An Efficient Multidimensional Fusion Algorithm for IoT Data Based on Partitioning

Jin Zhou  
College of Computer Science and Technology, Jilin University, Changchun 130012, China

Liang Hu  
College of Computer Science and Technology, Jilin University, Changchun 130012, China

Feng Wang  
College of Computer Science and Technology, Jilin University, Changchun 130012, China

Huimin Lu  
College of Software, Changchun University of Technology, Changchun 130012, China.

Kuo Zhao  
College of Computer Science and Technology, Jilin University, Changchun 130012, China

Follow this and additional works at: https://tsinghuauniversitypress.researchcommons.org/tsinghua-science-and-technology

Part of the Computer Sciences Commons, and the Electrical and Computer Engineering Commons

Recommended Citation
An Efficient Multidimensional Fusion Algorithm for IoT Data Based on Partitioning

Jin Zhou, Liang Hu, Feng Wang, Huimin Lu, and Kuo Zhao*

Abstract: The Internet of Things (IoT) implies a worldwide network of interconnected objects uniquely addressable, via standard communication protocols. The prevalence of IoT is bound to generate large amounts of multsource, heterogeneous, dynamic, and sparse data. However, IoT offers inconsequential practical benefits without the ability to integrate, fuse, and glean useful information from such massive amounts of data. Accordingly, preparing us for the imminent invasion of things, a tool called data fusion can be used to manipulate and manage such data in order to improve process efficiency and provide advanced intelligence. In order to determine an acceptable quality of intelligence, diverse and voluminous data have to be combined and fused. Therefore, it is imperative to improve the computational efficiency for fusing and mining multidimensional data. In this paper, we propose an efficient multidimensional fusion algorithm for IoT data based on partitioning. The basic concept involves the partitioning of dimensions (attributes), i.e., a big data set with higher dimensions can be transformed into certain number of relatively smaller data subsets that can be easily processed. Then, based on the partitioning of dimensions, the discernible matrixes of all data subsets in rough set theory are computed to obtain their core attribute sets. Furthermore, a global core attribute set can be determined. Finally, the attribute reduction and rule extraction methods are used to obtain the fusion results. By means of proving a few theorems and simulation, the correctness and effectiveness of this algorithm is illustrated.

Key words: Internet of Things; data fusion; multidimensional data; partitioning; rough set theory

1 Introduction

The Internet of Things (IoT)[1, 2] has definitely received considerable interest from both academia and industry; this technology aims at developing and accomplishing the future Internet. The IoT refers to things having identities and virtual personalities operating within smart spaces that use intelligent interfaces to connect and communicate within social, environmental, and user contexts. There are two main integral factors in IoT: Internet and Thing. The Internet can be defined as the worldwide network of interconnected computer networks based on a standard communication protocol—the Internet suite (TCP/IP), while the Thing is an object that is not precisely identifiable. The IoT will create a range of potentially new products and services in many different domains, such as smart homes, e-health, automotive, transport and logistics, and environmental monitoring[3, 4].

We are preparing for an imminent invasion of things. With the development of IoT, more and more interconnected physical objects and devices (referred to as Things) and their virtual representations will be seamlessly integrated. The primary goal of interconnecting devices and collecting and processing...
data from them is to create situation awareness and enable applications, machines, and human users to better understand their surrounding environments, to make intelligent decisions’ and to better interact with the dynamics of their environments. However, IoT offers inconsequential practical benefits if it does not have the ability to integrate, fuse, and glean useful information from such massive amounts of data generated by a world of interconnected devices. Therefore, the fusing and mining high-dimensional data sets derived from the IoT proves to be a formidable challenge.

In the 1980s, rough set theory\[5, 6\], presented by the polish mathematician Prof. Pawlak, deals with the uncertainty and vagueness in data. Further, it can be used to effectively analyze large amounts of data without prior knowledge. It is a good tool to process data with missing values[7]. In rough set theory, two important processes are involved in data fusion. One is attribute reduction in which the basic knowledge expressions are acquired from information systems by eliminating redundancy attributes without modifying the classification accuracy of the original knowledge. The other process is rule extraction in which category representations that match certain probabilistic qualities are mined from multidimensional data.

Therefore, on the basis of the rough set theory, we propose an efficient fusion algorithm for multidimensional IoT data based on partitioning. The basic idea of this algorithm is that a large data set with higher dimensions can be transformed into relatively smaller data sets that can be easily processed. Therefore, firstly, we partition the high-dimensional data set into certain blocks of lower-dimensional data sets. Then, we compute the core attribute set of each block of data. Thereafter, we take the advantage of the core attribute sets of all data subset to determine a global core attribute set. Finally, based on this global core attribute set, we compute the reduction and mine the correlations among the multidimensional measurement data and certain interesting states with regard to the facilities or humans.

2 Data Fusion in IoT

The IoT is expected to usher in a world where physical objects are seamlessly integrated into information networks in order to provide advanced and intelligent services to human beings. Since it is associated with a large number of wireless sensor devices, IoT generates a large amount of data, which is massive, from multiple sources, heterogeneous, dynamic, and sparse. Accordingly, data fusion is an important tool for the manipulation and management of this data in order to improve processing efficiency and provide advanced intelligence.

The general definition of data fusion\[8, 9\] is that it is a formal framework that contains expressed means and tools for the alliance of data originating from different sources. It aims at obtaining information of greater quality: the exact definition of greater quality depends on the application. In the IoT environment, data fusion is also a framework that comprises theories, methods, and algorithms for interoperating and integrating multisource heterogeneous data from sensor measurements or other sources, combining and mining the measurement data from multiple sensors and related information obtained from associated databases, and achieving improved accuracy and more specific inferences than that obtained by using only a single sensor.

Recently, one of the most popular research topics in data fusion for IoT is the interoperability and integration\[5, 6\] of multisource heterogeneous data, including IoT data abstraction\[10, 11\] and access, linked sensor data\[12\], resource/service search and discovery\[13\], semantic Web technologies. Another popular research topic is big data management and mining\[14-17\] for gleaning useful information from the massive amount of data generated by such networks. These studies are mainly based on semantic Web technologies. Another popular research topic is data fusion theory and algorithm and the distributed information system technology\[18\]. In this paper, the proposed efficient fusion algorithm for multidimensional IoT data based on partitioning is related to a fusion method for big data. This algorithm focuses on the manner of improving the computational efficiency of data with higher dimensions. The fusion results will be discussed in future works.

3 Preliminaries of Rough Set Theory

Definition 1 An information system is the pair $S = (U, A)$, where $U$ is a non-empty finite set of objects, $A$ is a non-empty finite set of attributes, and for every $a \in A$, there is a mapping $a, a : U \rightarrow V_a$, where $V_a$ is...
called the value set of \( a \). An information system \( S = (U, C \cup D) \), where \( D \cap C = \emptyset \) is usually called a decision table. The elements of \( C \) are called the conditional attributes and \( D \) is the decision attribute set.

It may happen that some of the attribute values for an object are missing. To distinguish such a situation, a so-called null value, denoted by “*”, is usually assigned to such attributes. If \( V_a \) contains a null value for at least one attribute \( a \in A \), then \( S \) is called an incomplete information system; otherwise, it is a complete one.

**Definition 2** Let \( S = (U, A) \) be an information system and \( P \) be an attribute set, where \( P \subseteq A \). We define the following tolerance relation on \( U \).

\[
\text{SIM}(P) = \{(u, v) \in U \times U \mid \forall a \in P, (a(u) = a(v)) \lor (a(u) = * \land a(v) = *)\}
\]  

This tolerance relation is reflexive and symmetric, but not necessarily transitive.

\( S_P(u) \) is called a tolerance class of \( u \) under \( P \), which is the maximal set of objects that are possibly indistinguishable with \( P \) with \( u \).

\[
S_P(u) = \{v \in U \mid (u, v) \in \text{SIM}(P)\}
\]  

\( U/\text{SIM}(P) \) is the classification of \( U \) or the knowledge of \( U \) induced by \( P \).

\[
U/\text{SIM}(P) = \{S_P(u) \mid u \in U\}
\]

It should be noted that the tolerance classes in \( U/\text{SIM}(P) \) do not necessarily yield a partition of \( U \). They form a cover of \( U \) in general.

**Definition 3** Let \( S = (U, A) \) be an incomplete information system; \( X \), a subset of \( U \); and \( P \subseteq A \), an attribute set. In the rough set theory model, on the basis of the tolerance relation, \( X \) can be characterized as \( \text{SIM}(P)X \) and \( \text{SIM}(P)^X \), which are called the lower and upper approximations, respectively. Here,

\[
\left\{ \begin{array}{l}
\text{SIM}(P)X = \bigcup\{Y \in U/\text{SIM}(P) \mid Y \subseteq X\}, \\
\text{SIM}(P)^X = \bigcup\{Y \in U/\text{SIM}(P) \mid Y \cap X \neq \emptyset\}
\end{array} \right.
\]

We can redefine the \( P \)-lower and \( P \)-upper approximations of \( X \) by using the tolerance classes on \( U \).

\[
\left\{ \begin{array}{l}
\text{SIM}(P)X = \{u \in U \mid S_P(u) \subseteq X\}, \\
\text{SIM}(P)^X = \{u \in U \mid S_P(u) \cap X \neq \emptyset\}
\end{array} \right.
\]  

**Definition 4** Let \( S = (U, C \cup D) \) be an incomplete decision table, and the objects in \( U \) be partitioned into \( r \) mutually exclusive crisp subsets by the decision attribute set \( D \) and \( U/\text{ind}(D) = \{X_1, X_2, \ldots, X_r\} \). Given any subset \( P \subseteq C \) and the tolerance relation \( \text{SIM}(P) \) induced by \( P \), we can define the lower and upper approximations of the decision attribute set \( D \) as follows:

\[
\left\{ \begin{array}{l}
\text{SIM}(P)D = \{\text{SIM}(P)X_1, \text{SIM}(P)X_2, \ldots, \text{SIM}(P)X_r\}, \\
\text{SIM}(P)^D = \{\text{SIM}(P)X_1, \text{SIM}(P)X_2, \ldots, \text{SIM}(P)X_r\}
\end{array} \right.
\]

Let \( \text{POS}_P(D) = \bigcup_{i=1}^r \text{SIM}(P)X_i \), which is called the positive region of \( D \) with respect to the condition attribute set \( P \).

**Definition 5** Let \( S = (U, C \cup D) \) be an incomplete decision table, and \( P \subseteq C \). When \( \partial_P(x) = \{i \mid i = D(y), y \in S_P(x)\} \), then \( \partial_P \) is the generalized decision function in \( S \). If for any \( x \in U \), we always get \( |\partial_P(x)| = 1 \), then \( S \) is consistent; otherwise, it is inconsistent.

**Definition 6** Let \( S = (U, C \cup D) \) be an incomplete decision table, and \( P \subseteq C \). We say that \( D \) depends on \( P \) in degree \( k \), denoted by \( P \Rightarrow_k D \), if 

\[
k = \gamma(P, D) = |\text{POS}_P(D)| / |U|
\]

where \( |U| \) denotes the cardinality of \( U \).

If \( \exists(a \in P)(\gamma(P, D) = \gamma(P - \{a\}, D)) \), then the attribute \( a \) is unnecessary with respect to the decision attribute \( D \); otherwise, attribute \( a \) is necessary.

We say that \( P(P \subseteq C) \) is a \( D \)-reduction (reduction with respect to \( D \)) of \( C \), if \( \forall(a \in P)(\gamma(P, D) > \gamma(P - \{a\}, D)) \) and \( \text{POS}_C(D) = \text{POS}_P(D) \).

**Definition 7** The intersection of all the attribute reductions of \( C \) relative to the decision attribute set \( D \) is known as the core of \( C \) relative to \( D \) and is denoted as \( \text{Core}_C \).

Let \( S = (U, C \cup D) \) be an incomplete decision table, and \( P \subseteq C \). The discernibility matrix \(^{[20]}\) based on the tolerance relationship is defined as

\[
M_P = \begin{bmatrix}
m_P(i, j), & \text{min}\{|\partial_P(x_i)|, |\partial_P(x_j)|\} = 1; \\
\emptyset, & \text{else}
\end{bmatrix}_{n \times n}
\]

where

\[
m_P(i, j) = \begin{cases}
\{b \mid (b \in P) \land (b(x_i) \neq *) \land (b(x_j) \neq *) \land (b(x_i) \neq b(x_j))\}, & D(x_i) \neq D(x_j); \\
\emptyset, & \text{else}
\end{cases}
\]

and \( 1 \leq i, j \leq n = |U| \).

**Theorem 1** Let \( S = (U, C \cup D) \) be an incomplete decision table, \( P \subseteq C \), and \( M_P \) be the discernibility matrix based on the tolerance relationship. If any \( m_P(i, j) \in M_P \) consists of only one attribute \( a \), then the attribute \( a \) is one of the core attributes of \( C \) relative
to $D$.

There are two representative reduction criterions. One is to maintain the positive region of the target decision unchanged, and the other is to maintain the conditional entropy of the target decision unchanged.

**Definition 8** Let $S = (U, C \cup D)$ be an incomplete decision table. The conditional entropy is defined as

$$E(D \mid C) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |S_C(x_i) - |S_C(x_i) \cap S_D(x_i)|$$

where $S_C(x_i) \in U/SIM(C)$ and $S_D(x_i) \in U/SIM(D)$.

**Theorem 2** Let $S = (U, C \cup D)$ be an incomplete decision table. If and only if the attribute $a \in C$ is a core attribute of $C$, then under the tolerance relation, we have $POS^U_{C-\{a\}}(D) \subset POS^U_C(D)$, $\gamma(C, D) > \gamma(C - \{a\}, D)$ and $E(D \mid C - \{a\}) - E(D \mid C) > 0$.

**Definition 9** Let $S = (U, C \cup D)$ be an incomplete decision table, $P \subseteq C$, and $\forall a \in (C - P)$. Then, the significance measure of attribute $a$ in $P$ is defined as

$$\text{Sig}(a, P, D, U) = \frac{\gamma_{P,U}(a)}{|P|} - \gamma_P(D)$$

(11)

where $\gamma_P(D) = |POS^U_P(D)| / |U|$.

**Definition 10** Let $S = (U, C \cup D)$ be an incomplete decision table. Then, the number $\text{Supp}_x(C, D) = |C(x) \cap D(x)|$ is called the support of the decision rule $C \rightarrow_x D$ and the number

$$\sigma_x(C, D) = \text{Supp}_x(C, D) / |U|$$

(12)

is referred to as the strength of the decision rule $C \rightarrow_x D$.

With every decision rule $C \rightarrow_x D$, we associate the certainty factor of the decision rule, denoted by $\text{Cer}_x(C, D)$, which can be defined as follows.

$$\text{Cer}_x(C, D) = \frac{|C(x) \cap D(x)|}{|C(x)|}$$

(13)

If $\text{Cer}_x(C, D) = 1$, then $C \rightarrow_x D$ is called a certain decision rule; if $0 < \text{Cer}_x(C, D) < 1$, the decision rule is referred to as an uncertain decision rule.

Besides, we will also use a coverage factor of the decision rule, denoted as $\text{Cov}_x(C, D)$ and defined as

$$\text{Cov}_x(C, D) = \frac{|C(x) \cap D(x)|}{|D(x)|}$$

(14)

If $C \rightarrow_x D$ is a decision rule, then $D \rightarrow_x C$ is called an inverse decision rule. The inverse decision rule can be used to give an explanation (reason) for a decision.

## 4 Core Attributes in Positive Granulation World

A variety of attribute sets comprise granulation world. Further, such granulations can comprise a global knowledge space. Generally, the quality of knowledge in different granulation world is not equal. If the computation on global knowledge space can be transformed into computation on granulations that comprise global knowledge space, it is likely to process easily. Therefore, it is interesting in the relationships between the granulations and the global knowledge space.

**Definition 11** Let $S = (U, A)$ be an incomplete information system, $P \subseteq A$, and $Q \subseteq A$. We define a partial relation $\leq$ (or $\geq$) on $2^A$ as follows: We say that $Q$ is coarser than $P$ (or $P$ is finer than $Q$), denoted by $P \leq Q$ (or $P \geq Q$), respectively, if and only if $S_P(u_i) \subseteq S_Q(u_i)$ for $i \in \{1, 2, \cdots, |U|\}$. If $P \leq Q$ and $P \neq Q$, we say that $Q$ is strictly coarser than $P$ (or $P$ is strictly finer than $Q$) and denoted by $P < Q$ or $Q < P$. In fact, we can deduce that $(P \leq Q) \Leftrightarrow (S_P(u_i) \subseteq S_Q(u_i))$ and $(P < Q) \Leftrightarrow (S_P(u_i) \subset S_Q(u_i))$.

**Definition 12** Let $S = (U, A)$ be an incomplete information system, $P_1 \subseteq A$, and $P = \{P_1, P_2, \cdots, P_n\}$ denote a family of attribute sets with $P_1 \supseteq P_2 \supseteq \cdots \supseteq P_n$, where $P$ is called a positive granulation world.

Let $[P_i]$ denote the attribute set in which the attributes are contained by $P_i$.

**Theorem 3** Let $S = (U, C \cup D)$ be an incomplete decision table, and $P = \{P_1, P_2, \cdots, P_n\}$ denote a family of attribute sets with $P_1 \supseteq P_2 \supseteq \cdots \supseteq P_n$. If $\exists a \in [P_i] (1 \leq i \leq n)$ and attribute $a \notin \text{Core}_{P_i}^{C_U}$, then the attribute $a \notin \text{Core}_{P_k}^{U}(i + 1 \leq k \leq n)$.

**Proof** The classification accuracy of the family of attribute sets $P_i = \{P_1, P_2, \cdots, P_n\}$ equals the classification $[P_i]$. Suppose $\exists a \in [P_i] (1 \leq i \leq n)$, attribute $a \notin \text{Core}_{P_i}^{C_U}$, and $a \in \text{Core}_{P_k}^{U} (i + 1 \leq k \leq n)$, we have that $\text{POS}_{[P_k]}^U - \text{POS}_{[P_{i-1}]}^U(D) \subset \text{POS}_{[P_k]}^U(D)$ and $E(D \mid [P_k] - \{a\}) - E(D \mid [P_k]) > 0$.

(a) Let $x \in \text{POS}_{[P_k]}^U - \text{POS}_{[P_{i-1}]}^U(D)$. Then, $\exists y \in U$. $D(x) \neq D(y)$, $y \in S_{[P_{i-1}]}(x)$, and $y \notin S_{[P_k]}(x)$. That is to say, for $\forall b \in [P_{k}] - \{a\}$, $b(x) = b(y)$ and $a(x) \neq a(y)$. Further, as $[P_i] \subseteq [P_k]$, we have $\forall b \in [P_i] - \{a\}, b(x) = b(y)$.
and \(a(x) \neq a(y)\). According to Theorem 1, the attribute \(a \in \text{Core}_{P_i}^U\). Contradict.

(b) Since \(E(D \mid [P_k] - \{a\}) - E(D \mid [P_k]) > 0\), according to the definition of conditional entropy, we have that \(\sum_{j=1}^{[U]} |S_{P_j}(x_j)| < \sum_{j=1}^{[U]} |S_{[P_j] - \{a\}}(x_j)|\). That is to say, \(\exists x \in U\) and \(\exists y \in U, D(x) \neq D(y), y \in S_{[P_j] - \{a\}}(x), and y \notin S_{P_j}(x)\). We know that \(|S_{P_j}(x)| < |S_{[P_j] - \{a\}}(x)|\) and \(|S_{P_j}(x) \cap S_D(x)| < |S_{[P_j] - \{a\}}(x) \cap S_D(x)|\). Therefore, \(E(D \mid [P_k] - \{a\}) - E(D \mid [P_k]) > 0\). That is to say, for \(\forall b \in [P_k] - \{a\}, b(x) = b(y)\) and \(a(x) \neq a(y)\).

Further, as \([P_i] \subseteq [P_j]\), we have that \(\forall b \in [P_i] - \{a\}, b(x) = b(y)\) and \(a(x) \neq a(y)\). According to Theorem 1, the attribute \(a \in \text{Core}_{P_i}^U\). Contradict.

Consider (a) and (b), we can conclude that the attribute \(a \notin \text{Core}_{P_i}^U\). This completes the proof.

This theorem illustrates that if an attribute is not the core attribute of one granulation in a positive granulation world, it will not be the core attribute of any other finer granulation, including the conditional attribute set \(C\).

Theorem 4 Let \(S = (U, C \cup D)\) be an incomplete decision table and let \(P = \{P_1^*, P_2^*, \ldots, P_n^*\}\) denote a family of attribute sets with \(P_1^* \supseteq P_2^* \supseteq \cdots \supseteq P_n^*\) and \(P_1^* = \{P_1^*, P_2^*, \ldots, P_n^*\}\). According to Theorem 1, we get \(\text{Core}_{P_1}^U \subseteq \text{Core}_{P_2}^U \subseteq \ldots \subseteq \text{Core}_{P_n}^U\).

Proof

(a) Suppose the attribute \(a \in \text{Core}_{P_i}^U\) and \(a \notin \text{Core}_{P_j}^U\). That is to say, the attribute \(a \in [P_i]\) and \(a \notin [P_j]\). As proven in Theorem 3, the attribute \(a \notin \text{Core}_{P_j+1}^U\). Contradict. We can conclude that \(\text{Core}_{P_i+1}^U \subseteq \text{Core}_{P_i}^U \cup E\).

(b) As \(E \subseteq P_i^*, \) then \(E \supseteq P_i^* \supseteq [P_i+1]\). According to Theorem 3, we have \(\text{Core}_{P_i+1}^U \cap (E - \text{Core}_{P_i}^U) = \emptyset\). As \(\text{Core}_{P_i+1}^U \subseteq \text{Core}_{P_i}^U \cup E\), then \(\text{Core}_{P_i+1}^U \subseteq \text{Core}_{P_i}^U \cup \text{Core}_{P_i}^U\).

This completes the proof.

This theorem illustrates that any core attribute of a finer granulation in a positive granulation world belongs to the core attribute set of either the coarser granulation or the difference set between them.

Theorem 5 Let \(S = (U, C \cup D)\) be an incomplete decision table, and \(C = \{E_1, E_2, \ldots, E_n\}\) be a partition of the conditional attribute set \(C\). \(P = \{P_1^*, P_2^*, \ldots, P_n^*\}\) be the family of attribute sets with \(P_i^* = \{E_1, E_2, \ldots, E_i\}\). Then, for \(i = 1, 2, \ldots, n\), given that \(P_i = \{P_1^*, P_2^*, \ldots, P_n^*\}\), we have
\[
\text{Core}_{P_{i+1}}^U \subseteq (\text{Core}_{P_i}^U \cup \text{Core}_{E_{i+1}}^U)
\]

Typically,
\[
\text{Core}_{C}^U \subseteq (\text{Core}_{E_1}^U \cup \text{Core}_{E_2}^U \cup \ldots \cup \text{Core}_{E_n}^U)
\]

Proof As \(E = \{E_1, E_2, \ldots, E_n\}\) is a partition of the conditional attribute set \(C\). That is to say, \(\forall E_i, E_j \in E, E_i \cap E_j = \emptyset and \{E_1, E_2, \ldots, E_n\} = C\). As \(P_i^* = \{E_1, E_2, \ldots, E_i\}\), then \(P_i^* \supseteq P_j^* \supseteq \cdots \supseteq P_n^*\). Further, \([P_{i+1}] - [P_i] = E_{i+1}, i \in \{1, 2, \ldots, n - 1\}\). According to Theorem 4, we have \(\text{Core}_{P_{i+1}}^U \subseteq (\text{Core}_{P_i}^U \cup \text{Core}_{E_{i+1}}^U)\). This completes the proof.

This theorem illustrates that the core attribute set of the conditional attribute set \(C\) is the subset of the union of the core attribute sets of a partition of \(C\).

Theorem 6 Let \(S = (U, C \cup D)\) be an incomplete decision table and \(C = \{E_1, E_2, \ldots, E_n\}\) be a partition of the conditional attribute set \(C\); \(M_{E_i}\) is the discernibility matrix. Then, for every attribute \(a \in \bigcup_{i=1}^n \text{Core}_{E_i}^U\) if \(m(j, k)^{E_i}\) distinguishes \(a \in \text{Core}_{E_i}^U\) and for all \(m(j, k)^{E_i} = (1 \leq l \leq n, l \neq i)\), then \(\text{Core}_{C}^U = \bigcup \{a\}\).

Proof

(a) As \(m(j, k)^{E_i}\) distinguishes \(a \in \text{Core}_{E_i}^U\), i.e., \(m(j, k)^{E_i} = \{a\}\). Further, as all \(m(j, k)^{E_i}\) = \((1 \leq l \leq n, l \neq i)\), i.e., there is only one attribute \(a\) in the conditional attribute set \(C\) that can distinguish the objects \(x_j\) and \(x_k\). According to Theorem 1, we get \(a \in \text{Core}_{C}^U\).

(b) As proven in Theorem 5, we know that \(\text{Core}_{C}^U \subseteq \bigcup_{i=1}^n \text{Core}_{E_i}^U\). Hence, examining every attribute \(a \in \bigcup_{i=1}^n \text{Core}_{E_i}^U\), we will definitely get \(\text{Core}_{C}^U = \bigcup \{a\}\). This completes the proof.

This theorem illustrates that through the core attributes of the data subset based on the partition of the conditional attribute set \(C\), we can get the core attribute set of \(C\) relative to \(D\).

5 Fusion Algorithm for Multidimensional Data in IoT

Since a large number of different types of objects (such as sensors) would be interconnected in the IoT, data and information with high diversity and in large volumes would be generated. For elucidating the intelligence of IoT, it would be necessary to integrate and combine this multidimensional (multifaceted) data\(^{[16, 21]}\) in order...
to fuse and mine knowledge. This knowledge can provide specific advice and guide either our or machines actions. Therefore, it is imperative to process this deluge of multidimensional data and information in order to demonstrate the intelligence of IoT.

5.1 Modeling of the IoT data

We model the following problem: a specific amount of interesting data is obtained from \( N \) sensors deployed through certain sensor networks and the Web. Suppose the multisource heterogeneous information or data has been transformed and interoperated semantically. For sensor \( i \) (\( 1 \leq i \leq N \)) has types of measurement data. These \( N \) sensors, in total, have \( M \) types of measurement data. For each type of measurement data for a single sensor, there are \( T \) measured values collected at \( T \) different times \( t_i \) (\( i = 1, 2, \ldots, T \)). Figure 1 shows the structure of the multidimensional measurement data. We use the \( T \times \sum_{i=1}^{N} M_i \times N \) matrix \( R \) to express the \( \sum_{i=1}^{M} M_i \) types of measured values derived from \( N \) sensors at \( T \) different times.

In order to determine certain relationships between these multifaceted (multidimensional) measurement data and certain interesting states of the facilities or humans, types of data measured by all sensors can be expressed as attributes \( a_1, a_2, \ldots, a_n \), and the states can be expressed as the decision attribute set \( D \). Then, as shown in Figs. 2 and 3, we obtain the information system that this model intends to represent.

5.2 Multidimensional data fusion algorithm

In order to improve the computational efficiency of processing high-dimensional data in IoT, we propose an efficient multidimensional fusion method based on partitioning, as shown in Fig. 4. This algorithm is based on the idea that high-dimensional data can be divided into smaller pieces (granulations) that are easier to process and prior knowledge is unnecessary. Further, data with missing values can be processed.

**Level 1** The data preprocessing block is responsible for two tasks. One is to normalize the data in the information system \( S \), which can yield data that are easily comparable. The other is to replace the missing values with “*” to successfully process the missing data in the fusion algorithm. This level depends mostly on the data characters.

**Level 2** The multidimensional data partition block, based on the partition of attributes ruled by certain principles, enables the division of high-dimensional data into smaller blocks of lower-dimensional data sets.

As shown in Fig. 2, for one sensor, we use the \( M_i \times T \) matrix \( W \) to express the measured values derived from \( M_i \) types of measurements at \( T \) different times.
As shown in Fig. 3, we get \( M \) types of measurement data at the same time from \( N \) sensors, which are expressed by the matrix \( Q \).

\[
S = (W_1, W_2, \cdots, W_N)
\]  

(18)

Therefore, we ultimately get the relationships (fusion results) implied in the multidimensional data.

**Algorithm 1** Computation of core attribute set based on partitioning

| Input: Information system of Fig. 3 |
| Output: Core\(_C^U\). |
| (1) Begin |
| (2) For \( 1 \leq j \leq M \) |
| (3) \{For \( 1 \leq i \leq \left| U \right| \} |
| (4) \{For \( 1 \leq k \leq \left| U \right| \} |
| (5) \{Compute \( m(i,k)^Q_j \) \} ; \text{// Determine every discernibility matrix of the data subsets} \( M_{Q_j} \). |
| (6) If any \( m(i,k)^Q_j = \{a\} \) \{There is single one element in \( m(i,k)^Q_j \). |
| (7) \( Core_{Q_j}^U = Core_{Q_j}^U \cup \{a\} \} \} ; \text{// Determine every core attribute set of the data subsets} |
| (8) For every attribute \( a \in \bigcup_{j=1}^{M} Core_{Q_j}^U \) |
| (9) \{For every \( m(i,k)^Q_j = \{a\} \} |
| (10) \{For \( 1 \leq l \leq M, l \neq j \} |
| (11) \{If \( \bigcup_{l=1}^{M} m(i,k)^Q_j = \varnothing \} \} \} ; \text{// Determine the core attribute set of the global data from the union of every core attribute set of the data subsets} |
| (12) \( Core_{C}^U = Core_{C}^U \cup \{a\} \} \} ; \text{//} |
| (13) Return Core\(_C^U\); |
| (14) End |

**Algorithm 2** Computation of attribute reduction based on core attribute set which is gotten in Algorithm 1

| Input: Core\(_C^U\) |
| Output: Reduction |
| (1) Begin |
| (2) Reduction = Core\(_C^U\); |
| (3) Compute \( POS_U^C(D) \) and \( POS_{Reduction}^U(D) \); |
| (4) While \( POS_{Reduction}^U(D) \neq POS_U^C(D) \) do |
| (5) \{For every attribute \( b \in (C – Reduction) \} |
| (6) \{Compute Sig\(_b(C, Core_{C}^U, D, U)\); \} |
| (7) Reduction = Reduction \( \cup \{b_0\} \} ; \text{// where} \( b_0 = Max\{Sig\(_b(C, Core_{C}^U, D, U)\}; \} |
| (8) Compute \( POS_{Reduction}^U(D) \}; |
| (9) Return Reduction; |
| (10) End |
5.3 Computational efficiency of algorithms

The basic idea of the fusion algorithm used for multidimensional data based on rough set theory is that by applying certain type of partition rule to a conditional attribute set, a high-dimensional data set can be divided into a certain number of data subsets with relatively lower dimensions. These lower-dimensional data subsets can be easily processed and computed. In effect, we take advantage of all these sub-computations to achieve a global computation.

In the original method, the computation of the core attribute set for a global conditional attribute set is implemented by simply computing the global discernibility matrix. Its computational complexity is $O(\frac{C^2 |U|^2}{n})$. However, in the proposed method, we segment the conditional attribute set $C$ into several sections and compute the discernibility matrixes of these sections. Taking advantage of these results, the global core attribute set can then be computed. Therefore, its computational complexity is $O\left(\frac{|C|^2 |U|^2}{n}\right)$.

When compared to the original method, by means of serial processing, the proposed method may exhibit a little—or even no—improvement in the computational complexity. This is because the algorithm efficiency is determined by a number of factors, namely, the number of objects $U$, the number of conditional attributes $C$, and the manner of partitioning. These factors are interrelated, and therefore, influence the algorithm efficiency.

However, by means of parallel processing, the computation of the core attribute set for a global data set can be channeled to different computational nodes for processing. Generally, the computation of a discernibility matrix for big multidimensional data set is fairly complicated. This is because the more the dimensions and terms, the larger the discernibility matrix. This exerts a tremendous load on the computational resources. Therefore, a divide-and-rule strategy is a good idea for manipulating big multidimensional data sets, which definitely improves the computational complexity.

For illustrating the efficiency of the proposed algorithm, we deployed a wireless sensor network in three rooms of our lab for measuring ambient environmental data. The network employed a total of 20 sensors that included temperature sensors, light sensors, humidity sensors, and voltage sensors. Besides, air pressure sensors, carbon dioxide sensors, proximity sensors, triaxial accelerometer sensors, and so on can also be used as the wireless sensor nodes in the deployed network. Our experiment was designed to determine the type of indoor environment that is comfortable for certain people. This determination is complicated because a comfortable environment depends on the season, the time of day, weather, indoor facilities, personal preferences, and so on. Therefore, we integrate the sensor data, data from the Web (seasonal, weather), location information, and state of the facilities (e.g., air-conditioning, lights, windows, and curtains). In order to mine the relationships between the attributes, the sensor data and Web information are set as the condition attributes and states of the facilities and persons are set as the decision attributes. In this paper, we focus on the computational efficiency of a fusion algorithm for multidimensional data.

Algorithm 3 Computation of attribute reduction based on core attribute set

| Input: Information system Reduction $\cup D$ based on Algorithm 2 |
| Output: Rules: Reduction $\rightarrow_i D$ |
| (1) Begin |
| (2) For $1 \leq i \leq |U|$ |
| (3) $\{\text{Compute } \sigma_{t_i} (\text{Reduction}, D), \text{Cer}_{t_i} (\text{Reduction}, D), \text{ and Cov}_{t_i} (\text{Reduction}, D)\}$ |
| (4) If $\sigma_{t_i} (\text{Reduction}, D) \geq \alpha$ and $\text{Cer}_{t_i} (\text{Reduction}, D) \geq \beta$ and $\text{Cov}_{t_i} (\text{Reduction}, D) \geq \varepsilon$ |
| (5) Return Reduction $\rightarrow_i D$ |
| (6) End |

As shown in Fig. 5, the proposed novel fusion algorithm for multidimensional data performs better than the original algorithm in a simulated distributed computing environment.

6 Conclusions

In order to determine an acceptable quality of intelligence, diverse and voluminous data have to be combined, integrated, and fused. The computation of attribute reduction is proven to be a non-deterministic polynomial-time hard (NP-hard) problem. Therefore, IoT offers a formidable challenge in the computation and fusion of high-dimensional big data generated by the participating networks. In this paper, we propose an efficient computational method for multidimensional
Fig. 5 Comparison of computational times for fusing multidimensional measurement data.

The advantage of this algorithm is that it transforms a big problem into smaller problems that are easier to process; this substantially reduces the temporal and spatial computational complexities. Several theorems have been presented in order to illustrate the correctness of the proposed algorithm. Further, we perform a simulation to enumerate the better efficiency and effectiveness of the proposed algorithm. In a future study, the fusion results of the measurement data will be presented. The relationships between the number of dimensions, number of partitions, and volume of objects and their influence on the computation efficiency will be discussed. Furthermore, the effect of these relationships on the fusion algorithm and algorithm optimization will be investigated.

Acknowledgements

This work was supported in part by the National High-Tech Research and Development (863) Program of China (No. 2011AA010101), the National Natural Science Foundation of China (Nos. 61103197, 61073009, and 61240029), the Science and Technology Key Project of Jilin Province (No. 2011ZDGG007), the Youth Foundation of Jilin Province of China (No. 201101035), the Fundamental Research Funds for the Central Universities of China (No. 200903179), China Postdoctoral Science Foundation (No. 2011M500611), the 2011 Industrial Technology Research and Development Special Project of Jilin Province (No. 2011006-9), the 2012 National College Students’ Innovative Training Program of China, and European Union Framework Program: MONICA Project under the Grant Agreement Number PIRSES-GA-2011-295222.

References


[16] N. Pham and R. Pagh, A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional Data, in Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’12), Beijing, China, 2012, pp. 877-885.


Jin Zhou received his BS degree in computer science and technology from Liaoning University of Technology in 2004, followed by his MEng degree in computer applications technology in 2010 from same university. He is currently a PhD candidate in the College of Computer Science and Technology, Jilin University. His research interests are in information fusion, semantic web, and the Internet of Things.

Liang Hu received his MEng and PhD degrees in computer science from Jilin University in 1993 and 1999. Currently, he is a professor and doctoral supervisor at the College of Computer Science and Technology, Jilin University, China. His research areas are network security and distributed computing, including theories, models and algorithms of PKI/IBE, IDS/IPS, and grid computing. He is a member of the China Computer Federation.

Huimin Lu received her PhD degree in computer science and technology from Xian Jiaotong University, China in 2010. Currently, she is a working post doctor in computer science and technology research station, Jilin University, China. In addition, she is an associate professor at Changchun University of Technology, China. Her research interests include intelligent information processing and artificial intelligence.

Feng Wang received his BEng degree in information security from Harbin Institute of Technology in 2009, followed by his MEng degree in network and information security in 2012 from Jilin University. He is currently a PhD candidate in the College of Computer Science and Technology, Jilin University. His research interests are in computer networks, information security, the Internet of Things and Cyber Physical Systems.

Kuo Zhao received his BEng degree in computer software from Jilin University in 2001, followed by his MS degree in computer architecture in 2004 and PhD in computer software and theory from the same university in 2008. He is currently an associate professor in the College of Computer Science and Technology, Jilin University. His research interests are in operating systems, computer networks, and information security.