entropy

MDPI

Article

Information Fusion in a Multi-Source Incomplete Information System Based on Information Entropy

Mengmeng Li † and Xiaoyan Zhang *,†

School of Sciences, Chongqing University of Technology, Chongqing 400054, China; cqutmmli@126.com

* Correspondence: zxy19790915@163.com; Tel.: +86-150-2300-2286

† These authors contributed equally to this work.

Received: 14 August 2017; Accepted: 19 October 2017; Published: 17 November 2017

Abstract: As we move into the information age, the amount of data in various fields has increased dramatically, and data sources have become increasingly widely distributed. The corresponding phenomenon of missing data is increasingly common, and it leads to the generation of incomplete multi-source information systems. In this context, this paper's proposal aims to address the limitations of rough set theory. We study the method of multi-source fusion in incomplete multi-source systems. This paper presents a method for fusing incomplete multi-source systems based on information entropy; in particular, by comparison with another method, our fusion method is validated. Furthermore, extensive experiments are conducted on six UCI data sets to verify the performance of the proposed method. Additionally, the experimental results indicate that multi-source information fusion approaches significantly outperform other approaches to fusion.

Keywords: incomplete information system; information entropy; multi-source information fusion; rough set theory

1. Introduction

Information fusion is used to obtain more accurate and definite inferences from the data provided by any single information source by integrating multiple information sources; several definitions have been proposed in the literature [1–9]. The theory of information fusion was first used in the military field; it is defined as a multi-level and multi-aspect process that handles problems. In fact, data fusion can be broadly summarized as such a process; namely, to synthesize comprehensive intelligence from multi-sensor data and information according to established rules and analysis methods, and on this basis, to provide the user-required information, such as decisions, tasks, or tracks. Therefore, the basic purpose of data fusion is to obtain information that is more reliable than data from any single input. Along with the progress of time, information fusion technology has become increasingly important in the field of information service. Multi-source information fusion is one of the most important parts of information service in the age of big data, and many productive achievements have been made. Many scholars have conducted research on multi-source information fusion. For example, Hai [10] investigated predictions of formation drillability based on multi-source information fusion. Cai et al. [11] researched multi-source information fusion-based fault diagnosis of a ground-source heat pump using a Bayesian network. Ribeiro et al. [12] studied an algorithm for data information fusion that includes concepts from multi-criteria decision-making and computational intelligence, especially fuzzy multi-criteria decision-making and mixture aggregation operators with weighting functions. Some relative papers have studied entropy measure with other fuzzy extensions. For instance, Wei et al. [13] proposed uncertainty measures of extended hesitant fuzzy linguistic term sets. Based on interval-valued intuitionistic fuzzy soft sets, Liu et al. [14] proposed a theoretical development on the entropy. Yang et al. [15] proposed cross-entropy measures of linguistic hesitant intuitionistic fuzzy systems.

An information system is the main expression of an information source and the basic structure underlying information fusion. An information system is a data table that describes the relationships among objects and attributes. There is a great deal of uncertainty in the process of information fusion. Rough set theory is usually used to measure the uncertainty in an information table. Rough set theory—which was introduced by Pawlak [16–20]—is an extension of classical set theory. In data analysis, it can be considered a mathematical and soft computational tool to handle imprecision, vagueness, and uncertainty. This relatively new soft computing methodology has received a great deal of attention in recent years, and its effectiveness has been confirmed by successful applications in many science and engineering fields, including pattern recognition, data mining, image processing, and medical diagnosis [21,22]. Rough set theory is based on the classification mechanism, and the theory is classified as an equivalence relation in a specific universe, and this equivalence relation constitutes a partition of the universe. A concept (or more precisely, the extension of a concept) is represented by a subset of a universe of objects, and is approximated by a pair of definable concepts in a logic language. The main idea of rough set theory is the use of known knowledge in a knowledge base to approximate inaccurate and uncertain knowledge. This seems to be of fundamental importance to artificial intelligence and cognitive science. An information system is the basic structure underlying information fusion, and rough set theory is usually used to measure the uncertainty in an information system. Therefore, it is feasible to use rough set theory for information fusion. Some scholars have conducted research in this field. For example, Grzymala-Busse [23] presented and compared nine different approaches to missing attribute values. For testing both naive classification and new classification techniques of LERS (Learning from Examples based on Rough Sets) were used. Dong et al. [24] researched the processing of information fusion based on rough set theory. Wang et al. [25] investigated multi-sensor information fusion based on rough sets. Huang et al. [26] proposed a novel method for tourism analysis with multiple outcome capability based on rough set theory. Luo et al. [27] studied incremental update of rough set approximation under the grade indiscernibility relation. Yuan et al. [28] considered multi-sensor information fusion based on rough set theory. In addition, Khan et al. [29,30] used views of the membership of objects to study rough sets and notions of approximates in multi-source situations. Md et al. [31] proposed a modal logic for multi-source tolerance approximation spaces based on the principle of considering only the information that sources have about objects. Lin et al. studied an information fusion approach based on combining multi-granulation rough sets with evidence theory [32]. Recently, Balazs and Velásquez conducted a systematic study of opinion mining and information fusion [33].

However, these methods of information fusion are all based on complete information systems; a smaller amount of research has been conducted for incomplete information systems $(IISs)$. Jin et al. [34] studied feature selection in incomplete multi-sensor information systems based on positive approximation in rough set theory. IISs occur as a result of the ability to acquire data, the production environment, and other factors that result in the presence of original data with unknown values of attributes. As science has developed, people have found many ways to obtain information. An information box [35] can have multiple information sources, and every information source can be used to construct an information system. If all information sources are incomplete, then they can be used to construct multiple incomplete information systems. Therefore, the motivation for this paper is shown as follows: From the current research situation, most methods of information system fusion are all based on complete information systems. In order to broaden the research background of information fusion, we study the method of incomplete information system fusion. In order to reduce the amount of information loss in the process of information system fusion, we proposed the method which used information entropy to fuse incomplete information systems. In particular, by comparison with another method, our fusion method is validated. In this paper, we discuss the multi-source fusion of incomplete information tables based on information entropy. It is concluded that the method proposed here is more effective after comparing it with the mean value fusion method.

This rest of this paper is organized as follows: Some relevant notions are reviewed in Section 2. In Section 3, we define conditional entropy in a multi-source decision system, propose a fusion method based on conditional entropy, and design an algorithm for creating a new information table from a multi-source decision table based on conditional entropy. In Section 4, we download some data sets from UCI to prove the validity and reliability of our method; furthermore, we analyze the results of the experiment. The paper ends with conclusions in Section 5.

2. Preliminaries

In this section, we simply review some basic concepts relating to rough set theory, incomplete information systems, incomplete decision systems, and conditional entropy (CE) in incomplete decision systems. More details can be found in the literature [16,36–39].

2.1. Rough Sets

In rough set theory, let $S = (U, AT, V, f)$ be an information system. The $U = (x_1, x_2, \ldots, x_n)$ is the object set. The $AT = (a_1, a_2, \ldots, a_m)$ is the attribute set. The $V = (v_1, v_2, \ldots, v_m)$ is a set of corresponding attribute values. The $f : U \to V$ is a mapping function.

Let $P \subseteq R$ and $P \neq \phi$, the intersection of all the equivalence relations in P is called the equivalence relation on P or the indistinguishable relation is defined by $IND(P)$.

Let X be a subset of U. Then, x is an object of U, the equivalence class of x about R is defined by

$$[x]_R = \{y \in U | xRy\},$$

which represents the equivalence class that contains x.

When a set X expresses a union of equivalence classes, the set X can be precisely defined; otherwise, the set X can only be approximated; in rough set theory, upper and lower approximation sets are used to describe the set X. Given a finite nonzero set, U, which is called the domain, that R is an equivalence relation in the universe U and $X \subseteq U$, the upper and lower approximations of X are defined by

$$\overline{R}(X) = \{x \in U | [x]_R \cap X \neq \emptyset\},$$
$$\underline{R}(X) = \{x \in U | [x]_R \subseteq X\}.$$

The R positive region, negative region, and the boundary region of X are defined as follows, respectively.

$$pos_R(X) = \underline{R}(X), neg_R(X) = \sim \overline{R}(X) \text{ and } bn_R(X) = \overline{R}(X) - \underline{R}(X)$$

The approximation accuracy and roughness of the concept X in an attribute set, A, are defined as follows:

$$\alpha_A(X) = \frac{|\underline{A}(X)|}{|\overline{A}(X)|}, \rho_A(X) = 1 - \alpha_A(X),$$

respectively. They are often used for measuring uncertainty in rough set theory. $|X|$ refers to the cardinality of the set X.

The approximation accuracy for rough classification was proposed by Pawlak [19] in 1991. By employing the attribute set R, the approximation accuracy provides the percentage of possibly correct decisions when classifying objects.

Let $DS = (U, AT \cap D, V, f)$ be a decision system, $U/D = \{Y_1, Y_2, \ldots, Y_m\}$ be a classification of universe U, and R be an attribute set satisfying $R \subseteq AT$. Then, the R-lower and R-upper approximations of U/D are defined as

$$\underline{R}(U/D) = \underline{R}(Y_1) \cup \underline{R}(Y_2) \cup \cdots \cup \underline{R}(Y_m)$$

$$\overline{R}(U/D) = \overline{R}(Y_1) \cup \overline{R}(Y_2) \cup \cdots \cup \overline{R}(Y_m).$$

The approximation accuracy of U/D for R is defined as

$$\alpha_R(U/D) = \frac{\sum_{Y_i \in U/D} |\underline{R}(Y_i)|}{\sum_{Y_i \in U/D} |\overline{R}(Y_i)|}.$$

Recently, Dai and Xu [40] extended this to incomplete decision systems; i.e.,

$$\alpha_B(U/D) = \frac{\sum_{Y_i \in U/D} |\underline{T_B}(Y_i)|}{\sum_{Y_i \in U/D} |\overline{T_B}(Y_i)|}.$$

The corresponding approximation roughness of U/D for R is defined as

$$Roughness_R(U/D) = 1 - \alpha_R(U/D).$$

2.2. Incomplete Information System

A quadruple $IS = (U, AT, V, f)$ is an information system. U is a nonempty finite set of objects, AT is a nonempty finite set of attributes, $V = \bigcup_{a \in A} V_a$, where V_a is the domain of a, and $f : U \times AT \to V$ is an information function such that $f(x, a) \in V_a$ for each $a \in AT$ and $x \in U$. A decision system, (DS), is a quadruple $DS = (U, AT \cup DT, V, f)$, where C is the condition attribute set, D is the decision attribute set, and $C \cap D = \phi$, V is the union of the attribute domain.

If there exists $a \in AT$ and $x \in U$ such that $f(a, x)$ is equal to a missing value (denoted "$*$"), then the information system is an incomplete information system (IIS). Otherwise, the information system is a complete information system (CIS). If $* \notin V_{DT}$ but $* \in V_{AT}$, then we call the decision system an incomplete decision system (IDS). If $* \notin V_{DT}$ and $* \notin V_{AT}$, then the information system is a complete decision system (CDS).

Because there are missing values, the equivalence relation is not suitable for incomplete information systems. Therefore, Kryszkiewicz [36,37] defined a tolerance relation for incomplete information systems. Given an incomplete information system, $IIS = (U, AT, V, f)$, for any attribute subset $B \subseteq AT$, let $T(B)$ denote the binary tolerance relation between objects that are possibly indiscernible in terms of B. $T(B)$ is defined as

$$T(B) = \{ (x, y) | \forall a \in B, f(a, x) = f(a, y) \text{ or } f(a, x) = * \text{ or } f(a, y) = * \}$$

The tolerance class of object x with reference to an attribute set B is denoted $T_B(x) = \{y | (x, y) \in T(B)\}$. For $X \subseteq U$, the lower and upper approximations of X with respect to B are defined as

$$\overline{T_B}(X) = \{x \in U | T_B(x) \cap X \neq \emptyset\},$$
$$\underline{T_B}(X) = \{x \in U | T_B(x) \subseteq X\}.$$

3. Multi-Source Incomplete Information Fusion

With the development of science and technology, people have access to increasing numbers of channels from which to obtain information. The diversity of the channels has produced a large number of incomplete information sources—that is, a multi-source incomplete information system. Investigating some special properties of this system and fusing the information are the focus of the information technology field. In this section, we present a new fusion method for multi-source incomplete information systems and compare our fusion method with the mean value fusion method in a small experiment.

3.1. Multi-Source Information Systems

Let us consider the scenario in which we obtain information regarding a set of objects from different sources. Information from each source is collected in the above information system, and thus,

a family of the single information systems with the same domain is obtained; it is called a multi-source information system [41].

Definition 1. *(see [32]) A multi-source information system can be defined as*

$$MS = \{IS_i | IS_i = (U, AT_i, \{(V_a)_{a \in AT_i}\}, f_i)\},$$

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

In particular, a multi-source decision information system is given by $MS = \{IS_i | IS_i = (U, AT_i, \{(V_a)_{a \in AT_i}\}, f_i, D, g)\}$, where D is a finite non-empty set of decision attributes and $g_d : U \to V_d$ for any $d \in D$, where V_d is the domain of decision attribute d. The multi-source information system includes s single information sources. Let the s overlapping pieces of single-source information system form an information box with s levels, as shown Figure 1, which comes from our previous study [35].

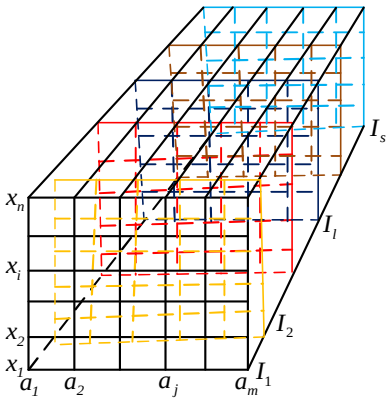

Figure 1. A multi-source information box.

3.2. Multi-Source Incomplete Information System

Definition 2. *A multi-source incomplete information system (MIIS) is defined as $MIIS = \{IIS_i | IIS_i = (U, AT_i, \{(V_a)_{a \in AT_i}\}, f_i)\}$, where*

1. *IIS_i is the incomplete information system of subsystem i;*
2. *U is a finite non-empty set of objects;*
3. *AT_i is the finite non-empty set of attributes for subsystem i;*
4. *$\{V_a\}$ is the value of attribute $a \in AT_i$;*
5. *$f_i : U \times AT_i \to \{(V_a)_{a \in AT_i}\}$ such that for all $x \in U$ and $a \in AT_i$, $f_i(x, a) \in V_a$.*

In particular, a multi-source decision information system is given by $MIIS = \{IIS_i | IIS_i = (U, AT_i, \{(V_a)_{a \in AT_i}\}, f_i, D, g)\}$, where D is a finite non-empty set of decision attributes and $g_d : U \to V_d$ for any $d \in D$, where V_d is the domain of decision attribute d.

3.3. Multi-Source Incomplete Information Fusion

Because the information box in each table is not complete, we propose a new fusion method.

Definition 3. *Let I be an incomplete information system (IIS) and $U = \{x_1, x_2, \ldots, x_n\}$. $\forall a \in AT$, $x_i, x_j \in U$, we define the distance between any two objects in U with attribute a as follows.*

$$dis_a(x_i, x_j) = \begin{cases} 0, & if\ f(x_i, a) = * \ or\ f(x_j, a) = *; \\ |f(x_i, a) - f(x_j, a)| & else. \end{cases}$$

Definition 4. *Given an incomplete information system $IIS = (U, AT, V, f)$, for any attribute $a \in AT$, let $T(a)$ denote the binary tolerance relation between objects that are possibly indiscernible in terms of a. $T(a)$ is defined as*

$$T(a) = \{ (x, y) \mid dis_a(x, y) L_a \},$$

where L_a indicates the threshold associated with attribute a. The tolerance class of object x with reference to attribute a is denoted by $T_a(x) = \{y | (x, y) \in T(a)\}$.

Definition 5. *Given an incomplete information system $IIS = (U, AT, V, f)$, for any attribute subset $B \subseteq AT$, let $T(B)$ denote the binary tolerance relation between objects that are possibly indiscernible in terms of B. $T(B)$ is defined as*

$$T(B) = \bigcap_{a \in B} T(a).$$

The tolerance class of object x with respect to an attribute set B is denoted by $T_B(x) = \{y | (x, y) \in T(B)\}$.

In the literature [39], Dai et al. proposed a new conditional entropy to evaluate the uncertainty in an incomplete decision system. Given an incomplete decision system $IDS = (U, AT \cup DT, V, f)$, $U = \{u_1, u_2, \ldots, u_n\}$. $B \subseteq AT$ is a set of attributes, and $U/D = \{Y_1, Y_2, \ldots, Y_m\}$. The conditional entropy of D with respect to B is defined as

$$H(D|B) = -\sum_{i=1}^{|U|} \sum_{j=1}^{m} \frac{|T_B(u_i) \cap Y_j|}{|U|} \log \frac{|T_B(u_i) \cap Y_j|}{|T_B(u_i)|}.$$

Because the conditional entropy is monotonous and because the attribute set B increases in importance as the conditional entropy decreases, we have the following definitions:

Definition 6. *Let $I_1, I_2, \ldots, I_s$ be s incomplete information systems and $U = \{u_1, u_2, \ldots, u_n\}$. $\forall a \in AT$, $U/D = \{Y_1, Y_2, \ldots, Y_m\}$. The uncertainty of the information sources in D with respect to I_q $(q = 1, 2, \ldots, s)$ for attribute a is defined as*

$$H_a(D|I_q) = -\sum_{i=1}^{|U|} \sum_{j=1}^{m} \frac{\left|T_a^q(u_i) \cap Y_j\right|}{|U|} \log \frac{\left|T_a^q(u_i) \cap Y_j\right|}{\left|T_a^q(u_i)\right|},$$

where $T_a^q(u_i)$ is the tolerance class of the information sources in D with respect to I_q $(q = 1, 2, \ldots, s)$ for attribute a.

Because the conditional entropy of Dai [39] is monotonous, $H_a(D|I_q)$ $(q = 1, 2, \ldots, s)$ for attribute a is also monotonous, and for attribute a, the smaller the conditional entropy is, the more important the information source is. We have the following Definition 7:

Definition 7. *Let $I_1, I_2, \ldots, I_s$ be s incomplete information system. We define the l^{th} $(l = 1, 2, \ldots, s)$ incomplete information system, which is the most important for attribute a, as follows:*

$$l^a = \underset{q \in \{1,2,\cdots,s\}}{arg\,min}\ (H_a(D|I_q)),$$

where l^a represents the l^{th} information source, which is the most important for attribute a.

Example 1. *Let us consider a real medical examination issue at a hospital. When diagnosing leukemia, there are 10 patients, $x_i (i = 1, 2, \ldots, 10)$, to be considered. They undergo medical examinations at four hospitals, which test 6 indicators, $a_i (i = 1, 2, \ldots, 6)$, where a_1–a_6 are, respectively, the "hemoglobin count", "leukocyte count", "blood fat", "blood sugar", "platelet count", and "Hb level". Tables 1–4 are incomplete evaluation tables based on the medical examinations performed at the four hospitals; the symbol "$*$" means that an expert cannot determine the level of a project.*

Table 1. Information source I_1.

U	a_1	a_2	a_3	a_4	a_5	a_6
x_1	143	11	250.3	150	79	60.1
x_2	160.8	11.1	160.2	115.9	88	43
x_3	127.3	4	118.2	$*$	114	80.2
x_4	130.2	5.6	120.5	98.5	150	77.9
x_5	132.6	$*$	115.7	72.8	177	89.3
x_6	200.1	15.4	230	120.5	76	44.9
x_7	125	5.8	111	80	$*$	77.3
x_8	167	16.7	225	120	80	40
x_9	$*$	$*$	222.5	133.4	77	55.3
x_{10}	135	8.1	116	100	210	99

Table 2. Information source I_2.

U	a_1	a_2	a_3	a_4	a_5	a_6
x_1	$*$	11.2	249.9	149.8	78	59
x_2	161	11	$*$	115	87	45.5
x_3	132.3	3.7	120.5	88	115	81
x_4	127.8	$*$	120.5	99	152	78
x_5	129.8	6.3	117	$*$	175	89
x_6	197.3	15	269.7	$*$	75	45.2
x_7	130.5	5.5	$*$	80.3	181	77.2
x_8	$*$	16.7	222.9	121	81	40.9
x_9	178.9	13.3	222.8	133	76	55
x_{10}	132.1	7.9	116.1	101.1	211	$*$

Table 3. Information source I_3.

U	a_1	a_2	a_3	a_4	a_5	a_6
x_1	140.1	$*$	250	150.1	79	$*$
x_2	165	12.3	160.9	114.8	88	45
x_3	$*$	4.2	120.5	87.5	115	81
x_4	130	5.1	121	$*$	151	77.9
x_5	130.6	6.9	117.9	73	176	88.8
x_6	$*$	16.8	$*$	119.9	75	$*$
x_7	127.7	5.2	111.2	79.6	181	77
x_8	166	$*$	221.3	119.9	81	40.8
x_9	173.8	13.4	223	132.9	77	54.5
x_{10}	133.5	8	$*$	100.2	$*$	100.1

Table 4. Information source I_4.

U	a_1	a_2	a_3	a_4	a_5	a_6
x_1	142.5	11	$*$	150	78	60
x_2	163.2	12.2	160.3	114	86	$*$
x_3	133.3	4	117.8	88.1	115	81
x_4	$*$	5	$*$	99	150	77.9
x_5	131.8	$*$	116.5	72.9	$*$	89.2
x_6	200	16.3	$*$	150	74	45
x_7	129	5	111	$*$	181	77
x_8	$*$	16.2	221	120.2	81	$*$
x_9	172	13	$*$	$*$	77	55
x_{10}	134	8.2	$*$	100	210	99.8

Suppose $V_D = \{Leukemia\ patient, Non\ leukemia\ patient\}$ and $U/D = \{Y_1, Y_2\}$, where $Y_1 = \{x_1, x_2, x_6, x_8, x_9\}$, $Y_2 = \{x_3, x_4, x_5, x_7, x_{10}\}$. Then, the conditional entropy of the information sources of D with respect to I_q $(q = 1, 2, 3, 4)$ for attribute a_i $(i = 1, 2, \ldots, 6)$ is as follows:

Because the conditional entropy can be used to evaluate the importance of information sources for attribute a, we can determine the importance of all attributes for all information sources by using Definition 7 and Table 5. The smaller the conditional entropy is, the more important the information sources are for attribute a. Therefore, I_1 is the most important for a_1 and a_6, I_2 is the most important for a_3 and a_5, and I_4 is the most important for a_2 and a_4. I_3 is not the most important for any attribute. A new information system, (NIS) is established by part of each table. Furthermore, we take I_1 for the value of a property for a_1 and a_6, I_2 for the property's value for a_3 and a_5, and I_4 for the property's value for a_2 and a_4. That is, $NIS = (V_{a_1}^{I_1}, V_{a_2}^{I_4}, V_{a_3}^{I_2}, V_{a_4}^{I_4}, V_{a_5}^{I_2}, V_{a_6}^{I_1})$, where $V_{a_i}^{I_q}$ $(q = 1, 2, 3, 4; i = 1, 2, \ldots, 10)$ represents the range of attribute a_i under I_q, and we obtain the new information system (NIS) after fusion. The new information system, (NIS), after fusion is shown in Table 6.

Table 5. The conditional entropy of information sources for different attributes.

U	I_1	I_2	I_3	I_4
a_1	2.5141	2.5467	3.0103	2.6553
a_2	2.4615	2.3810	2.2310	1.9983
a_3	2.5467	2.3583	2.6966	3.0103
a_4	2.8029	2.8741	2.7936	2.7256
a_5	2.1759	1.6443	2.2084	2.0198
a_6	2.7936	3.0103	2.8741	2.9453

Table 6. The result of multi-source information fusion.

U	a_1	a_2	a_3	a_4	a_5	a_6
x_1	143	11	249.9	150	78	60.1
x_2	160.8	12.2	$*$	114	87	43
x_3	127.3	4	120.5	88.1	115	80.2
x_4	130.2	5	120.5	99	152	77.9
x_5	132.6	$*$	117	72.9	175	89.3
x_6	200.1	16.3	269.7	150	75	44.9
x_7	125	5	$*$	$*$	181	77.3
x_8	167	16.2	222.9	120.2	81	40
x_9	$*$	13	222.8	$*$	76	55.3
x_{10}	135	8.2	116.1	100	211	99

The fusion process is shown in Figure 2. Suppose that there is a multi-source information system $MS = \{I_1, I_2, \cdots, I_s\}$ that contains s information systems and that there are n objects and m attributes in each information system $I_i (i = 1, 2, \ldots, s)$. We calculate the conditional entropy of each attribute by using Definition 6. Then, we determine the minimum of the conditional entropy for each attribute of the values using Definition 7. For example, we use different colors of rough lines to express the corresponding attributes to select a source. Then, the selected attribute values are integrated into a new information system.

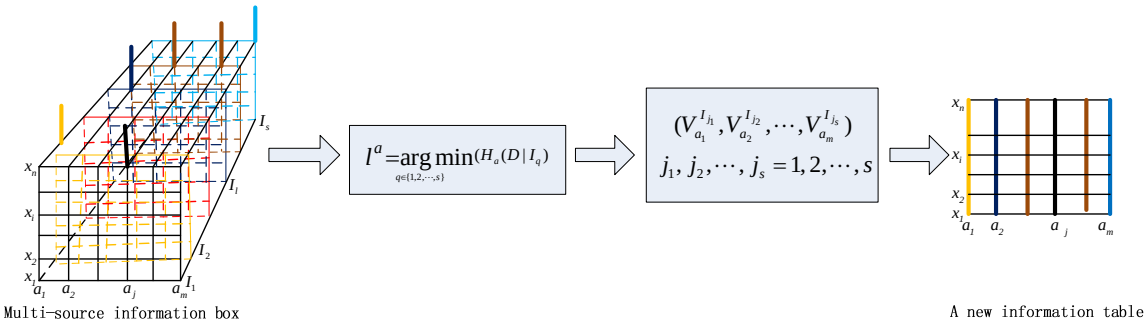

Figure 2. The process of multi-source information fusion.

In practical applications, the mean value fusion method is one of the common fusion methods. We compare this type of method with conditional entropy fusion based on approximation accuracy. The results of two types of fusion method are presented in Tables 6 and 7.

Table 7. The result of mean value fusion of multiple information sources.

U	a_1	a_2	a_3	a_4	a_5	a_6
x_1	141.8667	11.0667	250.0667	149.975	78.5	59.7
x_2	162.5	11.65	160.4667	114.925	87.25	44.5
x_3	130.9667	3.975	119.25	87.8667	114.75	80.8
x_4	129.3333	5.2333	120.6667	98.8333	150.75	77.925
x_5	131.2	6.6	116.775	72.9	176	89.075
x_6	199.1333	15.875	249.85	130.1333	75	45.0333
x_7	128.05	5.375	111.0667	79.9667	181	77.125
x_8	166.5	16.5333	222.55	120.275	80.75	40.5667
x_9	174.9	13.2333	222.7667	133.1	76.75	54.95
x_{10}	133.65	8.05	116.05	100.325	210.3333	99.6333

Using Tables 6 and 7, we compute the approximation accuracy of the results of the two fusion methods and compare their approximation accuracy. Please see Table 8.

Table 8. The approximation accuracies of two fusion methods.

	Multi-Source Fusion	**Mean Value Fusion**
Approximation accuracy	0.42857	0.33333

By comparing the approximation accuracies, we see that multi-source fusion is better than mean value fusion. Therefore, we design a multi-source fusion algorithm (Algorithm 1) and analyze its computational complexity.

The given algorithm (Algorithm 1) is a new approach to multi-source information fusion. Its approximation accuracy is better than that of mean value fusion in the result of example Section 3.3. First, we can calculate all the similarity classes $T_a^q(x)$ for any $x \in U$ for attribute a. Then, the conditional entropy, $H_a(D|I_q)$, is computed for information source q and attribute a. Finally, the minimum of the

conditional entropy of the information source is selected for attribute a, and the results are spliced into a new table. The computational complexity of Algorithm 1 is shown in Table 9.

Algorithm 1: An algorithm for multi-source fusion.

Input : A multi-source information system $MS = \{I_1, I_2, \ldots, I_s\}$ and a classification $U/D = \{Y_1, Y_2, \ldots, Y_m\}$;
Output : A new information table.

1 **begin**
2 **for** $q = 1 : s$ **do**
3 **for** *each* $a \in AT$ **do**
4 **for** *each* $x_i \in U$ **do**
  compute: $T_a^q(x_i)$; // compute all $T_a^q(x_i)$, for any $x_i \in U$ for attribute a;
5 **end**
6 $HCE \leftarrow 0$;
7 **for** $i = 1 : |U|$ **do**
8 **for** $j = 1 : m$ **do**
9 **if** $|T_a^q(x_i \cap Y_j| > 0$ **then**
10 $HCE \leftarrow HCE - \frac{|T_a^q(x_i \cap Y_j|}{|U|} log \frac{|T_a^q(x_i \cap Y_j|}{|T_a^q(x-i|}$;
11 **end**
12 **end**
13 **end**
14 $H_a(D|I_q) \leftarrow HCE$; // record CE for attribute a and information source q;
15 **end**
16 **end**
17 **for** *each* $a \in AT$ **do**
18 $minCE \leftarrow +\infty$;
19 **for** $q = 1 : s$ **do**
20 **if** $H_a(D|I_q) < minCE$ **then**
21 $minCE \leftarrow H_a(D|I_q)$;
22 $l^a \leftarrow q$;
23 **end**
24 **end**
25 **end**
26 **return**: $(V_{a_1}^{l^{a_1}}, V_{a_2}^{l^{a_2}}, \cdots, V_{a_{|AT|}}^{l^{|AT|}})$.
27 **end**

Table 9. Computational complexity of Algorithm 1.

Steps 4–5	$O(\|U\|^2)$
Steps 6–14	$O(\|U\| \times m^2)$
Steps 1–16	$O(s \times \|AT\| \times (\|U\|^2 + \|U\| \times m^2))$
Steps 17–25	$O(\|AT\| \times s)$
Step 26	$O(\|U\| \times \|AT\|)$
Total	$O(s \times \|AT\| \times (\|U\|^2 + \|U\| \times m^2) + \|AT\| \times s + \|U\| \times \|AT\|)$

In steps 4 and 5 of Algorithm 1, we compute all $T_a^q(x)$ for any $x \in U$ for attribute a. Steps 6–14 calculate the conditional entropy for information source q and attribute a. Steps 17–26 are to find the minimum of the conditional entropy of the corresponding source for any $a \in AT$. Finally, the results are returned.

4. Experimental Evaluation

In this section, to further illustrate the correctness of the conclusions of the previous example, we conduct a series of experiments to explain why the approximate precision of conditional entropy fusion

is generally higher than that of the mean value fusion based on standard data sets from the machine learning data repository of the University of California at Irvine (http://archive.ics.uci.edu/ml/datasets.html) called "Statlog (Vehicle Silhouettes)", "Letter Recognition", "Phishing Websites", "Robot Execution Failures", "Semeion Handwritten Digit", and "SPECTF Heart" in Table 10. The experimental program is running on a personal computer with the hardware and software described in Table 11.

Table 10. Experimental data sets.

No.	Data Set Name	Abbreviation	Objects	Attributes	Decision Classes	Number of Sources	Elements
1	Wholesale Customers	WC	440	9	4	10	39,600
2	Statlog (Vehicle Silhouettes)	S (VS)	846	19	4	10	160,740
3	Airfoil Self-Noise	AS-N	1503	7	5	10	105,210
4	Image Segmentation	IS	2310	20	7	10	462,000
5	Statlog (Landsat Satellite)	S (LS)	6435	37	6	10	2,380,950
6	EEG Eye State	EES	14,980	15	2	10	2,247,000

Table 11. Description of the experimental environment.

Name	Model	Parameters
CPU	Intel i3-370	2.40 GHz
Memory	Samsung DDR3	2 GB; 1067 MHz
Hard Disk	West Data	500 GB
System	Windows 7	32 bit
Platform	V C + +	6.0

To build a real multi-source incomplete information system, we propose a method for obtaining incomplete data from multiple sources. First, to obtain incomplete data, a complete data set with some data randomly deleted is used as the original incomplete data set. Then, a multi-source incomplete decision table is constructed by adding Gaussian noise and random noise to the original incomplete data set.

Let $MIIS = \{I_1, I_2, \ldots, I_s\}$ be a multi-source incomplete decision table constructed using the original incomplete information table, I.

First, s numbers $(g_1, g_2, \ldots, g_s)$ that have an $N(0, \sigma)$ distribution, where σ is the standard deviation, are generated. The method of adding Gaussian noise is as follows:

$$I_i(x,a) = \begin{cases} I(x,a) + g_i & \text{if } (I(x,a) \neq *) \\ * & \text{else} \end{cases},$$

where $I(x,a)$ is the value of object x with attribute a in the original incomplete information table and $I_i(x,a)$ represents object x with attribute a in the i-th incomplete information source.

Then, s random numbers $(e_1, e_2, \ldots, e_s)$ between $-e$ and e, where e is a random error threshold, are generated. The method of adding random noise is as follows:

$$I_i(x,a) = \begin{cases} I(x,a) + e_i & \text{if } (I(x,a) \neq *) \\ * & \text{else} \end{cases},$$

where $I(x,a)$ represents the value of object x for attribute a in the original incomplete information table and $I_i(x,a)$ represents object x for attribute a in the i-th incomplete information source.

Next, 40% of the objects are randomly selected from the original incomplete information table, I, and Gaussian noise is added to these objects. Then, 20% of the objects are randomly selected from the rest of the original incomplete information table, I, and random noise is added to these objects.

Finally, a multi-source incomplete decision table, $MIIS = \{I_1, I_2, \ldots, I_s\}$, can be created.

5. Related Works and Conclusion Analysis

In different fields of science, the standard deviation of Gaussian noise and the random error threshold of random noise may differ. In this paper, we conducted 20 experiments for each data set and set the standard deviation σ and the random error threshold e to values from 0 to 2, with an increase of 0.1 in each experiment. For CE fusion and mean value fusion, the approximation accuracy of U/D for each data set is displayed in Table 12 and Figures 3–8. CE and M stand for CE fusion and mean value fusion, respectively.

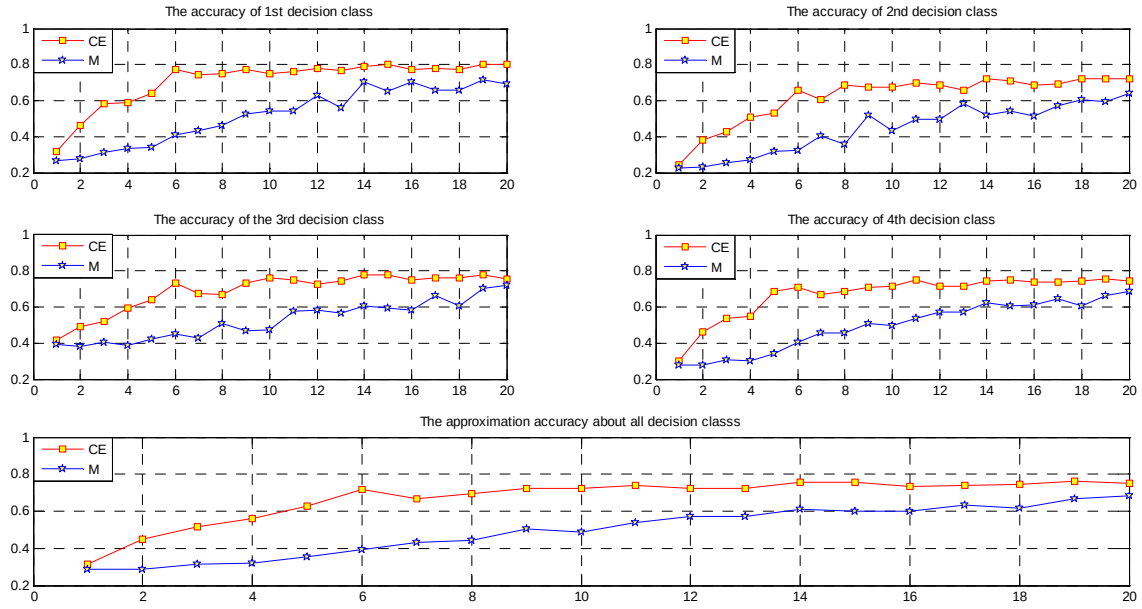

Figure 3. Approximation accuracies for the decision classes in data set WC.

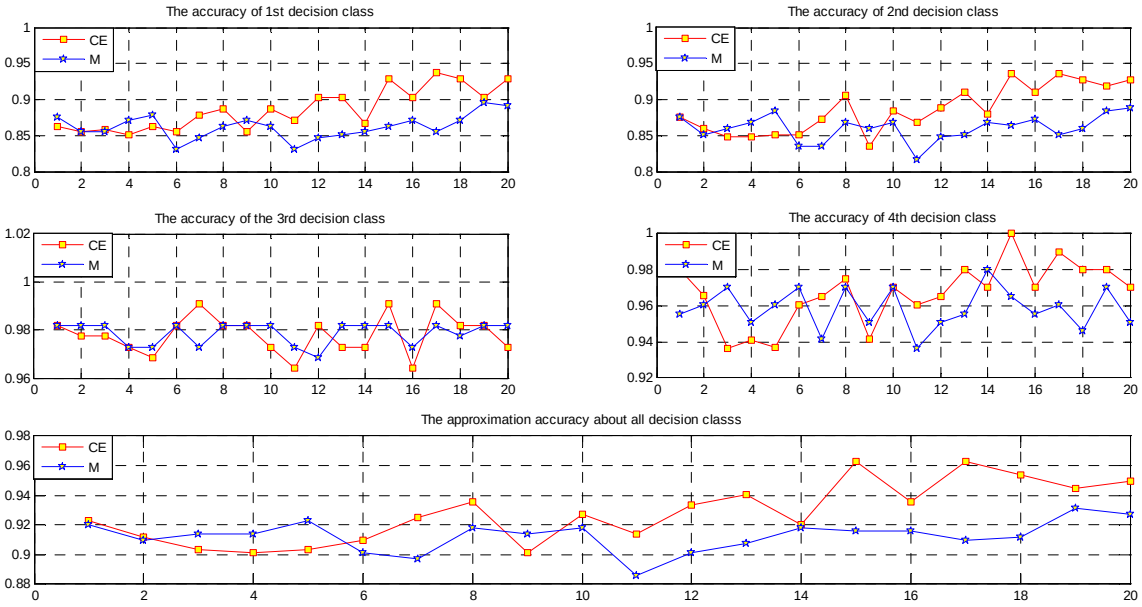

Figure 4. Approximation accuracies for the decision classes in data set S (VS).

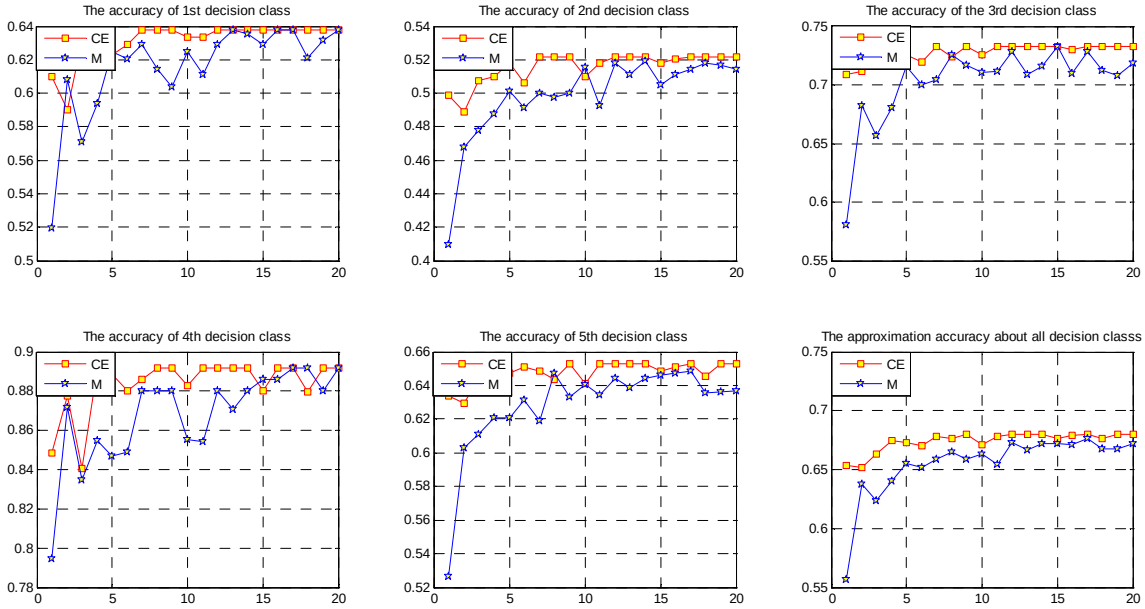

Figure 5. Approximation accuracies for the decision classes in data set AS-N.

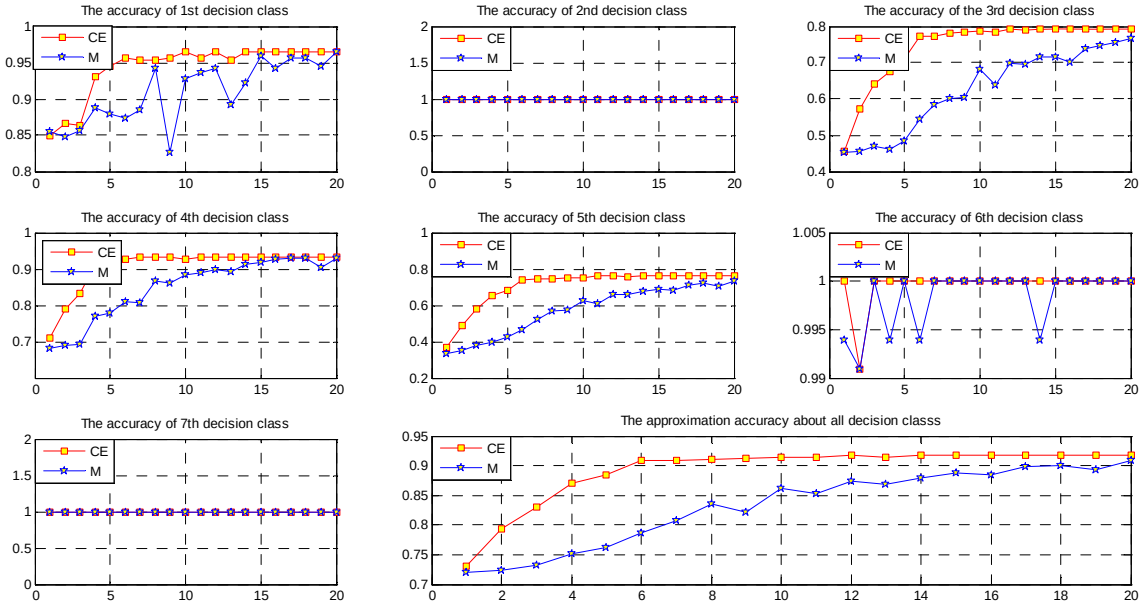

Figure 6. Approximation accuracies for the decision classes in data set IS.

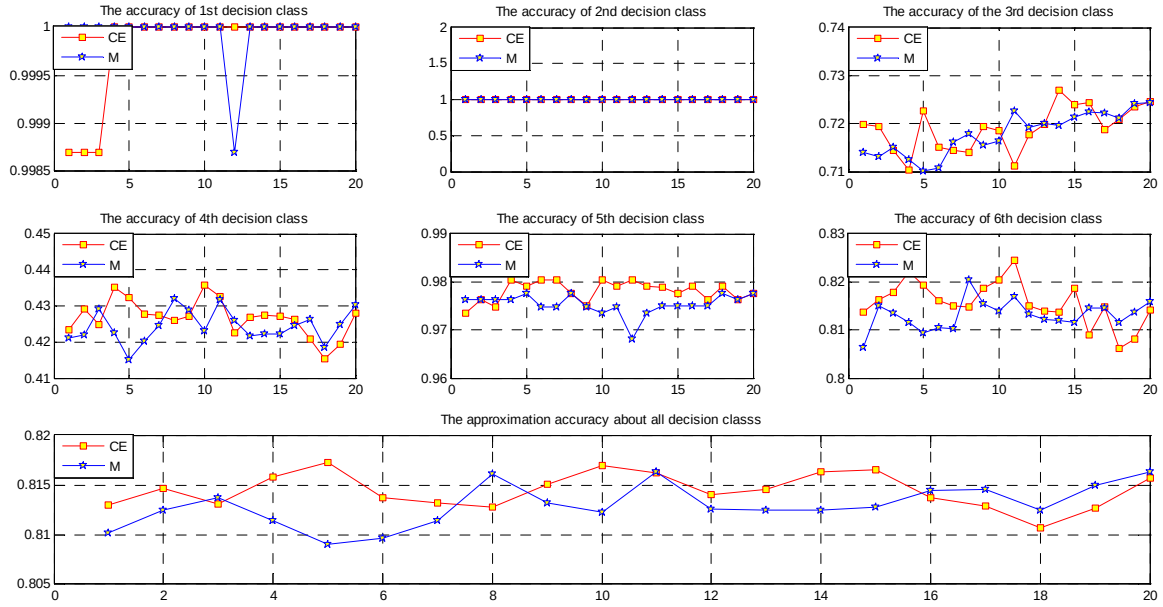

Figure 7. Approximation accuracies for the decision classes in data set S (LS).

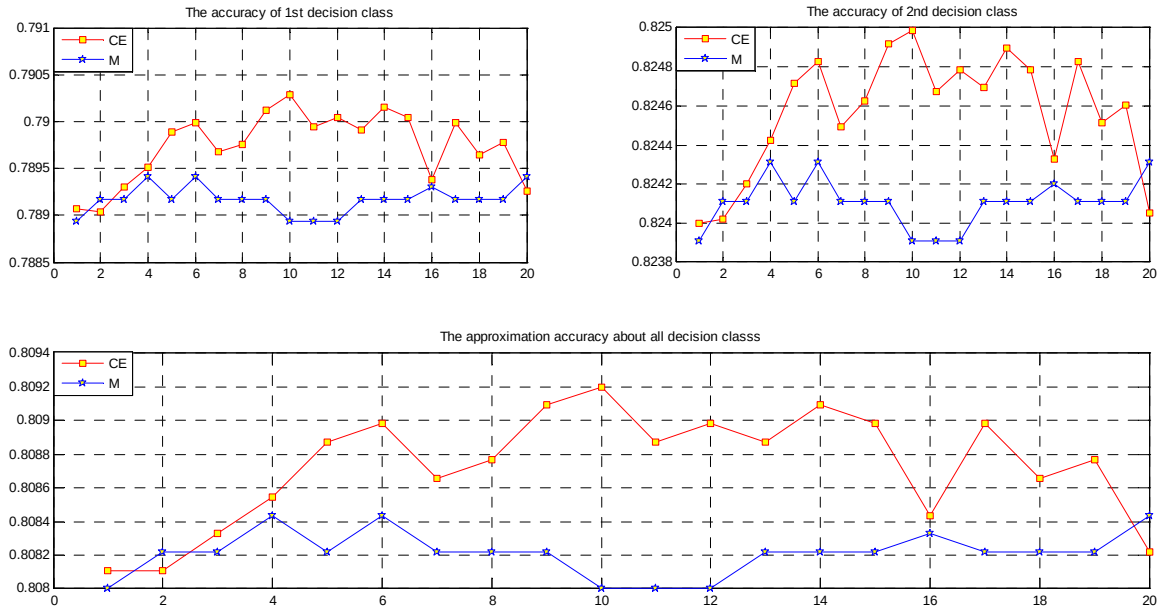

Figure 8. Approximation accuracies for the decision classes in data set EES.

We can easily see from Figures 3–8 and Table 12 that when the noise is small, in most cases, the approximation accuracy of CE fusion is slightly higher than that of mean value fusion. In a certain range, as the noise increases, the approximation accuracy of CE fusion becomes much better than that of mean value fusion.

By observing the approximation accuracies of the extensions of concepts of CE and mean value fusion for the six data sets, we find that in most cases, the approximation accuracy of CE fusion is higher than that of mean value fusion. In a certain range, as the amount of noise increases, the accuracies of the extensions of concepts of CE and mean value fusion trend upward, but they are not strictly monotonic.

Table 12. Approximation accuracies of conditional entropy fusion (CE) and mean value fusion (M) for each data set.

No.	WC		S (VS)		AS-N		IS		S (LS)		EES	
	CE	M.	CE	M.	CE	M.	CE	M.	CE	M.	CE	M.
1	0.316602	0.285538	0.922727	0.920545	0.653333	0.556552	0.730996	0.719472	0.813005	0.810172	0.80811	0.808001
2	0.449165	0.28801	0.911864	0.909707	0.65141	0.637566	0.792931	0.723264	0.814686	0.812448	0.80811	0.808219
3	0.516181	0.316947	0.903262	0.914027	0.663276	0.623302	0.830129	0.73321	0.813092	0.81371	0.808328	0.808219
4	0.559664	0.321196	0.901124	0.914027	0.674797	0.640402	0.870916	0.75169	0.8159	0.811491	0.808547	0.808437
5	0.628114	0.352861	0.903262	0.922727	0.673337	0.655633	0.88357	0.762622	0.817306	0.809001	0.808874	0.808219
6	0.71673	0.39515	0.909707	0.901124	0.670086	0.651709	0.90823	0.786262	0.813727	0.809616	0.808983	0.808437
7	0.669118	0.432635	0.924915	0.896861	0.678436	0.658445	0.908979	0.807603	0.813283	0.811405	0.808656	0.808219
8	0.696226	0.445619	0.935927	0.918367	0.676614	0.665231	0.910552	0.835308	0.812831	0.816204	0.808765	0.808219
9	0.720532	0.504039	0.901124	0.914027	0.680261	0.658811	0.912129	0.821331	0.815166	0.813283	0.809093	0.808219
10	0.72447	0.486781	0.927107	0.918367	0.671166	0.663441	0.913708	0.861613	0.817015	0.812335	0.809202	0.808001
11	0.74031	0.536304	0.914027	0.886288	0.678436	0.654011	0.914498	0.853094	0.816292	0.816329	0.808874	0.808001
12	0.725	0.569728	0.933714	0.901124	0.680261	0.672983	0.916873	0.874697	0.814041	0.812587	0.808983	0.808001
13	0.720307	0.569966	0.940367	0.907554	0.680261	0.666667	0.914569	0.867871	0.814634	0.8125	0.808874	0.808219
14	0.754902	0.611408	0.920545	0.918367	0.680261	0.672069	0.916873	0.880065	0.816406	0.812474	0.809093	0.808219
15	0.75835	0.59792	0.962877	0.916195	0.676614	0.672078	0.916873	0.888618	0.816608	0.812839	0.808983	0.808219
16	0.736434	0.599647	0.935927	0.916195	0.678979	0.671174	0.916873	0.884146	0.813762	0.814495	0.808437	0.808328
17	0.741748	0.634234	0.962877	0.909707	0.680261	0.676614	0.916873	0.897667	0.812944	0.814582	0.808983	0.808219
18	0.748047	0.618705	0.953811	0.911864	0.676614	0.66792	0.916873	0.900779	0.810762	0.8125	0.808656	0.808219
19	0.761811	0.667286	0.944828	0.931507	0.680261	0.667385	0.916873	0.893469	0.812717	0.815001	0.808765	0.808219
20	0.753425	0.684701	0.949309	0.927107	0.680261	0.672255	0.916873	0.908193	0.815734	0.816406	0.808219	0.808437

6. Conclusions

In this paper, we studied multi-source information fusion in view of the conditional entropy. There are many null information sources in the age of big data. To solve the problem of integrating multiple incomplete information sources, we studied an approach based on multi-source information fusion. We transformed a multi-source information system into an information table by using this fusion method. Furthermore, we used rough set theory to investigate the fused information table, and compared the accuracy of our fusion method with that of the mean value fusion method. According to the accuracies, CE fusion is better than mean value fusion under most conditions. In this paper, we constructed six multi-source information systems, each containing 10 single information sources. Based on these data sets, a series of experiments was conducted; the results showed the effectiveness of the proposed fusion method. This study will be useful for fusing uncertain information in multi-source information systems. It provides valuable selections for data processing in multi-source environments.

Acknowledgments: The authors wish to thank the anonymous reviewer. This work is supported by the Natural Science Foundation of China (No. 61105041, No. 61472463 and No. 61402064), the National Natural Science Foundation of CQ CSTC (No. cstccstc2015jcyjA1390), the Graduate Innovation Foundation of Chongqing (No. CYS16217) and the Graduate Innovation Foundation of Chongqing University of Technology (No. YCX2016227).

Author Contributions: Mengmeng Li is the principal investigator of this work. He performed the simulations and wrote this manuscript. Xiaoyan Zhang contributed to the data analysis work and checked the whole manuscript. All authors revised and approved the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bleiholder, J.; Naumann, F. Data fusion. *ACM Comput. Surv.* **2008**, *41*, 1–41.
2. Lee, H.; Lee, B.; Park, K.; Elmasri, R. Fusion techniques for reliable information: A survey. *Int. J. Digit. Content Technol. Appl.* **2010**, *4*, 74–88.
3. Khaleghi, B.; Khamis, A.; Karray, F.O. Multisensor data fusion: A review of the state of the art. *Inf. Fusion* **2013**, *14*, 28–44.
4. Han, C.Z.; Zhu, H.Y.; Duan, Z.S. *Multiple-Source Information Fusion*; Tsinghua University Press: Beijing, China, 2010.
5. Peng, D. *Theory and Application of Multi Sensor Multi Source Information Fusion*; Thomson Learning Press: Beijing, China, 2010.
6. Schueremans, L.; Gemert, D.V. Benef it of splines and neural networks in simulation based structural reliability analysis. *Struct. Saf.* **2005**, *27*, 246–261.
7. Pan, W.K.; Liu, Z.D.; Ming, Z.; Zhong, H.; Wang, X.; Xu, C.F. Compressed Knowledge Transfer via Factorization Machine for Heterogeneous Collaborative Recommendation. *Knowl. Based Syst.* **2015**, *85*, 234–244.
8. Wang, X.Z.; Huang, J. Editorial: Uncertainty in Learning from Big Data. *Fuzzy Sets Syst.* **2015**, *258*, 1–4.
9. Wang, X.Z.; Xing, H.J.; Li, Y.; Hua, Q.; Dong, C.R.; Pedrycz, W. A Study on Relationship between Generalization Abilities and Fuzziness of Base Classifiers in Ensemble Learning. *IEEE Trans. Fuzzy Syst.* **2015**, *23*, 1638–1654.
10. Hai, M. Formation drillability prediction based on multisource information fusion. *J. Pet. Sci. Eng.* **2011**, *78*, 438–446.
11. Cai, B.; Liu, Y.; Fan, Q.; Zhang, Y.; Liu, Z.; Yu, S.; Ji, R. Multi-source information fusion based fault diagnosis of ground-source heat pump using Bayesian network. *Appl. Energy* **2014**, *114*, 1–9.
12. Ribeiro, R.A.; Falcão, A.; Mora, A.; Fonseca, J. FIF: A fuzzy information fusion algorithm based on multi-criteria decision. *Knowl. Based Syst.* **2014**, *58*, 23–32.
13. Wei, C.P.; Rodrguez, R.M.; Martnez, L. Uncertainty Measures of Extended Hesitant Fuzzy Linguistic Term Sets. *IEEE Trans. Fuzzy Syst.* **2017**, 1, doi:10.1109/TFUZZ.2017.2724023.
14. Liu, Y.Y.; Luo, J.F.; Wang, B.; Qin, K. A theoretical development on the entropy of interval-valued intuitionistic fuzzy soft sets based on the distance measure. *Int. J. Comput. Intell. Syst.* **2017**, *10*, 569.

15. Yang, W.; Pang, Y.F.; Shi, J.R. Linguistic hesitant intuitionistic fuzzy cross-entropy measures. *Int. J. Comput. Intell. Syst.* **2017**, *10*, 120.
16. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356.
17. Pawlak, Z. Rough set theory and its applications to data analysis. *Cybern. Syst.* **1998**, *29*, 661–688.
18. Pawlak, Z.; Skowron, A. Rough sets: Some extensions. *Inf. Sci.* **2007**, *177*, 28–40.
19. Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning About Data*; Kluwer: Boston, MA, USA, 1991.
20. Pawlak, Z. Vagueness and uncertainty: A rough set perspective. *Comput. Intell.* **1995**, *11*, 227–232.
21. Li, H.L.; Chen, M.H. Induction of multiple criteria optimal classification rules for biological and medical data. *Comput. Biol. Med.* **2008**, *38*, 42–52.
22. Liu, J.F.; Hu, Q.H.; Yu, D.R. A weighted rough set based method developed for class imbalance learning. *Inf. Sci.* **2008**, *178*, 1235–1256.
23. Grzymala-Busse. J.W.; Hu, M. A Comparison of Several Approaches to Missing Attribute Values in Data Mining. In Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing, Banff, AB, Canada, 16–19 October 2000; Springer-Verlag: Berlin, Germany, 2000; pp. 378–385.
24. Dong, G.; Zhang, Y.; Dai, C.; Fan, Y. The Processing of Information Fusion Based on Rough Set Theory. *J. Instrum. Meter China* **2005**, *26*, 570–571.
25. Wang, J.; Wang, Y. Multi-sensor information fusion based on Rough set. *J. Hechi Univ.* **2009**, *29*, 80–82.
26. Huang, C.C.; Tseng, T.L.; Chen, K.C. Novel Approach to Tourism Analysis with Multiple Outcome Capability Using Rough Set Theory. *Int. J. Comput. Intell. Syst.* **2016**, *9*, 1118–1132.
27. Luo, J.F.; Liu, Y.Y.; Qin, K.Y.; Ding, H. Incremental update of rough set approximation under the grade indiscernibility relation. *Int. J. Comput. Intell. Syst.* **2017**, *10*, 212.
28. Yuan, X.; Zhu, Q.; Lan, H. Multi-sensor information fusion based on rough set theory. *J. Harbin Inst. Technol.* **2006**, *38*, 1669–1672.
29. Khan, M.A.; Banerjee, M. A study of multiple-source approximation systems. *Lect. Notes Comput. Sci.* **2010**, *12*, 46–75.
30. Khan, M.A.; Banerjee, M. A perference-based multiple-source rough set model. *Lect. Notes Comput. Sci.* **2010**, *6068*, 247–256.
31. Md, A.K.; Ma, M.H. A modal logic for multiple-source tolerance approximation spaces. *Lect. Notes Comput. Sci.* **2011**, *6521*, 124–136.
32. Lin, G.P.; Liang, J.Y.; Qian, Y.H. An information fusion approach by combining multigranulation rough sets and evidence theory. *Inf. Sci.* **2015**, *314*, 184–199.
33. Balazs, J.A.; Velásquez, J.D. Opinion Mining and Information Fusion: A survey. *Inf. Fusion* **2016**, *27*, 95–110.
34. Zhou, J.; Hu, L.; Chu, J.; Lu, H.; Wang, F.; Zhao, K. Feature Selection from Incomplete Multi-Sensor Information System Based on Positive Approximation in Rough Set Theory. *Sens. Lett.* **2013**, *11*, 974–981.
35. Yu, J.H.; Xu, W.H. Information fusion in multi-source fuzzy information system with same structure. In Proceedings of the 2015 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 12–15 July 2015; pp. 170–175.
36. Kryszkiewicz, M. Rough set approach to incomplete information systems. *Inf. Sci.* **1998**, *112*, 39–49.
37. Kryszkiewicz, M. Rules in incomplete information systems. *Inf. Sci.* **1999**, *113*, 271–292.
38. Pawlak, Z.; Skowron, A. Rudiments of rough sets. *Inf. Sci.* **2007**, *177*, 3–27.
39. Dai, J.; Wang, W.; Xu, Q. An Uncertainty Measure for Incomplete Decision Tables and Its Applications. *IEEE Trans. Cybern.* **2013**, *43*, 1277–1289.
40. Dai, J.; Xu, Q. Approximations and uncertainty measures in incomplete information systems. *Inf. Sci.* **2012**, *198*, 62–80.
41. Khan, M.A.; Banerjee, M. Formal reasoning with rough sets in multiple-source approximation systems. *Int. J. Approx. Reason.* **2008**, *49*, 466–477.

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).