Information Fusion 93 (2023) 192–208

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

A multi-source information fusion model for outlier detection

Pengfei Zhang, Tianrui Li *, Guoqiang Wang, Dexian Wang, Pei Lai, Fan Zhang **

School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, PR China
Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 611756, PR China
National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 611756, PR China
Manufacturing Industry Chains Collaboration and Information Support Technology Key Laboratory of Sichuan Province, Southwest Jiaotong University, Chengdu 611756, PR China

ARTICLE INFO

Keywords:
Multi-source information fusion
Information set
Knowledge granule
Granular computing
Fuzzy knowledge measure
Outlier detection

ABSTRACT

Multi-source information fusion (MSIF) is a useful strategy for combining complimentary data from numerous information sources to produce an overall precise description, which can help with effective decision-making, prediction, and categorization, etc. In order to find the objects that are different from the expected ones after fusion, i.e., anomalies, or outliers, an MSIF model is put forward for outlier detection. This is a two-stage model that includes fusion of multiple information sources and outlier detection of fused data. The first stage uses information sets to construct uncertainty criteria for information source values and combines multiple information sources into a single information source based on the minimum uncertainty strategy. The second stage uses the Gaussian kernel method for possibility modeling based on the fused data to construct knowledge granules. From the perspective of granular computing, outliers in the fused data can be assigned to each knowledge granule. Then, we can find all outliers just by evaluating these knowledge granules. Inspired by this, the fuzzy knowledge measure (FKM) is proposed to evaluate the knowledge granule. Moreover, several metrics are induced on the basis of FKM to describe outliers in knowledge granules and an FKM-based outlier detection algorithm (FKMOD) is designed. Finally, we conduct the experiments on sixteen open access outlier detection datasets. The experimental results show that the proposed FKMOD method has more accurate detection performance than nine classical methods.

1. Introduction

1.1. Research background

Multi-source information fusion (MSIF) is the technique of combining and merging information or data from multiple sources in order to form a unified result [1]. A single data source can only obtain part of the information segment of the object, while the information of multiple data sources can perfectly and accurately reflect the overall information of the objects after fusion. The fused information thereby has the characteristics of abnormality, complementarity, coordination, real-time and low cost [2].

MSIF techniques can be broadly divided into three levels, i.e., data level-based fusion, feature level-based fusion, and decision level-based fusion [3]. The necessary information must be correlated and aligned regardless of the kind of data level, feature level, or decision level with the difference of the sequence in which the data are correlated and mutually matched. The advantages of these three types of fusion technologies at different fusion stages depend on actual needs [4]. Theoretically, the advantage of data level fusion is that a large amount of original information can be retained to provide the target with as detailed information as possible, and to obtain the most accurate fusion effect possible. In this paper, we mainly focus on data level fusion approaches [5], which can efficiently deal with homogeneous data (mainly numerical data) collected by multiple information sources.

MSIF algorithms are the basis and important elements of fusion processing [6]. At present, most algorithms involved in MSIF are combined with various theories, e.g. Bayes estimator [7,8], fuzzy set theory [9,10], possibility theory [11], D–S theory [3,12,13], rough set theory [2,14], etc. In addition, it is also a good way to combine different theories with each other [15–17]. One of the main challenges of MSIF is the problem of uncertainty, which involves the uncertainties of the data itself and fusion process. Also, the information loss of data after fusion has to be considered, as well as the occurrence of outliers. Therefore, finding the outliers in the fused information system is a problematic

* Corresponding author at: School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, PR China.
** Corresponding author at: Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 611756, PR China.
E-mail addresses: feifeihappy55@163.com (P. Zhang), trli@swjtu.edu.cn (T. Li), gqwang18@163.com (G. Wang), wangdexian@my.swjtu.edu.cn (D. Wang), peilai@my.swjtu.edu.cn (P. Lai), fan.zhang@swjtu.edu.cn (F. Zhang).

https://doi.org/10.1016/j.inffus.2022.12.027
Received 8 July 2022; Received in revised form 27 November 2022; Accepted 28 December 2022
Available online 3 January 2023

topic that deserves attention, which is beneficial for improving the quality of fusion.

1.2. Related work

In recent years, related researches based on MSIF have attracted more and more attention. For example, Yager [18] provided a general fusion framework for multi-source data. However, the framework emphasizes that the fusion process involves the use of data supplied by information sources and other knowledge but does not indicate which data type to be fused. Lin et al. proposed an information fusion method based on granular computing theory and D–S theory, which can combine the granular structures with both reliability and conflict from multiple information sources [17]. But how to construct the belief function is challenging in their approach, and they cannot cope with complex applications. Xu et al. considered the information entropy approach to fuse multiple data sources in incomplete information systems [19]. Their approach necessitates the use of labels as a guide during the fusion process and may result in the fusion of incomplete data, which lowers the quality of the fusion. Che et al. employed D–S theory and probability theory to fuse the numerical characterization of uncertain data in a multi-source information system [16]. Nevertheless, the conditional probability of the target set needs to be calculated in their algorithm. If the sample is increased, it will cause an increase in the information uncertainty and affect the quality of fusion as well. Yang et al. put forward a multigranulation method for information fusion in multi-source decision information systems [20]. They adopted a pessimistic and optimistic fusion mechanism, which is too tolerant or restrictive when solving practical problems and is also not conducive to the fusion of multiple information sources. Huang et al. used a dynamic fusion mechanism to merge multiple interval-valued data based on the idea of fuzzy granularity [21]. The focus of this paper is on the effective fusion of multi-source interval-valued data with dynamic data source updating, which includes the addition of new sources and the removal of old sources. Zhang et al. concentrated on efficient fusion of multi-source homogeneous information system with a data-level fusion model, which fully considered the further processing of the fused data from the perspective of attribute reduction [5]. This method is beneficial to improving the quality of fusion and reducing the redundant features after fusion. Nonetheless, their method is sensitive to the neighborhood radius, which also affects the accuracy of fusion.

The literature cited above examines the mechanisms of multi-source information fusion from various perspectives. However, the methods and models involved will increase the uncertainty of information during the fusion process, thus affecting the quality of fusion. In addition, few studies consider the repulsiveness of the fused data, that is, there may be outliers in the fused data. Outlier detection (OD) is also can be called anomaly detection, which is a key data mining task with copious applications [22], including intrusion detection [23], rare disease detection [24], fraud detection [25], etc. It is generally recognized that the technological or environmental limitations of the gathering prevent the data from each information source (e.g. sensor) from being fused as intended. For instance, the fusion process may produce noise or some anomalous objects, which may affect the process of decision-making and the accuracy of the fusion method. Thereby, it is meaningful to perform outlier detection in the fused data, which not only helps to enhance the quality of fusion, but also aids to discover the valuable information. In summary, it is crucial to consider both the lack of fusion accuracy caused by the uncertainty in the fusion process and the problem of outlier detection in the fused data.

1.3. Proposed work

Regarding some of the previously mentioned fusion methods and consideration of the situation of outliers after fusion, this paper proposes a unified multi-source information fusion model for outlier detection, as shown in Fig. 1. This model can be divided into two stages, which are described as follows.

1.3.1. The first stage

The goal of this stage is to fuse multiple information sources. The previous review discussed that one of the most critical factors affecting fusion accuracy is the problem of uncertainty. The main characteristics of uncertainty problems are randomness, fuzziness, incompleteness, and instability, among which randomness and fuzziness are the essential characteristics of uncertainty. In order to portray the uncertainty of information in data sources, the original data needs to be transformed into fuzzy sets by possibility modeling. This is because uncertainty in fuzzy set theory is described by the membership function, which maps the information source values (ISVs) to the degrees of association to the set in the interval [0, 1]. Regarding a group of attribute values $X = \{x_1, x_2, \ldots, x_n\}$, a fuzzy set F is a set of ordered pairs, i.e., $\{(x_1, \mu(x_1)), (x_2, \mu(x_2)), \ldots, (x_n, \mu(x_n))\}$. For any $x_k \in X$, The $\mu(x_k))$ gives the "membership degree" of x_k in fuzzy set F. However, the following disadvantages exist with this fuzzy set representation of uncertainty: (1) It treats the values of membership function separate from the ISVs. (2) It does not have a mechanism to link them together into a single entity. (3) The overall degree of ambiguity of the fuzzy set is not taken into account in its individual elements. (4) The predefined regular shape of membership function is restrictive. In other words, the interaction between the ISVs and the membership degrees is not recorded. Additionally, the individual membership degree does not capture the overall uncertainty linked to the vague concept itself. Therefore, to address these problems, Aggarwal and Hanmandlu introduced the concept of information set (ISet) [26]. It is a general framework for uncertainty representation by explicitly considering the actual ISVs of any type, such as probabilistic, possibilistic, and other actual attribute values.

From Fig. 1, a multi-source numerical information system (MsNIS) is constructed, which is composed of m sub-numerical information systems, i.e., NIS_1, NIS_2, $\cdots$, NIS_m. First, it is necessary to translate the original data from the sub-numerical information systems into the actual ISVs because the uncertainty of the data itself is not easy to measure. Then, we develop such a method by means of the generalized Hanman–Anirban entropy function [27] that is information-theoretic and can connect the ISVs and the value of gain function into a single entropy value with its free parameters. Herein, the gain function can be called as an "agent" that perceives the uncertainty in the ISVs. In this paper, the Gaussian membership degree (GMD) plays the role of an agent in the case where the ISV forms a fuzzy set because it is a particular form of the Hanman–Anirban entropy function that is readily available. Subsequently, the information set (ISet) is used to measure the ISVs in each sub-numerical information system. The fundamental idea enshrined in the ISet is to unravel uncertainty by the parametric gain function by the values of the information source. It can also give the uncertainty in the possibility distribution (provided by membership function) by capturing this distribution through the information gain. Next, two other ISet-based metrics are introduced to better characterize the uncertainty of ISVs and agents, i.e., Shannon source transform (SST) and Shannon inverse transform (SIST). The total uncertainty measure (TUM) of an ISV is determined by calculating the sum of SST and SIST, which is beneficial for us to perform MSIF and solve the uncertainty problem caused by information source itself and information gain function. Finally, based on the principle of minimum uncertainty [5], a new single-source numerical information system is produced by fusing m sub-numerical information systems.

1.3.2. The second stage

The previous stage is to consider the fusion of multiple information sources, and the task of this stage is mainly to investigate the outlier problem of the fused data. In fact, the first stage is the fusion of attributes. The attribute with the minimum uncertainty is selected as the attribute of the new fused information system. This process may cause some objects to be unsuitable or exclusive in the new information system. Therefore, the purpose of this stage is to find out these "unsocial" objects (so-called outliers) to obtain the final fusion result.

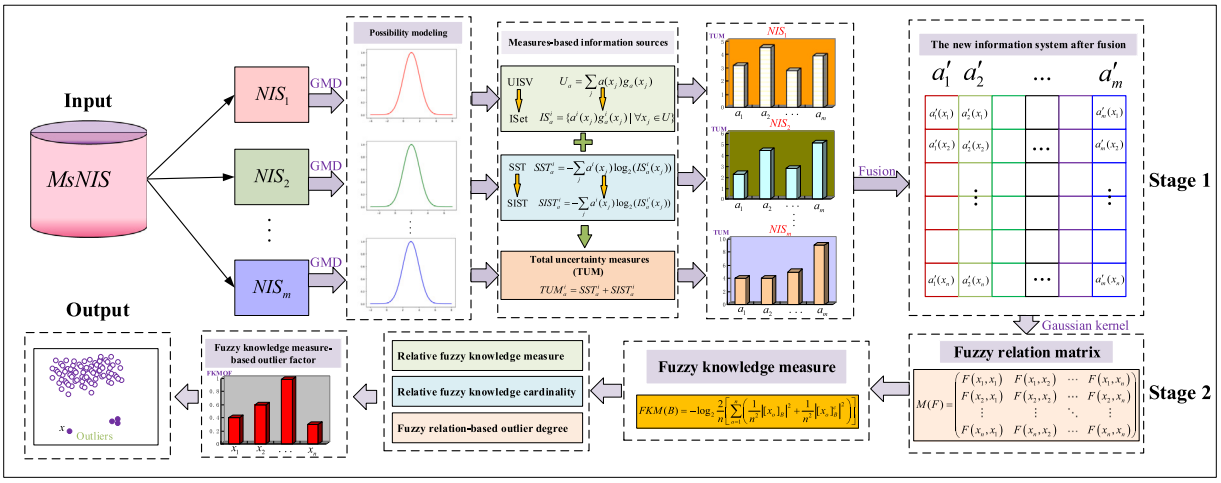
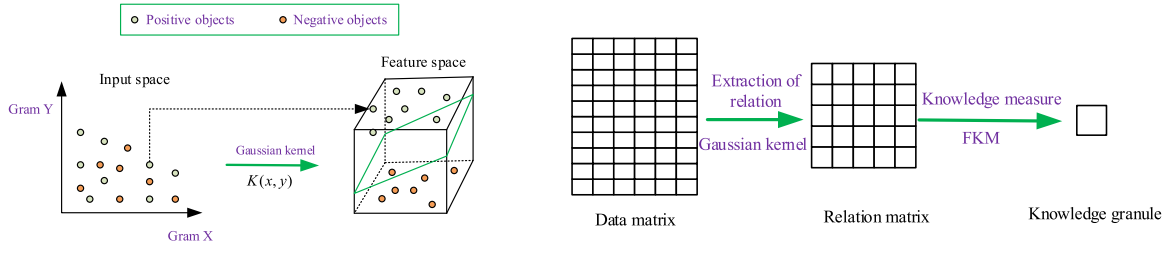


Fig. 1. A unified framework of a multi-source information fusion model for outlier detection.



(a) Possibility modeling (b) Extraction of relations

Fig. 2. (a) shows the effect of Gaussian kernel functions on data separation. This is a process of possibility modeling. (b) reflects the relationships between objects, and the granular structure of the universe. It is also a schematic diagram of the granular computing.

In the 1990s, Vapnik et al. systematically established statistical learning theory and successfully introduced the support vector machine (SVM) algorithm, which kicked off the research of kernel methods [28]. The basic principle of the kernel method is to map the input space to a high-dimensional (even infinite-dimensional) feature space through the nonlinear mapping implicitly defined by the kernel function and to indirectly realize the non-linearity of the original input space by implementing a linear algorithm in the feature space. According to the T. M. Cover theorem, data mapping into a high-dimensional feature space will result in more linearly separable data (see Fig. 2(a)). The Mercer condition turns the corresponding optimization problem into a convex problem, so there are no local minima. Moreover, it has been proved that if the input space is finite, a high-dimensional feature space must make the samples separable [29]. Table 1 lists four commonly used kernel functions. Contrary to linear kernels, Gaussian kernel can handle situations where the relationship between class labels and features is non-linear because they map objects non-linearly to higher dimensional space. In [30], linear kernel is proved to be a particular case of Gaussian kernel since linear kernel with penalty parameters have the same performance as Gaussian kernel. The polynomial kernel has more hyperparameters than the Gaussian kernel, and the number of parameters affects the model complexity. In addition, the literature [31] has shown that the performance of the Sigmoid kernel is almost the same as that of the Gaussian kernel under certain parameters, and the kernel matrix calculated by the Sigmoid kernel is not necessarily a positive-definite matrix. Meanwhile, the SVM model constructed by it is usually less accurate than the Gaussian kernel. Therefore, this paper chooses the Gaussian kernel function to model the fused data.

On the other hand, the Gaussian kernel function is used to calculate the fuzzy relation between objects of the fused data. According to

Table 1
The four commonly used kernel functions, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ denote the output vectors, σ indicates the slope, d refers to the degree of polynomial, and r the constant term.

Kernel functions	Formulas
Linear kernel function	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$
Polynomial kernel function	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\sigma \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + r)^d, \sigma > 0$
Gaussian kernel function	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma}\right), \sigma > 0$
Sigmoid kernel function	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\sigma \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + r)$

Moser's work, any kernel satisfying reflexivity and symmetry is at least T_{cos}-transitive [32]. In [33], Hu et al. proved that Gaussian kernel satisfy reflexive, symmetric and T_{cos}-transitive, which can induce the fuzzy T_{cos}-equivalence relation. From the data-driven perspective, the information in the original information system can be seen as a matrix composed of data, that is, a data matrix. A fuzzy relation matrix that captures the relationship between objects can be created using the fuzzy T_{cos}-equivalence relation. Simultaneously, the fuzzy relation matrix expresses the granular structure of the universe from the idea of granular computing (see Fig. 2(b)). The granular structure comprises each knowledge (information) granule, which is a collection of objects aggregated by fuzzy T_{cos}-equivalence relation. Thus a fuzzy set about objects can be created by the knowledge granule.

Considering the inclusion relationship between objects and the knowledge granule, it is necessary to find a measure to distinguish objects in the knowledge granule to mine the outlier objects. According to granular computing, a fuzzy set can also be identified as the knowledge granule containing objects [34]. As a result, the criteria for evaluating fuzzy sets also apply to knowledge granules. Generally speaking, fuzzy entropy is used to estimate the degree of uncertainty, disorder and

irregularity between fuzzy sets [35]. However, an entropy measure cannot capture all uncertainties in fuzzy sets. To solve this problem, the term named fuzzy knowledge measure (FKM) is proposed, which can be viewed as a dual measure of fuzzy entropy or uncertainty for a fuzzy set. In contrast to fuzzy entropy, a FKM can be used to measure the degree of certainty, order and regularity between fuzzy sets [36]. A FKM appears that the less entropy may always accompany the greater amount of knowledge [37]. Thus, the FKM is employed to evaluate the knowledge granules to find the outliers in this paper. Meanwhile, three new fuzzy knowledge measures are derived on the basis of FKM, which are used to describe the degree of outliers in new fused information system. Finally, FKM-based an algorithm for outlier detection (FKMOD) is designed.

1.4. Contribution

In order to reduce the uncertainty in the data itself in MSIF and the existence of outliers in the fused data, we propose a unified multi-source information fusion model for outlier detection. It is noteworthy that our fusion model is suitable for multi-source homogeneous information system, especially for numerical data. The fusion model can be divided into two stages. The first stage involves the fusing of data from multiple information sources, while the second stage involves the detection of outliers in the fused data. The advantage of this proposed model is to reduce the uncertainty of information sources and improve the quality of fusion. According to aforementioned descriptions, our main contributions are as follows.

(1) A unified multi-source information fusion model is proposed for outlier detection, which can better handle the fusion of multiple homogeneous information sources and the outliers of the fused data.
(2) ISet is used for the first time to deal with uncertainty in the fusion of multiple information sources, which can simultaneously measure the uncertainty of ISVs and agent.
(3) Several fuzzy knowledge measures about the outlier degrees of objects are proposed, which can effectively find the outliers of the fused data.
(4) The experimental results show the effectiveness of our proposed FKMOD and its superiority compared to the existing some classical algorithms.

1.5. Organization

The rest of this paper is organized as follows. Section 2 gives the main definitions of multi-source numerical information systems, information sets and uncertainty measures of ISV. Section 3 comes up with a multi-source numerical data fusion algorithm based on infimum-measure approach. Gaussian kernel based-fuzzy knowledge measure for outlier detection algorithm is designed in Section 4. In Section 5, suitable experiments are presented to test the proposed algorithms performance, in comparison with some classical methods. Conclusions are provided in Section 6.

2. Preliminaries

In this section, we formally summarize some notations and definitions used in this paper, such as fuzzy set, fuzzy relation, single-source information systems and multi-source information systems [38–43].

In fuzzy sets, given a non-empty finite sample set $U = \{x_1, x_2, \dots, x_n\}$ and $|U| = n$, $F : U \to [0,1]$ expresses a mapping of U to $[0,1]$ where F is called the fuzzy set on U [38]. $\forall\, x \in U$, $F(x)$ is known as membership function of F, or the membership degree of x for F. The fuzzy set is indicated as $F = (f(x_1), f(x_2), \dots, f(x_n))$ or $F = \sum_{i=1}^{n} f(x_i)/x_i$. The cardinality of F can be denoted as $|F| = \sum_{i=1}^{n} F(x_i)$. In addition, $\mathcal{F}(U)$ denotes the set of all fuzzy sets on U. A fuzzy relation is a fuzzy set based on the Cartesian product of two universes, which is based on the Cartesian product of U and itself, indicated by $\mathcal{F}(U \times U)$ [43]. The fuzzy relation F can be conveniently represented by a fuzzy relation matrix, which is often denoted as

$$M(F) = \begin{pmatrix} F(x_1,x_1) & F(x_1,x_2) & \cdots & F(x_1,x_n) \\ F(x_2,x_1) & F(x_2,x_2) & \cdots & F(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ F(x_n,x_1) & F(x_n,x_2) & \cdots & F(x_n,x_n) \end{pmatrix}, \tag{1}$$

where $F(x_i, x_j) \in [0,1]$. In addition, fuzzy relations satisfy three properties: $\forall\, x, y, z \in U$, we have (1) $F(x,x) = 1$ (reflexive), (2) $F(x,y) = F(y,x)$ (symmetric), and (3) $F(x,z) \geq F(x,y) \wedge F(y,z)$ (transitive). If F satisfies reflexive and symmetric, then F is regarded to as fuzzy similarity relation on U; if F satisfies reflexive, symmetric and transitive, then F is regarded to as fuzzy equivalence relation. The fuzzy relation is a special kind of the fuzzy set, including the following operations: union, intersection, and inverse.

Definition 1 ([42]). Given $F_1, F_2 \in \mathcal{F}(U)$, five operations are defined as:

(1) $F_1 = F_2 \Leftrightarrow F_1(x,y) = F_2(x,y)$;
(2) $F = F_1 \cup F_2 \Leftrightarrow F = \max\{F_1(x,y), F_2(x,y)\} = F_1(x,y) \vee F_2(x,y)$;
(3) $F = F_1 \cap F_2 \Leftrightarrow F = \min\{F_1(x,y), F_2(x,y)\} = F_1(x,y) \wedge F_2(x,y)$;
(4) $F_1 \subseteq F_2 \Leftrightarrow F_1(x,y) \leq F_2(x,y)$;
(5) $F^c(x,y) = 1 - F(x,y)$.

Definition 2 ([39]). Let (U, A, V) be a single-source information system, where U is a non-empty finite sample set and A is a non-empty finite attribute set. V is called an union of attribute domain, namely, $V = \sum_{a \in A} V_a$, where V_a is the attribute domain of the attribute a. An information function $a : U \times A \longrightarrow V$ that satisfies any $a \in A$ and $x \in U$, with $a(x) \in V_a$.

Definition 3 ([5]). Let $MsNIS = \{NIS_i | NIS_i = (U, A, V_i), i = 1, 2, \dots, m\}$ be a multi-source numerical information system, where (1) U is a set of objects (non-empty finite set); (2) NIS_i is the ith numerical data table in the MsNIS and m is the number of information source; (3) A is a set of attributes of the ith NIS_i (non-empty finite set); (4) V_a is the value of the attribute $a \in A$; and (5) $U \times A \to V_i$ is an information function for each $x \in U$ and $a \in A$, $a(x) \in V_i$ in the ith NIS_i where $a(x)$ is the information value with respect to x under the numeric attribute a. Moreover, the MsNIS can also be expressed by

$$(U, MsNIS) = \{NIS_1, NIS_2, \dots, NIS_m\}. \tag{2}$$

Example 1. Table 2 represents an MsNIS where $MsNIS = \{NIS_1, NIS_2, NIS_3, NIS_4\}$. $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ denotes six samples (objects) and $A = \{a_1, a_2, a_3, a_4\}$ denotes the attribute set.

2.1. Information sets in an MsNIS

For capturing uncertainty in information sources, Agarwal et al. established the concept of information set (ISet), which took into consideration both values of information source and membership degrees [26,44].

Definition 4. Let $MsNIS = \{NIS_i | NIS_i = (U, A, V_i), i = 1, 2, \dots, m\}$ be an MsNIS. For any $a \in A$, a collection of information source values (ISV) of the attribute a in an MsNIS can be denoted as

$$ISV_a = \{a(x_j) | \forall\, x_j \in U\}, \tag{3}$$

where $a(x_j)$ represents the individual information source value (or attribute value).

Table 2
An MsNIS.

U	NIS_1				NIS_2				NIS_3				NIS_4			
	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
x_1	2596	233	33.6	0.7	2606	237	31	0.9	2742	248	27.1	0.9	2504	214	43.3	0.9
x_2	2590	286	26.6	10.9	2605	225	35.3	0.9	2609	213	30.1	1.1	2503	220	34.6	0.9
x_3	2804	229	23.3	7.3	2617	230	30.5	0.9	2503	224	25.8	0.7	2501	230	39.3	2.2
x_4	2785	250	28.1	1	2612	221	0	0.7	2495	224	30	0.6	2880	206	35.4	2.9
x_5	2595	204	43.1	3.9	2612	219	37.6	0.6	2610	216	45.8	1.8	2768	252	39.8	6.8
x_6	2579	236	25.6	1.8	2886	243	38	2.7	2517	228	29.6	1.6	2511	225	29	1.9

In [26], Agarwal and Hanmandlu used Hanman–Anirban entropy function to represent the uncertainty of information source values. In the context of fuzzy set, for any $a \in A$, the uncertainty of ISV can denoted by

$$U_a = \sum_j a(x_j) g_a(x_j), x_j \in U, \tag{4}$$

where

$$g_a(x_j) = e^{-\alpha(a(x_j)^3+\beta(a(x_j))^2+\gamma(a(x_j))+\delta)^\lambda} \tag{5}$$

is called information gain corresponding to x_j. The presence of parameters in the information gain function causes the uncertainty representation to be adaptable [44].

Definition 5. Let $MsNIS = \{NIS_i | NIS_i = (U, A, V_i), i = 1, 2, \dots, m\}$ be an MsNIS. For any $a \in A$, the ISet corresponding to ISV and information gain in ith NIS_i is defined by

$$IS_a^i = \{a^i(x_j) g_a^i(x_j) | \forall\ x_j \in U\} \tag{6}$$

According to Eq. (6), for any $x_j \in U$, $IS_a^i(x_j) = a^i(x_j) g_a^i(x_j)$ is called an information value.

2.2. Uncertainty measures of ISV in an MsNIS

The ISet can evaluate the values of information sources based on their distributions, which can include membership degree values, possibility values, probability values and other attribute values. In this subsection, several uncertainty measures based on Shannon entropy are proposed to measure the fuzziness of ISV in an MsNIS.

Definition 6. Let $MsNIS = \{NIS_i | NIS_i = (U, A, V_i), i = 1, 2, \dots, m\}$ be an MsNIS. For any $a \in A$, the Shannon source transform (SST) of ISVs of attribute a with respect to U in ith NIS_i is defined by

$$SST_a^i = -\sum_j a^i(x_j) \log_2(IS_a^i(x_j)), x_j \in U, \tag{7}$$

where $IS_a^i(x_j) = a^i(x_j) g_a^i(x_j)$.

From Definition 6, the SST_a^i can evaluate the ISV via the information value $IS_a^i(x_j)$. Namely, it gives the uncertainty in the ISVs of attribute a with respect to U.

Definition 7. Let $MsNIS = \{NIS_i | NIS_i = (U, A, V_i), i = 1, 2, \dots, m\}$ be an MsNIS. For any $a \in A$, the Shannon inverse source transform (SIST) of ISVs of attribute a with respect to U in ith NIS_i is defined by

$$SIST_a^i = -\sum_j a^i(x_j) \log_2(IS_a^{i'}(x_j)), x_j \in U, \tag{8}$$

where $IS_a^{i'}(x_j) = a^i(x_j)(1 - g_a^i(x_j))$ is the complementary information value given by the complement information gain $1 - g_a(x_j)$.

The SIST gives the higher form of the complementary information, which is proved to be meaningful and useful in many practical applications [26].

Definition 8. Let $MsNIS = \{NIS_i | NIS_i = (U, A, V_i), i = 1, 2, \dots, m\}$ be an MsNIS. For any $a \in A$, the total uncertainty measures (TUM) of ISVs of the attribute a with respect to U in ith NIS_i can be denoted as

$$TUM_a^i = SST_a^i + SIST_a^i \tag{9}$$

Definition 8 revels the measures of total uncertainty in the ISVs, which is useful for evaluating the fuzziness of ISVs.

3. Multi-source information fusion in an MsNIS

In this section, we will use the previously mentioned uncertainty measures to fuse multiple information sources. The combination of multiple information sources enables complementary information and improves the accuracy of data classification [2]. Also, it can reveal the relationship between information sources [18]. However, there are two major issues with multi-source information fusion: (1) The variety of ways to merge information sources; and (2) the uncertainty of information sources. Since numerical data is frequently used in practical applications, this paper focuses on demonstrating the effectiveness of the proposed method by fusing multiple sources of numerical data.

3.1. Infimum-measure fusion method

According to Definition 8, the larger the value of TUM_a, the greater the fuzziness of ISVs, i.e., the greater the uncertainty of ISVs. Generally speaking, we wish to reduce the uncertainty of the ISVs in the multi-source information fusion process to acquire deterministic information. In this context, the infimum-measure function [5] will be used to fuse multi-source numerical data as follows.

Definition 9. Let $MsNIS = \{NIS_i | NIS_i = (U, A, V_i), i = 1, 2, \dots, m\}$ be an MsNIS. For any $a_k \in A$, the kth attribute of a new information system after fusion is denoted as

$$\mathrm{Inf}(\mathscr{F}(NIS_1(a_k)), \mathscr{F}(NIS_2(a_k)), \dots, \mathscr{F}(NIS_m(a_k))) \xrightarrow{select} V_{a_k}^{NIS_i}, \tag{10}$$

where $\mathscr{F}$ is called the infimum-measure function where $\mathscr{F} = TUM_a^i$.

From Definition 9, the multi-source information fusion scheme adopts the minimum uncertainty method based on infimum-measure function. The goal of this strategy is to minimize the total uncertainty of ISVs of the corresponding attributes in the new information system after fusion. Furthermore, the infimum-measure method is a type of multi-source homogeneous information fusion that has the advantage of retaining data from the original information sources while integrating data from multiple sources [5]. This can reduce the conflict and redundancy of multi-source information, and achieve information complementarity.

In terms of the concept of ISet, the information gain needs to be determined in the process of MSIF. The information gain $g_a(x_j)$ gives the gain in information corresponding to each ISV $a(x_j)$. Depending on different practical applications, the functions of information gain are also diverse, e.g. Gaussian membership degree function, sigmoid membership degree function, trapezoid membership degree function, triangular membership function and so on [26].

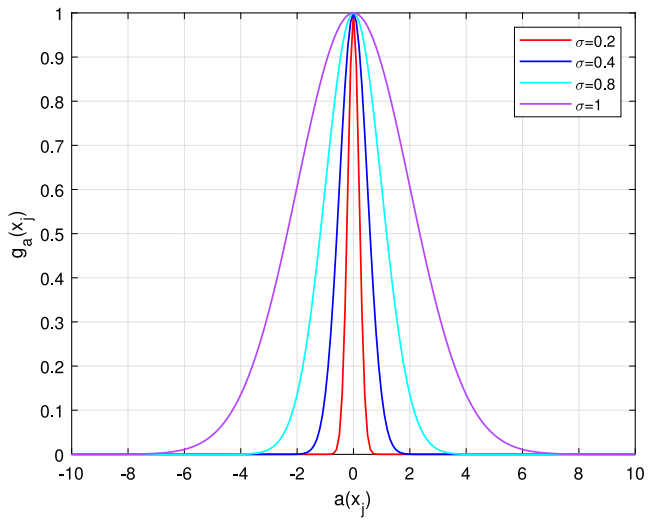


Fig. 3. The red, green, blue and purple lines represent the different Gaussian membership degree functions generated by $g_a(x_j) = e^{-(\frac{a(x_j)-a_{\text{mean}}}{\sqrt{2}\sigma_a})^2}$.

In this paper, we use the Gaussian membership degree as the information gain function, which is special form of Hanman–Anirban entropy function. Let $MsNIS = \{NIS_i | NIS_i = (U, A, V_i), i = 1, 2, \dots, m\}$ be an MsNIS where $U = \{x_1, x_2, \dots, x_n\}$ and $A = \{a_1, a_2, \dots, a_l\}$. For any $a \in A$, the mean IVSs can be denoted as $a_{\text{mean}} = \frac{1}{|U|}\sum_{j=1}^{n} a(x_j)$, $x_j \in U$. σ_a is the standard deviation of IVSs under the attribute a. On the basis of Eq. (5), if $\alpha = 0$, $\beta = 0$, $\gamma = \frac{1}{\sqrt{2}\sigma_a}$, $\delta = -\frac{a_{\text{mean}}}{\sqrt{2}\sigma_a}$, then we have

$$g_a(x_j) = e^{-(\frac{a(x_j)-a_{\text{mean}}}{\sqrt{2}\sigma_a})^{\lambda}}, \tag{11}$$

where $\lambda = 2$, different values of a_{mean} correspond to various Gaussian membership degree functions when σ_a is fixed, as shown in Fig. 3.

3.2. A specific fusion algorithm

There are generally discrepancies in order of magnitude and dimensionality in real-world datasets. As a result, the numerical attributes of the original information sources are normalized before data processing. There are many common normalization processing techniques, including $\min - \max$ normalization, z score normalization and decimal standardization, etc. In this paper, the original data sources are first normalized using the $\min - \max$ method, which is computed as follows.

$$N(a(x_j)) = \frac{a(x_j) - \min_{a(x_j)}}{\max_{a(x_j)} - \min_{a(x_j)}}, \tag{12}$$

where $\max_{a(x_j)}$ and $\min_{a(x_j)}$ are the maximum and minimum ISVs of the attribute a on U, respectively.

The next step is to perform fuzzy modeling on the normalized data by Gaussian membership degree function, which allows mapping from fuzzy sets to ISet. Based on this, the ISet of all ISVs under each information source can be calculated. Thereby the Gaussian membership degree functions act as the role of information gain or agent. By Definitions 7, the TUM of ISVs of each information source can be quantified. Finally, the minimum TUM of all information sources about attribute a is selected as the attribute of the new information system by Eq. (10). According to the aforementioned descriptions, we give the fusion algorithm of MsNIS as follows.

Example 2 (*Continued from Examples* 1)**.** According to Eqs. (11) and (12), multi-source numerical data of Table 2 are normalized and fuzzed as shown in Table 3. Then, by Definition 5, all information values of the MsNIS are calculated as shown in Table 4. The uncertainty measure for normalized ISVs are calculated as follows. It is worth noting that the 0 values have a very small $\epsilon = 0.0001$ in order to avoid problems with the logarithmic function. For the attribute a_1, by Definition 6, then

$SST^1_{a_1} = -(0.0756 * \log_2(0.0615) + 0.0489 * \log_2(0.0381) + 1 * \log_2(0.3208) + 0.9156 * \log_2(0.3875) + 0.0711 * \log_2(0.0575) + 0 * \log_2(0.0001)) \approx 3.7200$;

$SST^2_{a_1} = -(1 * \log_2(0.1721) + 0.4615 * \log_2(0.4353) + 0.0324 * \log_2(0.0220) + 0 * \log_2(0.0001) + 0.4656 * \log_2(0.4374) + 0.0891 * \log_2(0.0688)) \approx 4.1707$;

$SST^3_{a_1} = -(0.0036 * \log_2(0.0032) + 0 * \log_2(0.0001) + 0.0427 * \log_2(0.0392) + 0.0249 * \log_2(0.0227) + 0.0249 * \log_2(0.0227) + 1 * \log_2(0.0824)) \approx 4.1017$;

$SST^4_{a_1} = -(0.0079 * \log_2(0.0062) + 0.0053 * \log_2(0.0041) + 0 * \log_2(0.0001) + 1 * \log_2(0.2177) + 0.7045 * \log_2(0.4193) + 0.0264 * \log_2(0.0214)) \approx 3.3289$.

By Definition 7, then $SIST^1_{a_1} \approx 1.2213$; $SIST^2_{a_1} \approx 1.0494$; $SIST^3_{a_1} \approx 0.1282$; $SIST^4_{a_1} \approx 0.9076$. Hence, according to Definition 8, we have $TUM^1_{a_1} = 3.7200 + 1.2213 = 4.9413$; $TUM^2_{a_1} = 4.1707 + 1.0494 = 5.2201$; $TUM^3_{a_1} = 4.1017 + 0.1282 = 4.2299$; $TUM^4_{a_1} = 3.3289 + 0.9076 = 4.2365$.

Finally, on the basis of Definition 9, $\text{Inf}(4.9413, 5.2201, 4.2299, 4.2365) = 4.2299 \xrightarrow{select} V^{NIS_3}_{a_1}$. Therefore, attribute a_1 of the third NIS_3 can be selected as the attribute a_1 of a new information system. In the same way, the TUMs of remaining three attributes (a_2, a_3 and a_4) in all information sources are calculated as follows.

$\text{Inf}(6.1920, 5.7074, 5.7516, 5.9823) = 5.7074 \xrightarrow{select} V^{NIS_2}_{a_2}$; $\text{Inf}(5.5840, 5.4493, 11.9492, 5.1912) = 5.1912 \xrightarrow{select} V^{NIS_4}_{a_3}$, and $\text{Inf}(5.1843, 5.7465, 5.1912, 5.1213) = 5.1213 \xrightarrow{select} V^{NIS_4}_{a_4}$.

Let (U, A', V) be a new information system, where $A' = (V^{NIS_3}_{a_1}, V^{NIS_2}_{a_2}, V^{NIS_4}_{a_3}, V^{NIS_4}_{a_4}) = \{a'_1, a'_2, a'_3, a'_4\}$. It is shown in Table 5. According to Table 5, it is clear that the attributes in the new information system integrate the advantages of the attributes in NIS_2, NIS_3, and NIS_4. This also fully reflects the diversity and complementary characteristics of MSIF. As a result, the data in the new fused information system is more comprehensive and unified, facilitating further processing of the data and improving the efficiency and accuracy of fusion.

Algorithm 1: Multi-source numerical data fusion (MsNDF)

Input: A MsNIS $MsNIS = \{NIS_i | NIS_i = (U, A, V_i), i = 1, 2, \cdots, m\}$, where $|U| = n, |A| = l$.
Output: A new single-source information system (U, A', V).

```
for i ← 1 to m and j ← 1 to n do
    for ∀ a ∈ A do
        Normalize the original data sources by Eq. (12);
        Compute the Gaussian membership degrees by Eq. (11);
         // Gaussian membership degree acts as the
         information gain g_a(x_j).
        Compute information sets by Eq. (6);
        Compute SST by Eq. (7);
        Compute SIST by Eq. (8);
        Compute TUM by Eq. (9).
    end
    for k ← 1 to l and ∀ a_k ∈ A do
        Compute the TUM of the k-th attribute of all information
         sources by Eq. (10);
        A' ← ∅, A' ← A' ∪ {a_k};   // Select the attribute with
         minimal TUM as the k-th attribute of the new
         fused information system.
    end
    Return A'.
end
Obtain a new single-source information system (U, A', V).
```

Table 3
The upper table and lower table are normalized IVSs and Gaussian membership degrees of the MsNIS, respectively.

U	NIS_1				NIS_2				NIS_3				NIS_4			
	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
x_1	0.0756	0.3537	0.5202	0.0000	1.0000	1.0000	0.0650	0.2500	0.0036	0.7500	0.8158	0.1429	0.0079	0.1739	1.0000	0.0000
x_2	0.0489	1.0000	0.1667	1.0000	0.4615	0.0000	0.2150	0.4167	0.0000	0.2500	0.9289	0.1429	0.0053	0.3043	0.3916	0.0000
x_3	1.0000	0.3049	0.0000	0.6471	0.0324	0.3143	0.0000	0.0833	0.0427	0.4583	0.8026	0.1429	0.0000	0.5217	0.7203	0.2203
x_4	0.9156	0.5610	0.2424	0.0294	0.0000	0.3143	0.2100	0.0000	0.0249	0.0833	0.0000	0.0476	1.0000	0.0000	0.4476	0.3390
x_5	0.0711	0.0000	1.0000	0.3137	0.4656	0.0857	1.0000	1.0000	0.0249	0.0000	0.9895	0.0000	0.7045	1.0000	0.7552	1.0000
x_6	0.0000	0.3902	0.1162	0.1078	0.0891	0.4286	0.1900	0.8333	1.0000	1.0000	1.0000	1.0000	0.0264	0.4130	0.0000	0.1695
U	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
x_1	0.8134	0.9646	0.8666	0.6308	0.1721	0.1368	0.8106	0.8882	0.8870	0.6586	0.9853	0.9554	0.7849	0.7690	0.3744	0.6985
x_2	0.7801	0.1752	0.8735	0.2031	0.9431	0.5412	0.9810	0.9993	0.8828	0.8886	0.8832	0.9554	0.7813	0.9529	0.8808	0.6985
x_3	0.3208	0.9118	0.5958	0.7165	0.6787	0.9912	0.7003	0.6451	0.9294	0.9953	0.9910	0.9554	0.7741	0.9305	0.8710	0.9803
x_4	0.4232	0.9170	0.9577	0.6794	0.6231	0.9912	0.9780	0.5097	0.9112	0.6352	0.0927	0.8448	0.2177	0.4425	0.9475	0.9889
x_5	0.8079	0.3562	0.1444	0.9951	0.9394	0.7014	0.0948	0.3076	0.9112	0.4949	0.7974	0.7716	0.5952	0.1651	0.8173	0.1119
x_6	0.7153	0.9892	0.7985	0.8022	0.7723	0.9757	0.9639	0.5545	0.0824	0.2719	0.7809	0.0876	0.8093	0.9994	0.2238	0.9410

Table 4
All information values of the MsNIS.

U	NIS_1				NIS_2				NIS_3				NIS_4			
	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
x_1	0.0615	0.3411	0.4508	0.0000	0.1721	0.1368	0.0527	0.2221	0.0032	0.4940	0.8038	0.1365	0.0062	0.1337	0.3744	0.0000
x_2	0.0381	0.1752	0.1456	0.2031	0.4353	0.0000	0.2109	0.4164	0.0000	0.2221	0.8204	0.1365	0.0041	0.2900	0.3449	0.0000
x_3	0.3208	0.2780	0.0000	0.4636	0.0220	0.3115	0.0000	0.0538	0.0397	0.4562	0.7954	0.1365	0.0000	0.4855	0.6273	0.2160
x_4	0.3875	0.5144	0.2322	0.0200	0.0000	0.3115	0.2054	0.0000	0.0227	0.0529	0.0000	0.0402	0.2177	0.0000	0.4240	0.3352
x_5	0.0575	0.0000	0.1444	0.3122	0.4374	0.0601	0.0948	0.3076	0.0227	0.0000	0.7890	0.0000	0.4193	0.1651	0.6173	0.1119
x_6	0.0000	0.3860	0.0928	0.0865	0.0688	0.4182	0.1831	0.4620	0.0824	0.2719	0.7809	0.0876	0.0214	0.4128	0.0000	0.1595

Table 5
The new fused single-source information system (U, A', V).

U	a'_1	a'_2	a'_3	a'_4
x_1	2742	237	43.3	0.9
x_2	2609	225	34.6	0.9
x_3	2503	230	39.3	2.2
x_4	2495	221	35.4	2.9
x_5	2610	219	39.8	6.8
x_6	2517	243	29	1.9

4. Gaussian kernel based-fuzzy knowledge measure for outlier detection in the new single-source information system

According to Moser's work, any kernel satisfying reflexivity and symmetry is at least T_{cos}-transitive [32]. Let U be the universe. A real-valued function $K : U \times U \to \mathbf{R}$ is called a kernel if it is symmetric and positive-semidefinite, i.e., $\forall\ x, y \in U, K(x, y) = K(y, x)$. Then, for any kernel $K : U \times U \to [0, 1]$ with $K(x, x) = 1$ is at least T_{cos}-transitive, where $T_{cos}(x, y) = \max(ab - \sqrt{1-a^2}\sqrt{1-b^2}, 0)$ [45]. If some kernel functions satisfy reflexive, symmetric and T_{cos}-transitive, then the fuzzy T_{cos}-equivalence relations are then calculated with these kernel functions [33]. In this paper, the Gaussian kernel is used to obtain fuzzy T_{cos}-equivalence relation in the new fused information system.

Definition 10. Let (U, A', V) be a new fused information system and $B \subseteq A'$. For any $x_i, x_j \in U$ and $a \in B$, the Gaussian kernel is used to extract the fuzzy T_{cos}-equivalence relation between x_i and x_j w.r.t the attribute subset B, which can be denoted as

$$K_G^B(x_i, x_j) = e^{-\frac{\|x_i - x_j\|_B^2}{2\sigma}}, \tag{13}$$

where $\|x_i - x_j\|_B^2$ is Euclidean distance between x_i and x_j w.r.t attribute a. $\sigma \in (0, 1]$ is a tolerance parameter, which can adjust the granularity of the fuzzy approximation space by Gaussian kernel function.

Given the fused information system (U, A', V) where $U = \{x_1, x_2, \dots, x_n\}$ and $A' = \{a'_1, a'_2, \dots, a'_m\}$. For any $B \subseteq A'$, then a fuzzy relation matrix $M(K_G^B) = (K_G^B(x_i, x_j))_{n\times n}$ can be induced by Gaussian kernel T_{cos}−equivalence relation. In fuzzy relation matrix $M(K_G^B) = (K_G^B(x_i, x_j))_{n\times n}$, each row $((K_G^B(x_i, x_1), K_G^B(x_i, x_2), \dots, K_G^B(x_i, x_n))$ denotes a fuzzy set. Namely, for any $x_i \in U$, the fuzzy set induced by B can be denoted as

$$\begin{aligned}[x_i]_B &= \frac{K_G^B(x_i, x_1)}{x_1} + \frac{K_G^B(x_i, x_2)}{x_2} + \cdots + \frac{K_G^B(x_i, x_n)}{x_n} \\ &= (K_G^B(x_i, x_1), K_G^B(x_i, x_2), \dots, K_G^B(x_i, x_n)).\end{aligned} \tag{14}$$

From the perspective of granular computing, $[x_i]_B$ can also be identified as the fuzzy knowledge (information) granule containing x_i. A collection of all knowledge granules is called the knowledge granular structure (KGS) [46]. Undoubtedly, $[x_i]_B(x_j) = K_G^B(x_i, x_j)$, $|[x_i]_B| = \sum_j^n K_G^B(x_i, x_j)$, and $0 \le |[x_i]_B| \le n$. $K_G^B(x_i, x_j) = 1$, it means that x_j belongs entirely to $[x_i]_B$; $K_G^B(x_i, x_j) = 0$, it means that x does not belong to $[x_i]_B$ at all. In this paper, the degree of membership $K_G^B(x_i, x_j)$ is determined by conjunction method, which is often used in many literature [33,42,47]. Namely, for any $B = \{a'_1, a'_2, \dots, a'_h\} \subseteq A'$ $(h \le m)$, then $K_G^B(x_i, x_j) = \bigwedge_{k=1}^h K_G^{a_k}(x_i, x_j)$.

4.1. Fuzzy knowledge measure

The knowledge granules mentioned in the previous section are obtained from fuzzy sets induced by fuzzy T_{cos}-equivalence relations. In fuzzy sets, the fuzzy entropy provides the average amount of ambiguity. Similarly, the average amount of knowledge present in a knowledge granule can be considered. This kind of knowledge measure is called dual measure of fuzzy entropy. Until now, some representative fuzzy measures have been proposed, as shown in Table 6.

In [36], several examples verify that $H_K(F)$, $H_{\mathrm{PP}}(F)$, $H_{LL}(F)$, $H_{HY}(F)$, and $H_{\alpha'}^{\beta'}(F)$ measures (see Table 6) are not applicable in some cases. In addition, the formulas for $H_{\mathrm{PP}}(F)$, $H_{HY}(F)$, and $H_{\alpha'}^{\beta'}(F)$ are relatively complex, and the efficiency of the calculation is greatly reduced. In view of this, Arya and Kumar proposed an improved fuzzy knowledge measure for fuzzy sets, i.e., $H_{AK}(F)$, which complies with the extended idea of De Luca and Termini axioms. Although many fuzzy measures have been proposed and traditionally used for multiple attribute decision-making or evaluating the uncertainty of fuzzy sets,

Table 6
Some representative fuzzy measures. Let F be a fuzzy set, $F = \{(x, \mu_F(x))|x \in U\}$ where $U = \{x_1, x_2, \ldots, x_n\}$.

References	Measures
Ygaer [48]	$H_Y(F) = 1 - \frac{d_p(F,F^c)}{n^{\frac{1}{p}}}, F^c = 1 - \mu_F(x)$
Kosko [49]	$H_K(F) = \frac{d_p(F,F_{\text{near}})}{d_p(F,F_{\text{far}})}$
Pal and Pal [50]	$H_{\text{PP}}(F) = \frac{1}{n}\sum_{i=1}^{n}\left[\mu_F(x_i)\, e^{1-\mu_F(x_i)} + (1-\mu_F(x_i))\, e^{\mu_F(x_i)}\right]$
Li and Liu [51]	$H_{LL}(F) = \sum_{i=1}^{n} S\left(cr\left(\xi_F = x_i\right)\right)$
HWang and Yang [52]	$H_{HY}(F) = \frac{1}{1-e^{-\frac{1}{2}}}\sum_{i=1}^{n}\left[\left(1-e^{-\mu_{F^c}(x_i)}\right) I_{\left[\mu_F(x_i)\geq\frac{1}{2}\right]} + \left(1-e^{-\mu_F(x_i)}\right) I_{\left[\mu_F(x_i)<\frac{1}{2}\right]}\right]$
Joshi and Kumar [53]	$H^{\beta'}_{\alpha'}(F) = \frac{\alpha'\times\beta'}{n(\alpha'-\beta')}\left[\sum_{i=1}^{n}\left\{\left(\mu_F(x_i)^{\beta'} + (1-\mu_F(x_i))^{\beta'}\right)^{\frac{1}{\beta'}} - \left(\mu_F(x_i)^{\alpha'} + (1-\mu_F(x_i))^{\alpha'}\right)^{\frac{1}{\alpha'}}\right\}\right]$
Arya and Kumar [36]	$H_{AK}(F) = \log_2\left[\frac{2}{n}\sum_{i=1}^{n}\left(\mu_F^2(x_i) + (1-\mu_F(x_i))^2\right]\right]$

few studies have used them for outlier detection. Driven by this motivation, we improve the fuzzy knowledge measure $H_{AK}(F)$ for capturing outliers in the knowledge granule as follows.

Definition 11. Let (U, A', V) be the fused information system and $|U| = n$. For any $B \subseteq A'$ and $x_o \in U$, the fuzzy knowledge measure (FKM) of the knowledge granule $[x_o]_B$ w.r.t the attribute subset B can be defined as

$$FKM(B) = -\log_2 \frac{2}{n}\left[\sum_{o=1}^{n}\left(\frac{1}{n^2}\left|[x_o]_B\right|^2 + \frac{1}{n^2}\left|[x_o]_B^c\right|^2\right)\right], \tag{15}$$

where $[x_o]_B^c(x) = 1 - [x_o]_B(x)$.

Proposition 1. *Let (U, A', V) be the fused information system and $|U| = n$. For any $B \subseteq A'$. If $|[x_o]_B| \in [\frac{n}{2}, n]$, then $FKM(B)$ is an increasing function with respect to attribute subset B; If $|[x_o]_B| \in [0, \frac{n}{2})$, then $FKM(B)$ is a decreasing function with respect to attribute subset B.*

Proof. To prove the monotonicity of $FKM(B)$, it only needs to prove the monotonicity of $\frac{2|[x_o]_B|^2+n^2-2n|[x_o]_B|}{n^2}$. Let $|[x_o]_B| = \mathcal{X}$, then we have $f(\mathcal{X}) = \frac{2}{n^2}\mathcal{X}^2 - \frac{2}{n}\mathcal{X} + 1$. $f'(x) = \frac{4}{n^2}\mathcal{X} - \frac{2}{n} = 0$. Hence, $\mathcal{X} = \frac{n}{2}$. When $\mathcal{X} \geq \frac{n}{2}$, then $f(\mathcal{X})$ is an increasing function with respect to $\mathcal{X}$. When $\mathcal{X} < \frac{n}{2}$, then $f(\mathcal{X})$ is a decreasing function with respect to $\mathcal{X}$. It is obvious $f(\mathcal{X})$ is an increasing function of $\mathcal{X}$ in $[\frac{n}{2}, n]$ and a decreasing function of $\mathcal{X}$ in $[0, \frac{n}{2})$. Since $B_1 \subseteq B_2$, we have $|[x]_{B_1}| \geq |[x]_{B_2}|$. Hence, $\mathcal{X}$ is a decreasing function with respect to attribute subset B. Thereby $f(\mathcal{X})$ can be regarded as a compound function with respect to B. When $\mathcal{X} \in [\frac{n}{2}, n]$, the compound function $f(\mathcal{X})$ is a decreasing function with respect to B, however, when $\mathcal{X} \in [0, \frac{n}{2})$, it is an increasing function. Therefore, $FKM(B)$ is an increasing function with respect to attribute subset B when $|[x_o]_B| \in [\frac{n}{2}, n]$, and $FKM(B)$ is a decreasing function with respect to attribute subset B when $|[x_o]_B| \in [0, \frac{n}{2})$.

Proposition 2. *Let (U, A', V) be the fused information system and $|U| = n$. For any $B \subseteq A'$. $FKM(B)$ has a minimum value is $-\log_2 \frac{1}{n}$.*

Proof. According to Proposition 1, $FKM(B)$ has the minimum value when $|[x_o]_B| = \frac{n}{2}$. Then, we have $FKM(B) = -\log_2 \frac{2}{n}\left[\left(\frac{1}{n^2}(\frac{n}{2})^2 + \frac{1}{n^2}(n-\frac{n}{2})^2\right)\right] = -\log_2 \frac{1}{n}$.

By Propositions 1 and 2, the proposed FKM has minimum value, the size of which depends on the number of objects in the knowledge granule. Each knowledge granule has its inherent knowledge amount (or information entropy). Moreover, there will always be unknown factors that cause information redundancy before we fully understand it [54]. From the perspective of information theory, reducing information redundancy can be achieved by minimizing the information entropy because of the more significant the information entropy, the greater the uncertainty. Therefore, as long as we approach the minimum value of FKM as much as possible, the information entropy can be reduced, thereby increasing the amount of knowledge. Fig. 4 illustrates the role of FKM in outlier detection. From Fig. 4, suppose that (U, A', V) is the fused information system where $U = \{x_1, x_2, \ldots, x_n\}$ and $A' = \{a'_1, a'_2, \ldots, a'_m\}$. Given a subset $B = \{a'_1, a'_2, \ldots, a'_h\} \subseteq A'$, $\forall x_i \in U$, then a knowledge granule (KG_i, i.e., $[x_i]_B$) can be obtained (see Eq. (14)). All knowledge granules can form a knowledge granular structure (KGS) with respect to U, that is, $KGS = \{KG_1, KG_2, \ldots, KG_n\} = \{[x_1]_B, [x_2]_B, \ldots, [x_n]_B\}$. Moreover, using FKM and its extended measures to evaluate the objects of each knowledge granule in the KGS, and then ranking the outlier factors corresponding to all the objects, all outliers in the universe of discourse U can be found. On the basis of this, we designed the schemes for mining outliers in knowledge granules based on the proposed FKM as follows.

4.2. Schemes of outlier detection

In this subsection, the fuzzy knowledge relative measure is defined for establishing outlier detection schemes in the fused information system.

Definition 12. Let (U, A', V) be the fused information system and $|U| = n$. For any $B \subseteq A'$ and $x_i \in U$, if $x_{o,o\neq i}$ is the deleted object in U, then the relative fuzzy knowledge measure of x_o on attribute set B is defined by

$$RFKM_B(x_o) = \begin{cases} 1 - \frac{FKM_{x_o}(B)}{FKM(B)}, & FKM_{x_o}(B) < FKM(B); \\ 0, & \text{otherwise}, \end{cases} \tag{16}$$

where $FKM_{x_o}(B) = -\log_2 \frac{2}{n-1}\left[\sum_{i=1}^{n-1}\left(\frac{1}{(n-1)^2}\left|[x_o]_B\right|^2 + \frac{1}{(n-1)^2}\left|[x_o]_B^c\right|^2\right)\right]$ refers to the fuzzy knowledge measure after deleting x_o. Moreover, $\{U - \{x_o\}\}/B = \{[x_{i_1}]_B, [x_{i_2}]_B, \ldots, [x_{i_{n-1}}]_B\}$.

From Definition 12, the larger the value $RFKM_B(x_o)$ of x_o and the greater the uncertainty of x_o, the more likely it is to be an outlier.

Definition 13. Let (U, A', V) be the fused information system and $|U| = n$. For any $B \subseteq A'$ and $x_o \in U$, the relative fuzzy knowledge cardinality of x_o w.r.t. B can be defined as

$$RFKC_B(x_o) = |[x_o]_B| - \frac{\sum_{r=1}^{n-1}|[x_{i_r}]_B|}{n-1}, x_{i_r} \in U, i \neq o. \tag{17}$$

Definition 13 shows whether x_o belongs to majority classes, i.e., $RFKC_B(x_o) \geq 0$, then x_o can be considered to be in the majority classes; $RFKC_B(x_o) < 0$, then x_o can be considered to be in minority classes. According to [42], the outlier degree based on fuzzy relation of each object is defined as follows.

Definition 14. Let (U, A', V) be the fused information system and $|U| = n$. For any $B \subseteq A'$ and $x_o \in U$, the fuzzy relation-based outlier degree of x_o can be denoted by

$$FROD_B(x_o) = \begin{cases} RFKM_B(x_o) \times \left(\frac{n - RFKC_B(x_o)}{2n}\right), & RFKC_B(x_o) \geq 0 \\ RFKM_B(x_o) \times \sqrt{\frac{n + abs(RFKC_B(x_o))}{2n}}, & RFKC_B(x_o) < 0 \end{cases} \tag{18}$$

where $abs(RFKC_B(x_o))$ refers to the absolute value of $RFKC_B(x_o)$.

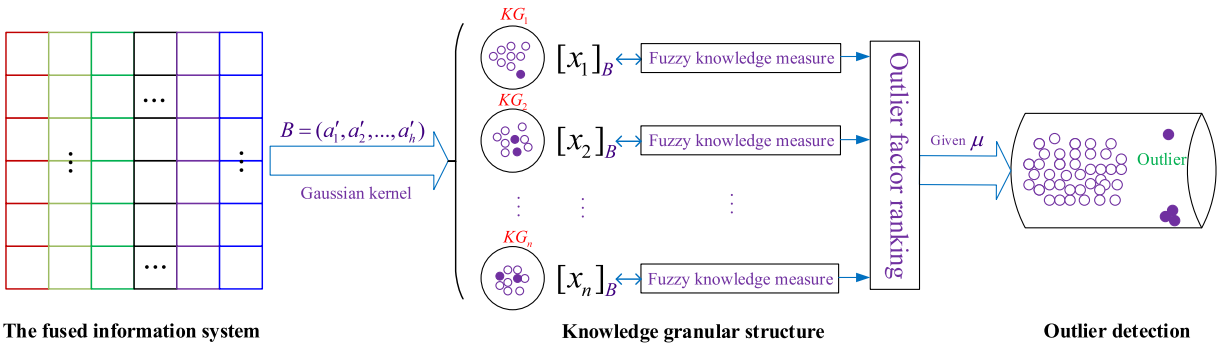


Fig. 4. Fuzzy knowledge measures evaluate knowledge granules to identify outliers.

From Definition 14, $FROD_B(x_o)$ can be used to evaluate the degree of outliers of object x_o. The value of $FROD_B(x_o)$ is determined by the attribute subset B, which has generality. There are $2^{|A'|}$ attribute subsets in A'. As a result, calculating the outlier degrees of fuzzy relations for $2^{|A'|}$ attribute subsets is an open problem. In what follows, the attribute sequence and attribute subset sequence schemes are used to overcome this problem.

Definition 15. Let (U, A', V) be the fused information system and $A' = \{a_1, a_2, \dots, a_l\}$.

(1) If $FKM(a'_h) \geq FKM(a'_{h+1})$ for any $h \in \{1, 2, \dots, h, \dots, l-1\}$, then the attribute sequence (AS) is denoted as $AS =< a'_1, a'_2, \dots, a'_l >$.

(2) If $B_1 = \{a'_1\}$, $B_2 = B_1 \cup \{a'_2\}$, and $B_{h+1} = B_h \cup \{a'_{h+1}\}$ for any $h \in \{1, 2, \dots, h, \dots, l-1\}$, then the attribute subset sequence (ASS) is denoted as $ASS =< B_1, B_2, \dots, B_l >$.

According to Definition 15, starting with feature set $\{a'_1\}$, the feature with the highest fuzzy knowledge measure is added one by one until the set containing the whole feature is acquired. Following the preceding descriptions and definitions, we arrive at the definition of the outlier factor of fuzzy knowledge measure as follows.

Definition 16. Let (U, A', V) be the fused information system and $|U| = n$. For any $x_o \in U$, the fuzzy knowledge measure-based outlier factor $FKMOF(x_o)$ of x_o can be defined as

$$FKMOF(x_o) = 1 - \frac{\sum_{h=1}^{l}\left(1 - FROD_{a'_h}(x_o)\right)\omega_{a'_h}(x_o) + \sum_{h=1}^{l}\left(1 - FROD_{B_h}(x_o)\right)\omega_{B_h}(x_o)}{2n}, \tag{19}$$

where $\omega_B(x_o) = \sqrt{\frac{|[x_o]_B|}{n}}$ is a weight function for any $B \subseteq A'$.

Definition 16 shows the $FKMOF(x_o)$) is inversely proportional to its weight function. Namely, the smaller $\omega_B(x_o)$ of x_o is, the more likely x_o is an outlier.

Definition 17. Let (U, A', V) be the fused information system. Given a parameter μ. For any $x_o \in U$, if $FKMOF(x_o) > \mu$, then x_o is regarded to as an outlier based on fuzzy knowledge measure in U.

4.3. Fuzzy knowledge measure-based an algorithm for outlier detection

This subsection proposes an outlier detection algorithm based on a fuzzy knowledge measure (FKMOD) as follows.

According to Algorithm 2, Steps 2–4 are to compute the fuzzy relation matrix of each attribute, whose time complexity is $O(n \times n)$. Then, for all attributes, the time complexity is $O(l \times n \times n)$. Steps 7–9 focus on computing the fuzzy knowledge measure of attribute subset B_h, whose time complexity is $O(l)$. The time complexity of Steps 11–22 are $O(n \times l)$. Based on this, the total time complexity of Algorithm 2 is $O(l \times n \times n + l + n \times l)$. As a result, the time complexity of Algorithm 2 is $O(ln^2)$ in the worst case.

Algorithm 2: FKMOD

Input: A new fused single-source information system (U, A', V) on the basis of $MsNIS$, parameters $\sigma, \gamma \in [0, 1]$, where $A' = \{a'_1, a'_2, \cdots, a'_l\}$ and $|A'| = l$
Output: Outlier Set (OS)

1 $OS \leftarrow \emptyset$;
2 **for** $h \leftarrow 1$ *to* l *and* $\sigma \in ([0, 1]$ **do**
3 By Definition 10, compute $M(K_G^{a'_h})$;
4 By Definition 11, compute $FKM(a'_h)$;
5 **end**
6 **for** $h \leftarrow 1$ *to* l **do**
7 By Definition 15, obtain $AS =< a'_1, a'_2, \cdots, a'_l >$, and construct $ASS =< B_1, B_2, \cdots, B_l >$;
8 Compute $FKM(B_h)$;
9 **end**
10 **for** $o = 1$ *to* n **do**
11 **for** $h \leftarrow 1$ *to* l **do**
12 By Definition 12, compute $RFKM_{\{a'_h\}}(x_o)$ and $RFKM_{B_h}(x_o)$;
13 By Definition 13, compute $RFKC_{\{a'_h\}}(x_o)$ and $RFKC_{B_h}(x_o)$;
14 By Definition 14, compute $FROD_{\{a'_h\}}(x_o)$, $FROD_{B_h}(x_o)$, $\omega_{\{a'_h\}}(x_o)$ and $\omega_{B_h}(x_o)$;
15 **end**
16 By Definition 16, compute $FKMOF(x_o)$.
17 **end**
18 **if** $FKMOF(x_o) > \mu$ **then**
19 Outlier $(O) \leftarrow O = O \cup \{x_o\}$;
20 **end**
21 Return OS.

Example 3 (*Continued from Example 2*)**.** In Table 5, it is a new single-source information system after fusion. According to Algorithm 2, we perform the calculation by the following steps.

(1) We first compute the Gaussian kernel similarity relation matrix of each attribute $(a_{h,h=1,2,3,4})$ as follows.

$$M(K_G^{a'_1}) = \begin{pmatrix} 1.0000 & 1.0000 & 0.9999 & 0.3737 & 0.6153 & 0.9997 \\ 1.0000 & 1.0000 & 1.0000 & 0.3718 & 0.6131 & 0.9996 \\ 0.9999 & 1.0000 & 1.0000 & 0.3679 & 0.6085 & 0.9993 \\ 0.3737 & 0.3718 & 0.3679 & 1.0000 & 0.9165 & 0.3874 \\ 0.6153 & 0.6131 & 0.6085 & 0.9165 & 1.0000 & 0.6310 \\ 0.9997 & 0.9996 & 0.9993 & 0.3874 & 0.6310 & 1.0000 \end{pmatrix};$$

Table 7
The calculation results of $FKM_{x_o}(\{a'_h\})$, $RFKM_{\{a'_h\}}(x_o)$ and $RFKC_{\{a'_h\}}(x_o)$.

U	$FKM_{x_o}(\{a'_h\})$				$RFKM_{\{a'_h\}}(x_o)$				$RFKC_{\{a'_h\}}(x_o)$			
	$\{a'_1\}$	$\{a'_2\}$	$\{a'_3\}$	$\{a'_4\}$	$\{a'_1\}$	$\{a'_2\}$	$\{a'_3\}$	$\{a'_4\}$	$\{a'_1\}$	$\{a'_2\}$	$\{a'_3\}$	$\{a'_4\}$
x_1	5.0516	5.8501	6.2644	5.8539	0.2027	0.1846	0.1121	0.1812	1.2305	1.1128	0.2750	1.0368
x_2	5.0541	5.7090	5.7330	5.8629	0.2023	0.2042	0.1874	0.1799	1.2248	1.3447	1.2042	1.0212
x_3	5.0596	5.8091	5.6747	5.6668	0.2014	0.1903	0.1957	0.2073	1.2123	1.2516	1.2510	1.3983
x_4	6.3526	6.2140	5.6582	5.7469	0.0000	0.1338	0.1980	0.1961	−0.9691	0.5112	1.3062	1.3345
x_5	5.7445	7.0220	5.7335	7.3963	0.0933	0.0212	0.1873	0.0000	0.3848	−0.7032	1.1594	−1.3614
x_6	5.0343	5.7071	6.6748	5.6680	0.2054	0.2045	0.0539	0.2072	1.2703	1.3717	−0.3500	1.3832

Table 8
The calculation results of $RFKM_{B_h}(x_o)$, $RFKC_{B_h}(x_o)$, $FROD_{a'_h}(x_o)$ and $FROD_{B_h}(x_o)$.

U	$RFKM_{B_h}(x_o)$				$RFKC_{B_h}(x_o)$				$FROD_{a'_h}(x_o)$				$FROD_{B_h}(x_o)$			
	B_1	B_2	B_3	B_4	B_1	B_2	B_3	B_4	$\{a'_1\}$	$\{a'_2\}$	$\{a'_3\}$	$\{a'_4\}$	B_1	B_2	B_3	B_4
x_1	1.0000	1.0000	1.0000	1.0000	1.1128	1.1925	0.3978	0.5701	0.0806	0.0752	0.0535	0.0749	0.4073	0.4006	0.4669	0.4525
x_2	1.0000	1.0000	1.0000	1.0000	1.3447	1.2999	1.3093	1.2710	0.0805	0.0792	0.0749	0.0746	0.3879	0.3917	0.3909	0.3941
x_3	1.0000	1.0000	1.0000	1.0000	1.2516	1.2828	1.3080	1.4361	0.0804	0.0753	0.0774	0.0795	0.3957	0.3931	0.3910	0.3803
x_4	1.0000	1.0000	1.0000	1.0000	0.5112	0.7390	1.1257	−0.5781	0.0000	0.0612	0.0775	0.0763	0.4574	0.4384	0.4062	0.7404
x_5	1.0000	1.0000	1.0000	1.0000	−0.7032	−1.4588	−0.8248	−0.1433	0.0437	0.0159	0.0756	0.0000	0.7474	0.7884	0.7541	0.7155
x_6	1.0000	1.0000	1.0000	1.0000	1.3717	1.4543	0.5599	0.6385	0.0810	0.0789	0.0392	0.0797	0.3857	0.3788	0.4533	0.4468

$$M(K_G^{a'_2}) = \begin{pmatrix} 1.0000 & 0.9829 & 0.8859 & 0.9709 & 0.5037 & 0.9434 \\ 0.9829 & 1.0000 & 0.9541 & 0.9122 & 0.6153 & 0.9879 \\ 0.8859 & 0.9541 & 1.0000 & 0.7632 & 0.7941 & 0.9887 \\ 0.9709 & 0.9122 & 0.7632 & 1.0000 & 0.3679 & 0.8430 \\ 0.5037 & 0.6153 & 0.7941 & 0.3679 & 1.0000 & 0.7087 \\ 0.9434 & 0.9879 & 0.9887 & 0.8430 & 0.7087 & 1.0000 \end{pmatrix};$$

$$M(K_G^{a'_3}) = \begin{pmatrix} 1.0000 & 0.6897 & 0.9206 & 0.7374 & 0.9422 & 0.3679 \\ 0.6897 & 1.0000 & 0.9016 & 0.9967 & 0.8749 & 0.8586 \\ 0.9206 & 0.9016 & 1.0000 & 0.9325 & 0.9981 & 0.6020 \\ 0.7374 & 0.9967 & 0.9325 & 1.0000 & 0.9095 & 0.8181 \\ 0.9422 & 0.8749 & 0.9981 & 0.9095 & 1.0000 & 0.5647 \\ 0.3679 & 0.8586 & 0.6020 & 0.8181 & 0.5647 & 1.0000 \end{pmatrix};$$

$$M(K_G^{a'_4}) = \begin{pmatrix} 1.0000 & 1.0000 & 0.9520 & 0.8892 & 0.3717 & 0.9652 \\ 1.0000 & 1.0000 & 0.9498 & 0.8860 & 0.3679 & 0.9632 \\ 0.9520 & 0.9498 & 1.0000 & 0.9855 & 0.5501 & 0.9989 \\ 0.8892 & 0.8860 & 0.9855 & 1.0000 & 0.6536 & 0.9764 \\ 0.3717 & 0.3679 & 0.5501 & 0.6536 & 1.0000 & 0.5218 \\ 0.9652 & 0.9632 & 0.9989 & 0.9764 & 0.5218 & 1.0000 \end{pmatrix}.$$

(2) According to Definition 11, the fuzzy knowledge measure of single attribute can be computed, i.e., $FKM(a'_1) \approx 6.3359$, $FKM(a'_2) \approx 7.1742$, $FKM(a'_3) \approx 7.0551$ and $FKM(a'_4) = 7.1491$. Thus, we have $AS =< a'_2, a'_4, a'_3, a'_1 >$; $ASS =< \{a'_2\}, \{a'_2, a'_4\}, \{a'_2, a'_4, a'_3\}, \{a'_2, a'_4, a'_3, a'_1\} >$. Let $ASS =< \{a'_2\}, \{a'_2, a'_4\}, \{a'_2, a'_4, a'_3\}, \{a'_2, a'_4, a'_3, a'_1\} > =< B_1, B_2, B_3, B_4 >$, where $B_1 = \{a'_2\}, B_2 = \{a'_2, a'_4\}, B_3 = \{a'_2, a'_4, a'_3\}$, $B_4 = \{a'_2, a'_4, a'_3, a'_1\}$.

(3) According to Definitions 12 and 13, the fuzzy knowledge granularity after deleting an object can be computed. For instance, since $FKM_{x_1}(a'_1) \approx 5.0516$, then $FKM_{x_1}(a'_1) < FKM(a'_1)$, thus $RFKM_{a_1}(x_1) = 1 - \frac{5.0516}{6.3359} \approx 0.2027$. In addition, we can obtain $RFKC_{a'_1}(x_1) \approx 1.2305$. All the calculation results of $FKM_{x_o}(\{a'_h\})$, $RFKM_{\{a'_h\}}(x_o)$ and $RFKC_{\{a'_h\}}(x_o)$ $(h = 1, 2, 3, 4; o = 1, 2, \ldots, 6)$ w.r.t four single attribute sets are shown in Table 7. Similarly, the calculation results of $RFKM_{B_h}(x_o)$, $RFKC_{B_h}(x_o)$, $FROD_{a'_h}(x_o)$ and $FROD_{B_h}(x_o)$ $(h = 1, 2, 3, 4; o = 1, 2, \ldots, 6)$ are shown in Table 8.

(4) According to Definition 16, the fuzzy knowledge measure-based outlier factor of $x_{o,o=1,2,\ldots,6}$ can be calculated as follows.

$FKMOF(x_1) = 0.3281$; $FKMOF(x_2) = 0.2952$; $FKMOF(x_3) = 0.2925$; $FKMOF(x_4) = 0.3706$; $FKMOF(x_5) = 0.5007$ and $FKMOF(x_6) = 0.3174$.

(5) Given $\mu = 0.5$, then $FKMOF(x_1), FKMOF(x_2), FKMOF(x_3), FKMOF(x_4) < \mu$, however, $FKMOF(x_5) > \mu$. By Definition 17, the outlier set (OS) can be obtained, i.e., $OS = \{x_5\}$.

Table 9
The summary of datasets.

Datasets	Rename	Attributes	Outliers (%)	Normal
Wine	D1	13	10 (7.7%)	119
Cardio	D2	21	176 (9.6%)	1655
Diabetes	D3	8	26 (4.9%)	500
Ecoli	D4	7	9 (2.6%)	327
Ionosphere	D5	34	24 (9.64%)	225
Glass	D6	9	9 (4.2%)	205
Pageblocks	D7	10	258 (5.0%)	4913
Wdbc	D8	31	39 (9.85%)	357
Yeast	D9	8	5 (0.44%)	1136
Vowels	D10	12	50 (3.4%)	1406
Thyroid	D11	6	93 (2.5%)	3679
Musk	D12	166	97 (3.2%)	2965
Satimage-2	D13	36	71 (1.2%)	5732
Optdigits	D14	64	150 (3%)	5066
Annthyroid	D15	6	534 (7.42%)	6666
Pendigits	D16	16	156 (2.27%)	6714

5. Experimental analysis

5.1. Description of datasets

In this section, sixteen multi-dimensional point datasets (There is one record per data point, and each record contains several attributes) are selected from the outlier detection datasets (ODDS)[1] and github,[2] which are shown in Table 9.

5.2. Multi-source numerical information fusion

Given that multi-source numerical information systems are not yet available in the current public datasets. We refer to the approach in [5] and treat the datasets in Table 9 as follows.

(1) The min – max normalization is performed on the original numerical data. Then, multiple information sources can be obtained by adding Gaussian noise and random noise methods [20]. In order to facilitate the experiment, each original dataset generates three information sources, i.e., $(U, MsNIS) = (NIS_1, NIS_2, NIS_3)$.

(2) Mean value (MV) fusion: Given an $MsNIS = \{NIS_1, NIS_2, \ldots, NIS_m\}$ where $U = \{x_1, x_2, \ldots, x_n\}$ and $A = \{a_1, a_2, \ldots, a_l\}$. For any

[1] http://odds.cs.stonybrook.edu/
[2] https://github.com/BElloney/Outlier-detection

$x_j \in U$ and $a_k \in A$, the MV fusion method can be denoted as follows.

$$MeanNIS(a_k(x_j)) = \sum_{i=1}^{m} \frac{NIS_i(a_k(x_j))}{m}, \tag{20}$$

where $MeanNIS(a_k(x_j))$ refers to the fused data and $NIS_i(a_k(x_j))$ refers to the initial ISV of ith information source.

Generally, the classification accuracy is used as the evaluation index to evaluate the accuracy of multi-source information fusion. A better fusion method reveals that the classification accuracy after fusion is not lower than that before fusion.

We compare the classification performance of the fusion methods by using three classifiers, i.e., NaiveBayes (NB), k-nearest neighbor (kNN, k=3), and support vector machine (SVM). All classification experiments were derived from ten-fold cross validation. It randomly divides a sample into ten subsets. Nine of them are treated as the training set, and the remaining one is regarded as the test set. After 10 rounds, the final performance is obtained by calculating the average and standard deviations of the classification accuracy. There are four methods are compared, namely, the average classification accuracy of multiple information sources (MIS), the average value of multiple information sources (MV), neighborhood information amount (NIA) [5], and our MsNDF method. The classification experiments of different fusion methods are compared as shown in Fig. 5. From Fig. 5, the classification accuracy of these methods is similar on the sixteen datasets, and our method is slightly better than other methods on the whole. This shows that our method is effective without loss of accuracy during the multi-source information fusion process.

5.3. Outlier detection in a new fused information system

To evaluate the superiority of the proposed FKMOD algorithm, nine classical algorithms are compared, i.e., local outlier factor (LOF) [55], local correlation integral (LOCI) [56], k-nearest neighbor (kNN) [57], corresponding fuzzy rough granules-based outlier detection (FRGOD) [58], the relevant fuzzy information entropy-based outlier detection (FIEOD) [42], DIStance-based (DIS) [59], connectivity-based outlier factor (COF) [60], cluster-based local outlier factor (CBLOF) [61] and angle-based outlier detection (ABOD) [62].

In all experiments, for kNN and LOF algorithms, the optimal values of their respective parameters k and MinPts in the range of [10, 50] with step size 10. The two parameters α and β required by CBLOF algorithm can be set to 90% and 5, respectively. The optimal similarity threshold s can be obtained in [0, 10] with steps size 1 [61]. The Euclidean distance is adopted for kNN, DIS and LOF algorithms. For FRGOD and FIEOD algorithms, the optimal parameters for λ and δ can be obtained in $[0.1, 1]$ with step size 0.1. After extensive experiments, the optimal threshold $\sigma = 0.01$ of our proposed algorithm FKMOD performs the best in most cases.

In this paper, we use precision (P), recall (R), and receiver operating characteristic (ROC) curves to evaluate the effectiveness of FKMOD algorithm. Considering that most outlier detection methods finally output the value of the outlier factor based on each sample. Then, the values of outlier factor corresponding to these samples can be arranged in a descending order, and numbered from 1. It is recorded as $t = 1$. Given a sequence number t, the samples greater than or equal to t can be regarded as outliers. It should be noted that if t is too small, the real outliers may not be found. On the contrary, if t is too large, some normal points may be misjudged as outliers. To avoid these two situations, $P(t)$ and $R(t)$ are used to find a suitable t. For any t, we have

$$P(t) = \frac{|OS(t) \cap OS^{\circ}|}{|OS(t)|} \times 100\%; \ R(t) = \frac{|OS(t) \cap OS^{\circ}|}{|OS^{\circ}|} \times 100\%, \tag{21}$$

where $OS(t)$ refers to the outlier set detected by the given t, which can be known as a function of t; OS° refers to the true outlier set. In addition, $P(t)$ and $R(t)$ indicate the ratio of outliers detected at a given t to true outliers and all true outliers, respectively.

The false positive rate (FPR) is the abscissa on the ROC curve, while the true positive rate (TPR) is the ordinate. Then, $FPR(t)$ and $TPR(t)$ can be calculated as follows.

$$\mathrm{FPR}(t) = \frac{|\mathrm{OS}(t) - OS^{\circ}|}{|U - OS^{\circ}|} \times 100\%; \ \ \mathrm{TPR}(t) = R(t) = \frac{|\mathrm{OS}(t) \cap OS^{\circ}|}{|OS^{\circ}|} \times 100\%. \tag{22}$$

Given a sequence number t, the larger the values of $P(t)$, $R(t)$ and $TPR(t)$, the better performance of outlier detection. On the contrary, the smaller the value of $FPR(t)$, the better the performance of outlier detection. If $t = |OS^{\circ}|$, then we have $P(t) = R(t)$.

According to Eq. (21), Table 10 show the comparison results of $P(t)$ and $R(t)$ varying with t for ten algorithms on the fused sixteen datasets. The best results are highlighted in bold and the second best results are underlined "_". From Table 10, the following findings can be drawn.

(1) The proposed algorithm FKMOD outperforms the other nine algorithms on the D2, D3, D6, D8, D9, D11, D12, D14, D15 and D16 datasets. The FKMOD, LOF, kNN and COF algorithms perform better on D5 dataset. In addition, FKMOD, FRGOD, FIEOD and DIS algorithms also perform better than other four outlier methods on D9 dataset. This shows that our method performs effectively on most datasets.
(2) kNN algorithm performs better on D4, D7 and D10 datasets, while FKMOD, LOF and COF have comparable performance on D1 dataset. It is also clear from the most datasets that our algorithm uses t that is less than other algorithms when $R(t) = 100\%$.
(3) Taking D8 dataset as an example, the accuracy $P(t)$ of the FKMOD algorithm is 96.15% when $t = 26$, while the accuracy of LOF, LOCF, kNN, FRGOD, FIEOD, DIS, COF, CBLOF and ABOD are 92.31%, 38.46%, 92.31%, 92.31%, 88.46%, 92.31%, 92.31%, 71.08%, 7.69% and 7.69%, respectively. When $t = 41$, the $R(t)$ of FKMOD and LOF reaches 100% for the first time, and the remaining other algorithms cannot make the recall reach 100% within t less than or equal to 50 except RGOD and FIEOD algorithms. This shows that FKMOD can detect 39 real outliers when $t = 41$, while other algorithms may detect all outliers only when the value of t is greater than 41.

To further demonstrate the detection performance of the proposed algorithm, Friedman's test [63] and Nemenyi's test [64] are performed to indicate the statistical significance of the experimental results. The Friedman's test can be used to determine whether the performance of the comparative approaches is identical. Prior to using Friedman's test, we sorted the $P(t)$ and $R(t)$ of each method on sixteen datasets from high to low, i.e., 1, 2, ⋯. If two algorithms have the same $P(t)$ or $R(t)$, then the order values are their mean ranking. Friedman's test is denoted as follows.

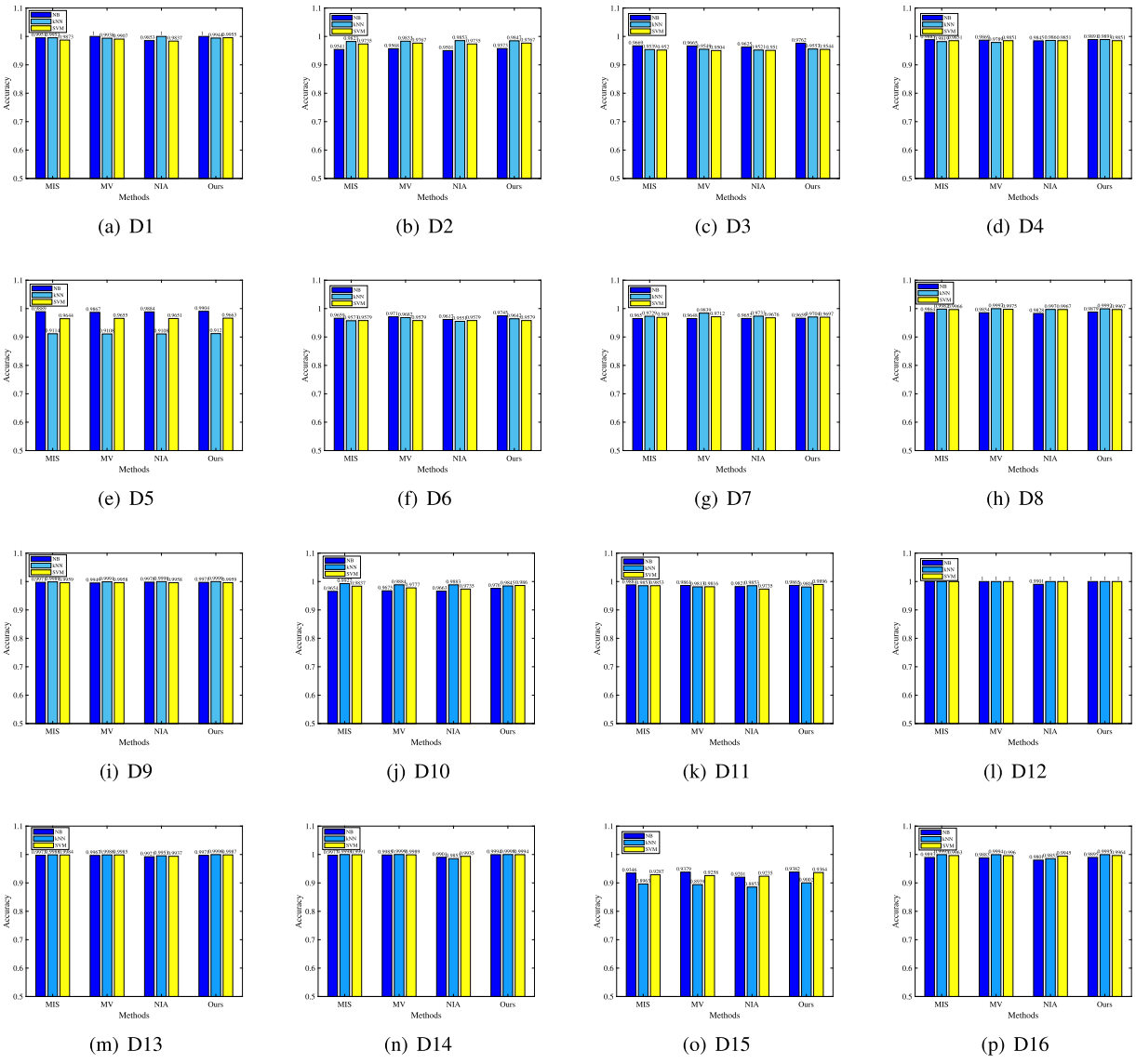
$$\chi_F^2 = \frac{12N}{A(A+1)}\left(\sum_{i=1}^{A} r_i^2 - \frac{A(A+1)^2}{4}\right), \ F_F = \frac{(N-1)\chi_F^2}{N(A-1)-\chi_F^2} \tag{23}$$

where A and N are the number of algorithms and datasets, respectively. r_i refers to the mean ranking of the ith algorithm. The variable χ_F^2 obeys the χ^2 distribution with $A-1$ degrees of freedom. Furthermore, F_F obeys the F distribution with $A-1$ and $(A-1)(N-1)$ degrees of freedom. Then, Nemenyi's test is defined as follows.

$$CD = q_{\alpha}\sqrt{\frac{A(A+1)}{6N}}, \tag{24}$$

where q_{α} refers to the number of comparison algorithms and α expresses the significance level.

As a result, we have $A = 10$ and $N = 16$, the F distribution has 9 and 135 degrees of freedom. Each dataset is sorted according to the test performance from high to low, and the mean ranking can be assigned. From Table 10, Friedman's test is achieved by the comparison of FKMOD, LOF, LOCI, kNN, FRGOD, FIEOD, DIS, COF, CBLOF and ABOD. Moreover, the mean ranking of these algorithms are computed.



(a) D1 (b) D2 (c) D3 (d) D4

(e) D5 (f) D6 (g) D7 (h) D8

(i) D9 (j) D10 (k) D11 (l) D12

(m) D13 (n) D14 (o) D15 (p) D16

Fig. 5. Comparison of classification accuracy of different methods in the new fused information system based on NB, kNN and SVM classifiers.

Subsequently, the values of χ_F^2 and F_F can be easily obtained. The results of the mean ranking of ten algorithms and the values of χ_F^2 and F_F in terms of $P(t)$ and $R(t)$ are shown in Table 11.

By calculating, when the significance level $\alpha = 0.1$, each values of F_F on two algorithms is greater than the critical value 1.678 of $F_F(9, 135)$. Therefore, the null hypothesis that "all algorithms have the same performance" is rejected on $P(t)$ and $R(t)$. Nemenyi's test is commonly used to calculate the critical difference (CD), which is denoted by Eq. (24). Hence, the CD is calculated as CD=3.1257 when $\alpha = 0.1$. If the CD between algorithms is greater than 3.1257, then the performance of the two algorithms differs significantly. Finally, the figures of Nemenyi's test are shown in Fig. 6. If there is no obvious difference between these methods, then connect them with a horizontal line segments. As seen from Fig. 6(a) and Fig. 6(b), the proposed FKMOD obtains the lowest mean ranking, and its performance is significantly better than other nine algorithms when the CD is larger than 3.1257. The two Figures show that the performance of FKMOD is comparable to that of LOF, kNN, FRGOD, and FIEOD. It is worth noting that the mean ranking of our proposed algorithm is the lowest among all algorithms (see Table 11). In general, the proposed FKMOD outperforms the other nine methods in terms of $P(t)$ and $R(t)$.

Moreover, to show the rationality of the proposed FKMOD algorithm more graphically, Fig. 7 gives the ROC curves on these sixteen fused datasets. From Fig. 7, it is obvious that the ROC curves of the proposed algorithm are significantly closer to the upper left corner of first quadrant, and the area under the curve is the largest for most of the datasets, e.g. D1, D3, D4, D5, D7, D8, D10, D11, D12, D13, D14, D15 and D16 datasets. Although the area of the upper left corner of our method on D2, D6 and D9 datasets is slightly inferior to the COF, LOCI and LOF algorithms, it still has some advantages over the other six algorithms. In general, the performance of our proposed FKMOD is better than other outlier detection algorithms.

Finally, in order to visualize the problem of selecting outlier detection thresholds, D7, D8 and D9 datasets are used as examples for analysis. According to Definitions 16–17, the FKMOF of sample distributions of D7, D8 and D9 datasets are exhibited in Fig. 8. From Fig. 8, the correlational analysis are described as follows.

(1) The fused new D7 information system contains 5171 objects and 10 attributes. In addition, it has a total of 258 true outliers representing 5% of the total objects. Let $\mu = 0.65$. Although most of the outliers are detected, 0.65 is not an ideal value because some of the normal points in the detected samples are mistaken for outliers.

(2) The fused new D8 information system contains 396 objects and 31 attributes. Moreover, it has a total of 39 true outliers representing 9.85% of the total objects. When $\mu = 0.71$, all outliers are detected,

Table 10
Comparison of experimental results on sixteen datasets (%).

Datasets	t	FKMOD		LOF		LOCI		kNN		FRGOD		FIEOD		DIS		COF		CBLOF		ABOD	
		P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)
D1	5	20.00	10.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.00	10.00	0.00	0.00	0.00	0.00
	17	17.65	30.00	23.53	40.00	0.00	0.00	11.76	20.00	11.76	20.00	11.76	20.00	11.76	20.00	17.65	30.00	5.88	10.00	0.00	0.00
	22	18.18	40.00	22.73	50.00	0.00	0.00	13.64	30.00	9.09	20.00	13.64	30.00	9.09	20.00	18.18	40.00	4.55	10.00	0.00	0.00
	30	20.00	60.00	16.67	50.00	0.00	0.00	16.67	50.00	10.00	30.00	13.33	40.00	10.00	30.00	20.00	60.00	6.67	20.00	0.00	0.00
	58	12.07	70.00	13.79	80.00	3.45	20.00	12.07	70.00	13.79	80.00	10.34	60.00	12.07	70.00	12.07	70.00	12.07	70.00	5.17	30.00
	94	10.64	100.00	10.64	100.00	7.45	70.00	10.64	100.00	8.51	80.00	8.51	80.00	9.57	90.00	10.64	100.00	10.64	100.00	7.45	70
	Avg.	**16.42**	51.67	14.56	**53.33**	1.82	15.00	10.80	45.00	8.86	38.33	9.60	38.33	8.75	38.33	**16.42**	51.67	6.63	35.00	2.10	16.67
D2	10	100.00	5.68	90.00	5.11	50.00	2.84	100.00	5.68	90.00	5.11	70.00	3.98	100.00	5.68	50.00	2.84	0.00	0.00	0.00	0.00
	20	100.00	11.36	90.00	10.23	70.00	7.95	95.00	10.80	85.00	9.66	65.00	7.39	90.00	10.23	45.00	5.11	5.00	0.57	0.00	0.00
	100	69.00	39.20	51.00	28.98	54.00	30.68	64.00	36.36	68.00	38.64	58.00	32.95	70.00	39.77	28.00	15.91	6.00	3.41	2.00	1.14
	200	56.50	64.20	39.50	44.89	42.50	48.30	58.00	65.91	57.50	65.34	45.50	51.70	55.50	63.07	23.00	26.14	7.00	7.95	1.00	1.14
	300	49.00	83.52	37.00	63.07	35.33	60.23	48.67	82.95	48.33	82.39	41.00	69.89	48.67	82.95	21.00	35.80	9.67	16.48	0.67	1.14
	400	40.75	92.61	33.75	76.70	28.50	64.77	40.50	92.05	40.25	91.48	34.25	77.84	39.50	89.77	18.75	42.61	10.00	22.73	0.50	1.14
	500	34.20	97.16	30.60	86.93	24.40	69.32	34.00	96.59	34.00	96.59	30.00	85.23	33.80	96.02	16.20	46.02	9.60	27.27	0.40	1.14
	600	29.17	99.43	28.33	96.59	22.17	75.57	28.83	98.30	28.83	98.30	26.67	90.91	28.67	97.73	14.83	50.57	9.17	31.25	0.33	1.14
	700	25.14	100.00	24.57	97.73	19.86	78.98	25.00	99.43	25.00	99.43	23.71	94.32	24.71	98.30	13.29	52.84	8.00	31.82	0.29	1.14
	Avg.	**55.97**	**65.91**	47.19	56.69	38.53	48.74	54.89	65.34	52.99	65.21	43.79	57.13	54.54	64.84	25.56	30.87	7.16	15.72	0.58	0.88
D3	5	100.00	19.23	80.00	15.38	20.00	3.85	80.00	15.38	80.00	15.38	40.00	7.69	83.33	19.23	60.00	11.54	20.00	3.85	0.00	0.00
	9	88.89	30.77	66.67	23.08	33.33	11.54	66.67	23.08	66.67	23.08	44.44	15.38	66.67	23.08	44.44	15.38	33.33	11.54	0.00	0.00
	21	61.90	50.00	38.10	30.77	28.57	23.08	42.86	34.62	57.14	46.15	38.10	30.77	47.62	38.46	38.10	30.77	14.29	11.54	4.76	3.85
	54	46.30	96.15	35.19	73.08	14.81	30.77	35.19	73.08	35.19	73.08	29.63	61.54	31.48	65.38	24.07	50.00	11.11	23.08	1.85	3.85
	77	33.77	100.00	28.57	84.62	14.29	42.31	28.57	84.62	29.87	88.46	25.97	76.92	28.57	84.62	18.18	53.85	7.79	23.08	2.60	7.69
	80	32.50	100.00	30.00	92.31	13.75	42.31	28.75	88.46	28.75	88.46	25.00	76.92	30.00	92.31	17.50	53.85	7.50	23.08	2.50	7.69
	Avg.	**60.56**	**66.03**	46.42	53.21	20.79	25.64	47.01	53.21	49.60	55.77	33.86	44.87	47.95	53.85	33.72	35.90	15.67	16.03	1.95	3.85
D4	10	70.00	77.78	70.00	77.78	50.00	55.56	70.00	77.78	70.00	77.78	50.00	55.56	70.00	77.78	70.00	77.78	0.00	0.00	0.00	0.00
	20	35.00	77.78	35.00	77.78	35.00	77.78	35.00	77.78	35.00	77.78	35.00	77.78	35.00	77.78	35.00	77.78	5.00	11.11	5.26	11.11
	40	17.50	77.78	17.50	77.78	20.00	88.89	17.50	77.78	17.50	77.78	17.50	77.78	17.50	77.78	17.50	77.78	2.50	11.11	2.50	11.11
	60	11.67	77.78	11.67	77.78	13.33	88.89	13.33	88.89	11.67	77.78	11.67	77.78	11.67	77.78	11.67	77.78	1.67	11.11	1.64	11.11
	90	7.78	77.78	8.89	88.89	8.89	88.89	8.89	88.89	7.78	77.78	7.78	77.78	7.78	77.78	7.78	77.78	2.22	22.22	2.22	22.22
	100	8.00	88.89	8.00	88.89	8.00	88.89	8.00	88.89	7.00	77.78	7.00	77.78	7.00	77.78	8.00	88.89	2.00	22.22	3.00	33.33
	120	7.50	100.00	6.67	88.89	6.67	88.89	6.67	88.89	6.67	88.89	5.83	77.78	5.83	77.78	6.67	88.89	1.67	22.22	2.50	33.33
	Avg.	22.49	82.54	22.53	82.54	20.27	82.54	**22.77**	**84.13**	22.23	79.37	19.25	74.60	22.11	77.78	22.37	80.95	2.15	14.29	2.45	17.46
D5	10	100.00	41.67	100.00	41.67	80.00	33.33	100.00	41.67	100.00	41.67	100.00	41.67	100.00	41.67	100.00	41.67	0.00	0.00	10.00	4.17
	20	100.00	83.33	100.00	83.33	80.00	66.67	100.00	83.33	100.00	83.33	95.00	79.17	100.00	83.33	100.00	83.33	5.00	4.17	5.00	4.17
	30	80.00	100.00	80.00	100.00	70.00	87.50	80.00	100.00	76.67	95.83	73.33	91.67	76.67	95.83	80.00	100.00	3.33	4.17	3.33	4.17
	40	60.00	100.00	60.00	100.00	57.50	95.83	60.00	100.00	57.50	95.83	57.50	95.83	60.00	100.00	60.00	100.00	5.00	8.33	5.00	8.33
	50	48.00	100.00	48.00	100.00	46.00	95.83	48.00	100.00	48.00	100.00	46.00	95.83	48.00	100.00	48.00	100.00	4.00	8.33	4.00	8.33
	60	40.00	100.00	40.00	100.00	38.33	95.83	40.00	100.00	40.00	100.00	40.00	100.00	40.00	100.00	40.00	100.00	3.33	8.33	3.33	8.33
	70	34.29	100.00	34.29	100.00	32.86	95.83	34.29	100.00	34.29	100.00	34.29	100.00	34.29	100.00	34.29	100.00	2.86	8.33	4.29	12.50
	Avg.	**66.04**	**89.29**	**66.04**	**89.29**	57.81	81.55	**66.04**	**89.29**	65.21	88.10	63.73	86.31	65.56	88.69	**66.04**	**89.29**	3.36	5.95	4.99	7.14
D6	12	16.67	22.22	8.33	11.11	16.67	22.22	9.09	11.11	8.33	11.11	16.67	22.22	8.33	11.11	16.67	22.22	8.33	11.11	0.00	0.00
	26	11.54	33.33	7.69	22.22	7.69	22.22	11.54	33.33	7.69	22.22	7.69	22.22	7.69	22.22	11.54	33.33	3.85	11.11	0.00	0.00
	32	12.50	44.44	6.25	22.22	6.25	22.22	9.38	33.33	6.25	22.22	6.25	22.22	6.25	22.22	9.38	33.33	6.25	22.22	3.13	11.11
	52	9.62	55.56	9.62	55.56	3.85	22.22	13.46	77.78	7.69	44.44	9.62	55.56	7.69	44.44	11.54	66.67	3.85	22.22	1.92	11.11
	63	11.11	77.78	7.94	55.56	3.17	22.22	11.11	77.78	7.94	55.56	7.94	55.56	7.94	55.56	9.52	66.67	4.76	33.33	1.59	11.11
	74	12.16	100.00	8.11	66.67	2.70	22.22	10.81	88.89	6.76	55.56	6.76	55.56	6.76	55.56	8.11	66.67	4.05	33.33	1.35	11.11
	Avg.	**12.27**	**55.56**	7.99	38.89	6.72	22.22	10.90	53.70	7.44	35.19	9.15	38.89	7.44	35.19	11.13	48.15	5.18	22.22	1.33	7.41
D7	8	100.00	3.10	100.00	3.10	0	0	100.00	3.10	100.00	3.10	100.00	3.10	100.00	3.10	87.50	2.71	0.00	0.00	12.50	0.39
	17	100.00	6.59	100.00	6.59	5.88	0.39	100.00	6.59	100.00	6.59	94.12	6.20	100.00	6.59	94.12	6.20	0.00	0.00	17.65	1.16
	29	96.55	10.85	96.55	10.85	13.79	1.55	96.55	10.85	89.66	10.08	79.31	8.91	93.10	10.47	86.21	9.69	0.00	0.00	13.79	1.55
	80	77.50	24.03	80.00	24.81	7.50	2.33	83.75	25.97	67.50	20.93	48.75	15.12	73.75	22.87	62.50	19.38	2.50	0.78	6.25	1.94
	185	60.00	43.02	65.41	46.90	3.78	2.71	68.11	48.84	52.43	37.60	41.62	29.84	52.43	37.60	40.00	28.68	2.16	1.55	4.86	3.49
	302	50.00	58.53	52.98	62.02	2.98	2.71	54.97	64.34	51.66	60.47	44.04	51.55	50.00	58.53	30.79	36.05	6.62	7.75	5.63	6.59
	455	40.44	71.32	42.42	74.81	2.20	3.88	43.96	77.52	41.54	73.26	40.00	70.54	42.20	74.42	24.62	43.41	7.91	13.95	3.74	6.59
	1000	23.40	90.70	24.70	95.74	5.30	20.54	24.70	95.74	23.70	91.86	23.50	91.09	23.80	92.25	14.80	57.36	5.90	22.87	1.70	6.59
	2274	11.35	100.00	11.26	99.22	3.12	27.52	11.26	99.22	11.13	98.06	11.21	98.84	11.21	98.84	8.14	71.71	3.69	32.56	2.46	21.71
	Avg.	62.14	45.35	63.70	47.11	4.95	6.93	**64.81**	**48.02**	59.73	44.66	53.62	41.69	60.72	44.96	49.85	30.58	3.20	8.83	7.62	5.56
D8	5	100.00	12.82	80.00	10.26	60.00	7.69	80.00	10.26	83.33	12.82	80.00	10.26	80.00	10.26	100.00	12.82	20.00	2.56	20.00	2.56
	9	100.00	23.08	88.89	20.51	66.67	15.38	77.78	17.95	90.00	23.08	88.89	20.51	88.89	20.51	88.89	20.51	22.22	5.13	11.11	2.56
	19	94.74	46.15	89.47	43.59	52.63	25.64	89.47	43.59	94.74	46.15	94.74	46.15	94.74	46.15	78.95	38.46	10.53	5.13	10.53	5.13
	26	96.15	64.10	92.31	61.54	38.46	25.64	92.31	61.54	92.31	61.54	88.46	58.97	92.31	61.54	73.08	48.72	7.69	5.13	7.69	5.13
	37	97.30	92.31	94.59	89.74	29.73	28.21	91.89	87.18	91.89	87.18	81.08	76.92	91.89	87.18	64.86	61.54	8.11	7.69	5.41	5.13
	41	95.12	100.00	95.12	100.00	26.83	28.21	87.80	92.31	92.68	97.44	80.49	84.62	92.68	97.44	65.85	69.23	9.76	10.26	4.88	5.13
	50	78.00	100.00	78.00	100.00	22.00	28.21	78.00	100.00	78.00	100.00	76.00	97.44	78.00	100.00	68.00	87.18	8.00	10.26	4.0	5.13
	Avg.	**94.47**	**62.64**	88.34	60.81	42.33	22.71	85.32	58.97	88.99	61.17	84.24	56.41	88.36	60.44	77.09	48.35	12.33	6.59	9.09	4.40

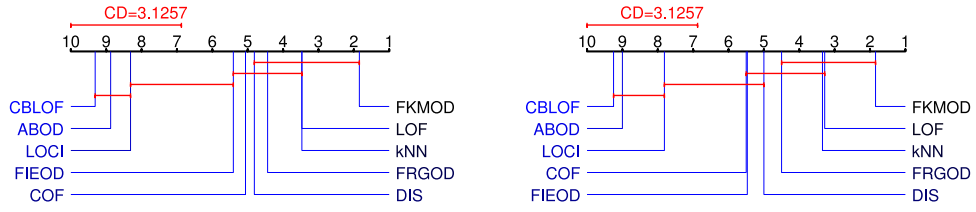
(*continued on next page*)

Table 10 (*continued*).

Datasets	t	FKMOD		LOF		LOCI		kNN		FRGOD		FIEOD		DIS		COF		CBLOF		ABOD	
		P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)	P(t)	R(t)
D9	5	40.00	40.00	40.00	40.00	20.00	20.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	20.00	20.00	20.00	20.00	20.00	20.00
	8	50.00	80.00	33.33	60.00	22.22	40.00	22.22	40.00	50.00	80.00	50.00	80.00	50.00	80.00	22.22	40.00	22.22	40.00	12.50	20.00
	14	35.71	100.00	35.71	100.00	21.43	60.00	35.71	100.00	35.71	100.00	35.71	100.00	35.71	100.00	35.71	100.00	14.29	40.00	7.14	20.00
	25	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	8.00	40.00	4.00	20.00
	38	13.16	100.00	13.16	100.00	13.16	100.00	13.16	100.00	13.16	100.00	13.16	100.00	13.16	100.00	13.16	100.00	5.26	40.00	2.63	20.00
	42	11.90	100.00	11.90	100.00	11.90	100.00	11.90	100.00	11.90	100.00	11.90	100.00	11.90	100.00	11.90	100.00	4.76	40.00	2.38	20.00
	50	10.00	100.00	10.00	100.00	10.00	100.00	10.00	100.00	10.00	100.00	10.00	100.00	10.00	100.00	10.00	100.00	4.00	40.00	4.00	20.00
	Avg.	**25.83**	**88.57**	23.44	85.71	16.96	74.29	21.86	82.86	**25.83**	**88.57**	**25.83**	**88.57**	**25.83**	**88.57**	19.00	80.00	11.22	37.14	7.52	22.86
D10	12	50.00	12.00	50.00	12.00	0.00	0.00	80.00	24.00	16.67	4.00	8.33	2.00	8.33	2.00	41.67	10.00	0.00	0.00	8.33	2.00
	27	51.85	28.00	51.85	28.00	3.70	2.00	62.96	34.00	11.11	6.00	7.41	4.00	11.11	6.00	44.44	24.00	3.70	2.00	11.11	6.00
	53	30.19	32.00	45.28	48.00	3.77	4.00	52.83	56.00	9.43	10.00	3.77	4.00	9.43	10.00	33.96	36.00	1.89	2.00	7.55	8.00
	80	23.75	38.00	36.25	58.00	6.25	10.00	42.50	68.00	7.50	12.00	5.00	8.00	7.50	12.00	28.75	46.00	5.00	8.00	5.00	8.00
	119	18.49	44.00	31.09	74.00	4.20	10.00	32.77	78.00	7.56	18.00	5.88	14.00	5.88	14.00	21.01	50.00	3.36	8.00	3.36	8.00
	163	15.34	50.00	23.93	78.00	3.07	10.00	25.15	82.00	6.75	22.00	5.52	18.00	6.75	22.00	17.50	56.00	2.45	8.00	2.45	8.00
	191	14.66	56.00	20.94	80.00	3.66	14.00	22.51	86.00	6.28	24.00	4.71	18.00	6.81	26.00	16.75	64.00	2.62	10.00	2.09	8.00
	500	8.80	88.00	9.60	96.00	2.80	28.00	9.60	96.00	4.60	46.00	4.20	42.00	4.40	44.00	8.40	84.00	1.20	12.00	1.20	12.00
	Avg.	26.63	43.50	33.62	59.25	3.43	9.75	**41.04**	**65.50**	8.74	17.75	5.60	13.75	7.53	17.00	26.56	46.25	2.53	6.25	5.14	7.50
D11	10	90.00	9.68	60.00	6.45	30.00	3.23	50.00	5.38	70.00	7.53	60.00	6.45	60.00	6.45	40.00	4.30	0.00	0.00	10.00	1.08
	20	95.24	21.51	35.00	7.53	25.00	5.38	30.00	6.45	35.00	7.53	45.00	9.68	30.00	6.45	35.00	7.53	5.00	1.08	10.00	2.15
	30	93.75	32.26	26.67	8.60	16.67	5.38	26.67	8.60	30.00	9.68	30.00	9.68	26.67	8.60	30.00	9.68	6.67	2.15	10.00	3.23
	50	96.15	53.76	34.00	18.28	16.00	8.60	20.00	10.75	22.00	11.83	20.00	10.75	22.00	11.83	24.00	12.90	4.00	2.15	8.00	4.30
	100	72.00	77.42	34.00	36.56	14.00	15.05	24.00	25.81	18.00	19.35	10.00	10.75	15.00	16.13	16.00	17.20	2.00	2.15	7.00	7.53
	200	42.00	90.32	27.00	58.06	14.00	30.11	26.00	55.91	18.00	38.71	6.00	12.90	12.00	25.81	16.00	34.41	3.00	6.45	6.50	13.98
	300	29.33	94.62	23.00	74.19	13.00	41.94	25.00	80.65	15.67	50.54	5.67	18.28	13.33	43.01	15.67	50.54	2.67	8.60	6.67	21.51
	395	23.54	100.00	19.24	81.72	11.90	50.54	20.76	88.17	13.92	59.14	4.81	20.43	11.65	49.46	15.44	65.59	2.28	9.68	6.08	25.81
	Avg.	**67.75**	**59.95**	32.36	36.42	17.57	20.03	27.80	35.22	27.82	25.54	22.68	12.37	23.83	20.97	24.01	25.27	3.20	4.03	8.03	9.95
D12	10	100.00	10.31	100.00	10.31	10.00	1.03	90.00	9.28	40.00	4.12	40.00	4.12	40.00	4.12	80.00	8.25	20.00	2.06	100.00	10.31
	20	100.00	20.62	100.00	20.62	20.00	4.12	65.00	13.40	35.00	7.22	35.00	7.22	25.00	5.15	50.00	10.31	10.00	2.06	100.00	20.62
	30	100.00	30.93	100.00	30.93	23.33	7.22	56.67	17.53	33.33	10.31	30.00	9.28	16.67	5.15	40.00	12.37	10.00	3.09	100.00	30.93
	60	100.00	61.86	96.67	59.79	20.00	12.37	53.33	32.99	31.67	19.59	26.67	16.49	18.33	11.34	23.33	14.43	11.67	7.22	100.00	61.86
	80	100.00	82.47	92.50	76.29	17.50	14.43	50.00	41.24	28.75	23.71	21.25	17.53	20.00	16.49	17.50	14.43	11.25	9.28	100.00	82.47
	90	100.00	92.78	91.11	84.54	16.67	15.46	48.89	45.36	30.00	27.84	21.11	19.59	17.78	16.49	15.56	14.43	11.11	10.31	95.56	88.66
	100	97.00	100.00	84.00	86.60	16.00	16.49	50.00	51.55	29.00	29.90	19.00	19.59	19.00	19.59	14.00	14.43	11.00	11.34	86.00	88.66
	110	88.18	100.00	78.18	88.66	15.45	17.53	46.36	52.58	29.09	32.99	20.00	22.68	22.73	25.77	12.73	14.43	10.00	11.34	78.18	88.66
	Avg.	**98.15**	**62.37**	92.81	57.22	17.37	11.08	57.53	32.99	32.11	19.46	26.63	14.56	22.44	13.02	31.64	12.89	11.88	7.09	94.97	59.02
D13	10	100.00	14.08	100.00	14.08	30.00	4.23	100.00	14.08	100.00	14.08	100.00	14.08	100.00	14.08	50.00	7.04	10.00	1.41	10.00	1.41
	20	100.00	28.17	100.00	28.17	25.00	7.04	76.92	28.17	100.00	28.17	100.00	28.17	100.00	28.17	35.00	9.86	10.00	2.82	5.00	1.41
	30	96.67	40.85	96.77	42.25	20.00	8.45	73.33	30.99	100.00	42.25	100.00	42.25	100.00	42.25	26.67	11.27	6.67	2.82	3.33	1.41
	50	82.00	57.75	94.00	66.20	16.00	11.27	56.00	39.44	100.00	70.42	100.00	70.42	100.00	70.42	18.00	12.68	8.00	5.63	4.00	2.82
	100	60.00	84.51	69.00	97.18	14.00	19.72	41.00	57.75	59.00	83.10	64.00	90.14	62.00	87.32	11.00	15.49	7.00	9.86	5.00	7.04
	200	33.50	94.37	34.50	97.18	11.00	30.99	29.00	81.69	31.00	87.32	33.50	94.37	32.00	90.14	6.50	18.31	6.50	18.31	6.00	16.90
	300	23.33	98.59	23.67	100.00	9.67	40.85	23.33	98.59	21.33	90.14	23.00	97.18	21.67	91.55	5.67	23.94	6.00	25.35	6.67	28.17
	500	14.20	100.00	14.20	100.00	8.20	57.75	14.20	100.00	13.20	92.96	13.80	97.18	13.20	92.96	4.20	29.58	5.60	39.44	5.80	40.85
	Avg.	63.71	64.79	66.52	**68.13**	16.73	22.54	51.72	56.34	65.57	63.56	**66.79**	66.73	66.11	64.61	19.63	16.02	7.47	13.20	5.73	12.50
D14	400	1.75	4.67	2.00	5.33	1.50	4.00	1.75	4.67	0.50	1.33	6.25	16.67	0.75	2.00	3.00	8.00	1.50	4.00	0.00	0.00
	600	4.83	19.33	2.00	8.00	1.00	4.00	2.33	9.33	1.00	4.00	5.50	22.00	1.17	4.67	3.83	15.33	1.17	4.67	0.17	0.67
	800	9.00	48.00	2.13	11.33	1.00	5.33	2.25	12.00	1.75	9.33	6.00	32.00	1.50	8.00	3.25	17.33	1.13	6.00	0.13	0.67
	1000	9.00	60.00	2.00	13.33	1.40	9.33	2.10	14.00	1.80	12.00	5.50	36.67	1.70	11.33	2.90	19.33	1.10	7.33	0.10	0.67
	1200	9.33	74.67	1.83	14.67	1.25	10.00	2.00	16.00	1.92	15.33	4.92	39.33	1.67	13.33	3.00	24.00	1.00	8.00	0.08	0.67
	1500	8.33	83.33	1.80	18.00	1.07	10.67	2.00	20.00	2.40	24.00	5.00	50.00	2.07	20.67	3.13	31.33	0.87	8.67	0.40	4.00
	Avg.	**7.04**	**48.33**	1.96	11.78	1.20	7.22	2.07	12.67	1.56	11.00	5.53	32.78	1.48	10.00	3.19	19.22	1.13	6.44	0.15	1.11
D15	10	100.00	1.87	60.00	1.12	10.00	0.19	60.00	1.12	80.00	1.50	50.00	0.94	70.00	1.31	30.00	0.56	0.00	0.00	10.00	0.19
	20	100.00	3.75	55.00	2.06	5.00	0.19	55.00	2.06	65.00	2.43	65.00	2.43	55.00	2.06	25.00	0.94	10.00	0.37	10.00	0.37
	30	100.00	5.62	36.67	2.06	3.33	0.19	40.00	2.25	50.00	2.81	50.00	2.81	40.00	2.25	20.00	1.12	10.00	0.56	13.33	0.75
	100	95.00	17.79	37.00	6.93	2.00	0.37	38.00	7.12	31.00	5.81	24.00	4.49	28.00	5.24	36.00	6.74	5.00	0.94	8.00	1.50
	200	76.12	28.65	30.50	11.42	1.50	0.56	31.50	11.80	23.50	8.80	15.50	5.81	20.50	7.68	30.00	11.24	5.50	2.06	6.50	2.43
	300	67.33	37.83	30.33	17.04	2.00	1.12	31.67	17.79	19.67	11.05	12.67	7.12	17.00	9.55	28.33	15.92	5.00	2.81	6.00	3.37
	600	39.33	44.19	25.83	29.03	3.00	3.37	24.83	27.90	16.67	18.73	10.17	11.42	14.00	13.11	20.83	23.41	5.00	5.62	5.67	6.37
	Avg.	**82.54**	**19.96**	39.33	9.95	3.83	0.86	40.14	10.01	40.83	7.30	32.48	5.00	34.93	5.89	27.17	8.56	5.79	1.77	8.50	2.14
D16	10	40.00	2.56	30.00	1.92	10.00	0.64	30.00	1.92	10.00	0.64	10.00	0.64	30.00	1.92	40.00	2.56	10.00	0.64	10.00	0.64
	20	60.00	7.69	20.00	2.56	15.00	1.92	15.00	1.92	10.00	1.28	10.00	1.28	20.00	2.56	25.00	3.21	10.00	1.28	10.00	1.28
	30	53.33	10.26	16.67	3.21	10.00	1.92	13.33	2.56	10.00	1.92	10.00	1.92	16.67	3.21	16.67	3.21	10.00	1.92	6.67	1.28
	50	48.00	15.38	14.00	4.49	10.00	3.21	14.00	4.49	12.00	7.69	12.00	3.85	20.00	6.41	10.00	3.21	8.00	2.56	6.00	1.92
	100	39.00	25.00	10.00	6.41	8.00	5.13	11.00	7.05	12.50	16.03	13.00	8.33	22.00	14.10	8.00	5.13	6.00	3.85	6.00	3.85
	200	26.00	33.33	7.00	8.97	7.00	8.97	7.00	8.97	13.67	26.28	13.50	17.31	25.00	32.05	5.50	7.05	6.00	7.69	5.50	7.05
	300	20.00	38.46	6.00	11.54	7.33	14.10	7.00	13.46	13.72	35.26	14.67	28.21	20.33	39.10	4.33	8.33	5.00	9.62	5.33	10.26
	500	17.60	56.41	5.80	18.59	6.40	20.51	6.00	19.23	13.20	42.31	14.00	44.87	17.00	54.49	4.00	12.82	4.40	14.10	5.00	16.03
	Avg.	**37.99**	**23.64**	13.68	7.21	9.22	7.05	12.92	7.45	11.89	16.43	12.15	13.30	21.38	19.23	14.19	5.69	7.43	5.21	6.81	5.29

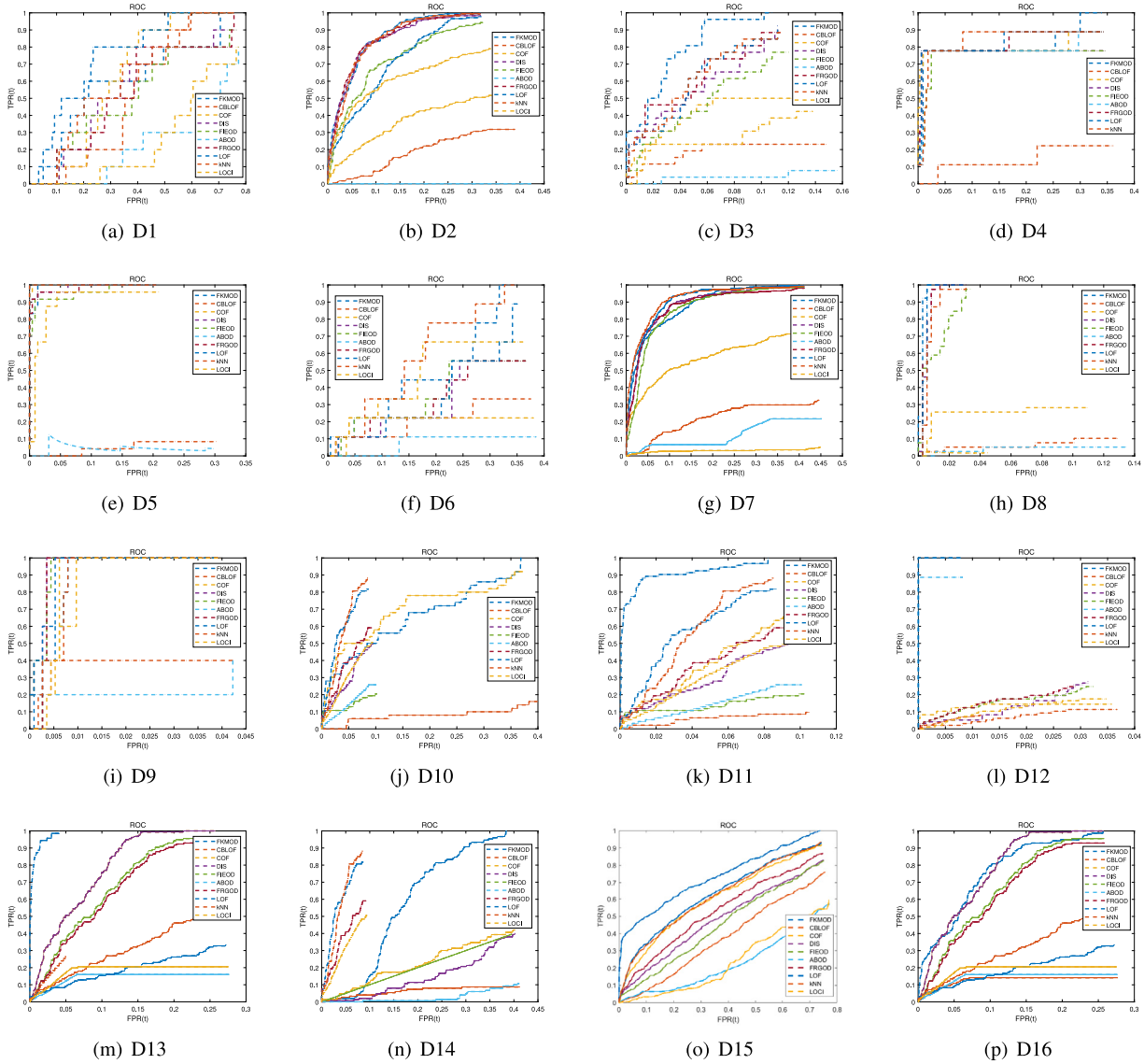
Table 11
The mean ranking and the values of χ_F^2 and F_F of the ten algorithms based on $P(t)$ and $R(t)$.

Algorithms	Mean ranking										χ_F^2	F_F
	FKMOD	LOF	LOCI	KNN	FRGOD	FIEOD	DIS	COF	CBLOF	ABOD		
P(t)	1.84	3.47	3.47	4.44	4.81	5.06	5.41	8.31	8.88	9.31	99.94	20.42
R(t)	1.84	3.28	3.34	4.50	5.00	5.47	5.50	7.81	9.00	9.25	97.49	18.86



(a) $P(t)$ (b) $R(t)$

Fig. 6. Comparison between FKMOD and the other nine algorithms under the Nemenyi's test.



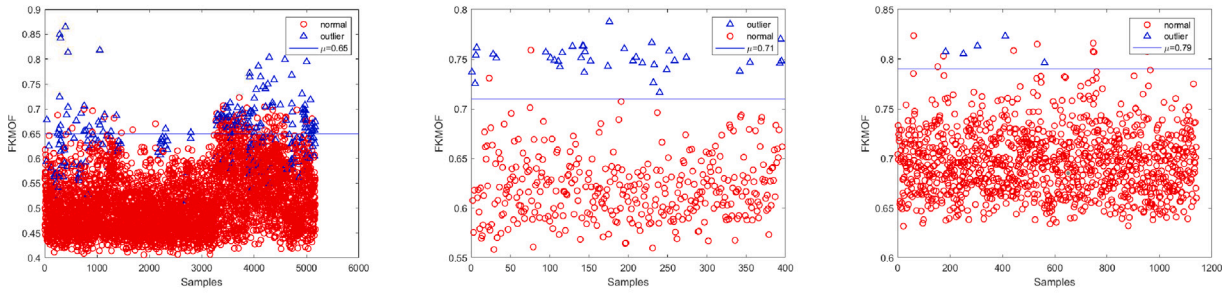
(a) D1 (b) D2 (c) D3 (d) D4

(e) D5 (f) D6 (g) D7 (h) D8

(i) D9 (j) D10 (k) D11 (l) D12

(m) D13 (n) D14 (o) D15 (p) D16

Fig. 7. The ROC curves for sixteen datasets.



(a) D7 (b) D8 (c) D9

Fig. 8. The FKMOF-based sample distributions in new fused D7, D8 and D9 information systems.

although two normal points are mistaken for outliers, but this is a normal phenomenon and within the tolerable error.

(3) The fused new D9 information system contains 1142 objects and 8 attributes. In addition, it has a total of 5 true outliers representing 0.44% of the total objects. Given $\mu = 0.79$, it is obvious that all outliers can be detected as well, but there are more normal points that are mistaken for outliers. This is due to the small percentage of true outliers, which is also one of the important challenges in practical applications. We can increase the value of μ appropriately, or adjust it further to reduce the error according to the actual situation.

According to the above descriptions, the determination of μ is an important step. In general, outlier detection methods only give the degree of outliers for each object. Before using a method to detect outliers, the user should give an empirical value of k as the expected number of outliers. The ten algorithms (FKMOD, LOF, LOCI, kNN, FRGOD, FIEOD, DIS, COF, CBLOF and ABOD) compared in this paper involve the setting of parameter μ, all depending on the value of k provided by the user. The principle of determining the value of k can be described as a three-step process: First, calculate the outlier factor for each object and sort the objects according to their outlier factor from largest to smallest. Then, determine the parameter μ. This step is to ensure that k objects are found to make the outlier degree in U higher than other objects. Finally, the k objects found will be returned to the user as outliers.

6. Conclusions

In this paper, a unified model of multi-source information fusion for outlier detection was proposed, which can be viewed as a process of two-stage data processing. The first stage is the fusion of various information sources, which offers a fusion approach with the least amount of uncertainty. The second stage involves advancing the fused data with the aim of discovering outliers, namely, objects that do not conform to expectations or are significantly different from the new fused information system. It is important to emphasize that the suggested unified model is effective and adaptable for homogeneous information sources (numerical data). This helps scholars adapt fusion techniques or outlier detection models to suit current needs. However, the proposed model is weak for nominal data or mixed data. In future work, we will extend the proposed model to fuse mixed data consisting of different types of data and tackle the problem of outlier detection for heterogeneous information sources.

CRediT authorship contribution statement

Pengfei Zhang: Conceptualization, Writing – original draft, Read and contributed to the manuscript. **Tianrui Li:** Supervision, Project administration, Read and contributed to the manuscript. **Guoqiang Wang:** Investigation, Read and contributed to the manuscript. **Dexian Wang:** Writing – review & editing, Read and contributed to the manuscript. **Pei Lai:** Visualization, Read and contributed to the manuscript. **Fan Zhang:** Methodology, Read and contributed to the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62176221, 62276215, 62076171), the Science and Technology Project of Sichuan Province (No. 2022JDRC0067), and the Fundamental Research Funds for the Central Universities (No. 2682021CX038).

References

[1] M. Muzammal, R. Talat, A.H. Sodhro, S. Pirbhulal, A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks, Inf. Fusion 53 (2020) 155–164.
[2] P. Zhang, T. Li, G. Wang, C. Luo, H. Chen, J. Zhang, D. Wang, Z. Yu, Multi-source information fusion based on rough set theory: A review, Inf. Fusion 68 (2021) 85–117.
[3] Z.-G. Liu, Y. Liu, J. Dezert, F. Cuzzolin, Evidence combination based on credal belief redistribution for pattern classification, IEEE Trans. Fuzzy Syst. 28 (4) (2019) 618–631.
[4] Q. Pan, Multi-Soure Information Fusion Theory and its Applications, Tsinghua University Press, 2013.
[5] P. Zhang, T. Li, Z. Yuan, L. Chuan, G. Wang, J. Liu, S. Du, A data-level fusion model for unsupervised attribute selection in multi-source homogeneous data, Inf. Fusion 80 (2022) 87–103.
[6] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Inf. Fusion 14 (1) (2013) 28–44.
[7] N. Xiong, P. Svensson, Multi-sensor management for information fusion: issues and approaches, Inf. Fusion 3 (2) (2002) 163–186.
[8] T. Li, J.M. Corchado, J. Bajo, S. Sun, J.F. De Paz, Effectiveness of Bayesian filters: An information fusion perspective, Inform. Sci. 329 (2016) 670–689.
[9] D. Dubois, H. Prade, On the use of aggregation operations in information fusion processes, Fuzzy Sets and Systems 142 (1) (2004) 143–161.
[10] Y. Fan, T.S. Ma, F.Y. Xiao, An improved approach to generate generalized basic probability assignment based on fuzzy sets in the open world and its application in multi-source information fusion, Appl. Intell. 51 (6) (2021) 3718–3735.
[11] S.A. Bouhamed, I.K. Kallel, R.R. Yager, É. Bossé, B. Solaiman, An intelligent quality-based approach to fusing multi-source possibilistic information, Inf. Fusion 55 (2020) 68–90.
[12] Y. Pan, L. Zhang, X. Wu, M.J. Skibniewski, Multi-classifier information fusion in risk analysis, Inf. Fusion 60 (2020) 121–136.
[13] Z. Huo, M. Martínez-García, Y. Zhang, L. Shu, A multisensor information fusion method for high-reliability fault diagnosis of rotating machinery, IEEE Trans. Instrum. Meas. 71 (2021) 1–12.
[14] W. Wei, J. Liang, Information fusion in rough set theory: An overview, Inf. Fusion 48 (2019) 107–118.
[15] Z.-G. Liu, L.-Q. Huang, K. Zhou, T. Denoeux, Combination of transferable classification with multisource domain adaptation based on evidential reasoning, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2020) 2015–2029.

[16] X.Y. Che, J.S. Mi, D.G. Chen, Information fusion and numerical characterization of a multi-source information system, Knowl.-Based Syst. 145 (2018) 121–133.
[17] G.P. Lin, J.Y. Liang, Y.H. Qian, An information fusion approach by combining multigranulation rough sets and evidence theory, Inform. Sci. 314 (2015) 184–199.
[18] R.R. Yager, A framework for multi-source data fusion, Inform. Sci. 163 (1–3) (2004) 175–200.
[19] M.M. Li, X.Y. Zhang, Information fusion in a multi-source incomplete information system based on information entropy, Entropy 19 (11) (2017) 570.
[20] L. Yang, W.H. Xu, X.Y. Zhang, B.B. Sang, Multi-granulation method for information fusion in multi-source decision information system, Internat. J. Approx. Reason. 122 (2020) 47–65.
[21] Y. Huang, T. Li, C. Luo, H. Fujita, S.-J. Horng, Dynamic fusion of multisource interval-valued data by fuzzy granulation, IEEE Trans. Fuzzy Syst. 26 (6) (2018) 3403–3417.
[22] G.S. Pang, C.H. Shen, L.B. Cao, A.V.D. Hengel, Deep learning for anomaly detection: A review, ACM Comput. Surv. 54 (2) (2021) 1–38.
[23] K.-H. Lai, D.C. Zha, J.J. Xu, Y. Zhao, G.C. Wang, X. Hu, Revisiting time series outlier detection: Definitions and benchmarks, in: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.
[24] Y. Tian, G.S. Pang, F.B. Liu, Y.H. Chen, S.H. Shin, J.W. Verjans, R. Singh, G. Carneiro, Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 128–140.
[25] W. Hilal, S.A. Gadsden, J. Yawney, A review of anomaly detection techniques and applications in financial fraud, Expert Syst. Appl. (2021) 116429.
[26] M. Aggarwal, M. Hanmandlu, Representing uncertainty with information sets, IEEE Trans. Fuzzy Syst. 24 (1) (2015) 1–15.
[27] M. Hanmandlu, A. Das, Content-based image retrieval by information theoretic measure, Def. Sci. J. 61 (5) (2011) 415.
[28] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[29] M. Girolami, Mercer kernel-based clustering in feature space, IEEE Trans. Neural Netw. 13 (3) (2002) 780–784.
[30] S.S. Keerthi, C.-J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, Neural Comput. 15 (7) (2003) 1667–1689.
[31] H.-T. Lin, C.-J. Lin, A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods, Neural Comput. 3 (1–32) (2003) 16.
[32] B. Moser, On the T-transitivity of kernels, Fuzzy Sets and Systems 157 (13) (2006) 1787–1796.
[33] Q. Hu, D. Yu, W. Pedrycz, D. Chen, Kernelized fuzzy rough sets and their applications, IEEE Trans. Knowl. Data Eng. 23 (11) (2010) 1649–1667.
[34] Q. Hu, D. Yu, Z. Xie, J. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Trans. Fuzzy Syst. 14 (2) (2006) 191–201.
[35] J.-L. Fan, Y.-L. Ma, Some new fuzzy entropy formulas, Fuzzy Sets and Systems 128 (2) (2002) 277–284.
[36] V. Arya, S. Kumar, Knowledge measure and entropy: a complementary concept in fuzzy theory, Granul. Comput. 6 (3) (2021) 631–643.
[37] K. Guo, Knowledge measure for Atanassov's intuitionistic fuzzy sets, IEEE Trans. Fuzzy Syst. 24 (5) (2015) 1072–1078.
[38] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1 (1978) 3–28.
[39] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (5) (1982) 341–356.
[40] M. Kryszkiewicz, Rough set approach to incomplete information systems, Inform. Sci. 112 (1–4) (1998) 39–49.
[41] Q.H. Hu, D.R. Yu, J.F. Liu, C.X. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inform. Sci. 178 (18) (2008) 3577–3594.
[42] Z. Yuan, H. Chen, T. Li, J. Liu, S. Wang, Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection, Fuzzy Sets and Systems 421 (2021) 1–28.
[43] W. Li, Y. Wei, W. Xu, General expression of knowledge granularity based on a fuzzy relation matrix, Fuzzy Sets and Systems 440 (2022) 149–163.
[44] M. Agarwal, M. Hanmandlu, K.K. Biswas, The properties and information measures for information sets, in: 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2014, pp. 412–417.
[45] B. Moser, On representing and generating kernels by fuzzy equivalence relations, J. Mach. Learn. Res. 7 (12) (2006) 2603–2620.
[46] Y.H. Qian, J.Y. Liang, W.Z. Wu, C.Y. Dang, Information granularity in fuzzy binary GrC model, IEEE Trans. Fuzzy Syst. 19 (2) (2010) 253–264.
[47] C. Wang, Y. Huang, M. Shao, D. Chen, Uncertainty measures for general fuzzy relations, Fuzzy Sets and Systems 360 (2019) 82–96.
[48] R.R. Yager, On the measure of fuzziness and negation part I: membership in the unit interval, Int. J. Gen. Syst. 8 (3) (1979) 338–353.
[49] B. Kosko, Fuzzy entropy and conditioning, Inform. Sci. 40 (2) (1986) 165–174.
[50] N.R. Pal, S.K. Pal, Higher order fuzzy entropy and hybrid entropy of a set, Inform. Sci. 61 (3) (1992) 211–231.
[51] P.K. Li, B.D. Liu, Entropy of credibility distributions for fuzzy variables, IEEE Trans. Fuzzy Syst. 16 (1) (2008) 123–129.
[52] C.-M. Hwang, M.-S. Yang, On entropy of fuzzy sets, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 16 (04) (2008) 519–527.
[53] R. Joshi, S. Kumar, A novel fuzzy decision-making method using entropy weights-based correlation coefficients under intuitionistic fuzzy environment, Int. J. Fuzzy Syst. 21 (1) (2019) 232–242.
[54] A.N. Redlich, Redundancy reduction as a strategy for unsupervised learning, Neural Comput. 5 (2) (1993) 289–304.
[55] M.M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander, LOF: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, pp. 93–104.
[56] S. Papadimitriou, H. Kitagawa, P.B. Gibbons, C. Faloutsos, Loci: Fast outlier detection using the local correlation integral, in: Proceedings 19th International Conference on Data Engineering (Cat. No. 03CH37405), IEEE, 2003, pp. 315–326.
[57] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, pp. 427–438.
[58] Z. Yuan, H.M. Chen, T.R. Li, B.B. Sang, S. Wang, Outlier detection based on fuzzy rough granules in mixed attribute data, IEEE Trans. Cybern. 2021 (2021).
[59] E.M. Knorr, R.T. Ng, V. Tucakov, Distance-based outliers: algorithms and applications, VLDB J. 8 (3) (2000) 237–253.
[60] J. Tang, Z.X. Chen, A.W.-C. Fu, D.W. Cheung, Enhancing effectiveness of outlier detections for low density patterns, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2002, pp. 535–548.
[61] Z.Y. He, X.F. Xu, S.C. Deng, Discovering cluster-based local outliers, Pattern Recognit. Lett. 24 (9–10) (2003) 1641–1650.
[62] H.-P. Kriegel, M. Schubert, A. Zimek, Angle-based outlier detection in high-dimensional data, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 444–452.
[63] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1) (1940) 86–92.
[64] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.