Expert Systems With Applications 112 (2018) 243–257

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures

Zhong Yuan[a,b], Xianyong Zhang[a,b,*], Shan Feng[a]

[a] College of Mathematics and Software Science, Sichuan Normal University, Chengdu 610066, China
[b] Institute of Intelligent Information and Quantum Information, Sichuan Normal University, Chengdu 610066, China

ARTICLE INFO

Article history:
Received 14 September 2017
Revised 30 May 2018
Accepted 6 June 2018
Available online 7 June 2018

Keywords:
Outlier detection
Neighborhood rough set
Neighborhood information entropy
Hybrid data-driving
Data mining

ABSTRACT

The outlier relies on its distinctive mechanism and valuable information to play an important role in expert and intelligent systems, and thus outlier detection has already been extensively applied in relevant fields including the fraud detection, medical diagnosis, public security, etc. The outlier detection methods of rough sets recently gain in-depth research, because they are data-driven and never require additional knowledge. However, classical rough set-based methods consider only categorical data; furthermore, neighborhood rough sets adhere to numeric and heterogeneous data, but their outlier detection is mainly restricted to numeric data now. According to the hybrid data-driving, this paper investigates outlier detection by the neighborhood information entropy and its developmental measures, and the applicable data sets widely concern categorical, numeric, and mixed data; as a result, the new method extends both the traditional distance-based and rough set-based methods to enrich outlier detection. Concretely, the neighborhood information system is first determined by the heterogeneous distance and self-adapting radius, the neighborhood information entropy is then defined to implement whole uncertainty measurement, three gradual information measures are further constructed to describe each single object, and finally the neighborhood entropy-based outlier factor (NEOF) is integratedly established to detect outliers; moreover, the NEOF-based outlier detection algorithm (called the NIEOD algorithm) is designed and applied. By virtue of UCI data experiments, the NIEOD algorithm is compared with six existing detection algorithms (including the NED, IE, SEQ, FindCBLOF, DIS, KNN algorithms), and the concrete results generally reflect the better effectiveness and adaptability of the new method.

1. Introduction

Knowledge discovery in databases (KDD), or data mining, is an important issue in the development of knowledge-based and data-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 (e.g., the traditional pattern recognition aims to construct a general pattern map to the majority of data), outlier detection targets to find the rare data whose behavior is very exceptional when compared with rest large amount of data. In fact, an outlier (also known as an anomaly) is a data point that significantly deviates from the rest data objects in data set (Hawkins, 1980), and it usually adheres to a new perspective or a specific mechanism to become more exciting than normal instances in knowledge discovery. As a result, the outlier relies on its distinctive mechanism and valuable information to play an important role in expert and intelligent systems, so outlier detection has already extensively applied in relevant fields including the intrusion detection, image processing, medical treatment, public security, etc (Han, Kamber, & Pei, 2011; Knorr & Ng, 1998). At present, outlier detection methods and their development exhibit both the theoretical significance and applied value in data mining, and thus this paper aims to establish a novel detection approach to process hybrid data which generally exist in practical systems.

Outlier detection concerns three traditional methods, i.e., the statistical method (Rousseeuw & Leroy, 1987), the proximity-based approach (Breunig, Kriegel, Ng, & Sander, 2000; Knorr & Ng, 1997; Knorr, Ng, & Tucakov, 2000), and the clustering-based method (Jain, 1999), to offer different features and advantages. The statistical method assures that normal data objects are generated by a statistical model, so abnormal points which never obey the model become outliers; this approach applies to data sets with the known

* Corresponding author.
E-mail addresses: yz2799@163.com (Z. Yuan), xianyongzh@sina.com.cn (X. Zhang), fengshanrq@sohu.com (S. Feng).

https://doi.org/10.1016/j.eswa.2018.06.013
0957-4174/

distribution and simplex attribute. The proximity-based approach emerges to improve the statistical way, and it usually adopts two basic strategies: the distance-based detection (Knorr et al., 2000) and the density-based detection (Breunig et al., 2000). Moreover, the clustering-based method mainly depends on different clustering ways to exhibit different effectiveness.

Most of above traditional detection methods require some additional information. Thus, detection methods based on rough sets recently gain in-depth research, because they are data-driven and never require additional knowledge. In fact, the traditional distance-based method computes an object distance to more apply to numeric data rather than categorical data, because the latter data never have a similar distance relationship. For this issue, rough sets are introduced into outlier detection to handle categorical data (Berna-Martinez & Ortega, 2015; Chen, Miao, & Wang, 2008; Jiang & Chen, 2015; Jiang, Sui, & Cao, 2008; 2009; 2010; 2011; Shaari, Bakar, & Hamdan, 2009); in particular, rough sets originate from the study of intelligent systems characterized by insufficient and incomplete information (Pawlak, 1982; 1991), and they have been successfully applied in machine learning, data mining, pattern recognition, etc. However, classical rough set-based detection methods consider only the basic equivalence to directly apply to categorical (or nominal) data rather than numeric data; in fact, numeric data can be discretized to follow the rough set way, but this preprocessing usually leads to time increase and information loss. Except the numeric and categorical data, both-combined mixed (or hybrid or heterogeneous) data universally exist in the real world, and their studies on outlier detection are undoubtedly required and challenging but there are rarely relevant reports.

To improve classical rough sets, neighborhood rough sets adopt the robust neighborhood to adhere to numeric and hybrid data, thus providing a more powerful platform. In early research, neighborhood spaces are thought to become more general topological spaces (Lin, 1988; 2008), neighborhood approximation properties are revealed (Wu & Zhang, 2002; Yao, 1998), and neighborhood rough sets are utilized for heterogeneous data reduction (Hu, Liu, & Yu, 2008; Hu, Yu, Liu, & Wu, 2008; Hu, Yu, & Xie, 2006). At present, neighborhood rough sets have been effectively and deeply applied in the attribute reduction, feature selection, classification recognition, and uncertainty reasoning, etc (Chen, Li, Cai, Luo, & Fujita, 2016; Chen, Zhang, Zheng, Ying, & Yu, 2017; Kumar & Inbarani, 2016; Liu et al., 2017; Wang, Shao, He, Qian, & Qi, 2016). However for outlier detection, the neighborhood rough sets-based ways never gain enough attentions, especially regarding mixed data processing. In fact, relevant neighborhood-based detection works are mainly restricted to numeric data (Chen, Miao, & Zhang, 2010; Li & Rao, 2012).

Against the above background, the hybrid data-driven outlier detection based on neighborhood rough sets becomes a valuable and novel work, and thus this paper mainly makes a preliminary study by virtue of information measure construction. In fact, the information entropy, proposed by Shannon (1948), establishes a fundamental mechanism of uncertainty measurement; it has been introduced into classical rough sets to make uncertainty representation via multiple entropy forms or information measures (Chen, Zhang et al., 2017; Düntsch & Gediga, 1998; Liang, Shi, Li, & Wierman, 2006; Liang, Wang, & Qian, 2009; Wang, Ma, & Yu, 2015; Zhang, Mei, Chen, & Li, 2016; Zhang & Miao, 2017), and the neighborhood entropy is particularly discussed by Chen, Wu, Chen, Tang, and Zhu (2014), Chen, Xue, Ma, and Xu (2017) and Li and Rao (2012). In this paper, the hybrid data-driven outlier detection is concretely investigated by the neighborhood information entropy and its developmental measures, and a corresponding detection algorithm (i.e., the neighborhood information entropy-based outlier detection (NIEOD) algorithm) is designed and applied. Concretely, the neighborhood information system is first determined by the heterogeneous distance and self-adapting radius, the neighborhood information entropy and its three in-depth measures are then mined to describe data objects by uncertainty measurement, and finally the neighborhood entropy-based outlier factor (NEOF) is integratedly established to provide the outlier detection and NIEOD algorithm. Based on relevant UCI data experiments, the NIEOD algorithm is compared with six main detection algorithms (including the NED, IE, SEQ, FindCBLOF, DIS, KNN algorithms), and the obtained results show that the new method generally has better effectiveness and adaptability. Regarding the contributions, the new method extends both the traditional distance-based and rough set-based methods to enrich outlier detection, and thus it extensively applies to categorical, numeric, and heterogeneous data.

The remainder of this paper is organized as follows. Section 2 reviews the neighborhood information system; Section 3 constructs the core outlier detection by developing information measures, and three subsections are provided to state the theoretical method, specific algorithm, and illustrative example; Section 4 makes data experiments and analyses via three typical UCI data sets; finally, Section 5 concludes the paper.

2. Neighborhood information system

Note that neighborhood rough sets have a basic formal background: the neighborhood information system. This fundamental system is reviewed in this section via several references (Hu, Liu et al., 2008; Hu, Yu, Liu et al., 2008; Hu, Yu, & Xie, 2008).

Usually, an information system is a basis of data mining, and can be written as a quadruple $IS = (U, A, V, f)$. Herein, universe $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects; A is a nonempty finite set of attributes; $V = \bigcup_{a \in A} V_a$ is the union of attribute domain V_a; f: $U \times A \to V$ is an information function:

$$f(x, a) \in V_a, \forall x \in U, \forall a \in A.$$

Furthermore, the information system can be specialized into a decision system $DS = (U, C \cup D, V, f)$ if $A = C \cup D$ and $C \cap D = \emptyset$, where

$$C = \{c_1, c_2, \ldots, c_m\} = \{c_j \mid j \in \{1, \ldots, m\}\}$$

and D denote the condition and decision attribute sets, respectively. Next, give $B \subseteq A$ and $B \subseteq C$ for the information and decision systems, respectively, and let

$$B = \{c_{j_1}, \ldots, c_{j_k}\} = \{c_{j_h} \mid h \in \{1, \ldots, k\}\} \ (k \in \{1, \ldots, m\}).$$

The distance is a basic tool in the two data systems to construct neighborhood rough sets. $\forall x, y, z \in U$, B-based distance function d_B : $U \times U \to \mathbf{R}^+$ ($\mathbf{R}^+$ is the set of nonnegative real numbers) satisfies three conditions:

(1) $d_B(x, y) \geq 0$, $d_B(x, y) = 0 \Leftrightarrow x = y$;
(2) $d_B(x, y) = d_B(y, x)$;
(3) $d_B(x, z) \leq d_B(x, y) + d_B(y, z)$.

For example, the Minkowski distance d_B is defined by

$$d_B^p(x, y) = \sqrt[p]{\sum_{h=1}^{k} |f(x, c_{j_h}) - f(y, c_{j_h})|^p}.$$

In particular, d_B^p becomes the Manhattan distance if $p = 1$, it becomes the Euclidean distance if $p = 2$, and it becomes the Chebychev distance if $p = \infty$. Moreover, a number of distance functions (Wilson & Martinez, 1997) are usually utilized, such as the Heterogeneous Euclidean-Overlap Metric (HEOM), the Value Difference Metric (VDM), the Heterogeneous Value Difference Metric (HVDM), and the Interpolated Value Difference Metric (IVDM).

Based on the distance function, the neighborhood radius is further introduced to granulate the universe, and thus the neighborhood, the neighborhood relation and knowledge are formed to construct the neighborhood information system.

Definition 1. The ε-neighborhood of $x \in U$ on B is

$$n_B^\varepsilon(x) = \{y \in U \mid d_B(x, y) \le \varepsilon\}, \tag{1}$$

where parameter $\varepsilon \ge 0$ is called the neighborhood radius. Furthermore, the $B - \varepsilon$ neighborhood relation on U emerges

$$nr_B^\varepsilon = \{(x, y) \in U \times U \mid y \in n_B^\varepsilon(x)\}, \tag{2}$$

while quotient set U/nr_B^ε constitutes the neighborhood covering or knowledge on U.

According to Definition 1, the neighborhood is determined and represented by both the distance function and neighborhood radius, and $n_B^\varepsilon(x)$ collects such elements whose B-based distance is not more than threshold ε. Hence, the neighborhood acts as a feature granule based on the distance. The neighborhood relation is a kind of similarity relations (with only reflexivity and symmetry) to describe objects' indistinguishability in terms of the distance. When $\varepsilon = 0$, nr_B^ε becomes the smallest equivalence relation to apply to categorical data; when $\varepsilon > 0$, nr_B^ε becomes the usual coarser similarity relation to apply to numeric data. Finally, the neighborhood knowledge constitutes a neighborhood covering of U to establish the underlying granular structure.

Definition 2. A neighborhood information system (NIS) and a neighborhood decision system (NDS) are, respectively, defined as the four tuples

$$\begin{aligned} NIS &= (U, NR_C^\varepsilon, V, f), \\ NDS &= (U, NR_C^\varepsilon \cup D, V, f), \end{aligned} \tag{3}$$

where $NR_C^\varepsilon = \{nr_B^\varepsilon \mid B \subseteq C\}$ means the set of all neighborhood relations on U.

The neighborhood information (or decision) system underlies development of neighborhood rough sets, as well as next research of outlier detection. They are mainly induced by different distance functions and neighborhood radii. In a special case $\varepsilon = 0$, both systems degenerate into the classical systems based on the equivalent relation, respectively, because then the neighborhood degenerates into the equivalence class, i.e., $n_B^\varepsilon(x) = [x]_B$. This case explains the expansion property of neighborhood rough sets (in the neighborhood information/decision system) for the classical rough sets (in the information/decision system).

3. Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures

Based on the neighborhood information (or decision) system, this section develops the neighborhood information entropy and its subsequent measures to gradually implement outlier detection, and there are three main parts to offer the theoretical method, specific algorithm, and illustrative example.

3.1. Theoretical method

By virtue of measure development, this subsection mainly establishes a new method of outlier detection, and the discussion items concretely include the data normalization preprocessing, neighborhood distance selection, adaptive neighborhood radius, outlier measure construction, and metric outlier discrimination.

At first, data processing in an information system usually concerns some difference regarding the order of magnitude or the dimension of quantity. To avoid the influence of different data, the original numeric data can be normalized before data processing to obtain accurate results. Several methods of data normalization are commonly used (Indurkhya & Nitin, 1998; Kennedy & Ruby, 1997), such as the min-max normalization, z-score normalization, and decimal scaling normalization. In this paper, the min-max normalization is mainly adopted for preprocessing, and the renewed formula of attribute values becomes

$$\begin{aligned} F(f(x_i, c_j)) = \frac{f(x_i, c_j) - min_{c_j}}{max_{c_j} - min_{c_j}} \in [0, 1] \\ (i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}), \end{aligned} \tag{4}$$

where max_{c_j} and min_{c_j} denote the maximum and minimum regarding attribute c_j in universe U, respectively.

The practical data generally concern the numeric and heterogeneous attributes, and the latter contain both numeric and categorical attributes. To well handle such complex data, the Heterogeneous Euclidean-Overlap Metric (HEOM) (Wilson & Martinez, 1997) is particularly utilized in our studies to represent the neighborhood distance.

Definition 3. The Heterogeneous Euclidean-Overlap Metric (HEOM) of $x, y \in U$ regarding B refers to

$$HEOM_B(x, y) = \sqrt{\sum_{h=1}^{k} d_{c_{j_h}}(x, y)^2}, \tag{5}$$

where $d_{c_{j_h}}(x, y) =$

$$\begin{cases} 1, & \text{if attribute values of } x \text{ and } y \text{ are unknown regarding attribute } c_{j_h}; \\ 0, & \text{if } c_{j_h} \text{ is a categorical attribute and } f(x, c_{j_h}) = f(y, c_{j_h}); \\ 1, & \text{if } c_{j_h} \text{ is a categorical attribute and } f(x, c_{j_h}) \ne f(y, c_{j_h}); \\ |f(x, c_{j_h}) - f(y, c_{j_h})|, & \text{if } c_{j_h} \text{ is a numeric attribute.} \end{cases} \tag{6}$$

The measure HEOM can deal with not only the numeric data but also the heterogeneous data; moreover, some unknown attribute values can be used. Therefore, HEOM becomes an effective distance in the neighborhood information system, and its inherent characteristic underlies the application superiority of the later outlier detection for mixed data. In particular, when $B = \{c_j\}$ and according to Eq. (5),

$$HEOM_{\{c_j\}}(x, y) = d_{\{c_j\}}(x, y). \tag{7}$$

produces to apply to single attribute c_j.

In general, the traditional neighborhood radius ε is fixed for all attributes c_j ($j \in \{1, \ldots, m\}$) and is given by expert experience, such as in the outlier detection references (Chen et al., 2010; Li & Rao, 2012). This setting mechanism has some limitation and subjectivity, so it easily causes parameter selection sensitivity of relevant algorithms. Therefore, introducing some objective information of single attribute c_j would lead to more adaptivity and reasonability. For this purpose, we particularly utilize the standard deviation – a sort of distributional and statistical information – to set up the neighborhood threshold as follows.

Definition 4. The neighborhood radius of $x \in U$ regarding attribute c_j ($j \in \{1, \ldots, m\}$) is determined by

$$\varepsilon_{c_j} = \varepsilon_{\{c_j\}} = \begin{cases} 0, & \text{if } c_j \text{ is a categorical attribute;} \\ \frac{std(c_j)}{\lambda}, & \text{if } c_j \text{ is a numeric attribute,} \end{cases} \tag{8}$$

where $std(c_j)$ is the standard deviation of attribute values regarding numeric attribute c_j while λ is a given parameter for radius adjustment.

Herein, the neighborhood radius ε_{c_j} gains a self-adapting feature regarding a single attribute, and it combines the objective statistical index $std(c_j)$ and subjective adjustable parameter λ.

(1) The standard deviation represents a sort of data dispersion degree for the average. A greater $std(c_j)$ corresponds to a more difference between majority values and their mean, while a smaller $std(c_j)$ implies that data values are closer to the average.
(2) λ is used to adjust the neighborhood size. If $\lambda < 1$, $\lambda = 1$, and $\lambda > 1$, then the neighborhood radius will be more than, equal to, and less than the standard deviation of attribute values, respectively.

In this paper, the standard deviation is considered as an important factor to adjust the neighborhood radius, and this strategy adds reasonable statistics, objectivity, and adaptivity, thus underling effectiveness and superiority of outlier detection. As a result, the radius setting deepens the fixed parameter strategy based on expert experience. In particular, ε_{c_j} regarding single attribute c_j will be mainly used later; meanwhile, based on the self-adapting generalization, ε_B regarding attribute subset B could also be constructed to implement generalized construction.

Based on the above three preliminary steps, i.e., the normalization preprocessing, the HEOM distance, and the adaptivity radius, we determine the neighborhood, neighborhood relation and neighborhood knowledge to establish the neighborhood information (or decision) system. Now, we transfer to the emphasis of outlier detection by measure development. Note that the classical information entropy can effectively measure uncertainty of classical rough sets and corresponding categorical data (Düntsch & Gediga, 1998). By virtue of the traditional style, the neighborhood information entropy is first defined to serve as the starting point of measure construction. Next, suppose neighborhood relation $nr_B^\varepsilon \in NR_C^\varepsilon$ induces neighborhood knowledge

$$U/nr_B^\varepsilon = \{N_1, N_2, \cdots, N_k\},$$

where B and ε are utilized for the generalized discussion, and $|.|$ denotes the set cardinality.

Definition 5. The neighborhood information entropy regarding neighborhood relation nr_B^ε is

$$NE^\varepsilon(B) = -\sum_{i=1}^{k} \frac{|N_i|}{|U|} \log_2 \frac{|N_i|}{|U|}, \tag{9}$$

where $\frac{|N_i|}{|U|}$ means the membership probability of object x for neighborhood N_i.

Aiming at neighborhood rough sets and hybrid data, the neighborhood information entropy (Eq. (9)) extends the basic style (Jiang, Sui, & Cao, 2010), which applies to only classical rough sets and categorical data. Moreover, it refers to the neighbour-based measure (Chen, Xue et al., 2017; Li & Rao, 2012), but it differs from the relevant formula (Chen, Xue et al., 2017)

$$-\sum_{i=1}^{k} \frac{|N_i|}{|U|^2} \log_2 \frac{1}{|N_i|}. \tag{10}$$

In essence, the neighborhood entropy simulates the classical form of information entropy to provide an overall uncertainty description for the neighborhood knowledge. This fundamental measure can be developed to present uncertainty of each object. Following this thought, we next construct a new notion to induce object uncertainty and underlie the outlier detection. For this purpose, deleting an object $x \in U$ is required and let

$$\{U - \{x\}\}/nr_B^\varepsilon = \{N_1', N_2', \cdots, N_{k'}'\}.$$

Definition 6. When removing x from U, suppose the surplus $U - \{x\}$ with relation nr_B^ε corresponds to the neighborhood information entropy

$$NE_x^\varepsilon(B) = -\sum_{i=1}^{k'} \frac{|N_i'|}{|U - \{x\}|} \log_2 \frac{|N_i'|}{|U - \{x\}|}. \tag{11}$$

Thus, the relative neighborhood entropy of $x \in U$ regarding neighborhood relation nr_B^ε is defined as

$$RNE_B^\varepsilon(x) = \begin{cases} 1 - \frac{NE_x^\varepsilon(B)}{NE^\varepsilon(B)}, & \text{if } NE_x^\varepsilon(B) < NE^\varepsilon(B); \\ 0, & \text{other cases.} \end{cases} \tag{12}$$

In Definition 6, the neighborhood information entropy $NE_x^\varepsilon(B)$ with $U - \{x\}$ (Eq. (11)) refers to the basic entropy $NE^\varepsilon(B)$ with U (Eq. (9)), so it corresponds to a sort of uncertainty without x (or of x). The entropy change between $NE_x^\varepsilon(B)$ and $NE^\varepsilon(B)$ can represent x's uncertainty to provide an outlier characteristic of objects. When deleting x, if the neighborhood information entropy regarding nr_B^ε decreases, then object x contains higher uncertainty regarding nr_B^ε to more tend to an outlier; if the entropy changes a bit and even increases, then x contains lower uncertainty to deviate from an outlier. On this basis, the change degree between $NE_x^\varepsilon(B)$ and $NE^\varepsilon(B)$ is described by the relative neighborhood entropy $RNE_B^\varepsilon(x)$ (Eq. (12)). Therefore, the relative neighborhood entropy can measure both the uncertainty and deviation degrees of object x, and concretely, the higher value of $RNE_B^\varepsilon(x)$ corresponds to the higher uncertainty and deviation of object x.

For outlier detection, the relative neighborhood entropy is fundamental but not necessarily sufficient, so other important elements are worth introducing to gain better effectiveness. Thus, the relative neighborhood cardinality is proposed to strengthen $RNE_B^\varepsilon(x)$ to a better index: the deviation degree.

Definition 7. The relative neighborhood cardinality of $x \in U$ regarding neighborhood relation nr_B^ε is

$$RNC(n_B^\varepsilon(x)) = |n_B^\varepsilon(x)| - \frac{|N_1'| + |N_2'| + \cdots + |N_{k'}'|}{k'}. \tag{13}$$

Definition 8. The neighborhood relation-based deviation degree of $x \in U$ regarding nr_B^ε is

$$NOD_B^\varepsilon(x) = \begin{cases} RNE_B^\varepsilon(x) \times \left(\frac{|U| - abs(RNC(n_B^\varepsilon(x)))}{2|U|}\right), \\ \qquad \text{if } RNC(n_B^\varepsilon(x)) > 0; \\ RNE_B^\varepsilon(x) \times \sqrt{\frac{|U| + abs(RNC(n_B^\varepsilon(x)))}{2|U|}}, \\ \qquad \text{if } RNC(n_B^\varepsilon(x)) \le 0, \end{cases} \tag{14}$$

where $abs(t)$ means the absolute value of t.

In outlier detection, an outlier factor is usually required to measure the deviation degree of data objects. In particular, the outlier factor based on the information entropy is constructed to apply to only categorical data (Jiang et al., 2010). Herein, aiming at numeric and mixed data in the neighborhood information system, we integrate the relative neighborhood entropy and cardinality into a fundamental index: the neighborhood relation-based deviation degree (Eq. (14)). In fact, the relative neighborhood cardinality $RNC(n_B^\varepsilon(x))$ (Eq. (13)) can judge whether an object belongs to the major class, which is the contrary of the rare class with outliers; by introducing the fundamental and absolute class information, the relative neighborhood entropy $RNE_B^\varepsilon(x)$ is improved to the neighborhood relation-based deviation degree $NOD_B^\varepsilon(x)$ (Eq. (14)). As a result, $NOD_B^\varepsilon(x)$ represents the deviation degree of object x regarding neighborhood relation nr_B^ε, thus underling the outlier factor.

In summary, $NOD_B^\varepsilon(x)$ becomes a good measure for outlier factor, but it concerns attribute subset B and neighborhood radius

ε (which can be replaced by self-adapting ε_B) to exhibit a universal property. In the most outlier detection, an index regarding B corresponds to the high-dimension and complexity, so each attribute c_j in the one-dimensional space is basically and integratedly utilized for simplicity and concreteness. Except the single attribute family, the attribute subset sequence is actually considered in data experiments, but the case which has bigger complexity of time and space exhibits almost the same effect; in other words, our research regarding the single attribute could achieve satisfactory experiment results. Thus, we adopt single attribute c_j and self-adapting radius ε_{c_j} (Eq. (8)) by setting up $B = \{c_j\}$ and corresponding $\varepsilon_B = \varepsilon_{\{c_j\}} = \varepsilon_{c_j}$, so $NOD_B^{\varepsilon}(x)$ is concretized to $NOD_{\{c_j\}}^{\varepsilon_{c_j}}(x)$ to represent the deviation degree regarding single relation $nr_{\{c_j\}}^{\varepsilon_{c_j}}$, where concrete forms

$$HEOM_{\{c_j\}},\ NE^{\varepsilon_{c_j}}(\{c_j\}),\ RNE_{\{c_j\}}^{\varepsilon_{c_j}}(x), RNC(n_{\{c_j\}}^{\varepsilon_{c_j}}(x)),\ NOD_{\{c_j\}}^{\varepsilon_{c_j}}(x)$$

are concerned. By virtue of the family function, multiple and specific deviation degrees, i.e., $NOD_{\{c_j\}}^{\varepsilon_{c_j}}(x)$ ($j \in \{1, ., m\}$), systematically emerge according to multiple neighborhood relations $nr_{\{c_j\}}^{\varepsilon_{c_j}}$ ($j \in \{1, ., m\}$), and by adding rational weights, they can be integrated to the following outlier factor to implement outlier detection.

Definition 9. The neighborhood entropy-based outlier factor (NEOF) of $x \in U$ is

$$NEOF(x) = 1 - \frac{\sum_{j=1}^{m}\left(1 - NOD_{\{c_j\}}^{\varepsilon_{c_j}}(x)\right)W_{\{c_j\}}^{\varepsilon_{c_j}}(x)}{2|C|}, \tag{15}$$

where the weight

$$W_{\{c_j\}}^{\varepsilon_{c_j}}(x) = \sqrt{\frac{|n_{\{c_j\}}^{\varepsilon_{c_j}}(x)|}{|U|}} \in (0, 1]. \tag{16}$$

Definition 10. Let μ be a given judgement threshold. $\forall x \in U$, if $NEOF(x) > \mu$, then x is called an outlier in U based on the neighborhood information entropy.

According to Eq. (15), the neighborhood entropy-based outlier factor $NEOF(x)$ appropriately integrates all neighborhood relation-based deviation degrees $NOD_{\{c_j\}}^{\varepsilon_{c_j}}(x)$ ($j \in \{1, ., m\}$) (regarding the single attribute) by introducing weights with neighborhood probability information. After the above in-depth development, $NEOF(x)$ gains practicability and efficiency to profoundly represent the outlier degree, so it naturally induces the outlier detection rule in Definition 10. For this detection approach, its practical validity will be later verified by a hybrid example and three data experiments.

Herein, the detection parameter μ and its setting are analyzed. In general, a method of outlier detection only gives an outlier degree for each object. Before using the method to detect outliers, the users should first input an empirical value (denoted by symbol *eon*) to denote the outlier number of their expectation. The *eon* value usually varies for different data sets, and its determination needs many trials even for a given data set. In our method, the μ setting could also depend on the *eon* value provided by the users.

(1) First, outlier factors $NEOF(x)$ ($\forall x \in U$) are calculated, and they are sorted in the descending order to correspond to an array $x'_1, x'_2, \ldots, x'_n$.
(2) Then, according to the above object sequence, we set up μ in range

$$NEOF(x'_{eon+1}) \le \mu < NEOF(x'_{eon}).$$

As a result, this setting strategy can guarantee that the *eon* objects $x'_1, x'_2, \ldots, x'_{eon}$ are necessarily found to hold higher outlier degrees than other objects in U, so the *eon* ones will be eventually returned to the users as the outliers.

Thus far, we have completed the theoretical construction of outlier detection, where the measure development plays an important role. Finally, let us summarize the measure construction. In fact, five basic measures are concerned to constitute a development clue:

the neighborhood information entropy $NE^{\varepsilon}(B)$
(Definition 5)
$\longrightarrow$ the relative neighborhood entropy $RNE_B^{\varepsilon}(x)$
(Definition 6)
$\longrightarrow$ the relative neighborhood cardinality $RNC(n_B^{\varepsilon}(x))$
(Definition 7)
$\longrightarrow$ the neighborhood relation-based deviation degree $NOD_B^{\varepsilon}(x)$
(Definition 8)
$\longrightarrow$ the neighborhood entropy-based outlier factor $NEOF(x)$
(Definition 9).

The relative neighborhood information entropy serves as the starting point of metric construction, and it has the whole uncertainty description; the middle three measures mainly utilize general neighborhood relation nr_B^{ε} to describe object x; the final outlier factor $NEOF(x)$ integrates related measures of x regarding neighborhood relation $nr_{\{c_j\}}^{\varepsilon_{c_j}}$ or single attribute c_j, where $j \in \{1, ., m\}$, and it is utilized to implement the outlier detection of x in universe U. To more clarify the measure development, Fig. 1 vividly presents the measure-structural evolution from a view of the single attribute.

(1) Regarding single attribute c_j, basic $NE^{\varepsilon_{c_j}}(\{c_j\})$ induces $RNE_{\{c_j\}}^{\varepsilon_{c_j}}(x)$, and by introducing $RNC(n_{\{c_j\}}^{\varepsilon_{c_j}}(x))$, the latter evolves to $NOD_{\{c_j\}}^{\varepsilon_{c_j}}(x)$.
(2) By introducing weight $W_{\{c_j\}}^{\varepsilon_{c_j}}(x)$ and considering all attributes, $NOD_{\{c_j\}}^{\varepsilon_{c_j}}(x)$ ($j \in \{1, \ldots, m\}$) areintegrated into $NEOF(x)$, which eventually becomes the outlier factor for detection.

As a result, the deeper measure, especially the final NEOF, more adheres to the outlier detection. Moreover, the neighborhood is determined by both the HEOM distance (Definition 3) and self-adapting radius ε_{c_j} (Definition 4), and two parameters λ and μ are needed to be set up.

3.2. Specific algorithm

Based on the gradual measure construction and final judgement rule, Section 3.1 provides the outlier detection method based on the neighborhood information entropy and its developmental measures. Furthermore, this subsection mainly designs a corresponding algorithm: the NIEOD algorithm (or Algorithm 2). As an invoking basis, a basic algorithm is first given to calculate the single-attribute neighborhood covering/knowledge, i.e., the SANC algorithm (or Algorithm 1).

The SANC algorithm aims to gain the neighbourhood covering $U/nr_{\{c_j\}}^{\varepsilon_{c_j}}$, and it particularly uses the data structure of a two-dimensional array. This algorithm improves the traditional one-by-one computing model, thus exhibiting better efficiency. Step 3 uses the heapsort method (Williams, 1964) to exhibit time complexity $O(n\log n)$, which improves the usual sorting time $O(n^2)$. Step 4 has frequency count n, Steps 5–24 also have frequency count n, so Steps 4–25 have frequency count $n \times n$. In the worst case, the SANC algorithm exhibits time complexity $O(n^2)$, which accords with that of the traditional algorithm (with the one-by-one comparison). However, the SANC algorithm fully utilizes the idea of

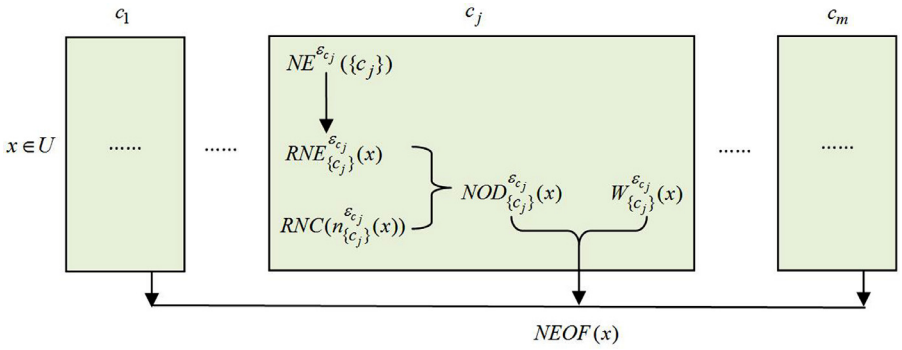

Fig. 1. Measure development figure based on the singe attribute.

Algorithm 1 Single-attribute neighborhood covering (SANC) algorithm.

Input: Information system $IS = (U, C, V, f)$ (with $|U| = n$ and fixed single-attribute $c_j \in C$) and parameter λ (determining radius ε_{c_j});

Output: Neighbourhood covering $U/nr^{\varepsilon_{c_j}}_{\{c_j\}}$ regarding single attribute c_j.

1: $U/nr^{\varepsilon_{c_j}}_{\{c_j\}} \leftarrow [1...n][]$;
/*For well storage, neighborhood knowledge $U/nr^{\varepsilon_{c_j}}_{\{c_j\}}$ is endowed with a two-dimensional array form. */
2: $N \leftarrow \emptyset$;
3: $[Rank, Index] \leftarrow$ Ascend_sort (U);
/*Ascend_sort(U) denotes that objects in U are sorted in an ascending order regarding attribute c_j. *Rank* is the first array to store the final ranking results, while *Index* is the second array to store original serial numbers before the ascending sorting. */
4: **for** $i \leftarrow 1$ to n **do**
5: $k \leftarrow i$;
6: **while** $k > 0$ **do**
7: **if** $HEOM_{\{c_j\}}(Rank[i], Rank[k]) \le \varepsilon_{c_j}$ **then**
8: $k \leftarrow k - 1$;
9: **else**
10: break;
11: **end if**
12: **end while**
13: $a \leftarrow k + 1$;
/*a records the low-limit serial number of object neighborhoods.*/
14: $k \leftarrow i + 1$;
15: **while** $k < n$ **do**
16: **if** $HEOM_{\{c_j\}}(Rank[i], Rank[k]) \le \varepsilon_{c_j}$ **then**
17: $k \leftarrow k + 1$;
18: **else**
19: break;
20: **end if**
21: **end while**
22: $b \leftarrow k - 1$;
/*b records the upper-limit serial number of object neighborhoods.*/
23: $N \leftarrow Rank[a...b]$;
24: $U/nr^{\varepsilon_{c_j}}_{\{c_j\}}[Index[i]][\,] \overset{N}{\leftarrow}$;
/*Neighborhood N is stored into the $Index[i]$ line of $U/nr^{\varepsilon_{c_j}}_{\{c_j\}}$.*/
25: **end for**
26: **return** $U/nr^{\varepsilon_{c_j}}_{\{c_j\}}$.

ordered binary and nearest neighborhood search to improve the traditional unordered one-by-one model. Accordingly, Steps 4–25 give the average comparison time: about n, which is much smaller than n^2: the comparison time of the traditional algorithm. Therefore, Algorithm 1 has actual time complexity $O(n\log n)$, which is better than $O(n^2)$ of the traditional algorithm.

The NIEOD algorithm first achieves the single-attribute neighbourhood covering by invoking the SANC algorithm; then, it gains a series of measures: the relative neighborhood entropy and cardinality, the deviation degree and outlier factor; finally, it makes the outlier discriminant to collect all identified outliers and thus outputs the entire outlier set. For the complexity analysis, Steps 2–5 have frequency count $m \times n \times \log n$ according to Algorithm 1, Steps 6–17 have frequency count $m \times n$, so Algorithm 2 exhibits frequency count

Algorithm 2 Neighborhood information entropy-based outlier detection (NIEOD) algorithm.

Input: Information system $IS = (U, C, V, f)$ (with $|U| = n$, $|C| = m$) and threshold μ;

Output: Neighborhood information entropy-based outlier set OS.

1: $OS \leftarrow \emptyset$;
2: **for** $j \leftarrow 1$ to m **do**
3: Obtain neighbourhood covering $U/nr^{\varepsilon_{c_j}}_{\{c_j\}}$ by Algorithm 1;
4: Calculate $NE^{\varepsilon_{c_j}}(\{c_j\})$;
5: **end for**
6: **for** $i \leftarrow 1$ to n **do**
7: **for** $j \leftarrow 1$ to m **do**
8: Calculate $RNE^{\varepsilon_{c_j}}_{\{c_j\}}(x_i)$;
9: Calculate $RNC(n^{\varepsilon_{c_j}}_{\{c_j\}}(x_i))$;
10: Calculate $NOD^{\varepsilon_{c_j}}_{\{c_j\}}(x_i)$;
11: Calculate $W^{\varepsilon_{c_j}}_{\{c_j\}}(x_i)$;
12: **end for**
13: Calculate $NEOF(x_i)$;
14: **if** $NEOF(x_i) > \mu$ **then**
15: $OS \leftarrow OS \cup \{x_i\}$;
16: **end if**
17: **end for**
18: **return** OS.

$m \times n \times \log n + m \times n$ (where $m \times n \ll m \times n \times \log n$).

As a result, the NIEOD algorithm offers overall time complexity $O(mn\log n)$ and space complexity $O(mn)$, respectively, thus becoming effective.

Table 1
Initial information system of Example 1.

U	c_1	c_2	c_3
x_1	D	4	0.7
x_2	B	7	0.4
x_3	D	1	0.6
x_4	B	2	0.3
x_5	B	8	0.5
x_6	C	10	0.8

Table 2
Standard information system of Example 1.

U	c_1	c_2	c_3
x_1	D	$\frac{1}{3}$	$\frac{4}{5}$
x_2	B	$\frac{2}{3}$	$\frac{1}{5}$
x_3	D	0	$\frac{3}{5}$
x_4	B	$\frac{1}{9}$	0
x_5	B	$\frac{7}{9}$	$\frac{2}{5}$
x_6	C	1	1

Table 3
Neighborhoods regarding each single attribute of Example 1.

U	c_1	c_2	c_3
x_1	$\{x_1, x_3\}$	$\{x_1, x_2, x_3, x_4\}$	$\{x_1, x_3, x_6\}$
x_2	$\{x_2, x_4, x_5\}$	$\{x_1, x_2, x_5, x_6\}$	$\{x_2, x_4, x_5\}$
x_3	$\{x_1, x_3\}$	$\{x_1, x_3, x_4\}$	$\{x_1, x_3, x_5\}$
x_4	$\{x_2, x_4, x_5\}$	$\{x_1, x_3, x_4\}$	$\{x_2, x_4\}$
x_5	$\{x_2, x_4, x_5\}$	$\{x_2, x_5, x_6\}$	$\{x_2, x_3, x_5\}$
x_6	$\{x_6\}$	$\{x_2, x_5, x_6\}$	$\{x_1, x_6\}$

3.3. Illustrative example

This subsection utilizes an example of the information system to illustrate the research contents of above two subsections, i.e., the theoretical method and specific algorithm of our outlier detection (based on the neighborhood information entropy and its developmental indexes).

Example 1. An information system $IS = (U, C, V, f)$ is provided in Table 1, where $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and $C = \{c_1, c_2, c_3\}$.

According to Table 1, this system concerns hybrid data, because the second column consists of categorical data while both the third and fourth columns consist of numeric data. At first, the original numeric data are standardized by the min-max normalization method (Eq. (4)), and the standard data table is offered by Table 2, which still exhibits the heterogeneous data feature.

Then, the self-adopting radius ε_{c_j} is considered. For standard Table 2, the second column has the categorical feature to produce $\varepsilon_{c_1} = 0$; the surplus numeric columns have standard deviations: $std(c_2) \approx 0.3610, std(c_3) \approx 0.3416$, so radii $\varepsilon_{c_2} \approx 0.3610$, $\varepsilon_{c_3} \approx 0.3416$ are achieved when let $\lambda = 1$ and consider Eq. (8). In short, we gain the radii regarding each single attribute:

$$\varepsilon_{c_1} = 0, \ \ \varepsilon_{c_2} \approx 0.3610, \ \ \varepsilon_{c_3} \approx 0.3416. \tag{17}$$

With the addition of the HEOM distance (Eq. (5)), the neighborhood information system $NIS = (U, NR_C^{\varepsilon}, V, f)$ is established. Regarding single attribute $c_j \in C$, the neighborhoods are given in Table 3, while the neighborhood knowledge/covering is further obtained as follows:

$$\begin{aligned} U/nr_{\{c_1\}}^{\varepsilon_{c_1}} &= \{\{x_1, x_3\}, \{x_2, x_4, x_5\}, \{x_6\}\}, \\ U/nr_{\{c_2\}}^{\varepsilon_{c_2}} &= \{\{x_1, x_2, x_3, x_4\}, \{x_1, x_2, x_5, x_6\}, \{x_1, x_3, x_4\}, \{x_2, x_5, x_6\}\}, \\ U/nr_{\{c_3\}}^{\varepsilon_{c_3}} &= \{\{x_1, x_3, x_6\}, \{x_2, x_4, x_5\}, \{x_1, x_3, x_5\}, \{x_2, x_4\}, \{x_2, x_3, x_5\}, \\ &\quad \{x_1, x_6\}\}; \end{aligned} \tag{18}$$

these results can be computed directly or by Algorithm 1.

Next, relevant measures of all single attributes are calculated, and then the outlier is finally detected. These processes concern the following six steps, which can be implemented directly or by Algorithm 2.

(1) According to Definition 5, the neighborhood information entropy of each attribute exhibits

$$\begin{aligned} NE^{\varepsilon_{c_1}}(\{c_1\}) &= -\sum_{i=1}^{3} \frac{|N_i|}{|U|} \log_2 \frac{|N_i|}{|U|} \\ &= -\left(\frac{2}{6}\log_2\frac{2}{6} + \frac{3}{6}\log_2\frac{3}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 1.4591, \end{aligned} \tag{19}$$

$$\begin{aligned} NE^{\varepsilon_{c_2}}(\{c_2\}) &= -\sum_{i=1}^{4} \frac{|N_i|}{|U|} \log_2 \frac{|N_i|}{|U|} \\ &= -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{4}{6}\log_2\frac{4}{6} + \frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) \\ &= 1.7800, \end{aligned} \tag{20}$$

$$\begin{aligned} NE^{\varepsilon_{c_3}}(\{c_3\}) &= -\sum_{i=1}^{6} \frac{|N_i|}{|U|} \log_2 \frac{|N_i|}{|U|} \\ &= -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6} + \frac{2}{6}\log_2\frac{2}{6}\right. \\ &\quad \left. + \frac{3}{6}\log_2\frac{3}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 3.0566. \end{aligned} \tag{21}$$

(2) The neighborhood information entropy after removing an object exhibits

$$\begin{aligned} &NE_{x_1}^{\varepsilon_{c_1}}(\{c_1\}) = NE_{x_3}^{\varepsilon_{c_1}}(\{c_1\}) \\ &= -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{3}{5}\log_2\frac{3}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 1.3710, \\ &NE_{x_2}^{\varepsilon_{c_1}}(\{c_1\}) = NE_{x_4}^{\varepsilon_{c_1}}(\{c_1\}) = NE_{x_5}^{\varepsilon_{c_1}}(\{c_1\}) \\ &= -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 1.5219, \\ &NE_{x_6}^{\varepsilon_{c_1}}(\{c_1\}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.9710; \end{aligned} \tag{22}$$

$$\begin{aligned} &NE_{x_1}^{\varepsilon_{c_2}}(\{c_2\}) = NE_{x_2}^{\varepsilon_{c_2}}(\{c_2\}) \\ &= -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.9710, \\ &NE_{x_3}^{\varepsilon_{c_2}}(\{c_2\}) = NE_{x_4}^{\varepsilon_{c_2}}(\{c_2\}) = NE_{x_5}^{\varepsilon_{c_2}}(\{c_2\}) = NE_{x_6}^{\varepsilon_{c_2}}(\{c_2\}) \\ &= -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{4}{5}\log_2\frac{4}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 1.6707; \end{aligned} \tag{23}$$

$$\begin{aligned} &NE_{x_1}^{\varepsilon_{c_3}}(\{c_3\}) = NE_{x_2}^{\varepsilon_{c_3}}(\{c_3\}) \\ &= -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) \\ &= 2.4063, \\ &NE_{x_3}^{\varepsilon_{c_3}}(\{c_3\}) = NE_{x_5}^{\varepsilon_{c_3}}(\{c_3\}) \\ &= -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 2.0285, \\ &NE_{x_4}^{\varepsilon_{c_3}}(\{c_3\}) = NE_{x_6}^{\varepsilon_{c_3}}(\{c_3\}) \\ &= -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5} + \frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) \\ &= 2.3841. \end{aligned} \tag{24}$$

Furthermore, according to Definition 6, the relative neighborhood information entropy of each attribute becomes

$$RNE^{\varepsilon_{c_1}}_{\{c_1\}}(x_1) = RNE^{\varepsilon_{c_1}}_{\{c_1\}}(x_3) = 0.0604,$$
$$RNE^{\varepsilon_{c_1}}_{\{c_1\}}(x_2) = RNE^{\varepsilon_{c_1}}_{\{c_1\}}(x_4) = RNE^{\varepsilon_{c_1}}_{\{c_1\}}(x_5) = 0,$$
$$RNE^{\varepsilon_{c_1}}_{\{c_1\}}(x_6) = 0.3346; \tag{25}$$

$$RNE^{\varepsilon_{c_2}}_{\{c_2\}}(x_1) = RNE^{\varepsilon_{c_2}}_{\{c_2\}}(x_2) = 0.4545,$$
$$RNE^{\varepsilon_{c_2}}_{\{c_2\}}(x_3) = RNE^{\varepsilon_{c_2}}_{\{c_2\}}(x_4) =$$
$$RNE^{\varepsilon_{c_2}}_{\{c_2\}}(x_5) = RNE^{\varepsilon_{c_2}}_{\{c_2\}}(x_6) = 0.0614; \tag{26}$$

$$RNE^{\varepsilon_{c_3}}_{\{c_3\}}(x_1) = RNE^{\varepsilon_{c_3}}_{\{c_3\}}(x_2) = 0.2128,$$
$$RNE^{\varepsilon_{c_3}}_{\{c_3\}}(x_3) = RNE^{\varepsilon_{c_3}}_{\{c_3\}}(x_5) = 0.3364,$$
$$RNE^{\varepsilon_{c_3}}_{\{c_3\}}(x_4) = RNE^{\varepsilon_{c_3}}_{\{c_3\}}(x_6) = 0.2200. \tag{27}$$

(3) According to Definition 7, the relative neighborhood cardinality gives

$$RNC(n^{\varepsilon_{c_1}}_{\{c_1\}}(x_1)) = RNC(n^{\varepsilon_{c_1}}_{\{c_1\}}(x_3)) = \frac{1}{3},$$
$$RNC(n^{\varepsilon_{c_1}}_{\{c_1\}}(x_2)) = RNC(n^{\varepsilon_{c_1}}_{\{c_1\}}(x_4)) = RNC(n^{\varepsilon_{c_1}}_{\{c_1\}}(x_5)) = \frac{4}{3},$$
$$RNC(n^{\varepsilon_{c_1}}_{\{c_1\}}(x_6)) = -\frac{3}{2}; \tag{28}$$

$$RNC(n^{\varepsilon_{c_2}}_{\{c_2\}}(x_1)) = RNC(n^{\varepsilon_{c_2}}_{\{c_2\}}(x_2)) = \frac{3}{2},$$
$$RNC(n^{\varepsilon_{c_2}}_{\{c_2\}}(x_3)) = RNC(n^{\varepsilon_{c_2}}_{\{c_2\}}(x_4)) =$$
$$RNC(n^{\varepsilon_{c_2}}_{\{c_2\}}(x_5)) = RNC(n^{\varepsilon_{c_2}}_{\{c_2\}}(x_6)) = 0; \tag{29}$$

$$RNC(n^{\varepsilon_{c_3}}_{\{c_3\}}(x_1)) = RNC(n^{\varepsilon_{c_3}}_{\{c_3\}}(x_2)) = \frac{4}{5},$$
$$RNC(n^{\varepsilon_{c_3}}_{\{c_3\}}(x_3)) = RNC(n^{\varepsilon_{c_3}}_{\{c_3\}}(x_5)) = \frac{3}{4},$$
$$RNC(n^{\varepsilon_{c_3}}_{\{c_3\}}(x_4)) = RNC(n^{\varepsilon_{c_3}}_{\{c_3\}}(x_6)) = -\frac{3}{5}. \tag{30}$$

(4) According to Definition 8, the neighborhood relation-based deviation degree regarding each attribute produces

$$NOD^{\varepsilon_{c_1}}_{\{c_1\}}(x_1) = NOD^{\varepsilon_{c_1}}_{\{c_1\}}(x_3) = 0.0285,$$
$$NOD^{\varepsilon_{c_1}}_{\{c_1\}}(x_2) = NOD^{\varepsilon_{c_1}}_{\{c_1\}}(x_4) = NOD^{\varepsilon_{c_1}}_{\{c_1\}}(x_5) = 0,$$
$$NOD^{\varepsilon_{c_1}}_{\{c_1\}}(x_6) = 0.2645; \tag{31}$$

$$NOD^{\varepsilon_{c_2}}_{\{c_2\}}(x_1) = NOD^{\varepsilon_{c_2}}_{\{c_2\}}(x_2) = 0.1704,$$
$$NOD^{\varepsilon_{c_2}}_{\{c_2\}}(x_3) = NOD^{\varepsilon_{c_2}}_{\{c_2\}}(x_4) =$$
$$NOD^{\varepsilon_{c_2}}_{\{c_2\}}(x_5) = NOD^{\varepsilon_{c_2}}_{\{c_2\}}(x_6) = 0.0434; \tag{32}$$

$$NOD^{\varepsilon_{c_3}}_{\{c_3\}}(x_1) = NOD^{\varepsilon_{c_3}}_{\{c_3\}}(x_2) = 0.0922,$$
$$NOD^{\varepsilon_{c_3}}_{\{c_3\}}(x_3) = NOD^{\varepsilon_{c_3}}_{\{c_3\}}(x_5) = 0.1472,$$
$$NOD^{\varepsilon_{c_3}}_{\{c_3\}}(x_4) = NOD^{\varepsilon_{c_3}}_{\{c_3\}}(x_6) = 0.1632. \tag{33}$$

(5) According to Definition 9, the neighborhood entropy-based outlier factor (NEOF) of each object can be calculated. As an example,

$$NEOF(x_1) = 1 - \frac{(1-0.0285)\sqrt{\frac{2}{6}} + (1-0.1704)\sqrt{\frac{4}{6}} + (1-0.0922)\sqrt{\frac{3}{6}}}{2\times 3} \approx 0.6866. \tag{34}$$

The total NEOF results are provided as follows:

$$NEOF(x_1) \approx 0.6866, NEOF(x_2) \approx 0.6623, NEOF(x_3) \approx 0.6933,$$
$$NEOF(x_4) \approx 0.6889, NEOF(x_5) \approx 0.6689, NEOF(x_6) \approx 0.7567. \tag{35}$$

(6) Let judgement threshold $\mu = 0.75$. According to Definition 10, all instances can be discriminated for outliers.

$$NEOF(x_1) < \mu, NEOF(x_2) < \mu, NEOF(x_3) < \mu,$$
$$NEOF(x_4) < \mu, NEOF(x_5) < \mu, NEOF(x_6) > \mu. \tag{36}$$

Hence, only x_6 has the high outlier factor which is greater than μ, so x_6 becomes the sole outlier based on the neighborhood information entropy and $OS = \{x_6\}$ is output in Algorithm 2. □

4. UCI data experiments and analyses

This section implements UCI data experiments and analyses to verify availability of the proposed method of outlier detection, especially the NIEOD algorithm (Algorithm 2).

In the concrete experiments, three UCI data sets (Bay, 1999) are mainly chosen, i.e., the Annealing data set (with hybrid attributes), Lymphography data set (with most categorical attributes), and Wisconsin Breast Cancer data set (with numeric attributes), and the NIEOD algorithm is compared to six main ways of outlier detection to reveal its effectiveness and performance. For simplification, an algorithm is represented by its core character, e.g., the NIEOD algorithm is replaced by simper NIEOD. All seven concerned algorithms and their strength-weakness comparison (on the capacity and applicability) are described in Table 4. As a result, NIEOD carries a more powerful mechanism of the neighborhood, radius, and information to widely apply to categorical, numeric, hybrid data, while its weakness mainly comes from measure construction based on single attributes. Herein, relevant algorithms implement the adaptability treatment, e.g., IE, SEQ adopt data discretization. Moreover, the experimental platform configuration concerns the processor Intel (R) core (TM) i5-2400, master frequency 3.10 GHz, memory 4G, operating system Windows 7, and programming environment Matlab R2015b.

To enhance comparability of experimental results, this paper mainly adopts the evaluation index system proposed by Aggarwal and Yu (2001) to evaluate the performance of each method of outlier detection. At present, this assessment system becomes the most commonly used way. Concretely, a detection method can be evaluated by its practical ratio between true and identified outliers, in a given data set where the belonging class is known for each object. If the actual outlier ratio is higher, then the performance of the algorithm is better. Under this strategy, NIEOD's parameter needs only λ, which adjusts the standard deviation to the neighborhood radius (Eq. (8)).

4.1. Annealing data set

The Annealing data set contains 798 objects, 37 condition attributes, and 1 decision attribute. The condition attributes include 30 categorical attributes and 7 numeric attributes. The 798 objects are partitioned into 5 classes, class 3 carries the largest number of objects, while the remaining classes are treated as rare classes

Table 4
Seven concerned algorithms and their strength-weakness comparison.

Naming (Reference)	Meaning or strategy	• Strength// ◦ Weakness
NED (Chen et al., 2010)	Neighborhood detection	• Powerful granulating ability, adaptability for categorical and numeric data// ◦ Complex neighborhood calculation, subjective radius selection
IE (Jiang et al., 2010)	Information entropy-based detection with rough sets	• Uncertainty measurement of information entropy, adaptability for categorical data// ◦ Inadaptability for numeric data
SEQ (Jiang, Sui, & Cao, 2009)	Sequence-based detection with rough sets	• Strong attribute impact on detection results, adaptability for categorical data// ◦ Discretization pretreatment for numeric data
Find-CBLOF (He, Xu, & Deng, 2003)	Finding cluster-based local outlier factor	• Good connection with classical clustering// ◦ Non-ideal detection effect
DIS (Knorr et al., 2000)	Distance-based detection	• Relative simplicity, adaptability for numeric data// ◦ Inadaptability for categorical data
KNN (Ramaswamy, Rastogi, & Shim, 2000)	K-nearest neighbor method	• Adaptability for numeric data// ◦ Inadaptability for categorical data
NIEOD (This article)	Neighborhood information entropy-based outlier detection	• Simple neighborhood calculation and rational radius setting, uncertainty measurement of neighborhood information entropy, adaptability for categorical, numeric, hybrid data// ◦ Information fusion of only single attributes

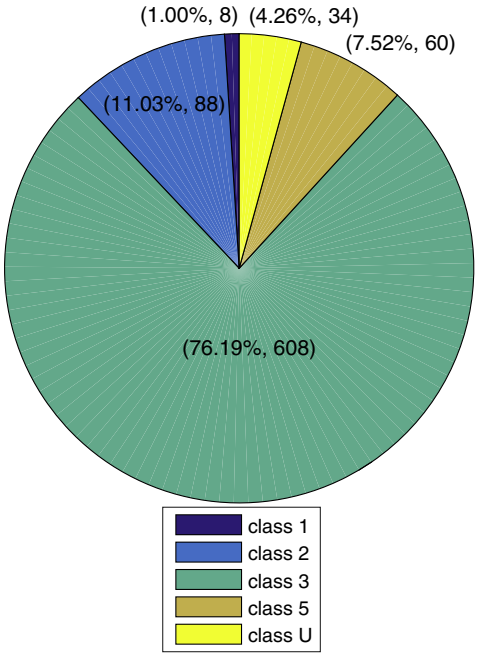

Fig. 2. Classification pie-chart of Annealing data set.

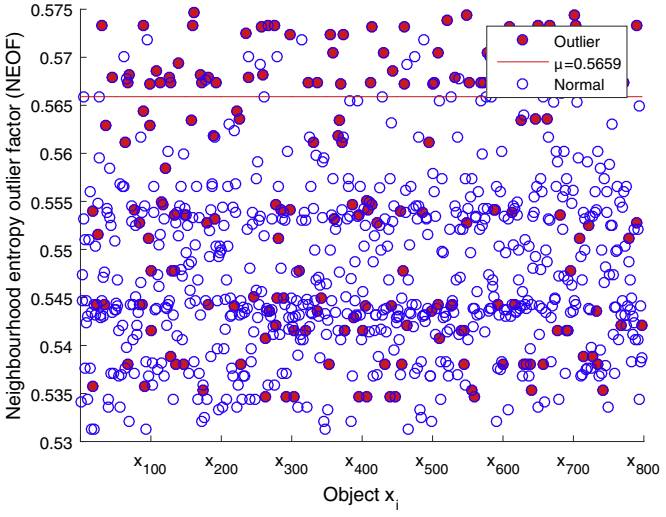

Fig. 3. NEOF-based object distribution in NIS_{An}.

to produce a total of 190 real outliers. The relevant class distribution is presented in Fig. 2, where the brackets carry the relative percentage and absolute amount. Herein, the relevant data induce a neighborhood information system $NIS_{An} = (U_{An}, NR^{\varepsilon}_{C_{An}}, V_{An}, f_{An})$, where *An* labels Annealing.

For NIEOD, let $\lambda_{An} = 0.2$. According to the neighborhood entropy-based outlier factor (NEOF) (Definition 9), the object distribution in NIS_{An} is presented in Fig. 3. Clearly, NIEOD exhibits a good detection effect to recognize most of all outliers, if the judgement threshold is set as $\mu = 0.5659$.

For NIS_{An}, experimental results of whole seven algorithms are provided in Table 5. The acquisition and interpretation of Table 5 are given as follows.

(1) For each algorithm, outlier factors or degrees of all objects are calculated and are sorted in the descending order, which naturally corresponds to an object sequence. According to the descending order or object sequence, "Top ratio (Number of objects)" describes the percentage (number) of the front objects; in other words, value $k\%$ (n_k) of "Top ratio (Number of objects)" corresponds to selecting the $k\%$ proportion (n_k amount) with high outlier values.

(2) Furthermore, two basic indexes can be established. "Number of rare classes included" denotes the number of selected objects that are real outliers (i.e., the number of outliers identified successfully), while "Coverage" denotes the ratio of "Number of rare classes included" to the whole number of U_{An}'s practical outliers (the latter is denoted by $|OS_{\text{true}}(U_{An})|$). Note that "Top ratio (Number of objects)" and "Number of rare classes included (Coverage)" act as the premise and result, respectively, to establish a systematical and dynamic estimation of accuracy.

(3) As is shown in Table 5, NIEOD has the better result than IE, DIS, KNN; for example, when the top ratio is 10.03% (the number of objects is 80), NIEOD gains true outlier number 64 (coverage $33.68\% \approx 64/190$), while IE, DIS, KNN gain only outlier number 34 (coverage 17.89%), outlier number 33 (coverage 17.37%), and outlier number 21 (coverage 11.05%), respectively. Now, NIEOD is compared to the surplus three algorithms: NED, SEQ, FindCBLOF. Regarding all levels of top ratio (number of objects), NIEOD is better than NED, SEQ,

Table 5
Experimental results in NIS_{An}.

Top ratio (%) (Number of objects)	Number of rare classes included (Coverage (%))						
	NIEOD	NED	IE	SEQ	FindCBLOF	DIS	KNN
10.03(80)	64(33.68)	51(26.84)	34(17.89)	36(18.95)	45(23.68)	33(17.37)	21(11.05)
13.16(105)	72(37.89)	67(35.26)	46(24.21)	51(26.84)	55(28.95)	44(23.16)	30(15.79)
17.54(140)	80(42.11)	81(42.63)	64(33.68)	71(37.37)	82(43.16)	61(32.11)	41(21.58)
21.93(175)	81(42.63)	84(44.21)	75(39.47)	86(45.26)	105(55.26)	77(40.53)	58(30.53)
26.19(209)	88(46.32)	92(48.42)	88(46.32)	96(50.53)	105(55.26)	84(44.21)	62(32.63)

NIS_{An} contains 798 objects and 190 outliers, denoted by $|U_{An}| = 798$ and $|OS_{\text{true}}(U_{An})| = 190$.

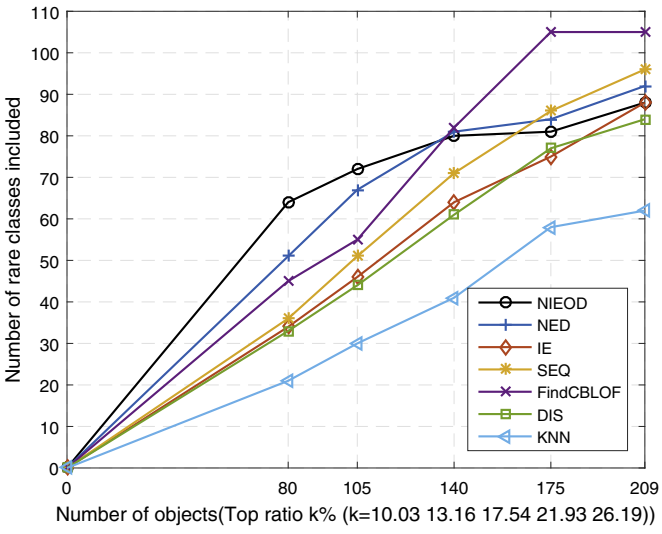

Fig. 4. Line chart of experimental results in NIS_{An}.

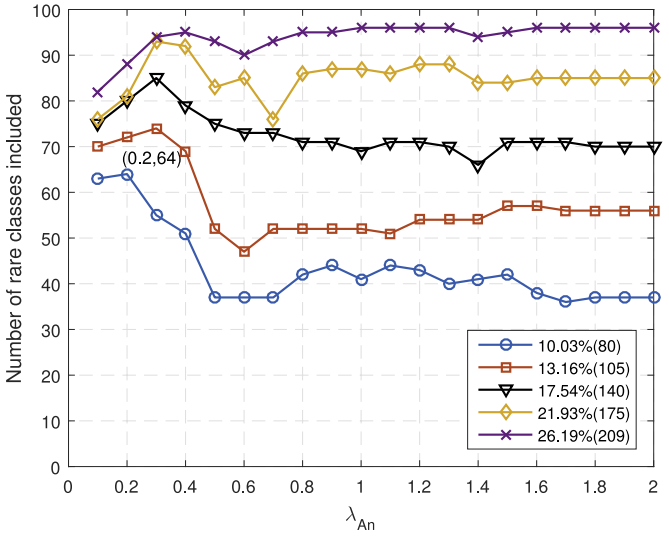

Fig. 5. Line chart of the number of rare classes included when λ_{An} changes.

FindCBLOF when the top ratio (the number of objects) become less, while NED, SEQ, FindCBLOF seem better in the other case.

For more visualization, the experimental results (i.e., Table 5) are also drawn in Fig. 4 by using broken lines. This line chart clearly reflects the mainstream accuracy superiority of NIEOD, when compared to the other six algorithms. Moreover, computational times are determined according to relevant algorithm implementations and average operation statistics, and NIEOD, NED, IE, SEQ, FindCBLOF, DIS, KNN generally spend 3006.32, 64.11, 3674.34, 1088.44, 1.17, 40.42, 25.38 seconds, respectively; thus, NIEOD's elapsed time is relatively much but is also practically feasible.

In the comparative experiment, NIEOD involves only one parameter λ_{An}, and NIEOD's result will be affected by λ_{An}'s fluctuation. Thus, the parameter change experiment is further made, and relevant results are given in Fig. 5, which also contains all five levels of "Top ratio (Number of objects)" in Table 5 and Fig. 4. By focusing on "Number of rare classes included", Fig. 5 can be utilized to implement parametric sensitivity analyses. As an example, the level of top ratio 10.03% (object number 80) is analyzed as follows.

(1) "Number of rare classes included (Coverage)" generally decreases when λ_{An} increases, i.e., the experimental result exhibits a weakening tendency accompanied by the parameter increase. However, the result becomes relatively stable if the parameter is relatively great (e.g., $\lambda_{An} > 0.5$).
(2) In all parametric cases, "Number of rare classes included" is greater than 36, which is better than results 34 of IE, 33 of DIS, and 21 of KNN at the same level. Moreover, when $\lambda_{An} = 0.2$, NIEOD achieves the best performance to provide 64 true outliers, which is better than results of the other six algorithms.

Similarly, the other levels can be analyzed by Fig. 5. In the general range of parameter λ_{An}, NIEOD becomes effective and optimum.

In summary, NIEOD exhibits the effectiveness and performance for the Annealing data set with heterogeneous attributes, when compared to the other six algorithms. As a conclusion, we see that NIEOD is effectively applied to hybrid data.

4.2. Lymphography data set

The Lymphography data set contains 148 objects, 18 condition attributes, and 1 decision attribute. The condition attributes have 3 numeric attributes and 15 categorical attributes. All 148 objects are partitioned into 4 classes: "normal find" (number 2 and proportion 1.35%), "metastases" (number 81 and proportion 54.73%), "malign lymph" (number 61 and proportion 41.22%), and "fibrosis" (number 4 and proportion 2.7%), whose pie-chart is provided in Fig. 6. Herein, "normal find" and "fibrosis" are regarded as rare classes with a total of 6 practical outliers, and related data underlie neighborhood information system $NIS_L = (U_L, NR^{\varepsilon}_{C_L}, V_L, f_L)$.

For NIEOD, set up $\lambda_L = 0.2$. The NEOF-based object distribution in NIS_L is presented in Fig. 7. NIEOD can recognize all 6 outliers, if judgement threshold $\mu = 0.6702$.

For the seven algorithms of NIEOD, NED, IE, SEQ, FindCBLOF, DIS, KNN, comparative experimental results are provided in Table 6, the corresponding line cart is described in Fig. 8. According to Table 6 and Fig. 8, we can easily conclude that NIEOD has (a bit) better superiority than or the same performance as the other six algorithms, regardless of which level of "Top ratio (Number of objects)" is considered. Moreover, NIEOD, NED, IE, SEQ, FindCBLOF, DIS, KNN generally spend 4.44, 1.93, 25.21, 19.03, 0.85, 2.28, 1.91 seconds, respectively, which all are quick.

Table 6
Experimental results in NIS_L.

Top ratio (%) (Number of objects)	Number of rare classes included (Coverage (%))						
	NIEOD	NED	IE	SEQ	FindCBLOF	DIS	KNN
4.73(7)	5(83.33)	4(66.67)	5(83.33)	5(83.33)	4(66.67)	5(83.33)	4(66.67)
5.41(8)	6(100.00)	4(66.67)	5(83.33)	5(83.33)	4(66.67)	5(83.33)	4(66.67)
6.08(9)	6(100.00)	4(66.67)	6(100.00)	5(83.33)	4(66.67)	6(100.00)	4(66.67)
8.11(12)	6(100.00)	4(66.67)	6(100.00)	6(100.00)	4(66.67)	6(100.00)	5(83.33)
10.14(15)	6(100.00)	5(83.33)	6(100.00)	6(100.00)	4(66.67)	6(100.00)	6(100.00)
13.51(20)	6(100.00)	6(100.00)	6(100.00)	6(100.00)	4(66.67)	6(100.00)	6(100.00)
20.27(30)	6(100.00)	6(100.00)	6(100.00)	6(100.00)	6(100.00)	6(100.00)	6(100.00)
27.03(40)	6(100.00)	6(100.00)	6(100.00)	6(100.00)	6(100.00)	6(100.00)	6(100.00)

NIS_L contains 148 objects and 6 outliers, denoted by $|U_L| = 148$ and $|OS_{\text{true}}(U_L)| = 6$.

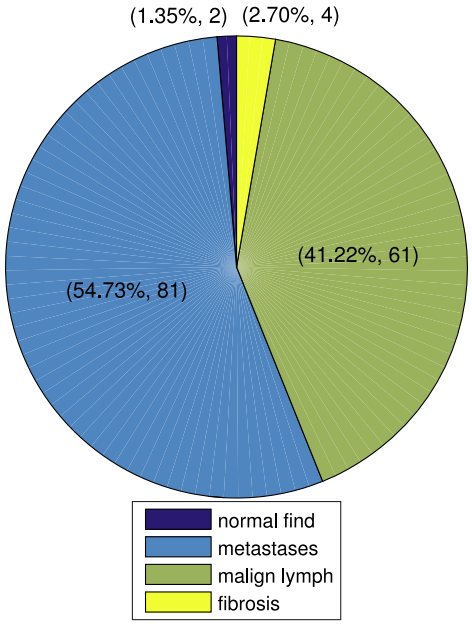

Fig. 6. Classification pie-chart of Lymphography data set.

For NIEOD, experimental results with the λ_L change are described in Fig. 9, which contains the first five standards of "Top ratio (Number of objects)" in Table 6 and Fig. 8. As an example, the level of top ratio 5.41% (object number 8) is focused on to make some sensitivity analyses as follows.

(1) When λ_L increases, "Number of rare classes included" monotonically decreases.
(2) However, its value changes slightly and stabilizes at two fixed values 5 and 6, so regarding the validity, NIEOD becomes better than or equal to other six algorithms.
(3) Moreover, $\lambda_L \in [0.1, 0.3]$ would functions best to identify all 6 outliers, and then NIEOD becomes optimal among all the seven algorithms.

By virtue of Fig. 9, the other four levels can be similarly analyzed to reveal NIEOD's relevant tendency and superiority regarding the λ_L parameter change.

In summary, when compared to the six algorithms (NED, IE, SEQ, FindCBLOF, DIS, KNN), NIEOD manifests its validity for the Lymphography data set, where most attributes are categorical. Thus, this subsection shows that NIEOD is availably applied to the categorical data, except for the heterogeneous data (in above Section 4.1).

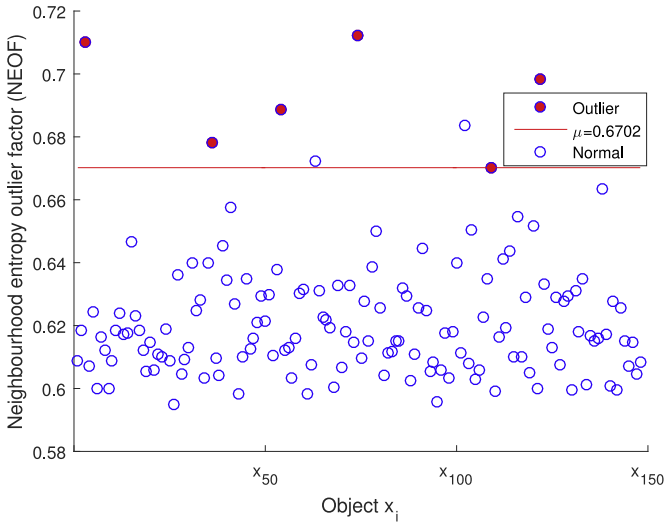

Fig. 7. NEOF-based object distribution in NIS_L.

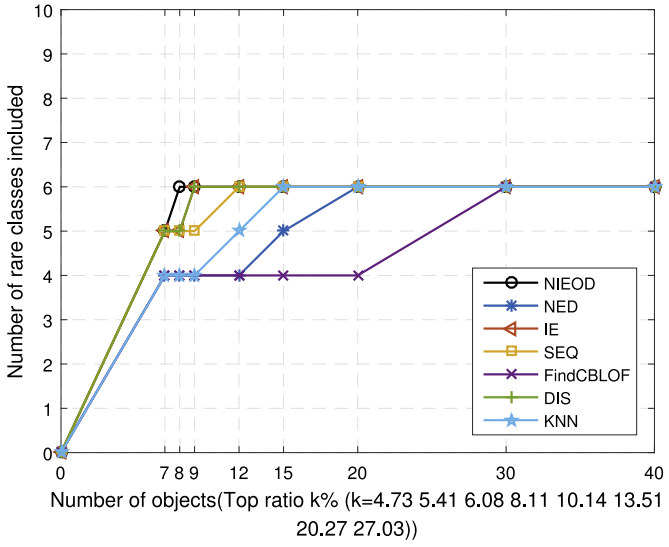

Fig. 8. Line chart of experimental results in NIS_L.

4.3. Wisconsin Breast Cancer data set

The Wisconsin Breast Cancer data set contains 699 objects, 9 numeric condition attributes, and 1 decision attribute. All instances are labeled to two cases: "benign" (number 458 and proportion 65.5%) and "malignant" (number 241 and proportion

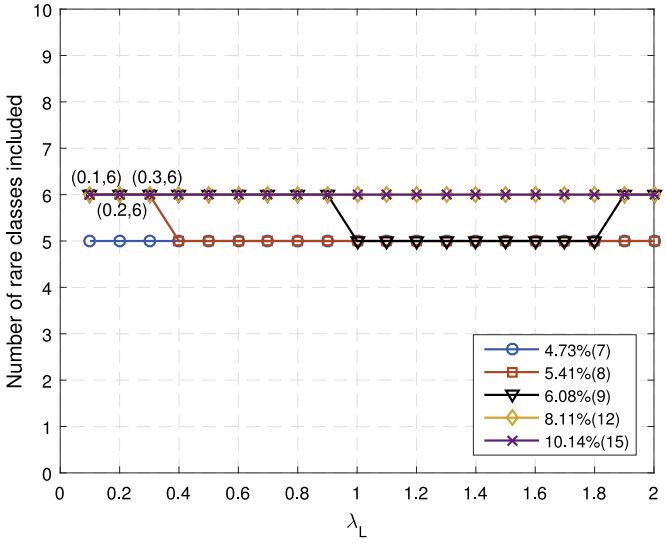

Fig. 9. Line chart of the number of rare classes included when λ_L changes.

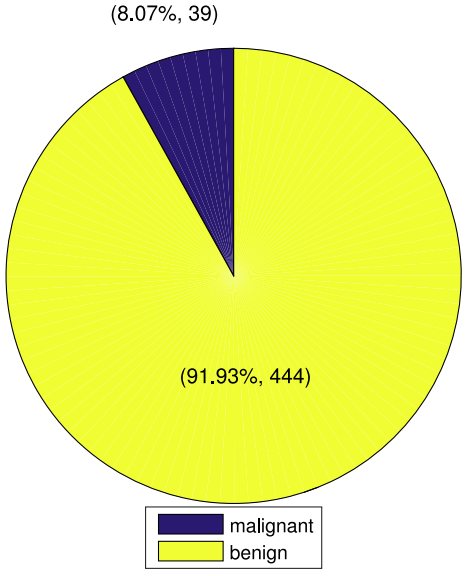

Fig. 10. Classification pie-chart of Wisconsin Breast Cancer data set.

34.5%). Herein, we mainly follow the experimental technique of Hawkins, He, Williams, and Baxter (2002) by removing some malignant instances to form a very unbalanced distribution. As a result, the resultant data set contains only 444 "benign" instances (with ratio 91.93%) and 39 "malignant" ones (with ratio 8.07%), whose pie-chart is given in Fig. 10, and "malignant" objects are deemed as outliers. Corresponding data induce neighborhood information system $NIS_W = (U_W, NR^{\varepsilon}_{C_W}, V_W, f_W)$.

For NIEOD, let $\lambda_W = 0.6$. The NEOF-based object distribution in NIS_W is shown in Fig. 11, and NIEOD can recognize all 39 outliers when $\mu = 0.5620$.

For the seven algorithms (NIEOD, NED, IE, SEQ, FindCBLOF, DIS, KNN), comparative experimental results and their corresponding line cart are provided in Table 7 and Fig. 12, respectively. For the performance, NIEOD is better than or equal to the other six algorithms, thus becoming optimal. Moreover, NIEOD, NED, IE, SEQ, FindCBLOF, DIS, KNN generally spend 118.71, 16.61, 193.74, 94.95, 0.89, 14.89, 10.08 seconds, respectively, and thus NIEOD's elapsed time is also quick and feasible.

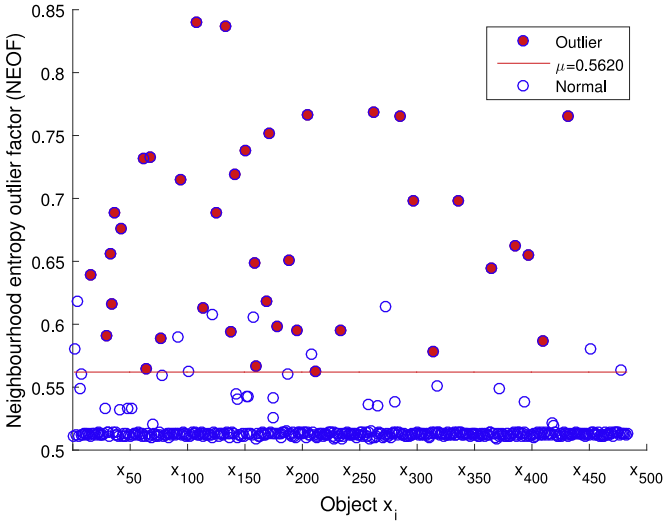

Fig. 11. NEOF-based object distribution in NIS_W.

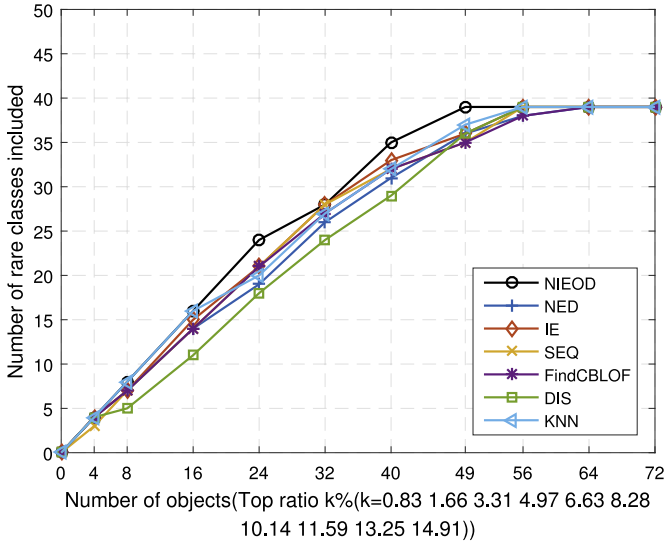

Fig. 12. Line chart of experimental results in NIS_W.

For NIEOD, experimental results with the λ_W change are described in Fig. 13, which contains five moderate standards of "Top ratio (Number of objects)" in Table 7 and Fig. 12. As an example, top ratio 10.14% (object number 49) is focused on to make some sensitivity analyses.

(1) When λ_W increases, "Number of rare classes included" almost increases.
(2) However, its value sharply changes when $\lambda_L \leq 0.4$, while generally stabilizes at points 37 and 38 when $\lambda_L \geq 0.4$. Because the stable values (37 and 38) are greater than other values achieved by the other algorithms, NIEOD has the better validity than the other algorithms when $\lambda_L \geq 0.4$.
(3) When $\lambda_W \geq 0.4$, NIEOD generally has stability and optimization. In particular, $\lambda_W = 0.4, 0.6, 1.8$ would offer the best performance to collect all 39 outliers, and then NIEOD maintains its optimal status among all algorithms.

The other four levels can be similarly analyzed to reveal NIEOD's stability and superiority in the parameter change.

In contrast with the other six algorithms (NED, IE, SEQ, FindCBLOF, DIS, KNN), NIEOD exhibits validity for the Wisconsin Breast Cancer data set, which carries only numeric attributes. Therefore,

Table 7
Experimental results in NIS_W.

Top ratio (%) (Number of objects)	Number of rare classes included (Coverage (%))						
	NIEOD	NED	IE	SEQ	FindCBLOF	DIS	KNN
0.83(4)	4(10.26)	4(10.26)	4(10.26)	3(7.69)	4(10.26)	4(10.26)	4(10.26)
1.66(8)	8(20.51)	7(17.95)	7(17.95)	7(17.95)	7(17.95)	5(12.82)	8(20.51)
3.31(16)	16(41.03)	14(35.90)	15(38.46)	14(35.90)	14(35.90)	11(28.21)	16(41.03)
4.97(24)	24(61.54)	19(48.72)	21(53.85)	21(53.85)	21(53.85)	18(46.15)	20(51.28)
6.63(32)	28(71.79)	26(66.67)	28(71.79)	28(71.79)	27(69.23)	24(61.54)	27(69.23)
8.28(40)	35(89.74)	31(79.49)	33(84.62)	32(82.05)	32(82.05)	29(74.36)	32(82.05)
10.14(49)	39(100.00)	36(92.31)	36(92.31)	35(89.74)	35(89.74)	36(92.31)	37(94.87)
11.59(56)	39(100.00)	38(97.44)	39(100.00)	39(100.00)	38(97.44)	39(100.00)	39(100.00)
13.25(64)	39(100.00)	39(100.00)	39(100.00)	39(100.00)	39(100.00)	39(100.00)	39(100.00)
14.91(72)	39(100.00)	39(100.00)	39(100.00)	39(100.00)	39(100.00)	39(100.00)	39(100.00)

NIS_W contains 483 objects and 39 outliers, denoted by $|U_W| = 483$ and $|OS_{\text{true}}(U_W)| = 39$.

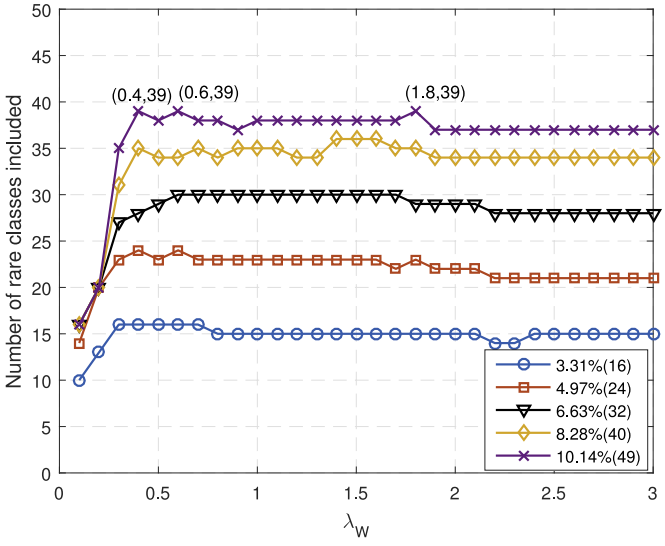

Fig. 13. Line chart of the number of rare classes included when λ_W changes.

NIEOD is also effectively applied to the numeric data, except for the previous hybrid and categorical data.

5. Conclusion

Outlier detection has extensive applications in expert and intelligent systems. However, the traditional distance-based detection method never effectively applies to categorical data, while the classical rough set-based method can not effectively handle numeric data and further mixed data. Against the hybrid data-driving of neighborhood information system, this paper researches the outlier detection based on the neighborhood information entropy and its developmental measures. The traditional information entropy is extended to the neighborhood information entropy, and the latter is further utilized to constructed serial information measures; as a result, the final outlier factor NEOF is established to effectively recognize outliers, and a relevant detection algorithm NIEOD is designed. Note that Fig. 1 provides the structural evolution diagram of all constructional measures. By virtue of the extended feature of neighborhood rough sets, the new detection approach and the corresponding NIEOD algorithm hold good adaptability, and their applicative data sets extensively concern categorical, numeric, and hybrid attributes.

Herein, the whole experimental outcomes of three UCI data sets are analyzed and summarized, and thus NIEOD effectively applies to all data types (including the hybrid, categorical, numeric types).

(1) Regarding the Annealing data set which is hybrid, NIEOD is always better than IE, DIS, KNN, while it is better than NED, SEQ, FindCBLOF for the less top ratio (or number of objects).
(2) Regarding the Lymphography data set which tends to the categorical, NIEOD has (a bit) better superiority than or the same performance as the other six algorithms: NED, IE, SEQ, FindCBLOF, DIS, KNN.
(3) Regarding the Wisconsin Breast Cancer data set which is numeric, NIEOD is better than or equal to the other six existing algorithms, thus becoming optimal.

Based on the relevant experiments and results, NIEOD generally achieves better performance than six main detection ways (including NED, IE, SEQ, FindCBLOF, DIS, KNN) from the accuracy viewpoint. For the computation time, NIEOD spends relatively much, which is mainly attributed to its pursuit of high detection accuracy (via the in-depth covering determination and multiple entropy calculation), but its time result well falls into the practical feasible range; in contrast, some other algorithms may have less elapsed time, but accordingly, they usually gain less detection accuracy. Moreover, NIEOD's parameter λ exhibits the similar monotonicity-change law for different top ratios (or numbers of objects), but its determination of optimal value needs sufficient experiments and relevant adjustments for each level of top ratios.

According to the theoretical construction and experimental verification, our detection measure and algorithm become robust and efficient, and the superiority mainly benefits from the distance selection of heterogeneous HEOM, the radius determination regarding self-adapting statistics, and the measure construction with in-depth integration. The relevant study improves and deepens the traditional distance-based and rough sets-based detection methods, such as those constructed by Jiang et al. (2010) and Chen et al. (2010). The obtained results enrich the fields of outlier detection and rough sets, especially from a new perspective of hybrid data-driving, thus holding the application prospect in data mining.

The new detection method has following two limitations and corresponding resolutions.

(1) The proposed detection method depends on the neighborhood information entropy and its developmental measures, and thus it more applies to data sets with some uncertainty mechanisms. For the relevant limitation, the new method needs to fully acquire their data law and adaptation environment by extensive practical applications in expert and intelligent systems.
(2) For the calculation simplicity, the information fusion in our modeling mainly adopts the integration technology of single attributes, and thus it may usually have a weaker effect or accuracy for outlier detection. Herein, a 2-dimensional synthetic example (i.e., Example 2) is particularly provided for relevant il-

Table 8
Initial information system and NEOF value of Example 2.

U	c_1	c_2	NEOF
x_1	0.1	0.9	0.7743
x_2	0.3	0.7	0.6524
x_3	0.4	0.6	0.6470
x_4	0.5	0.5	0.5748
x_5	0.6	0.4	0.6470
x_6	0.7	0.3	0.6524
x_7	0.9	0.1	0.7743
x_8	0.7	0.7	0.6047

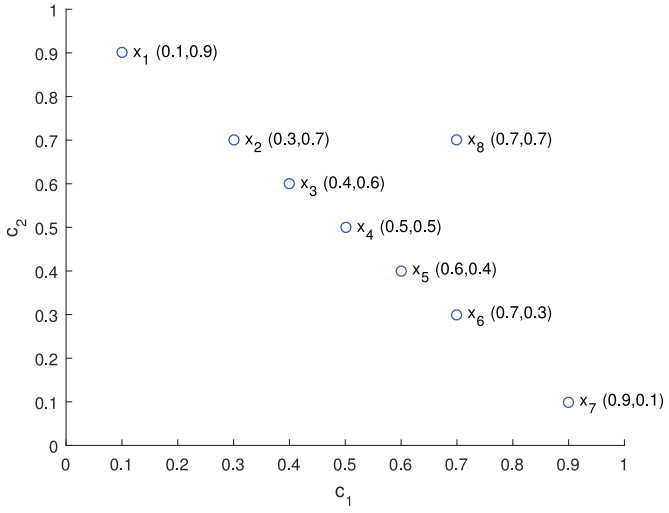

Fig. 14. Two-dimensional data distribution of Example 2.

lustrations. For this limitation, the integration strategy of nested attribute subsets can be further considered, and it tends to better results although it is accompanied with more complexities. Moreover, the overall or main attributes need more global considerations.

Example 2. An information system $IS = (U, C, V, f)$ is provided in Table 8, where $U = \{x_1, \ldots, x_8\}$ and $C = \{c_1, c_2\}$, and relevant data are vividly described in Fig. 14 to reflect the two-dimensional distribution. Let $\lambda = 1$, and Table 8 also gives relevant NIEOD's NEOF values (whose calculations are similar to those of Example 1). According to NIEOD, x_1, x_7 exhibit the highest NEOF value to naturally become outliers, such as for judgement threshold $\mu = 0.7$; on the other hand, x_8 offers the lowest outlier factor to never be viewed as an outlier. However, the data-distributional law (in Fig. 14) shows that both x_1, x_7 and x_8 largely tend to objective outliers, so NIEOD cannot discriminate x_8 to cause a weakness. Herein, outlier x_8 is actually hidden in one dimension but is clearly uncovered by two dimensions. Therefore, NIEOD leads to a detection failure of outlier x_8 because it examines only one local feature at a time, and thus a global detection strategy is required in addition. □

For the future work, we below propose four research directions which are insightful for expert systems with applications.

(1) The new detection method and its combination with existing methods or classical technologies need to be extensively applied to practical environments of data mining, such as the network intrusion detection and medical data processing. Multiple evaluation indexes and comprehensive assessment can be further adopted.
(2) On the basis of the information measures, other attribute measures (such as the attribute significance regarding dependency) can be mined to comprehensively represent the outlier degrees. In the detection process, the attribute subset sequence is worth solely or combinedly considering to achieve good performance.
(3) To reduce the computational complexity of relevant algorithms (such as NIEOD), the unsupervised feature selection or attribute reduction of data sets is worth performing before outlier detection. Based on neighborhood rough sets, both attribute reducts and non-reducts (Shaari et al., 2009) can be explored to hierarchically classify attributes into three characteristic parts: nuclear, marginal, and useless attributes.
(4) In the current scenario of big data, the new method and relevant outlier detection can adopt the parallel and distributed implementations by virtue of both the individual computing and joint integration of single attributes (or even attribute subsets). Moreover, incremental and dynamic outlier detection are worth deeply exploring based on data objects.

Acknowledgments

The authors thank all of the editors and reviewers for their valuable suggestions, which have substantially improved this paper.

This work was supported by National Natural Science Foundation of China (61673285 and 61203285), Sichuan Youth Science & Technology Foundation of China (2017JQ0046), and Scientific Research Project of Sichuan Provincial Education Department of China (15ZB0029).

References

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. *ACM Sigmod Record, 30*(2), 37–46.
Bay, S. D. (1999). The uci kdd repository (http://kdd.ics.uci.edu).
Berna-Martinez, J. V., & Ortega, M. A. A. (2015). Algorithm for the detection of outliers based on the theory of rough sets. *Decision Support Systems, 75*, 63–75.
Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). Lof: Identifying density-based local outliers. *ACM Sigmod Record, 29*(2), 93–104.
Chen, H. M., Li, T. R., Cai, Y., Luo, C., & Fujita, H. (2016). Parallel attribute reduction in dominance-based neighborhood rough set. *Information Sciences, 373*, 351–368.
Chen, Y. M., Miao, D. Q., & Wang, R. Z. (2008). Outlier detection based on granular computing. In *International conference on rough sets and current trends in computing* (pp. 283–292).
Chen, Y. M., Miao, D. Q., & Zhang, H. Y. (2010). Neighborhood outlier detection. *Expert Systems with Applications, 37*(12), 8745–8749.
Chen, Y. M., Wu, K. S., Chen, X. H., Tang, C. H., & Zhu, Q. X. (2014). An entropy-based uncertainty measurement approach in neighborhood systems. *Information Sciences, 279*, 239–250.
Chen, Y. M., Xue, Y., Ma, Y., & Xu, F. F. (2017). Measures of uncertainty for neighborhood rough sets. *Knowledge-Based Systems, 120*, 226–235.
Chen, Y. M., Zhang, Z. J., Zheng, J. Z., Ying, M., & Yu, X. (2017). Gene selection for tumor classification using neighborhood rough sets and entropy measures. *Journal of Biomedical Informatics, 67*, 59–68.
Düntsch, I., & Gediga, G. (1998). Uncertainty measures of rough set prediction. *Artificial Intelligence, 106*(1), 109–137.
Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. Elsevier..
Hawkins, D. M. (1980). *Identification of outliers*. Chapman and Hall.
Hawkins, S., He, H., Williams, G. J., & Baxter, R. A. (2002). *Outlier detection using replicator neural networks*. CiteSeer.
He, Z. Y., Xu, X. F., & Deng, S. C. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters, 24*(9–10), 1641–1650.
Hu, Q. H., Liu, J. F., & Yu, D. R. (2008). Mixed feature selection based on granulation and approximation. *Knowledge-Based Systems, 21*(4), 294–304.
Hu, Q. H., Yu, D. R., Liu, J. F., & Wu, C. X. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Information Sciences, 178*(18), 3577–3594.
Hu, Q. H., Yu, D. R., & Xie, Z. X. (2006). Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognition Letters, 27*(5), 414–423.
Hu, Q. H., Yu, D. R., & Xie, Z. X. (2008). Neighborhood classifiers. *Expert Systems with Applications, 34*(2), 866–876.
Indurkhya, & Nitin (1998). *Predictive data mining*. Morgan Kaufmann Publishers.
Jain, A. K. (1999). Data clustering: A review. *Acm Computing Surveys, 31*(3), 264–323.
Jiang, F., & Chen, Y. M. (2015). Outlier detection based on granular computing and rough set theory. *Applied Intelligence, 42*(2), 303–322.
Jiang, F., Sui, Y., & Cao, C. (2008). A rough set approach to outlier detection. *International Journal of General Systems, 37*(5), 519–536.
Jiang, F., Sui, Y., & Cao, C. (2009). Some issues about outlier detection in rough set theory. *Expert Systems with Applications, 36*(3), 4680–4687.
Jiang, F., Sui, Y., & Cao, C. (2010). An information entropy-based approach to outlier detection in rough sets. *Expert Systems with Applications, 37*(9), 6338–6344.

Jiang, F., Sui, Y., & Cao, C. (2011). A hybrid approach to outlier detection based on boundary region. *Pattern Recognition Letters, 32*(14), 1860–1870.
Kennedy, & Ruby, L. (1997). *Solving data mining problems through pattern recognition*. Prentice Hall PTR.
Knorr, E. M., & Ng, R. T. (1997). A unified notion of outliers: Properties and computation. In *International conference on knowledge discovery and data mining* (pp. 219–222).
Knorr, E. M., & Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *International conference on very large data bases* (pp. 392–403).
Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: Algorithms and applications. *The VLDB Journal, 8*(3), 237–253.
Kumar, S. U., & Inbarani, H. H. (2016). Pso-based feature selection and neighborhood rough set-based classification for bci multiclass motor imagery task. *Neural Computing & Applications*, 1–20.
Li, X. J., & Rao, F. (2012). Outlier detection using the information entropy of neighborhood rough sets. *Autoimmunity Reviews, 9*(9), 589–590.
Liang, J., Shi, Z., Li, D., & Wierman, M. J. (2006). Information entropy, rough entropy and knowledge granulation in incomplete information systems. *International Journal of General Systems, 35*(6), 641–654.
Liang, J. Y., Wang, J. H., & Qian, Y. H. (2009). A new measure of uncertainty based on knowledge granulation for rough sets. *Information Sciences, 179*(4), 458–470.
Lin, T. Y. (1988). Neighborhood systems and relational databases. In *Sixteenth ACM conference on computer science, Atlanta, Georgia, USA, February* (p. 725).
Lin, T. Y. (2008). Neighborhood systems: A qualitative theory for fuzzy and rough sets. In *Joint conference on information sciences* (pp. 257–260).
Liu, Y., Yang, J. J., Chen, Y. H., Tan, K. Z., Wang, L. G., & Yan, X. Z. (2017). Stability analysis of hyperspectral band selection algorithms based on neighborhood rough set theory for classification. *Chemometrics & Intelligent Laboratory Systems, 169*, 35–44.
Pawlak, Z. (1982). Rough set. *International Journal of Computer & Information Sciences, 11*(5).
Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*. Dordrecht: Kluwer Academic Publishers.
Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Record, 29*(2), 427–438.
Rousseeuw, P. J., & Leroy, A. M. (1987). Robust regression and outlier detection. *Journal of the American Statistical Association, 31*(2), 260–261.
Shaari, F., Bakar, A. A., & Hamdan, A. R. (2009). Outlier detection based on rough sets theory. *Intelligent Data Analysis, 13*(2), 191–206.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(4), 379–423.
Wang, C. Z., Shao, M. W., He, Q., Qian, Y. H., & Qi, Y. L. (2016). Feature subset selection based on fuzzy neighborhood rough sets. *Knowledge-Based Systems, 111*, 173–179.
Wang, G. Y., Ma, X. A., & Yu, H. (2015). Monotonic uncertainty measures for attribute reduction in probabilistic rough set model. *International Journal of Approximate Reasoning, 59*(C), 41–67.
Williams, J. W. J. (1964). Algorithm 232: Heapsort. *Communications of the ACM, 7*(6), 347–348.
Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research, 6*(1), 1–34.
Wu, W. Z., & Zhang, W. X. (2002). Neighborhood operator systems and approximations. *Information Sciences, 144*(1), 201–217.
Yao, Y. Y. (1998). Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences, 111*(1–4), 239–259.
Zhang, X., Mei, C. L., Chen, D. G., & Li, J. H. (2016). Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. *Pattern Recognition, 56*(1), 1–15.
Zhang, X. Y., & Miao, D. Q. (2017). Three-layer granular structures and three-way informational measures of a decision table. *Information Sciences, 412–413*, 67–86.