Information Sciences 378 (2017) 410–423

Contents lists available at ScienceDirect

Information Sciences

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

A novel approach to information fusion in multi-source datasets: A granular computing viewpoint

Weihua Xu*, Jianhang Yu

School of Mathematics and Statistics, Chongqing University of Technology, Chongqing 400054, PR China

ARTICLE INFO

Article history:
Received 31 August 2015
Revised 28 March 2016
Accepted 2 April 2016
Available online 8 April 2016

Keywords:
Granular computing
Information fusion
Multi-source information system
Triangular fuzzy granule

ABSTRACT

The advent of Big Data has seen both the sources and volumes of data increase rapidly. A multi-source information system can be used to represent information drawn from multiple sources. However, some of these sources are of less importance than others, and some are essentially worthless. Selecting the most valuable sources and efficiently fusing information are therefore core issues in the field of data science. To investigate this matter, we first propose internal-confidence and external-confidence degrees to estimate the reliability of each information source within a multi-source information system. A source selection principle is then constructed, allowing worthy and reliable information sources to be chosen. Furthermore, a new information fusion method is constructed by transforming the original information of each object into a triangular fuzzy information granule, and some uncertainty measures of this fusion process are studied. Finally, to interpret and comprehend the proposed theories and algorithm, extensive experiments are performed on six datasets to verify that our approach can deal with practical issues. The results indicate that the proposed triangular fuzzy granule fusion approach is efficient and effective for information fusion in multi-source datasets.

1. Introduction

With the continued growth of the Web 2.0 paradigm, the amount of freely available, user-generated data has reached an unprecedented volume. A number of information processing and data mining methods have been proposed to capitalize on the opportunities offered by this massive amount of data. The concept of Big Data, which was first identified in a *Nature* article in September 2008 [5], usually refers to massive, high-speed, and diverse information resources. Big Data is generated by everything around us at all times—produced by every digital process and social media exchange, and transmitted by all systems, sensors, and mobile devices. Big Data are produced by multiple sources with alarming velocity, volume, and variety.

In the Big Data age, the information technology industry, academia, and governments are making concerted efforts to determine the value of the data resources. To better respond to the challenges posed by this surfeit of data, joint efforts are required by researchers in the fields of storage technology, Web 2.0, processors, and computing paradigms [50]. Although studies on Big Data are currently very popular, different research fields tend to focus on different characteristics and processing methods. In the information age, complex data are often represented by multi-source information systems [12], in which data are taken from different information sources. Thus, multi-source information systems are constructed from

* Corresponding author. Tel.: +86 15998921583.
E-mail addresses: chxuwh@gmail.com (W. Xu), yujh2013@foxmail.com (J. Yu).

http://dx.doi.org/10.1016/j.ins.2016.04.009
0020-0255/

multi-source datasets, and can be regarded as one cause of the massive amount of data. It is therefore a challenge to identify which method is most effective in the processing of multi-source information systems or multi-source datasets. Chen and Zhang [9] argued that granular computing (GrC), a rapidly developing information processing paradigm [55], is the most useful method for Big Data research. The feasibility and advantages of applying GrC, as well as key problems in the GrC-based Big Data processing framework, have been investigated by Xu et al. [50]. In this study, we will focus on multi-source information fusion from the viewpoint of GrC.

The concept of GrC for information processing is based on Zadeh's "information granularity," which refers to theories, methodologies, techniques, and tools that make use of granules in the process of problem solving. First proposed in 1996 [64], information granularity is an important component of artificial intelligence and information processing [65,66]. GrC identifies the essential commonalities between the surprisingly diversified problems and technologies used in these fields, which could be cast into a unified framework known as a granular world. The outcome of GrC is achieved through the interaction of information granules and the external world at a granular or numeric level by collecting the necessary information granules [49]. An information granule is a clump of objects drawn together by their indiscernibility, similarity, and proximity of functionality [29,30]. Yu and Pedrycz regarded fuzzy sets as information granules [63], and employed the relationships among them to find solutions to various problems. Different granularity levels, which refer to the size of the information granules, are formed by decomposing and reforming granules. Rough set theory (RST) is another powerful mathematical tool for dealing with inexact, uncertain, or vague information [26,27]. The basic composition of RST relies on equivalence classes, namely, information granules. In recent years, the study of GrC and RST has progressed and a number of results have been reported. Qian et al. extended Pawlak's single-granulation case to a multigranulation rough set model [35,38,39] in terms of GrC. Many researchers have extended this multigranulation rough set. For example, Xu et al. developed a fuzzy multi-granulation rough set model [51], a generalized multigranulation rough set approach [52], and a multigranulation rough set model for ordered information systems [53]. Yang et al. studied the hierarchical structure properties of the multigranulation rough set [58] and considered multigranulation rough sets in incomplete information systems [59]. Lin et al. presented a neighborhood-based multigranulation rough set [16], while She and He explored the topological structures and properties of multigranulation rough sets [42]. Studies on the combination of GrC and RST are continuing [2,68]. GrC has become an effective framework for the design and implementation of efficient and intelligent information processing systems for various real-life decision-making applications [28]. Liu et al. investigated how the variation of attributes affects the granularity of the knowledge space, and proposed an incremental updating approximations algorithm for probabilistic rough sets [21]. The same researchers later utilized a hybrid information table to deal with a novel three-way decision model [22]. Liang systematically studied these three-way decision-theoretic rough sets [23,24], and Salehi et al. conducted a systematic mapping study of GrC [44]. GrC has proved to be very efficient for concept formation [56], data mining [1], and knowledge discovery [57]. Pedrycz et al. proposed the characterization of numeric data using a collection of information granules, enabling the key structure, topology, and essential relationships of the data to be described in the form of a family of fuzzy sets [33], and formulated a general framework of information granules for the description of data [34].

Multi-source information systems are used to represent information that comes from multiple sources. They consist of multiple datasets. A single-source information system is a special case of a multi-source information system [12]. Note that the "multi-source" is different from "multigranulation." A multi-source information system includes a family of single information systems with the same domain, whereas multigranulation refers to a single information system with multiple granular structures based on different granularities. According to the granulation approach, the objects in a multi-source information system can be granulated into multiple granular structures induced by a family of binary relations or a family of attribute sets. In each information subsystem, the objects are organized into a granular structure by an attribute set. It is natural to consider the fundamental issue of combining multiple granular structures from a multi-source information system [17]. Various studies have focused on communication between information systems [18,46]. Wang et al. investigated the homomorphisms between fuzzy information systems based upon the concepts of consistent functions and fuzzy relation mappings [47]. Following Wang's work, Zhu classified consistent functions as predecessor-consistent and successor-consistent, and then proceeded to present more properties of consistent functions [71]. Recently, Wang conducted research into fuzzy information systems and their homomorphisms [48]. Li utilized the fuzzy entropy fuzzy entropy to compress the feature vectors of a segmental point dataset to enhance the sensor models generalization power, and a decorrelated neural-net ensemble (DNNE) with random weights is employed to build the soft sensor [13]. Saoud's investigation improve both the source selection and the result merging process in distributed information retrieval systems [41]. In addition, for a multi-source information system that includes s single information sources, the overlapping sources can be considered to form an information box containing s levels [62].

Information fusion (initially called data fusion) originated in the processing of sonar signal systems in the United States in 1973. Information fusion technology is popular in such military applications. In particular, multi-sensor data fusion (MSDF) has occupied a pivotal position in the military field since the 1980s. In recent years, various information fusion techniques have become increasingly effective. Typical information fusion problems involve the integration of multi-source information for signal processing, image processing, knowledge representation, and inference, and these areas have been the objective of considerable research over recent years. Ma researched formation drillability prediction based on multi-source information fusion [25], and Cai et al. investigated multi-source information fusion for the fault diagnosis of ground-source heat pumps using Bayesian networks [8]. Ribeiro et al. studied an algorithm for information fusion that includes concepts such as fuzzy multi-criteria decision-making and mixture aggregation operators with weighting functions [3]. Lin et al. studied an

information fusion approach by combining multigranular rough sets and evidence theory [17]. Wang built occupant level estimation based on heterogeneous information fusion [45], Karim studied the data fusion in universal domain using dual semantic code [11]. And many new methods have been combined with information fusion, Yang investigated a driver fatigue recognition model based on information fusion and dynamic Bayesian network [61], Banerjee used the support vector machine to study the information fusion [7], Li estimated the effectiveness of Bayesian filters based on an information fusion perspective [14]. Li et al. researched the evidence supporting measure of similarity for reducing the complexity in information fusion [15]. Yager presented a general view of the multi-source data fusion process and described some of the considerations and information that must go into the development of a multi-source data fusion algorithm [54]. Recently, Balazs and Velásquez conducted a systematic study on opinion mining and information fusion [6]. Prior to this, we discussed information fusion in multi-source fuzzy information systems with the same structure, and proposed two information fusion approaches [62].

Although considerable progress has been achieved in the research and application of information fusion, little attention has been paid to information fusion based on GrC. As the scale and sources of data grow with the advent of the Big Data era, GrC represents a very efficient tool for data mining and information processing. We should therefore study multi-source information fusion based on GrC. Hence, in this paper, we propose the internal-confidence and external-confidence degrees to describe each single source in a multi-source information system. Based on these, a principle for information source selection is constructed, allowing the selected sources to be fused into a complex source in which every element is a triangular fuzzy information granule.

The remainder of this paper is organized as follows. In Section 2, some basic concepts of information systems, rough sets, and uncertainty measures are introduced. In Section 3, we study the validity of information sources and build the source selection criteria, then propose some principles of information fusion based on the selected information sources. To verify the information fusion and uncertainty measure output, Section 4 presents the results of a series of experiments carried out on several datasets. The paper ends with our conclusions in Section 5.

2. Preliminaries

For convenience, this section outlines some basic concepts of rough sets, multi-source information systems, and fuzzy sets. Further details can be found in the references cited in this section.

2.1. Rough sets and multi-source information systems

An information system is the basic description of some expression of information. In general, an information system [26] is defined as a quadruple $I = (U, AT, V, f)$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a finite non-empty set of objects (the universe of discourse), AT is a finite non-empty set of attributes, and V is the set of attribute values. For every attribute $a \in AT$, a set of values V_a is associated with the function f: $U \times AT \to V$ such that $f(x, a) \subseteq V_a$ for every $a \in AT$, $x \in U$. For simplicity, we usually abbreviate as $I = (U, AT, f)$. This is also called an approximation space or knowledge base. For any attribute set $A \subseteq AT$, there is an associated indiscernibility relation R_A that is defined as

$$IND(A) = \{(x, y) \in U \times U \mid \forall a \in A,\ f_a(x) = f_a(y)\} = R_A.$$

In an information system, there is generally a lot of redundant knowledge. In the process of knowledge processing, this extra knowledge produces unnecessary computation. To reduce the computational load, the theory of knowledge reduction (attribute reduction) was proposed. Let I be an information system. For any $B \subseteq A \subseteq AT$ and $a \in A$, we have the following definitions. If $IND(B) = IND(A - \{a\})$, then a is said to be not necessary, or redundant, in A; otherwise, a is a necessary attribute in A. The set of all necessary attributes in A is called the core, written as $Core(A)$. The attribute set is independent if, for any $a \in A$, a is necessary; otherwise, A is not independent. If B is an independent attribute set and $IND(A) = IND(B)$, we say B is a one-reduct of A. All reducts of A constitute the set $Red(A)$ [69,70].

According to the above definitions, one can easily determine the relationship between the core and reduct sets of A to be $Core(A) = \cap Red(A)$. The indiscernibility relation R_A, sometimes called the equivalence relation, divides the universe U into disjoint subsets. Such a partition is a quotient set of U, and is denoted by

$$U/R_A = \{[x]_{R_A} \mid x \in U\},$$

where $[x]_{R_A} = \{y \in U \mid (x, y) \in R_A\}$ is the equivalence class containing x with respect to R_A, also called the Pawlak information granule [17,40].

In view of GrC, U/R_A is a granular structure that can be represented by $K(R_A) = \{G_{R_A}(x_1), G_{R_A}(x_2), \ldots, G_{R_A}(x_n)\}$. Thus, a binary indiscernibility relation R_A is regarded as a granulation method for partitioning objects [36,37]. In particular, the finest granular structure on U is denoted as $K(\delta) = \{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}$, and the coarsest is denoted as $K(\omega) = \{\{x_1, x_2, \ldots, x_n\}\}$. Based on one information system, Pawlak [26] proposed the rough set theory in which, for any $X \in P(U)$ (where $P(U)$ represents the power set of U) representing a basic concept and an indiscernibility relation R induced by a subset of AT, one can characterize X by a pair of upper and lower approximations $\underline{R}(X) = \{x \in U \mid [x]_R \subseteq X\}$ and $\overline{R}(X) = \{x \in U \mid [x]_R \cap X \neq \emptyset\}$, respectively. Then, $pos(X) = \underline{R}(X)$, $neg(X) = \sim \overline{R}(X)$, $bn(X) = \overline{R}(X) - \underline{R}(X)$ are called the positive region, negative region, and boundary region of X.

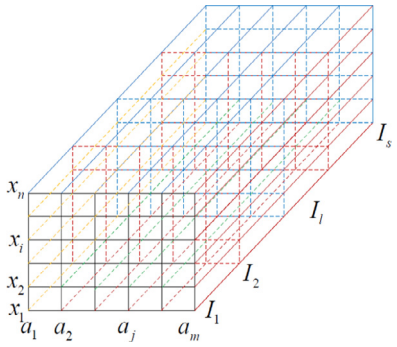

Fig. 1. A multi-source information box.

Table 1
A multi-source information system.

	I_1			I_2			I_3			I_4		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
x_1	1	2	1	2	2	1	2	2	2	2	2	1
x_2	0	2	1	1	1	2	2	2	1	2	1	0
x_3	1	1	2	1	0	1	2	2	2	2	1	0
x_4	1	1	2	1	0	1	2	2	1	2	0	1
x_5	1	1	0	1	1	2	2	2	1	2	1	0
x_6	0	2	2	0	1	1	0	1	0	2	0	1

With the rapid development of information science and technology, one can obtain information regarding a set of objects from different sources. Information from different sources is collected in the form of the information systems introduced above. Furthermore, a group of single information systems with the same domain is named a multi-source information system. This is represented as $MI = \{I_i \mid I_i = (U, AT_i, \{(V_a)_{a\in AT_i}, f_i\})\}$ in [17], where

- U is a finite non-empty set of objects;
- AT_i is a finite non-empty set of the attributes of each subsystem;
- $\{V_a\}$ is the value of the attribute $a \in AT_i$;
- $f_i : U \times AT_i \to \{(V_a)_{a\in AT_i}\}$ such that, for all $x \in U$ and $a \in AT_i$, $f_i(x, a) \in V_a$.

Let $MI = \{I_i \mid I_i = (U, AT_i, \{(V_a)_{a\in AT_i}, f_i\})\}$ be a multi-source information system that is composed of s single-source information systems, i.e., $|MI| = s$ (where $|\cdot|$ represents the cardinality of $\cdot$). In particular, if there exist some $i, j \in \{1, 2,..., s\}$, and $i \neq j$ such that $AT_i = AT_j$, then all of the information systems have the structure $MI = \{I_i \mid I_i = (U, AT, \{(V_a)_{a\in AT}, f_i\})\}$. In this study, we investigate multi-source information systems in which the sources have the same structure. If we let the s single-source information systems overlap, we can form an information box with s levels, as shown in Fig. 1 and considered in our previous study [62].

The information box is the basic structure used in the process of multi-source information fusion. More generally, there are many uncertainties in the information collection, such as equipment noise, missed data, collection time and technique, and so on. These two single information systems form a multi-source information system with a different structure. Moreover, we can employ a binary relation to granulate every single-source information system from a multi-source information system, and the s granular structures can be written as $K_1, K_2,..., K_s$. In terms of GrC, a multi-source information system can be represented by $MI = (U, K_1, K_2, \ldots, K_s)$.

Example 2.1. Let $MI = \{I_i \mid I_i = (U, AT, \{(V_a)_{a\in AT}, f_i\})\}$ be a multi-source information system, where the universe $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ consists of six patients. Suppose there are four hospitals ($I_i, i = 1, 2, 3, 4$) that provide information regarding attributes $a_j, j = 1, 2, 3$ of these patients, and a_{ij} denote three physical examination indicators. Every hospital can be regarded as an information source in this multi-source information system. The information provided by the four hospitals is presented in Table 1.

From this table, it is easy to calculate the granular structure for each information source:

$$K_1 = \{\{x_1\}, \{x_2\}, \{x_3, x_4\}, \{x_5\}, \{x_6\}\}, \quad K_2 = \{\{x_1\}, \{x_2, x_5\}, \{x_3, x_4\}, \{x_6\}\},$$
$$K_3 = \{\{x_1, x_3\}, \{x_2, x_4, x_5\}, \{x_6\}\}, \quad K_4 = \{\{x_1\}, \{x_2, x_3, x_5\}, \{x_4, x_6\}\}.$$

Therefore, the multi-source information system can be described as $MI = (U, K_1, K_2, K_3, K_4)$. In addition, we can compute the granular structure induced by all attributes, which is $K = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\} = K(\delta)$. We will utilize these granular structures in the next section.

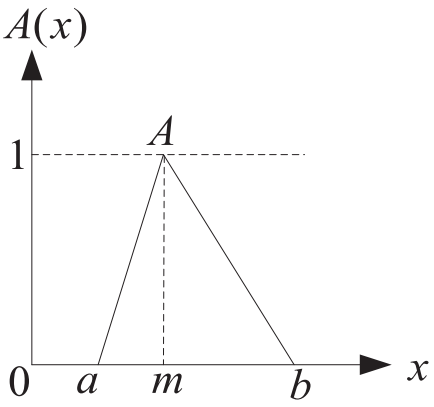

Fig. 2. A general triangular fuzzy number.

2.2. Fuzzy sets and triangular fuzzy numbers

Since Zadeh [67] first proposed the idea of fuzzy sets, in which partial membership of belonging to one or other sets is described by a membership function, fuzzy logic has made outstanding contributions to many fields. Information granules and information granulation [65] are intuitively appealing concepts that play an important role in human cognition, processing, and communication. The information granules are generic conceptual and computing objects of GrC. Information granulation is the basis of GrC.

A fuzzy set A on U is defined as a function assigning some value $A(x) \in [0, 1]$ to each element x of U. $A(x)$ is referred to as the membership degree of x with respect to the fuzzy set A. It can be described as follows:

$$A = \{< x, A(x) > \ | x \in U\}.$$

The support set of a fuzzy set A is defined as $supp(A) = \{x \in U | A(x) > 0\}$. For any $A, B \in F(U)$, we say that A is contained in B, denoted by $A \subseteq B$, if $A(x) \leq B(x)$ for all $x \in U$, and we say that $A = B$ if and only if $A \subseteq B$ and $A \supseteq B$. Given $A, B \in F(U)$ and $\forall x \in U$, the union and intersection of A and B are defined as:

$$(A \cup B)(x) = A(x) \vee B(x);$$
$$(A \cap B)(x) = A(x) \wedge B(x).$$

where $\vee$ and $\wedge$ denote the maximum and minimum operations, respectively. Triangular, Gauss, and trapezoidal fuzzy numbers are used in many energy fields. The triangular fuzzy number $A = (a, m, b)$ will be used later in this study (see Fig. 2, where a and b are the left and right boundary points, respectively, and m is the core of the triangular fuzzy number A).

Let $A = (a_1, m_1, b_1)$ and $B = (a_2, m_2, b_2)$ be two arbitrary triangular fuzzy numbers. The Hathaway distance [10] $d_h(A, B)$ of symmetric triangular fuzzy numbers A and B is defined as

$$d_h^2(A, B) = (a_1 - a_2)^2 + (m_1 - m_2)^2 + (b_1 - b_2)^2.$$

Yang et al. [60] identified a distance that can be defined for all LR-type fuzzy numbers. Let L (and R) be decreasing shape functions from $\mathbb{R}^+$ to [0, 1] with $L(0) = 1$, $L(x) < 1$ for all $x > 0$, $L(x) > 0$ for all $x < 1$, and $L(1) = 0$ (or $L(x) > 0$ for all x, $L(+\infty) = 0$). A fuzzy number A with the membership function

$$A(x) = \begin{cases} L\left(\dfrac{m_1 - x}{\alpha}\right), & \text{for } x < m; \\ 1, & \text{for } x = m; \\ R\left(\dfrac{x - m_2}{\beta}\right), & \text{for } x > m; \end{cases}$$

is called an LR-type triangular fuzzy number. Symbolically, A is denoted by $A = (a, m, b)_{LR}$, where $a > 0$ and $b > 0$ are called the left and right spreads, respectively. Given $A = (a_1, m_1, b_1)_{LR}$ and $A = (a_2, m_2, b_2)_{LR}$, Yang et al. defined the distance $d_{LR}(A, B)$ as $d_{LR}^2(A, B) = (m_1 - m_2)^2 + ((m_1 - la_1) - (m_2 - la_2))^2 + ((m_1 + rb_1) - (m_2 + rb_2))^2$, where $l = \int_0^1 L^{-1}(w)dw$ and $r = \int_0^1 R^{-1}(w)dw$. If L and R are linear, then $l = r = \frac{1}{2}$.

Pedrycz and Yu [31,32,63] introduced a general two-phase procedure for building information granules. In the first phase, a collection of segments of numeric data is used to establish a certain level of specificity when looking at the data. In the second phase, a granular representation of the data falling within the individual segments is formed. The fuzzy information granulation [59,63] aims to construct a fuzzy number on a given dataset D that includes s elements $x_1, x_2, \ldots, x_s$, where $a_i \in \mathbf{R}$. To build the triangular fuzzy information granules (fuzzy set A) on the given dataset, the following two aspects should be taken into account: i) fuzzy information granule A should embrace a sufficient amount of experimental data; ii)

fuzzy information granule A should be specific enough, which is accomplished by keeping its support as compact as possible. Based on these requirements, the following optimization problem can be obtained:

$$max \quad Q(A) = \frac{\sum_{x_i \in D} A(x_i)}{b-a},$$

where a and b are the left and right bounds of the support of a triangular fuzzy set A, respectively, $b-a$ is the width of the information granule, and D is the dataset including s elements. Furthermore, the parameters of the optimal fuzzy set A can be determined by the method in [4,63].

2.3. Information entropy and significance of attributes

Uncertainty is an important issue in information systems and RST. Existing measures of uncertainty include information granularity, entropy theory, and the significance degree. These measures have been successfully applied in many fields. The rough entropy and information entropy are often used as uncertainty measures in information processing.

Consider a given information system $I = (U, AT, V, f)$. For any $A \in AT$, $U/R_A = \{X_1, X_2, \ldots, X_m\}$, and the rough entropy of A is defined as [19,20]:

$$E_r(A) = -\sum_{i=1}^{m} \frac{|X_i|}{|U|} log_2 \frac{1}{|X_i|}.$$

It is obvious that there are minimum and maximum values of the rough entropy. When $U/R_A = K(\omega)$, the rough entropy of A has a minimum of 0; if $U/R_A = K(\delta)$, then the rough entropy of A has a maximum value of $log_2^{|U|}$. If there is some probability distribution $P_i = |X_i|/|U|$ on U, then the information entropy of information system I with respect to A can be written as:

$$H(A) = -\sum_{i=1}^{m} P_i log_2^{P_i}.$$

In particular, if there is some $P_i = 0$, we can set $0 \cdot log_2^0 = 0$. The entropy is a measure of the disorder of the system: the greater the entropy, the higher the disorder. To measure the uncertainty in the structure of the information system, Shannon took the concept of entropy in physics and applied it to information theory [43]. The Shannon entropy is related to the rough entropy by the expression $H(A) + E_r(A) = log_2^{|U|}$.

Each information system has many attributes, but some of these are redundant. To measure the significance of a single attribute, the relative and absolute significance of attributes have been developed [69,70]. Let $I = (U, AT, V, f)$ be an information system. For any $A \in AT$ and $a \in A$, $b \in (AT - A)$, the relative and absolute significance of attribute a in attribute set A are respectively defined as

$$Sig_{in}(a, A) = E_r(A - \{a\}) - E_r(A),$$
$$Sig_{out}(b, A) = E_r(A) - E_r(A \cup \{b\}).$$

In particular, when $A = \{a\}$, we have that $Sig_{in}(a, \{a\}) = E_r(\emptyset) - E_r(a) = |U| log_2^{|U|} - E_r(\{a\})$. According to this definition, the following properties hold for $Sig_{in}(a, A)$: 1) $0 \le Sig_{in}(a, A) \le |U| log_2^{|U|}$, 2) attribute a is necessary if and only if $Sig_{in}(a, A) > 0$, 3) $Core(A) = \{a \in A \mid Sig_{in}(a, A) > 0\}$.

For a given information system I, any attribute subset $A \subseteq AT$ and A is a reduct of information system I if $E_r(A) = E_r(AT)$ and, for any $a \in A$, $Sig_{in}(a, A) > 0$. These concepts are commonly used to study attribute reduction, and many useful heuristic algorithms for this purpose have been proposed for various types of information systems.

3. Granular computing approach to information fusion in multi-source datasets

The rapid development of information science and technology has given rise to an unprecedented volume of freely available, user-generated data. It is impossible for humans to make sense of the overall picture in a reasonable amount of time. Hence, efficient data mining has become a dominant theme within information science. There are numerous means of obtaining data, and the number of information sources is increasing sharply. At the same time, a large amount of this data is unreliable or invalid in real-life applications. The selection of reliable information sources is a key issue in the field of information technology research, and one that can greatly improve the efficiency of information processing. Consequently, we propose two numerical characteristics that measure the validity of information sources, and discuss some of their important properties. We then describe how to find an efficient information fusion method that takes advantage of the selected information sources, with multi-source information fusion investigated in detail in terms of GrC.

3.1. Selection of information sources

When identifying information sources, there are many unpredictable factors that affect the validity, authenticity, and reliability of the data. The accuracy of information acquired by physical instruments is restricted by the accuracy of the instrument, and will often contain some noise from the data collection process. Moreover, in the network transmission process (especially wireless network transmission), the accuracy of the information is affected by factors such as bandwidth, transmission delay, and energy. In some special fields where raw data cannot be obtained, such as biomedicine, military, aerospace, and other key areas of science and technology, approximated data are instead collected. Furthermore, information can be lost while the data are in storage. These factors act to decrease the validity of the acquired information source. To characterize the effectiveness of an information source, we define the two source quality metrics of the internal-confidence degree (*IC*) and external-confidence degree (*EC*). These metrics represent the credibility of the information source itself and the degree of mutual support between the sources in a multi-source information system.

A single information source is often composed of a number of items. For example, a dataset from a medical diagnosis may include a lot of physical examination items such as vital capacity, blood pressure, glucose level, and lipid level. For the rational use of medical resources and reduced stress for patients, unnecessary diagnoses should be removed. Suppose an attribute is given by the information source. Then, we can use the idea of attribute reduction from RST. For a multi-source information system, the reduction is $Red(MI) = \{Red(AT_1), Red(AT_2), \ldots, Red(AT_s)\}$. We define a measurement that characterizes the credibility of a source or the degree of reasonableness of a particular item. This is the internal-confidence (or self-confidence) degree, and is determined by the source itself.

Definition 3.1. Let $MI = \{I_i \mid I_i = (U, AT_i, \{(V_a)_{a\in AT_i}, f_i\})\}$ be a multi-source information system. For each single information source $I_i \in MI$, let $Red(AT_i)$ be the reduction of I_i. The internal-confidence degree of I_i can then be defined as follows:

$$IC(I_i) = \frac{|Red(AT_i)|}{|AT_i|}.$$

From the above definition, it is obvious that $0 \le IC(I_i) \le 1$, and that $IC(I_i)$ is the ratio of the cardinalities of $Red(AT_i)$ and AT_i. If $IC(I_i) > 0.5$, the majority of the attributes are useful and the source is reliable. In practical applications, we set different thresholds $IC(I_i) > \alpha$ in different fields according to the specific requirements. *IC* is an absolute quantity that is used to characterize one source in a multi-source information system, and is determined by a single information source's attributes and its reduct. There is no relationship between the *IC* values of each source in the multi-source information system. Based on the idea of the significance of attributes and Definition 3.1, we developed a heuristic algorithm for attribute reduction in multi-source information systems (see Algorithm 1).

Algorithm 1: Heuristic algorithm for computing the internal-confidence degree for each I_i in *MI*.

Input : A multi-source information system $MI = \{I_i \mid I_i = (U, AT_i, \{(V_a)_{a\in AT_i}, f_i\})\}$ consisting of $|MI| = s$ information sources.

Output : The set reduced attributes $Red(MI) = \{Red(AT_1), Red(AT_2), \ldots, Red(AT_s)\}$.

```
1  begin
2    for i = 1; i <= s; i++ do
       Initialize: Red(AT_i) ⟵ ∅ ;
3      for j = 1; j <= |AT_i|; j++ do
         Compute: Sig_in(a_j, AT_i);        // compute the absolute significance of a_j, as discussed in the Preliminaries;
4        if Sig_in(a_j, AT_i) > 0 then
5          Red(AT_i) = Red(AT_i) ∪ {a_j};   // update the reduction by entropy and significance of attribute;
6        end
7      end
8      if E_r(Red(AT)) ≠ E_r(AT_i) then
9        for each a ∈ AT_i − Red(AT_i) do
           Compute: a* = {a ∈ AT_i − Red(AT_i) | Sig_out(a*, Red(AT_i)) = max{Sig_out(a, Red(AT_i))}};
10       end
11       Red(AT_i) = Red(AT_i) ∪ {a*} then
12       Goto line 8;
13     else
         Compute: IC(I_i) = |Red(AT_i)|/|AT_i|;        // calculate IC(I_i) using Definition 3.1.
14       Return: The internal-confidence degree for each I_i ∈ MI.
15     end
16   end
17 end
```

Table 2
Internal-confidence degrees for each I_i.

	$E_r(a_{AT_i})$	$sig(a_{i1})$	$sig(a_{i2})$	$sig(a_{i3})$	$Red(AT_i)$	$IC(I_i)$
I_1	0.333	0.333	0	0.792	$\{a_{11}, a_{13}\}$	0.667
I_2	0.667	0	0	0	$\{a_{22}, a_{23}\}$	0.667
I_3	1.126	0	0	0.809	$\{a_{33}\}$	0.333
I_4	1.126	0	0.459	0	$\{a_{42}\}$	0.333

Example 3.1. (continued from Example 2.1). We now use the results of the previous calculation in Example 2.1. Algorithm 1 can be used to compute the internal-confidence degree $IC(I_i)$ for each $I_i \in MI$. We use the example of information source I_1; the calculation process for other sources is similar.

$$E_r(AT_1) = \frac{1}{6} \cdot log_2^1 + \frac{1}{6} \cdot log_2^1 + \frac{2}{6} \cdot log_2^2 + \frac{1}{6} \cdot log_2^1 + \frac{1}{6} \cdot log_2^1 = 0.333.$$

After deleting redundant attributes, the granular structure of I_1 is as follows:

$$U/(AT_1 - \{a_{11}\}) = \{\{x_1, x_2\}, \{x_3, x_4\}, x_5, x_6\};$$
$$U/(AT_1 - \{a_{12}\}) = \{\{x_1\}, \{x_2\}, \{x_3, x_4\}, x_5, x_6\};$$
$$U/(AT_1 - \{a_{13}\}) = \{\{x_1\}, \{x_2, x_6\}, \{x_3, x_4, x_5\}\}.$$

Hence, the absolute significance of each attribute $a_{1j} \in AT_1$ is $Sig_{in}(a_{11}) = 0.333 > 0$, $Sig_{in}(a_{12}) = 0$, $Sig_{in}(a_{13}) = 0.792$. Thus, the reduct of $Red(AT_1)$ is $\{a_{11}, a_{13}\}$. We also verify the equivalent entropy between $\{a_{11}, a_{13}\}$ and AT_1:

$$E_r(a_{11}, a_{13}) = E_r(AT_1) = 0.333.$$

Then, the minimal reduction of information source I_1 is $Red(AT_1) = \{a_{11}, a_{13}\}$, and the internal-confidence degree of I_1 is

$$IC(I_1) = \frac{|Red(AT_1)|}{|AT_1|} = 0.667.$$

The internal-confidence degrees of all $I_i \in MI$ are presented in Table 2.

After the attribute reduction and internal-confidence degree have been computed, each source's internal features can be captured. For effective information fusion, we further study the relationship between individual sources and explain the calculation of the difference and external-confidence degree among different sources.

Definition 3.2. Let $MI = \{I_i \mid I_i = (U, AT_i, \{(V_a)_{a \in AT_i}, f_i\})\}$ be a multi-source information system. For any two single information sources I_i and I_j in MI, the difference between them can be described as:

$$D(I_i, I_j) = \sum_{k=1}^{|U|} (|[x_k]_{AT_i} \cup [x_k]_{AT_j}| - |[x_k]_{AT_i} \cap [x_k]_{AT_j}|),$$

where $[x_k]_{AT_i}$ is an object set that is equivalent to x_k with respect to the attribute set AT (the equivalence class containing x_k with respect to R_{AT}).

For the difference between two information sources, we can state the following lemmas.

- For any $I_i, I_j \in MI$, $D(I_i, I_j) \geq 0$;
- For any $I_i, I_j \in MI$, $D(I_i, I_j) = D(I_j, I_i)$;
- Let I_i, I_j, I_k be three different information sources with attribute sets AT_i, AT_j and AT_k, respectively. If the relationship $AT_i \subseteq AT_j \subseteq AT_k$ holds, then $D(I_i, I_j) + D(I_j, I_k) = D(I_i, I_k)$.

Proof. Lemmas (1) and (2) can be proved directly by the above definition. Thus, we discuss Lemma (3). Because $AT_i \subseteq AT_j \subseteq AT_k$, then, for any $x \in U$, we have $[x]_{AT_k} \subseteq [x]_{AT_j} \subseteq [x]_{AT_i}$. Hence,

$$\begin{aligned} D(I_i, I_j) + D(I_j, I_k) &= \sum_{l=1}^{|U|}(|[x_k]_{AT_i} \cup [x_k]_{AT_j}| - |[x_k]_{AT_i} \cap [x_k]_{AT_j}|) + \sum_{l=1}^{|U|}(|[x_k]_{AT_j} \cup [x_k]_{AT_k}| - |[x_k]_{AT_j} \cap [x_k]_{AT_k}|) \\ &= \sum_{l=1}^{|U|}(|[x]_{AT_i}| - |[x]_{AT_j}|) + \sum_{l=1}^{|U|}(|[x]_{AT_j}| - |[x]_{AT_k}|) \\ &= \sum_{l=1}^{|U|}(|[x]_{AT_i}| - |[x]_{AT_k}|) \\ &= \sum_{l=1}^{|U|}(|[x_k]_{AT_i} \cup [x_k]_{AT_k}| - |[x_k]_{AT_i} \cap [x_k]_{AT_k}|) \\ &= D(I_i, I_k). \end{aligned}$$

This completes the proof. □

According to this definition, the minimum value of $D(I_i, I_j)$ is 0, and some maximum exists. If the single information source I_i has the finest granular structure $K(\delta) = \{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}$ and I_j has the coarsest structure $K(\omega) = \{\{x_1, x_2, \ldots, x_n\}\}$, then the difference between I_i and I_j reaches a maximum of $|U|^2 - 1$. Based on the definition of the difference between I_i and I_j, we can define the degree of association of information sources, namely, the external-confidence degree.

Definition 3.3. Let $MI = \{I_i \mid I_i = (U, AT_i, \{(V_a)_{a \in AT_i}, f_i\})\}$ be a multi-source information system that includes s single information sources. For any two single information sources I_i, $I_j \in MI$, the external correlation between I_i and I_j can be defined as:

$$ec(I_i, I_j) = 1 - \frac{D(I_i, I_j)}{|U|^2 - 1},$$

where $D(I_i, D_j)$ is the difference between I_i and I_j as calculated by Definition 3.2. From this definition, we can see that $ec(I_i, I_j) \in [0, 1]$, $ec(I_i, I_j) = ec(I_j, I_i)$, and $ec(I_i, I_i) = 1$. This metric can be used to describe the correlation between two sources in a multi-source information system. Furthermore, to demonstrate the external-confidence degree of a source in the entire information system, we define the ensemble external-confidence degree of I_i in MI as:

$$EC(I_i) = \frac{1}{s}\sum_{j=1}^{s}\left(1 - \frac{D(I_i, I_j)}{|U|^2 - 1}\right) = \frac{1}{s}\sum_{j=1}^{s} ec(I_i, I_j).$$

Similar to the internal-confidence degree, different thresholds $EC(I_i) > \beta$ may be applicable in different fields according to specific requirements. EC is a relative quantity that is used to characterize one source in a multi-source information system, and is determined by all of the information sources. To clarify the relationship between the external-confidence degree of I_i and I_j, after calculating all EC for all $I_i \in MI$, we can construct an external-confidence degree matrix M_{EC} as

$$M_{EC} = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1s} \\ e_{21} & e_{22} & \cdots & e_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ e_{s1} & e_{s2} & \cdots & e_{ss} \end{pmatrix}$$

where $e_{ij} = ec(I_i, I_j)$. This is an s-dimensional symmetric matrix, and each value on the main diagonal is 1. The sum of the ith row (or column) is $EC(I_i) = \frac{1}{s}\sum_j^s e_{ij}$.

Example 3.2. (continued from Example 2.1). This example demonstrates how to calculate the external-confidence degree for each $I_i \in MI$. From the results of Example 2.1 and Definition 3.3, the difference between I_i and I_j can be calculated as $D(I_1, I_2) = 2$, $D(I_1, I_3) = 10$, $D(I_1, I_4) = 10$, $D(I_2, I_3) = 8$, $D(I_2, I_4) = 8$, $D(I_3, I_4) = 12$. The external-confidence degree matrix M_{EC} can then be computed using Definition 3.2. The result is as follows:

$$M_{EC} = \begin{pmatrix} 1 & 0.943 & 0.714 & 0.714 \\ 0.943 & 1 & 0.771 & 0.771 \\ 0.714 & 0.771 & 1 & 0.657 \\ 0.714 & 0.771 & 0.657 & 1 \end{pmatrix}$$

Therefore, the external-confidence degrees for each $I_i \in MI\ (i = 1, 2, 3, 4)$ are $EC(I_i) = \{0.843, 0.871, 0.786, 0.786\}$, respectively.

We have defined *IC* and *EC*, two measures that represent the absolute and relative confidence in an information source. Applying these two indexes can lead to a different order in which we integrate information sources. The calculated results of Examples 3.1 and 3.2 suggest that information sources $I_1 \succeq_{IC} I_2 \succ_{IC} I_3 \succeq_{IC} I_4$ and $I_2 \succeq_{EC} I_1 \succ_{EC} I_3 \succeq_{EC} I_4$, respectively. Thus, establishing a universal source selection rule is a key issue in the study of multi-source information systems. We now construct scoring criteria for each source in a multi-source information system. According to Definitions 3.1 and 3.3, we can obtain *IC* and *EC* for any information source $I_i \in MI$. We can sort the sources according to these values, giving two vectors that represent the order of the sources. Furthermore, to describe the quality of a source more intuitively, we define a total score for each information source $I_i \in MI$ as follows:

$$TotalScore(I_i) = IC(I_i) + EC(I_i).$$

The total score for each information source in Table 1 can be calculated as $TotalScore(I_i) = (1.51, 1.538, 1.119, 1.119)$. Therefore, the quality of the information sources runs in the order $I_2 \succ I_1 \succ I_3 \succeq I_4$, where $I_2 \succ I_1$ denotes that information source I_2 is definitely better than I_1 and $I_3 \succeq I_4$ denotes that I_3 is not worse than I_4. Based on this source selection rule, high-credibility sources can be selected from a multi-source information system. Next, we discuss how to fuse information from the selected sources in view of GrC.

3.2. Granular computing approach to information fusion in multi-source datasets

After completing the information source selection procedure, a number of reliable sources have been identified. There are many approaches for information fusion. In this subsection, we focus on a fusion approach based on the fuzzy granulation of information. We use the selected source to build fuzzy information granules for each $x \in U$. In fact, multi-source information fusion essentially involves processing multiple single information sources. In this study, fuzzy information granules are established to replace the object descriptions in the multi-source environment. Thus, our multi-source information fusion approach constructs a fuzzy information table instead of the original information box, and each point of this table is a fuzzy number.

Let $MI = \{I_i \mid I_i = (U, AT_i, \{(V_a)_{a\in AT_i}, f_i\}), i = 1, 2, \ldots, s\}$ be a multi-source information system consisting of s single information sources. All of the information sources are selected by the rules described in the previous subsection. For any $x \in U$, the value of x under any one attribute $c \in AT$ can be described as $c(x) = (c^1(x), c^2(x), \ldots, c^s(x))$. The parameters of the optimal fuzzy set A can be determined by the method in [63]. The core of the information granules consists of those elements of A that are typical of the concept conveyed by the information granule. In a triangular fuzzy set, the core of A is formed by those m that minimize the sum of the absolute differences $\sum_{i=1}^{s} |m - c^i(x)|$. The solution to this optimization is a median point if s is odd, or the result is an interval from $a^{s/2}(x)$ to $c^{s/2+1}(x)$ if s is even. Consider the description of the core based on the following characterization:

- If s is odd, then m is equal to $c^i(x)$ $(m = c^i(x)$, where i satisfies $\sum_{j=1}^{i} s_j < s/2 < \sum_{j=1}^{i+1} s_j$;
- If s is even, then m is a real number between $a^i(x)$ and $c^{i+1}(x)$, where i satisfies $\sum_{j=1}^{i} s_j = k/2$, and $m = \frac{c^i(x)+c^{i+1}(x)}{2}$.

Once the modal point or core of A has been determined, the task of building a spread splits into two independent subproblems concerning the left and right spreads. These can be written as

$$\max\ Q(a) = \frac{\sum_{i=1}^{k_1} A(a^i(x)}{m-a},\ \text{and}\ \max\ Q(b) = \frac{\sum_{i=1}^{k_2} A(a^i(x)}{b-m}$$

for the left and right sections of A, respectively. Furthermore, because $A(x)$ is a triangular membership function, $Q(a)$ and $Q(b)$ attain maxima at $a = \frac{2p_1}{k_1} - m$, $b = \frac{2p_2}{k_2} - m$, where $p_1 = \sum_{i=1}^{k_1} x_i$ and $p_2 = \sum_{i=1}^{k_2} x_i$, and k_1, k_2 denote the number of data points located left and right of the modal value, respectively. Therefore, a fuzzy number can be obtained according to the data series $a_j(x) = (a_{j1}(x), a_{j2}(x), \ldots, a_{js}(x))$. Taking this transform operation for all attributes and each $x \in U$, the information box can be transformed into an information table, as shown in Fig. 3.

The small red triangle in Fig. 3 represents the attribute value of the i-th object and j-th attribute in the universe U. This is a fuzzy number of the form $A_{ij} = (a_{ij}, m_{ij}, b_{ij})$. Next, we define the distance and similarity degree between two objects, the rough entropy and information entropy in the fusion information system.

Definition 3.4. Let $MI = \{I_i \mid I_i = (U, AT_i, \{(V_a)_{a\in AT_i}, f_i\})\}$ be a multi-source information system. The number of attributes in each single information system is m. For any $x_k, x_l \in U$, the relative Minkowski distance d_M between x_k and x_l can be defined as:

$$d_M(x_l, x_k) = \frac{1}{m}\sum_{i=1}^{m}(|a_{li} - a_{ki}|^p + |m_{li} - m_{ki}|^p + |b_{li} - b_{ki}|^p)^{\frac{1}{p}}.$$

When $p = 2$, this is the Euclidean distance $d(x_l, x_k) = \frac{1}{m}\sum_{i=1}^{m}\sqrt{|a_{li} - a_{ki}|^2 + |m_{li} - m_{ki}|^2 + |b_{li} - b_{ki}|^2}$. According to this definition, we can easily obtain $d_M(x_l, x_k) = 0$ if and only if $a_{li} = a_{ki}$, $m_{li} = m_{ki}$ and $b_{li} = b_{ki}$ for $i = 1, 2, \ldots, m$. We use the

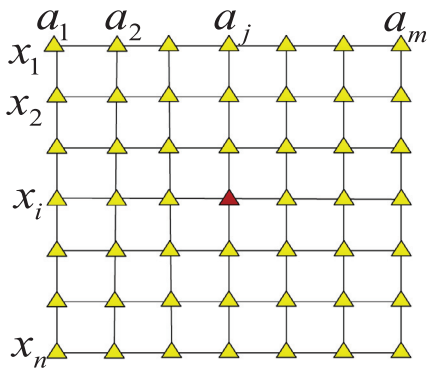

Fig. 3. Fuzzy information fusion table. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

Table 3
Result of multi-source information fusion.

	a_1	a_2	a_3
x_1	(1, 2, 2)	(2, 2, 2)	(1 ,1, 2)
x_2	(−0.5, 1.5, 2.5)	(1, 2, 2)	(0, 1, 2)
x_3	(0.5, 1.5, 2.5)	(0, 1, 2)	(−0.5, 1.5, 2.5)
x_4	(0.5, 1.5, 2.5)	(−0.5, 0.5, 2.5)	(1, 1, 2)
x_5	(0.5, 1.5, 2.5)	(1, 1, 2)	(−0.5, 0.5, 2.5)
x_6	(0, 0, 2)	(0, 1, 2)	(0, 1, 2)

Euclidean distance in this study, and define the similarity degree of x_l and x_k accordingly. For simplicity, we require the similarity degree $SI \in [0, 1]$, and so we normalize the distance as $d^*(x_l, x_k) = \frac{d(x_l, x_k)}{D}$, where $D = max(d(x_l, x_k)), k = 1, 2, \ldots, |U|$.

$$SI(x_l, x_k) = 1 - d^*(x_l, x_k).$$

It is obvious that $0 < S(x_l, x_k) \leq 1$. If $d = 0$, then x_l is equivalent to x_k. Given a threshold $\delta \in (0, 1]$, we define a new information granule with respect to the similarity degree $S(x_l, x_k)$ as:

$$[x_k]^{\delta} = \{x_l \in U \mid SI(x_l, x_k) \geq \delta\}.$$

According to this definition, these information granules constitute a covering on U. We can then define the information granularity, rough entropy, and information entropy, and take their respective uncertainly measures. For a multi-source information system $MI = \{I_i \mid I_i = (U, AT, \{(V_a)_{a \in AT}, f_i\})\}$, the rough entropy and information entropy can be obtained after the information fusion as $E_r^{\delta}(AT) = -\sum_{i=1}^{|U|} \frac{1}{|U|} log_2 \frac{1}{|[x_i]^{\delta}|}$ and $H^{\delta}(AT) = -\sum_{i=1}^{|U|} \frac{|1|}{|U|} log_2 \frac{|[x_i]^{\delta}|}{|U|}$, respectively. Similar to the classical information systems, we can easily find that $E_r^{\delta}(AT) + H^{\delta}(AT) = log_2^{|U|}$. The rough entropy and information entropy are important methods for measuring the uncertainty in a fused information system.

Example 3.3. (continued from Example 2.1). This example shows the result of multi-source information fusion. Suppose all of the sources are real and effective. The information fusion of this multi-source information system is presented in Table 3.

From Table 3, it is clear that the information fusion process results in the attribute value for each object becoming a triangular fuzzy number. Based on this table, the distance and similarity degree between two objects can be computed. Furthermore, a new binary relation between any two objects can be built. For example, the distances between x_1 and x_j are $d(x_1, x_j) = \{0, 1.22, 1.59, 1.3, 1.31, 1.82\}$, $j = 1, 2, \ldots, 6$. Thus, the similarity degrees are $SI(x_1, x_i) = \{1.00, 0.33, 0.13, 0.30, 0.28, 0\}$. The granular structure of the fusion result becomes apparent. Information processing methods such as RST can be used to analyze this information fusion table.

4. Experimental evaluations

In this section, to evaluate the performance of the proposed multi-source information fusion method, we report the results of a series of experiments based on standard datasets downloaded from the machine learning data repository of the University of California at Irvine (http://archive.ics.uci.edu/ml/datasets.html). These datasets are named “Energy efficiency,” “Airfoil Self-Noise,” “Wine Quality-red,” “Wine Quality-white,” “Spoken Arabic Digit,” and “Letter Recognition.” Some basic information about these datasets is outlined in Table 4. The final column of Table 4 gives the number of elements in the multi-source information system. These experiments were implemented using Visual C++ 6.0 and performed on a computer with an Intel Core i3-4150, 3.50 GHz CPU, 4.0 GB of memory, and 32-bit Windows 7 OS.

Table 4
Basic information about the datasets.

No.	Name	Abbreviation	Objects	Attributes	Elements of *MI*
1	*Energy efficiency*	EE	768	8	184320
2	*Airfoil Self – Noise*	AS	1503	6	270540
3	*Wine Quality – red*	WQ-r	1599	12	527670
4	*Wine Quality – white*	WQ-w	4898	12	1616340
5	*Spoken ArabicDigit*	SAD	8800	13	3432000
6	*Letter Recognition*	LR	20000	16	9600000

Table 5
Selected information sources for each multi-source information system.

Name	Selected information sources	Name	Selected information sources
EE	$I_1, I_2, I_4, I_6, I_{10}, I_{11}, I_{13}, I_{14}, I_{15}, I_{28}$	AS	$I_1, I_4, I_5, I_6, I_9, I_{14}, I_{18}, I_{25}, I_{26}, I_{27}$
WQ-r	$I_2, I_3, I_6, I_9, I_{10}, I_{12}, I_{14}, I_{19}, I_{27}, I_{29}$	WQ-w	$I_1, I_4, I_5, I_6, I_8, I_{10}, I_{14}, I_{18}, I_{29}, I_{30}$
LR	$I_1, I_3, I_7, I_{13}, I_{14}, I_{16}, I_{17}, I_{20}, I_{29}, I_{30}$	SAD	$I_5, I_8, I_9, I_{13}, I_{15}, I_{18}, I_{21}, I_{25}, I_{27}, I_{29}$

Table 6
Rough entropy and information entropy of the fusion of multi-source information systems.

δ	EE		AS		WQ-r		WQ-w		SAD		LR	
	E_r^δ	H^δ	E_r^δ	H^δ	E_r^δ	H^δ	E_r^δ	H^δ	E_r^δ	H^δ	E_r^δ	H^δ
0.55	8.141	1.444	10.245	0.309	10.502	0.141	12.237	0.022	11.194	1.909	12.446	1.843
0.60	7.920	1.665	10.192	0.362	10.449	0.194	12.220	0.038	10.890	2.213	11.640	2.648
0.65	7.823	1.762	10.095	0.458	10.378	0.265	12.188	0.070	10.477	2.626	10.628	3.660
0.70	7.617	1.968	10.008	0.546	10.280	0.363	12.126	0.132	9.866	3.237	9.391	4.896
0.75	7.376	2.209	9.869	0.684	10.137	0.506	12.011	0.247	8.938	4.165	7.923	6.364
0.80	7.209	2.376	9.629	0.924	9.930	0.713	11.807	0.451	7.541	5.562	6.276	8.012
0.85	6.563	3.022	9.423	1.131	9.604	1.039	11.450	0.808	5.516	7.587	4.482	9.805
0.90	6.000	3.585	9.051	1.503	9.025	1.618	10.778	1.480	2.905	10.198	2.631	11.656
0.95	6.000	3.585	8.234	2.319	7.682	2.961	9.105	3.153	1.198	11.905	1.302	12.985
1.00	0.000	9.585	0.000	10.554	0.311	10.332	0.418	11.840	0.823	12.280	1.101	13.186

Table 7
Rough entropy and information entropy of the original information system.

EE		AS		WQ-r		WQ-w		SAD		LR	
E_r	H	E_r	H	E_r	H	E_r	H	E_r	H	E_r	H
9.281	0.304	9.401	1.152	9.591	1.052	11.105	1.153	9.995	3.109	13.700	0.588

To build a real multi-source information system, we added noise to the original datasets and generated synthetic multi-source information systems in the following manner:

$$a_{ij}(x_k) = a_{0j}(x_k) \cdot (1 + \omega_i),$$

where $a_{ij}(x_k)$ denotes the kth object's attribute value under the jth attribute in the ith information source, $a_{0j}(x_k)$ is the value of the original source, and the disturbance ω_i obeys the normal distribution $\mathcal{N}(0, 0.1)$, $i = 1, 2, \ldots, 30$, namely, there are thirty information sources in our multi-source information system. The process of constructing the multi-source information used a single information source I_0 to produce a set including I_1, I_2,..., I_{30}, with I_0 being one of the original datasets described in Table 4. In our experiments, we chose the top ten information sources in *MI* according to *TotalScore*(I_i). The information selection methods were described in Section 3.1. The selected information sources for each multi-source information system are presented in Table 5.

According to this table, we can see that the selected information sources are distributed evenly in 1, 2,..., 30. This is because the noise ω_i follows the normal distribution $\mathcal{N}(0, 0.1)$. Based on these selected information sources for each multi-source information system, we used the proposed method to execute the information fusion process. Because the fusion results give numerous triangular fuzzy numbers for each multi-source information system, they are not shown in this document. Table 6 presents the rough entropy and information entropy computed for each fusion information system.

For comparison, the rough entropy and information entropy of the original information system were calculated (see Table 7). If the threshold is equal to 1, we can regard the systems as being equivalent.

In this experiment, we varied the threshold δ from 0.55 to 1 in intervals of 0.05. According to Table 6, as δ increases, the granularity decreases, that is, the rough entropy is decreasing but the information entropy is increasing. When $\delta < 1$, both the information entropy and rough entropy change slowly, indicating that, after the information fusion, the

multi-source information system has become a stable single composite of the system. Comparing Tables 6 and 7, we can see that the rough entropy of the fusion system is less than that of the original system, and the information entropy of the fusion system is greater than that in the original. This indicates that the granular information structure is more reasonable. In practice, it is possible to choose a different value of δ for the fusion of multi-source information system according to specific requirements.

5. Conclusions

In this study, we considered multi-source information fusion in terms of GrC. To filter unreliable and redundant information sources, we first proposed the internal-confidence and external-confidence degrees to estimate the significance of each information source. These metrics measure the absolute and relative quality of a single source, respectively. Based on these two measures, a total score was then defined for each single information source. After selecting the most appropriate sources using this score, we studied a novel approach for multi-source information fusion. By transforming the multi-source information system into an information table, the attribute value of each object was determined as a triangular fuzzy number. RST or similar methods could be used to investigate the fusion information table. In this paper, we constructed six multi-source information systems containing 30 single information sources. Based on these datasets, a series of experiments were conducted to demonstrate the effectiveness of the proposed fusion method. The results indicate that the proposed method is a useful approach for the fusion of multi-source information systems. It provides more options for information source selection and information fusion with respect to the multi-source environment. In the future, we will utilize the proposed source selection method and information fusion approach to solve some problems in practical application.

Acknowledgments

The author would like to thank the anonymous referees and Professor Witold Pedrycz, the editor in chief, for their valuable suggestions, which have improved the quality of the paper. This work is supported by the Natural Science Foundation of China (no. 61105041, no. 61472463, no. 61402064), the National Natural Science Foundation of CQ CSTC (no. cstc 2013jcyjA40051, cstc 2015jcyjA40053), the Graduate Innovation Foundation of Chongqing University of Technology (no. YCX2014236), and the Graduate Innovation Foundation of CQ (no. CYS15223).

References

[1] V.S. Ananthanarayana, M.M. Narasimha, D.K. Subramanian, Tree structure for efficient data mining using rough sets, Pattern Recognit. Lett. 24 (2003) 851–862.
[2] A. Albanese, S.K. Pal, A. Petrosino, Rough sets, kernel set, and spatiotemporal outlier detection, IEEE Trans. Knowl. Data Eng. 26 (2014) 194–207.
[3] R.A. Ribeiro, A. Falcão, A. Mora, J.M. Fonseca, FIF: a fuzzy information fusion algorithm based on multi-criteria decision making, Knowl. Based Syst. 58 (1) (2013) 23–32.
[4] A. Bargiela, W. Pedrycz, Granular Computing: An Introduction, Springer, Berlin, 2003.
[5] B. Bill, et al., Big data: the next google, interview by Duncan Graham-Rowe, Nature 455 (7209) (2008) 8–9.
[6] J.A. Balazs, J.D. Velásquez, Opinion mining and information fusion: a survey, Inf. Fusion 27 (2016) 95–110.
[7] T.P. Banerjee, S. Das, Multi-sensor data fusion using support vector machine for motor fault detection, Inf. Sci. 217 (2012) 96–107.
[8] B.P. Cai, Y.H. Liu, Q. Fan, et al., Multi-source information fusion based fault diagnosis of ground-source heat pump using Bayesian network, Appl. Energy 114 (2014) 1–9.
[9] C.L. Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on big data, Inf. Sci. 275 (2014) 314–347.
[10] R.J. Hathaway, J.C. Bezdek, W. Pedrycz, A parametric model for fusing heterogeneous fuzzy data, IEEE Trans. Fuzzy Syst. 4 (3) (1996) 270–281.
[11] M.S.A. Karim, K.S. Wong, Data fusion in universal domain using dual semantic code, Inf. Sci. 283 (2014) 123–141.
[12] M.A. Khan, M. Banerjee, Formal reasoning with rough sets in multiple-source approximation systems, Int. J. Approx. Reason 49 (2008) 466–477.
[13] W.T. Li, D.H. Wang, X.J. Zhou, T.Y. Chai, An improved multi-source based soft sensor for measuring cement free lime content, Inf. Sci. 323 (2015) 94–105.
[14] T.C. Li, J.M. Corchado, J. Bajo, S.D. Sun, J.F. De Paz, Effectiveness of Bayesian filters: an information fusion perspective, Inf. Sci. 329 (2016) 670–689.
[15] X.D. Li, J. Dezert, F. Smarandache, X.H. Huang, Evidence supporting measure of similarity for reducing the complexity in information fusion, Inf. Sci. 181 (10) (2011) 1818–1838.
[16] G.P. Lin, Y.H. Qian, J.J. Li, NMGRS: neighborhood-based multigranulation rough sets, Int. J. Approx. Reason 53 (7) (2012) 1080–1093.
[17] G.P. Lin, J.Y. Liang, Y.H. Qian, An information fusion approach by combining multigranulation rough sets and evidence theory, Inf. Sci. 314 (2015) 184–199.
[18] D.Y. Li, Y.C. Ma, Invariant characters of information systems under some homomorphisms, Inf. Sci. 129 (2000). 211–200.
[19] J.Y. Liang, Z.Z. Shi, The information entropy, rough entropy and knowledge granulation in rough set theory, Int. J. Uncertain Fuzzy Knowl. Based Syst. 12 (1) (2004) 37–46.
[20] J.Y. Liang, Y.H. Qian, Information granule and entropy theory in information system, Sci. Chin. (F) 51 (10) (2008) 1427–1444.
[21] D. Liu, T.R. Li, J.B. Zhang, Incremental updating approximations in probabilistic rough sets under the variation of attributes, Knowl. Based. Syst. 73 (2015) 81–96.
[22] D. Liu, D.C. Liang, C.C. Wang, A novel three-way decision model based on incomplete information system, Knowl. Based Syst. 91 (2016) 32–45.
[23] D.C. Liang, D. Liu, Systematic studies on three-way decisions with interval-valued decision-theoretic rough sets, Inf. Sci. 276 (2014) 186–203.
[24] D.C. Liang, D. Liu, Deriving three-way decisions from intuitionistic fuzzy decision-theoretic rough sets, Inf. Sci. 300 (2015) 28–48.
[25] H. Ma, Formation drillability prediction based on multi-source information fusion, J. Petrol. Sci. Eng. 78 (2011) 438–446.
[26] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[27] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.
[28] S.K. Pal, S.K. Meher, Natural computing: a problem solving paradigm with granular information processing, Appl. Soft Comput 13 (2013) 3944–3955.
[29] W. Pedrycz, Relational and directional aspects in the construction of information granules, IEEE Trans. Syst. Man Cybern. A Syst. Hum. 32 (5) (2002) 605–614.
[30] W. Pedrycz, A. Bargiela, Granular clustering: a granular signature of data, IEEE Trans. Syst. Man Cybern. B Cybern 32 (2) (2002) 212–224.
[31] W. Pedrycz, Knowledge-Based Clustering: From Data to Information Granules, Wiley Inter-Science, 2005.

[32] W. Pedrycz, Granular Computing: Analysis and Design of Intelligent Systems, CRC Press, Boca Raton, 2013.
[33] W. Pedrycz, A. Gacek, X.M. Wang, Clustering in augmented space of granular constraints: a study in knowledge-based clustering, Pattern Recognit. Lett. 67 (2015) 122–129.
[34] W. Pedrycz, G. Succi, A. Sillitti, J. Iljazi, Data description: a general framework of information granules, Knowl. Based Syst. 80 (2015) 98–108.
[35] Y.H. Qian, J.Y. Liang, Y.Y. Yao, C.Y. Dang, MGRS: a multi-granulation rough set, Inf. Sci. 180 (2010) 949–970.
[36] Y.H. Qian, J.Y. Liang, W.Z. Wu, C.Y. Dang, Knowledge structure, knowledge granulation and knowledge distance in a knowledge base, Int. J. Approx. Reason 50 (2009) 174–188.
[37] Y.H. Qian, J.Y. Liang, W.Z. Wu, Information granularity in fuzzy binary GRC model, IEEE Trans. Fuzzy Syst. 19 (2) (2011) 253–263.
[38] Y.H. Qian, J.Y. Liang, Rough set method based on multi-granulations, in: Proceedings of Fifth IEEE Conference on Cognitive Information, vol.1, 2006, pp. 297–304.
[39] Y.H. Qian, J.Y. Liang, C.Y. Pang, Incomplete multigranulation rough set, IEEE Trans. Syst. Man Cybern. A 20 (2010) 420–431.
[40] W.B. Qian, W.H. Shu, Mutual information criterion for feature selection from incomplete data, Neurocomputing 168 (2015) 210–220.
[41] Z. Saoud, S. Kechid, Integrating social profile to improve the source selection and the result merging process in distributed information retrieval, Inf. Sci. 336 (2016) 115–128.
[42] Y.H. She, X.L. He, On the structure of the multigranulation rough set model, Knowl. Based Syst. 36 (2012) 81–92.
[43] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423.
[44] S. Salehi, A. Selamat, H. Fujita, Systematic mapping study on granular computing, Knowl. Based Syst. 80 (2015) 78–97.
[45] H.T. Wang, Q.S. Jia, C. Song, R.X. Yuan, X.H. Guan, Building occupant level estimation based on heterogeneous information fusion, Inf. Sci. 272 (2014) 145–157.
[46] C.Z. Wang, C.X. Wu, D.G. Chen, Q.H. Hu, C. Wu, Communication between information systems, Inf. Sci. 178 (2008) 3228–3239.
[47] C.Z. Wang, D.G. Chen, L.K. Zhu, Homomorphisms between fuzzy information systems, Appl. Math. Lett. 22 (2009) 1045–1050.
[48] C.Z. Wang, D.G. Chen, Q.H. Hu, Fuzzy information systems and their homomorphisms, Fuzzy Sets Syst. 249 (2014) 128–138.
[49] L. Wang, X. Liu, W. Pedrycz, Y. Shao, Determination of temporal information granules to improve forecasting in fuzzy time series, Expert Syst. 41 (2014) 3134–3142.
[50] J. Xu, G.Y. Wang, H. Yu, Review of big data processing based on granular computing, Chin. J. Comput. 38 (8) (2015) 1497–1517.
[51] W.H. Xu, Q.R. Wang, X.T. Zhang, Multi-granulation fuzzy rough sets in a fuzzy tolerance approximation space, Int. J. Fuzzy Syst. 13 (4) (2011) 246–259.
[52] W.H. Xu, X.T. Zhang, Q.R. Wang, A generalized multi-granulation rough set approach, Lecture Notes Comput. Sci. 6840 (2012) 681–689.
[53] W.H. Xu, W.X. Sun, X.Y. Zhang, W.X. Zhang, Multiple granulation rough set approach to ordered information systems, Int. J. Gen. Syst. 41 (5) (2012) 475–501.
[54] R.R. Yager, A framework for multi-source data fusion, Inf. Sci. 163 (1–3) (2004) 175–200.
[55] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977–1989.
[56] Y.Y. Yao, Relational interpretations of neighborhood operators and rough set approximation operators, Inf. Sci. 101 (1998) 239–259.
[57] Y.Y. Yao, Perspectives of granular computing, in: Proceedings of 2005 IEEE International Conference on Granular Computing, vol. 1 (25), 2005, pp. 85–90.
[58] X.B. Yang, Y.H. Qian, J.Y. Yang, Hierarchical structures on multigranulation spaces, J. Comput. Sci. Technol. 27 (6) (2012) 1169–1183.
[59] X.B. Yang, X.N. Song, Z.H. Chen, J.Y. Yang, Multigranulation rough sets in incomplete information system, Incomplete Information System and Rough Set Theory, Springer, Berlin, Heidelberg, 2012, pp. 195–222.
[60] M.S. Yang, P.Y. Huang, et al., Fuzzy clustering algorithms for mixed feature variables, Fuzzy Sets Syst. 141 (2004) 301–317.
[61] G.S. Yang, Y.Z. Lin, P. Bhattacharya, A driver fatigue recognition model based on information fusion and dynamic Bayesian network, Inf. Sci. 180 (10) (2010) 1942–1954.
[62] J.H. Yu, W.H. Xu, Information fusion in multi-source fuzzy information system with the same structure, in: Proceedings of the 2015 International Conference on Machine Learning and Cybernetics, 2015, pp. 170–175.
[63] F.S. Yu, W. Pedrycz, The design of fuzzy information granules: tradeoffs between specificity and experimental evidence, Appl. Soft Comput. 9 (2009) 264–273.
[64] L. Zadeh, Fuzzy logic = computing with words, IEEE Trans. Fuzzy Syst. 4 (2) (1996) 103–111.
[65] L. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets Syst. 90 (2) (1997) 111–127.
[66] L. Zadeh, Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems, Soft Comput. 2 (1998) 23–25.
[67] L. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.
[68] X.Y. Zhang, D.Q. Miao, Quantitative information architecture, granular computing and rough set models in the double-quantitative approximation space of precision and grade, Inf. Sci. 268 (2014) 147–168.
[69] W.X. Zhang, W.Z. Wu, J.Y. Liang, et al., Rough Set Theory and Method, Science Press, Beijing, 2001.
[70] W.X. Zhang, Y. Liang, W.Z. Wu, Information Systems and Knowledge Discovery, Science Press, Beijing, 2003.
[71] P. Zhu, Q.Y. Wen, Homomorphisms between fuzzy information systems revisited, Appl. Math. Lett 24 (2011) 1548–1553.