire, 2003). However, in rough set community, there are few concerns on the problem of outlier detection. Therefore in this section, we discuss the issues of outlier definition and detection in information systems of rough sets. In the following subsection, we first give some definitions concerning IE-based outliers. Next, an example and an algorithm to find IE-based outliers are presented.

3.1. Definitions

Our definition for IE-based outliers in an information system follows the spirit of Hawkins' definition for outliers (Hawkins, 1980). That is, given an information system $IS = (U, A, V, f)$, for any object $x \in U$, if x has some characteristics that differ greatly

from those of most objects in U, in terms of the given attributes in A, then we may call object x an outlier in U with respect to IS.

As an effective measure of uncertainty, the information entropy, proposed by Shannon (1948), has been a useful mechanism for characterizing the information content in various modes and applications in many diverse fields (Liang & Shi, 2004). In order to measure the uncertainty in rough sets, many researchers have applied the information entropy to rough sets, and proposed different information entropy models in rough sets. Especially, Düntsch et al. presented a well-justified information entropy model for the measurement of uncertainty in rough sets, which can be defined as follows (Düntsch & Gediga, 1998).

Definition 3.1. Given an information system $IS = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes. For any $B \subseteq A$, let $IND(B)$ be the indiscernibility relation (or knowledge) on U determined by B, and $U/IND(B) = \{B_1, \ldots, B_m\}$ denote the partition of U induced by $IND(B)$. The information entropy $E(B)$ of knowledge $IND(B)$ is defined by

$$E(B) = -\sum_{i=1}^{m} \frac{|B_i|}{|U|} \log_2 \frac{|B_i|}{|U|},$$

where $|B_i|/|U|$ denotes the probability of any element $x \in U$ being in equivalence class B_i, $1 \leqslant i \leqslant m$. And $|M|$ denotes the cardinality of set M.

In the above definition, for any $B \subseteq A$, if $U/IND(B) = \{U\}$, then the information entropy $E(B)$ of knowledge $IND(B)$ achieves the minimum value 0. And if $U/IND(B) = \{\{x\} : x \in U\}$, then the information entropy $E(B)$ of knowledge $IND(B)$ achieves the maximum value $\log_2|U|$. That is, for any $B \subseteq A$, $0 \leqslant E(B) \leqslant \log_2|U|$.

In this paper, in order to detect outliers in rough sets, based on the concept of information entropy defined above, we propose a new concept – *relative entropy*, which gives a measure of uncertainty for every object in domain U.

Definition 3.2. Given an information system $IS = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes. For any $B \subseteq A$, let $U/IND(B) = \{B_1, \ldots, B_m\}$ be the partition of U induced by $IND(B)$. For any $x \in U$, let $E_x(B) = -\sum_{i=1}^{m-1} \frac{|B'_i|}{|U|-|[x]_B|} \log_2 \frac{|B'_i|}{|U|-|[x]_B|}$ denote the information entropy of knowledge $IND(B)$ when removing all objects in $[x]_B$ from U, where $[x]_B$ is the equivalence class of x under relation $IND(B)$, and $U/IND(B) - \{[x]_B\} = \{B'_1, \ldots, B'_{m-1}\}$. The relative entropy $RE_B(x)$ of object x under relation $IND(B)$ is defined by

$$RE_B(x) = \begin{cases} 1 - \frac{E_x(B)}{E(B)}, & \text{if } E(B) > E_x(B); \\ 0, & \text{otherwise}, \end{cases}$$

where $E(B)$ denotes the information entropy of knowledge $IND(B)$.

The meanings of the above definition can be described as follows. Given any $B \subseteq A$ and $x \in U$, when we delete all objects in equivalence class $[x]_B$ of x from U, if the information entropy of knowledge $IND(B)$ decreases greatly, then we may consider the uncertainty of object x under $IND(B)$ is high. On the other hand, if the information entropy of knowledge $IND(B)$ varies little or even increases, then we may consider the uncertainty of object x under $IND(B)$ is low or even equals 0. Therefore, the relative entropy $RE_B(x)$ of x under $IND(B)$ gives a measure for the uncertainty of x. The higher the relative entropy $RE_B(x)$ of x, the higher the uncertainty of x.

Since the aim of outlier detection is to find the small groups of objects in U who behave in an unexpected way or have abnormal properties. And uncertainty can be deemed as a kind of abnormal property. Therefore, in this paper, we may consider those objects in U whose relative entropies are always high as behaving in an unexpected way or featuring abnormal properties when comparing with other objects in U, and utilize the information contained within the relative entropy for outlier detection.

Especially, in the above definition, if $U/IND(B) = \{[x]_B, U - [x]_B\}$, then $E_x(B) = 0$. Correspondingly, $RE_B(x) = 1$. Therefore, it is easy to verify that for any object $x \in U$, $0 \leqslant RE_B(x) \leqslant 1$.

In Section 1, we have discussed that given an information system $IS = (U, A, V, f)$, in order to find outliers in U, we first divide all the objects of U into two categories: objects belonging to the minority groups in U and objects belonging to the majority groups in U, by virtue of a given standard. Next, we shall give a definition to characterize this standard.

Definition 3.3 (*Relative cardinality*). Given an information system $IS = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes. For any $B \subseteq A$, let $U/IND(B) = \{B_1, \ldots, B_m\}$ denotes the partition of U induced by relation $IND(B)$. For any object $x \in U$, let $[x]_B$ denote the equivalence class of x under relation $IND(B)$ and $U/IND(B) - \{[x]_B\} = \{B'_1, \ldots, B'_{m-1}\}$. We define the relative cardinality $RC([x]_B)$ of equivalence class $[x]_B$ as follows:

$$RC([x]_B) = |[x]_B| - \frac{|B'_1| + \cdots + |B'_{m-1}|}{m-1},$$

where $|M|$ denotes the cardinality of set M. In particular, if $[x]_B = U$, then we assume that $RC([x]_B) = |U|$.

From the above definition, it is easy to verify that for any $x \in U$ and $B \subseteq A$, $2 - |U| \leqslant RC([x]_B) \leqslant |U|$.

Given an information system $IS = (U, A, V, f)$, for any $B \subseteq A$ and $x \in U$, let $[x]_B$ be the equivalence class of x under relation $IND(B)$. If $RC([x]_B) > 0$, then we deem object x belonging to the majority groups in U. On the other hand, if $RC([x]_B) \leqslant 0$, then we deem object x belonging to the minority groups in U.

In the following, we construct two kinds of sequence – sequence of attributes and sequence of attribute subsets (Jiang, Sui, & Cao, 2009).

Definition 3.4 (*Sequence of attributes*). Let $IS = (U, A, V, f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_k\}$. We construct a *sequence* $S = \langle a'_1, a'_2, \ldots, a'_k \rangle$ *of attributes* in A, such that for each $1 \leqslant j < k$, $E(\{a'_j\}) \leqslant E(\{a'_{j+1}\})$, where $E(\{a'_j\})$ is the information entropy of knowledge $IND(\{a'_j\})$, for every singleton subset $\{a'_j\}$ of A.

Next, through decreasing the attribute set A gradually, we can determine a descending sequence of attribute subsets.

Definition 3.5 (*Descending sequence of attribute subsets*). Let $IS = (U, A, V, f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_k\}$. Let $S = \langle a'_1, a'_2, \ldots, a'_k \rangle$ be the sequence of attributes defined above. Given a sequence $AS = \langle A_1, A_2, \ldots, A_k \rangle$ of attribute subsets, where $A_1, A_2, \ldots, A_k \subseteq A$. If $A_1 = A$, $A_k = \{a'_k\}$ and $A_{j+1} = A_j - \{a'_j\}$ for every $1 \leqslant j < k$, then we call AS a *descending sequence of attribute subsets* in IS.

From the above definition, we can see that in $AS = \langle A_1, A_2, \ldots, A_k \rangle$, for every $1 \leqslant j < k$, A_{j+1} is the attribute subset transformed from A_j by removing the element a'_j from A_j, where a'_j is the jth element in sequence S. Therefore, given an information system $IS = (U, A, V, f)$, we can uniquely determine a sequence S of attributes and a descending sequence AS of attribute subsets.

Most current methods for outlier detection give a binary classification of objects (data records): is or is not an outlier. In real-life,

it is not so simple. For many scenarios, it is more meaningful to assign to each object a degree of being an outlier. Therefore, Breunig et al. proposed a method for identifying density-based local outliers (Breunig et al., 2000). He defined a *local outlier factor* (*LOF*) that indicates the degree of outlierness of an object using only the object's neighborhood. In this paper, similar to Breunig's method, we shall define a *entropy outlier factor* (*EOF*), which can indicate the degree of outlierness for every object of the universe in an information system. And before giving the definition of *entropy outlier factor*, we define another preliminary concept – *outlierness degree under indiscernibility relation*, which indicates the degree of outlierness for every object under a given indiscernibility relation (Jiang, Sui, & Cao, 2005, 2006, 2009).

Definition 3.6 (*Outlierness degree under indiscernibility relation*). Given an information system $IS = (U, A, V, f)$, for any object $x \in U$ and any subset $B \subseteq A$ of attributes, let $[x]_B$ denotes the equivalence class of x under indiscernibility relation $IND(B)$. We define the outlierness degree $OD_B(x)$ of object x under relation $IND(B)$ as follows:

$$OD_B(x) = \begin{cases} RE_B(x) \times \left(\frac{|U| - abs(RC([x]_B))}{2|U|}\right), & \text{if } RC([x]_B) > 0; \\ RE_B(x) \times \sqrt{\frac{|U| + abs(RC([x]_B))}{2|U|}}, & \text{if } RC([x]_B) \leqslant 0, \end{cases}$$

where $RE_B(x)$ is the relative entropy of x under relation $IND(B)$, $RC([x]_B)$ is the relative cardinality of equivalence class $[x]_B$. For any real number t, $abs(t)$ denotes the absolute value of t and $|M|$ denotes the cardinality of set M.

The above definition expresses such an idea that outlier detection always concerns the minority groups of objects in U, and objects belonging to the minority groups are more likely to be outliers than objects belonging to the majority groups. Therefore, if $RC([x]_B) \leqslant 0$, that is, x belongs to the minority groups in U, then x has a more possibility to be an outlier than those objects belonging to the majority groups in U. Furthermore, the higher the relative entropy (or uncertainty) of x, the more the likelihood of x to be an outlier.

Definition 3.7 (*Entropy outlier factor*). Let $IS = (U, A, V, f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_k\}$. Let $AS = \langle A_1, A_2, \ldots, A_k \rangle$ be the descending sequence of attribute subsets in IS, the *entropy outlier factor* $EOF(x)$ of object x in IS is defined as follows:

$$EOF(x) = 1 - \frac{\sum_{j=1}^{k}\left(1 - OD_{\{a_j\}}(x)\right) \times W_{\{a_j\}}(x) + \sum_{j=1}^{k}\left(1 - OD_{A_j}(x)\right) \times W_{A_j}(x)}{2 \times |A|},$$

where $OD_{A_j}(x)$ and $OD_{\{a_j\}}(x)$ are the outlierness degrees of object x, for every attribute subset $A_j \in AS$ and singleton subset $\{a_j\}$ of A, $1 \leqslant j \leqslant k$. For any $B \subseteq A$, $W_B : U \to (0,1]$ is a weight function such that for any $x \in U$, $W_B(x) = \sqrt{\frac{|[x]_B|}{|U|}}$. And $|M|$ denotes the cardinality of set M.

Table 1
Information system *IS*.

$U \setminus A$	a	b	c
u_1	0	0	0
u_2	1	2	1
u_3	0	2	2
u_4	2	2	0
u_5	0	2	1
u_6	1	1	2

Definition 3.8 (*Information entropy-based outliers*). Let $IS = (U, A, V, f)$ be an information system, and v be a given threshold value. For any object $x \in U$, if $EOF(x) > v$, then object x is called a *IE (information entropy)-based outlier* in IS, where $EOF(x)$ is the entropy outlier factor of x in IS.

3.2. An example

Example 1. Given an information system $IS = (U, A, V, f)$, where $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$, $A = \{a, b, c\}$, as shown in Table 1. Let threshold $v = 0.6$.

The partitions induced by all singleton subsets of A are as follows:

$$U/IND(\{a\}) = \{\{u_1, u_3, u_5\}, \{u_2, u_6\}, \{u_4\}\};$$
$$U/IND(\{b\}) = \{\{u_1\}, \{u_2, u_3, u_4, u_5\}, \{u_6\}\};$$
$$U/IND(\{c\}) = \{\{u_1, u_4\}, \{u_2, u_5\}, \{u_3, u_6\}\}.$$

From Definition 3.1,

$$E(\{a\}) = -\sum_{i=1}^{3}\frac{|B_i|}{|U|}\log_2\frac{|B_i|}{|U|} = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{2}{6}\log_2\frac{2}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 1.459;$$

$$E(\{b\}) = -\sum_{i=1}^{3}\frac{|B_i|}{|U|}\log_2\frac{|B_i|}{|U|} = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{4}{6}\log_2\frac{4}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 1.252;$$

$$E(\{c\}) = -\sum_{i=1}^{3}\frac{|B_i|}{|U|}\log_2\frac{|B_i|}{|U|} = -\left(\frac{2}{6}\log_2\frac{2}{6} + \frac{2}{6}\log_2\frac{2}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 1.585.$$

And from Definition 3.2,

$$E_{u_1}(\{a\}) = E_{u_3}(\{a\}) = E_{u_5}(\{a\}) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.918;$$

$$E_{u_2}(\{a\}) = E_{u_6}(\{a\}) = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = 0.811;$$

$$E_{u_4}(\{a\}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971;$$

$$E_{u_1}(\{b\}) = E_{u_6}(\{b\}) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 0.722;$$

$$E_{u_2}(\{b\}) = E_{u_3}(\{b\}) = E_{u_4}(\{b\}) = E_{u_5}(\{b\}) = 1;$$

$$E_{u_1}(\{c\}) = E_{u_4}(\{c\}) = E_{u_2}(\{c\}) = E_{u_5}(\{c\}) = E_{u_3}(\{c\}) = E_{u_6}(\{c\}) = 1.$$

Correspondingly, we can obtain that

$$RE_{\{a\}}(u_1) = RE_{\{a\}}(u_3) = RE_{\{a\}}(u_5) = 1 - \frac{0.918}{1.459} = 0.371;$$

$$RE_{\{a\}}(u_2) = RE_{\{a\}}(u_6) = 1 - \frac{0.811}{1.459} = 0.444;$$

$$RE_{\{a\}}(u_4) = 1 - \frac{0.971}{1.459} = 0.334;$$

$$RE_{\{b\}}(u_1) = RE_{\{b\}}(u_6) = 1 - \frac{0.722}{1.252} = 0.423;$$

$$RE_{\{b\}}(u_2) = RE_{\{b\}}(u_3) = RE_{\{b\}}(u_4) = RE_{\{b\}}(u_5) = 1 - \frac{1}{1.252} = 0.201;$$

$$RE_{\{c\}}(u_1) = RE_{\{c\}}(u_2) = RE_{\{c\}}(u_3) = RE_{\{c\}}(u_4) = RE_{\{c\}}(u_5) = RE_{\{c\}}(u_6) = 1 - \frac{1}{1.585} = 0.369.$$

In addition, from Definition 3.3,

$$RC([u_1]_{\{a\}}) = 1.5, \quad RC([u_2]_{\{a\}}) = 0, \quad RC([u_4]_{\{a\}}) = -1.5;$$
$$RC([u_1]_{\{b\}}) = -1.5, \quad RC([u_2]_{\{b\}}) = 3, \quad RC([u_6]_{\{b\}}) = -1.5;$$
$$RC([u_1]_{\{c\}}) = 0, \quad RC([u_2]_{\{c\}}) = 0, \quad RC([u_3]_{\{c\}}) = 0.$$

Next, we construct two sequences from $E(\{a\}), E(\{b\})$, and $E(\{c\})$. From Definition 3.4, the sequence of attributes is $S = \langle b, a, c\rangle$. And the descending sequence of attribute subsets is $AS = \langle A_1, A_2, A_3\rangle = \langle\{a, b, c\}, \{a, c\}, \{c\}\rangle$.

For $A_1 \in AS,\ U/IND(A_1) = \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5\}, \{u_6\}\}$;

For $A_2 \in AS,\ U/IND(A_2) = \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5\}, \{u_6\}\}$;

For $A_3 \in AS,\ U/IND(A_3) = \{\{u_1, u_4\}, \{u_2, u_5\}, \{u_3, u_6\}\}$.

Analogously, we can obtain that

$$RE_{A_1}(u_1) = RE_{A_1}(u_2) = RE_{A_1}(u_3) = RE_{A_1}(u_4) = RE_{A_1}(u_5) = RE_{A_1}(u_6) = 1 - \frac{\log_2 5}{\log_2 6} = 0.102;$$

$$RE_{A_2}(u_1) = RE_{A_2}(u_2) = RE_{A_2}(u_3) = RE_{A_2}(u_4) = RE_{A_2}(u_5) = RE_{A_2}(u_6) = 1 - \frac{\log_2 5}{\log_2 6} = 0.102;$$

$$RE_{A_3}(u_1) = RE_{A_3}(u_2) = RE_{A_3}(u_3) = RE_{A_3}(u_4) = RE_{A_3}(u_5) = RE_{A_3}(u_6) = 1 - \frac{1}{1.585} = 0.369.$$

$$RC([u_1]_{A_1}) = RC([u_2]_{A_1}) = RC([u_3]_{A_1}) = RC([u_4]_{A_1}) = RC([u_5]_{A_1}) = RC([u_6]_{A_1}) = 0;$$

$$RC([u_1]_{A_2}) = RC([u_2]_{A_2}) = RC([u_3]_{A_2}) = RC([u_4]_{A_2}) = RC([u_5]_{A_2}) = RC([u_6]_{A_2}) = 0;$$

$$RC([u_1]_{A_3}) = RC([u_2]_{A_3}) = RC([u_3]_{A_3}) = RC([u_4]_{A_3}) = RC([u_5]_{A_3}) = RC([u_6]_{A_3}) = 0.$$

For object $u_1 \in U$, from Definition 3.6, we can obtain that

$$OD_{\{a\}}(u_1) = 0.371 \times \frac{6 - 1.5}{12} = 0.139;$$

$$OD_{\{b\}}(u_1) = 0.423 \times \sqrt{\frac{6 + 1.5}{12}} = 0.334;$$

$$OD_{\{c\}}(u_1) = 0.369 \times \sqrt{\frac{6}{12}} = 0.261;$$

$$OD_{A_1}(u_1) = 0.102 \times \sqrt{\frac{6}{12}} = 0.072;$$

$$OD_{A_2}(u_1) = 0.102 \times \sqrt{\frac{6}{12}} = 0.072;$$

$$OD_{A_3}(u_1) = 0.369 \times \sqrt{\frac{6}{12}} = 0.261.$$

Hence, the entropy outlier factor of u_1 is as follows

$$EOF(u_1) = 1 - \frac{(1 - 0.139) \times \sqrt{\frac{3}{6}} + (1 - 0.334) \times \sqrt{\frac{1}{6}} + (1 - 0.261) \times \sqrt{\frac{2}{6}}}{2 \times 3} - \frac{(1 - 0.072) \times \sqrt{\frac{1}{6}} + (1 - 0.072) \times \sqrt{\frac{1}{6}} + (1 - 0.261) \times \sqrt{\frac{2}{6}}}{2 \times 3} \approx 0.5847 < v.$$

Therefore, u_1 is not a IE-based outlier in *IS*.

Analogously, we can obtain that $EOF(u_2) \approx 0.5262 < v$, $EOF(u_3) \approx 0.5008 < v$, $EOF(u_4) \approx 0.5522 < v$, $EOF(u_5) \approx 0.5008 < v$, and $EOF(u_6) \approx 0.6202 > v$. Therefore, u_6 is an IE-based outlier in *IS*. Other objects in U are all not IE-based outliers in *IS*.

3.3. Algorithm for detecting IE-based outliers

Algorithm 1.

Input: information system $IS = (U, A, V, f)$, where $A = \{a_1, a_2, \ldots, a_k\}$, $|U| = n$; thresholds v
Output: a set E of IE-based outliers in *IS*
Initialization: Let $E = \emptyset$

(1) For every $a \in A$
(2) {
(3) Sort all objects from U according to a given order (e.g. the lexicographical order) on domain V_a of attribute a;
(4) Determine the partition $U/IND(\{a\})$;
(5) Calculate $E(\{a\})$, the information entropy of $IND(\{a\})$
(6) }
(7) Determine the sequence $S = \{a'_1, a'_2, \ldots, a'_k\}$ of attributes in A, where for each $1 \leqslant j < k,\ E(\{a'_j\}) \leqslant E(\{a'_{j+1}\})$;
(8) Construct the descending sequence $AS = \{A_1, \ldots, A_k\}$ of attribute subsets from sequence S;
(9) For $1 \leqslant j \leqslant k$
(10) {
(11) Sort all objects from U according to a given order (e.g. the lexicographical order) on domain V_{A_j} of attribute subset A_j;
(12) Determine the partition $U/IND(A_j)$;
(13) Calculate $E(A_j)$, the information entropy of $IND(A_j)$
(14) }
(15) For every $x \in U$
(16) {
(17) For $1 \leqslant j \leqslant k$
(18) {
(19) Calculate $RE_{\{a_j\}}(x)$ and $RE_{A_j}(x)$, the relative entropies of x under relations $IND(\{a_j\})$ and $IND(A_j)$, respectively;
(20) Calculate $RC([x]_{\{a_j\}})$ and $RC([x]_{A_j})$, the relative cardinal- ities of equivalence classes $[x]_{\{a_j\}}$ and $[x]_{A_j}$, respectively;
(21) Calculate $OD_{\{a_j\}}(x)$ and $OD_{A_j}(x)$, the outlierness degrees of x under $IND(\{a_j\})$ and $IND(A_j)$, respectively;
(22) Assign two weights $W_{\{a_j\}}(x)$ and $W_{A_j}(x)$ to x, respectively
(23) }
(24) Calculate $EOF(x)$, the entropy outlier factor of object x;
(25) If $EOF(x) > v$ then $E = E \cup \{x\}$
(26) }
(27) Return E.

In Algorithm 1, we use a method proposed by Nguyen and Nguyen (1996) which can calculate the partition induced by an indiscernibility relation $IND(B)$ in $O(|B| \times nlogn)$ time, where $|B|$ and n are the cardinalities of B and U, respectively.

In the worst case, the time complexity of algorithm 1 is $O(k^2 \times nlogn)$, and its space complexity is $O(k \times (n + k))$, where k, n are the cardinalities of A and U, respectively.

4. Experimental results

To evaluate IE-based method for outlier detection, we ran our algorithm on two real-life data sets (*lymphography* and *cancer*) obtained from the UCI Machine Learning Repository (Bay, 1999). In

this section, we shall use these two data sets to demonstrate the performance of IE-based method against traditional distance-based method (Knorr & Ng, 1998; Knorr et al., 2000) and KNN algorithm (Ramaswamy, Rastogi, & Shim, 2000). In addition, on the *cancer* data set, we add the results of RNN-based outlier detection method for comparison, these results can be found in the work of Harkins, He, Willams, and Baxter (2002), Willams, Baxter, He, Harkins, and Gu (2002).

In our experiment, for the KNN algorithm, the results were obtained by using the 5th nearest neighbor (Ramaswamy et al., 2000) and the overlap metric in rough set theory (Jiang et al., 2009). Moreover, since in traditional distance-based outlier detection, being an outlier is regarded as a binary property, here we revise the definition of distance-based outlier detection by introducing a *distance outlier factor* (DOF) to indicate the degree of outlierness for every object in an information system (Jiang et al., 2005, Jiang, Sui, & Cao, 2006, 2009).

Definition 4.1 (*Distance outlier factor*). Given an information system $S = (U, A, V, f)$. For any object $x \in U$, the percentage of the objects in U lie greater than d' from x is called the *distance outlier factor of x* in S, denoted by

$$DOF(x) = \frac{|\{y \in U : dist(x,y) > d'\}|}{|U|},$$

where $dist(x,y)$ denotes the distance between objects x and y under a given distance metric (In our experiment, the overlap metric in rough set theory is adopted again (Jiang et al., 2009)), d' is a given parameter (In our experiment, we set $d' = |A|/2$), and $|M|$ denotes the cardinality of set M.

4.1. Lymphography data

The first is the lymphography data set, which can be found in the UCI machine learning repository (Bay, 1999). It contains 148 instances (or objects) with 19 attributes (including the class attribute). The 148 instances are partitioned into 4 classes: "normal find" (1.35%), "metastases" (54.73%), "malign lymph" (41.22%), and "fibrosis" (2.7%). Classes 1 and 4 ("normal find" and "fibrosis") are regarded as rare classes.

In (Aggarwal & Yu, 2001), Aggarwal et al. proposed a practicable way to test the effectiveness of an outlier detection method. That is, we can run the outlier detection method on a given data set and test the percentage of points which belonged to one of the rare classes (Aggarwal considered those kinds of class labels which occurred in less than 5% of the data set as rare labels (Aggarwal & Yu, 2001)). Points belonged to the rare class are considered as outliers. If the method works well, we expect that such abnormal classes would be over-represented in the set of points found.

In the lymphography data set, classes 1 and 4 ("normal find" and "fibrosis") should be regarded as rare class labels since they occur in less than 5% of the data set. In our experiment, data in the lymphography data set are input into an information system $IS_L = (U, A, V, f)$, where U contains all the 148 instances of lymphography data set and A contains 18 attributes of lymphography data set (not including the class attribute). We consider detecting outliers (rare classes) in IS_L. The experimental results are summarized in Table 2.

Table 2
Experimental Results in IS_L.

Top ratio (%) (number of objects)	Number of rare class included (coverage)		
	IE (%)	DIS (%)	KNN (%)
4(6)	5(83)	5(83)	4(67)
5(7)	5(83)	5(83)	4(67)
6(9)	6(100)	6(100)	4(67)
8(12)	6(100)	6(100)	5(83)
10(15)	6(100)	6(100)	6(100)

Table 3
Experimental results in IS_W.

Top ratio (%) (number of objects)	Number of rare class included (coverage)			
	IE (%)	DIS (%)	RNN (%)	KNN (%)
1(4)	4(10)	4(10)	3(8)	4(10)
2(8)	7(18)	5(13)	6(15)	8(21)
4(16)	15(38)	11(28)	11(28)	16(41)
6(24)	21(54)	18(46)	18(46)	20(51)
8(32)	28(72)	24(62)	25(64)	27(69)
10(40)	33(85)	29(74)	30(77)	32(82)
12(48)	36(92)	36(92)	35(90)	37(95)
14(56)	39(100)	39(100)	36(92)	39(100)
16(64)	39(100)	39(100)	36(92)	39(100)
18(72)	39(100)	39(100)	38(97)	39(100)
20(80)	39(100)	39(100)	38(97)	39(100)
28(112)	39(100)	39(100)	39(100)	39(100)

In Table 2, "IE", "DIS", "KNN" denote IE-based, traditional distance-based and KNN-based methods for outlier detection, respectively. For every object in U, the degree of outlierness is calculated by using the three outlier detection methods, respectively. For each outlier detection method, the "Top ratio (%) (number of objects)" denotes the percentage (number) of the objects selected from U whose degrees of outlierness calculated by the method are higher than those of other objects in U. And if we use $X \subseteq U$ to contain all those objects selected from U, then the "Number of rare class included" is the number of objects in X that belong to one of the rare classes. The "coverage" is the ratio of the "Number of rare class included" to the number of objects in U that belong to one of the rare classes (He, Deng, & Xu, 2005).

From Table 2, we can see that for the lymphography data set, IE-based and distance-based methods have the same performances, and they perform markedly better than KNN-based method.

4.2. Wisconsin breast cancer data

The Wisconsin breast cancer data set is found in the UCI machine learning repository (Bay, 1999). The data set contains 699 instances with 9 continuous attributes. Here, we follow the experimental technique of Harkins et al. by removing some of the *malignant* instances to form a very unbalanced distribution (Harkins et al., 2002; Willams et al., 2002). The resultant data set had 39 *malignant* instances and 444 *benign* instances. And the 9 continuous attributes in the data set are transformed into categorical attributes, respectively.[1] Here, *malignant* instances are deemed as outliers (He et al., 2005).

Similar to the treatment for the lymphography data set, data in the Wisconsin breast cancer data set are also input into an information system $IS_W = (U', A', V', f')$, where U' contains 483 instances of the data set and A' contains 9 categorical attributes of the data set. We detect outliers (*malignant* instances) in IS_W. The experimental results are summarized in Table 3.

Table 3 is similar to Table 2. From Table 3, we can see that for the Wisconsin breast cancer data set, the performances of IE-based and KNN-based methods are very close, and they perform better than the other two methods – RNN-based and distance-based methods, where the RNN-based method is the worst.

5. Conclusion

Outlier detection is becoming an important task for many data mining applications. In this paper, we presented a new method for

[1] The resultant data set is public available at: http://research.cmis.csiro.au/rohanb/outliers/breast-cancer/.

outlier detection, which exploits the information entropy model proposed by Düntsch et al. in rough sets. Given an information system, we first divided all objects of domain U into two categories: objects belonging to the minority groups and objects belonging to the majority groups, through calculating the relative cardinality. Then, we calculated the relative entropy for every object in U. Differing from the information entropy, relative entropy gives a measure of uncertainty for each object. And those objects that always belong to the minority groups and the relative entropies of which are always high have more likelihood to be an outlier than other objects in U. Experimental results on real data sets demonstrate the effectiveness of our method for outlier detection.

In the future work, to reduce the time complexity of IE-based outlier detection algorithm, we shall consider adopting some methods of attribute reduction in rough sets (Skowron & Rauszer, 1992) to have smaller number of attributes while preserving the performance of our algorithm.

Acknowledgements

This work is supported by the Natural Science Foundation(Grant Nos. 60802042, 60573063 and 60573064), the National 863 Program (Grant No. 2007AA01Z325).

References

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD international conference on managment of data* (pp. 37–46). California, USA.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. John Wiley and Sons.

Bay, S. D. (1999). The UCI KDD repository. Available online at: <http://kdd.ics.uci.edu>.

Beaubouef, T., Petry, F. E., & Arora, G. (1998). Information-theoretic measures of uncertainty for rough sets and rough relational databases. *Information Sciences, 109*, 535–563.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review (with discussion). *Statistical Science, 17*(3), 235–255.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 93–104). Dallas, USA.

Düntsch, I., & Gediga, G. (1998). Uncertainty measures of rough set prediction. *Artificial Intelligence, 106*, 109–137.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In D. Barbar et al. (Eds.), *Data mining for security applications*. Boston: Kluwer Academic Publishers.

Harkins, S., He, H. X., Willams, G. J., & Baxter, R. A. (2002). Outlier detection using replicator neural networks. In *Proceedings of the fourth international conference on data warehousing and knowledge discovery* (pp. 170–180). France.

Hawkins, D. (1980). *Identifications of outliers*. London: Chapman and Hall.

He, Z. Y., Deng, S. C., & Xu, X. F. (2005). An optimization model for outlier detection in categorical data. In *International conference on intelligent computing (ICIC(1) 2005)* (pp. 400–409). Hefei, China.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Jiang, F., Sui, Y. F., Cao, & C. G. (2005). Outlier detection using rough set theory. In *Proceedings of the 10th international conference on rough sets, fuzzy sets, data mining, and granular computing (RSFDGrC(2) 2005)* (pp. 79-87). Regina, Canada: LNAI 3642.

Jiang, F., Sui, Y. F., & Cao, C. G. (2006). Outlier detection based on rough membership function. In *Proceedings of the fifth international conference on rough set and knowledge technology (RSCTC 2006)* (pp. 388–397). Japan: LNAI 4259.

Jiang, F., Sui, Y. F., & Cao, C. G. (2009). Some issues about outlier detection in rough set theory. *Expert Systems with Applications, 36*(3), 4680–4687.

Johnson, T., Kwok, I., & Ng, R. T. (1998). Fast computation of 2-dimensional depth contours. In *Proceedings of the fourth international conference on knowledge discovery and data mining* (pp. 224–228). New York.

Knorr, E., & Ng, R. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th VLDB conference* (pp. 392–403). New York.

Knorr, E., Ng, R., & Tucakov, V. (2000). Distance-based outliers: Algorithms and applications. *VLDB Journal: Very Large Databases, 8*(3–4), 237–253.

Kovács, L., Vass, D., & Vidács, A. (2004). Improving quality of service parameter prediction with preliminary outlier detection and elimination. In *Proceedings of the second international workshop on inter-domain performance and simulation (IPS 2004)* (pp. 194–199). Budapest.

Lane, T., & Brodley, C. E. (1999). Temporal sequence learning and data reduction for anomaly detection. *ACM Transactions on Information and System Security, 2*(3), 295–331.

Liang, J. Y., & Shi, Z. Z. (2004). The information entropy, rough entropy and knowledge granulation in rough set theory. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 12*(1), 37–46.

Lin, T. Y., & Gereone, N. (1996). *Rough sets and data mining: Analysis of imprecise data*. Dordrecht: Kluwer Academic.

Nguyen, S. H., & Nguyen, H. S. (1996). Some efficient algorithms for rough set methods. In *IPMU'96* (pp. 1451–1456). Granada, Spain.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences, 11*, 341–356.

Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*. Dordrecht: Kluwer Academic Publishers.

Pawlak, Z., Grzymala-Busse, J. W., Slowinski, R., & Ziarko, W. (1995). Rough sets. *Communications of the ACM, 38*(11), 89–95.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large datasets. In *Proceedings of the ACM SIGMOD conference on management of data* (pp. 427–438). Dallas.

Shannon, C. E. (1948). The mathematical theory of communication. *Bell System Technical Journal, 27*(3–4), 373–423.

Skowron, A., & Rauszer, C. (1992). The discernibility matrices and functions in information systems. *Handbook of applications and advances of rough set theory* (Vol. 11, pp. 331–362). Dordrecht: Kluwer Academic Publishers.

Sui, Y. F., Xia, Y. M., & Wang, J. (2003). The information entropy of rough relational databases. In *Proceedings of the ninth international conference on rough sets, fuzzy sets, data mining, and granular computing* (pp. 320–324). China.

Wierman, M. J. (1999). Measuring uncertainty in rough set theory. *International Journal of General Systems, 28*(4), 283–297.

Willams, G. J., Baxter, R. A., He, H. X., Harkins, S., & Gu, L. F. (2002). A comparative study of RNN for outlier detection in data mining. In *Proceedings of the 2002 IEEE international conference on data mining* (pp. 709–712). Japan.

Yao, Y. Y., Zhao, Y., & Maguire, R. B. (2003). Explanation oriented association mining using rough set theory. In *Proceedings of the ninth international conference on rough sets, fuzzy sets, data mining, and granular computing* (pp. 165–172). China.

Zhong, N., Yao, Y. Y., & Ohshima, M. (2003). Peculiarity oriented multi-database mining. *IEEE Transactions on Knowlegde and Data Engineering, 15*(4), 952–960.