International Journal of scientific research and management (IJSRM)
||Volume||2||Issue||12||Pages||1777-1780||2014||
Website: www.ijsrm.in ISSN (e): 2321-3418

Granular Computing Based Data Mining In the View of Rough Set

[1]Fernandez raj D, [2]Gunasekaran G

1Research Scholar Department of Computer Science Engineering,
St. Peter's University, Avadi, Chennai, Tamil Nadu, India.
ferophd2014@gmail.com
2Professor and Principal, Meenakshi College of Engineering, K.K. Nagar,
Chennai, Tamil Nadu, India.

ABSTRACT
Granular computing (GRC) is an umbrella term to cover any theories, methodologies, techniques, and tools that make use of granules(i.e., subsets of a universe) in problem solving. The philosophy of granular computing has appeared in many fields, and it is likely playing a more and more important role in data mining. Rough set theory is a very important paradigms of granular computing, are often used to process vague information in data mining. In this chapter, based on the opinion of data is also a format for knowledge representation, a new understanding for data mining, domain-oriented data-driven data mining (3DM), is introduced at first. Its main idea is that data mining is a process of knowledge transformation. Then, the relationship of 3DM and GrC, especially from the view of rough set is discussed. Then, some examples are used to illustrate how to solve real problems in data mining using granular computing. Combining rough set theory, a flexible way for processing incomplete information systems is introduced firstly. Then, a high efficient attribute reduction algorithm is developed by translating set operation of granules into logical operation of bit strings with bitmap technology. Finally, two rule generation algorithms are introduced, and experiment results show that the rule sets generated by these two algorithms are simpler than other similar algorithms.

1. INTRODUCTION

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. It is necessary to acquire useful knowledge from large quantity of data. Traditionally, data mining is considered as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. That is to say, knowledge is generated from data. But in our opinion, knowledge is originally existed in the data, but just not understandable for human. In a data mining process, knowledge existed in a database is transformed from data format into another human understandable format like rule.

The philosophy of granular computing has appeared in many fields, and it is likely playing a more and more important role in data mining. Rough set theory and fuzzy set theory are two very important paradigms in granular computing.In this chapter, a new understanding for data mining, domain-oriented data-driven data mining (3DM), will be proposed.

2. BASIC CONCEPTS OF RELATED THEORIES

Domain-Oriented Data-Driven Data Mining (3DM)

Data mining (also known as Knowledge Discovery in Databases - KDD) is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. There are many commonly used techniques in data mining like artificial neural networks, fuzzy sets, rough sets, decision trees, genetic algorithms, nearest neighbor method, statistics based rule induction, linear regression and linear predictive coding.

Unfortunately, most data mining researchers pay much attention to technique problems for developing data mining models and methods, while little to basic issues of data mining. What is data mining? Our answer would be "data mining is a process of knowledge transformation". It is consistent with the process of human knowledge understanding. In our opinion, data is also a format for knowledge representation. The knowledge we mined from data was originally stored in data. Unfortunately, we cannot read, understand, or use it, since we cannot understand data.In a domain-driven data mining process, user's interesting, constraint, and prior domain knowledge are very important. An interaction between user and machine is needed. This is so called domain-oriented data-driven data mining (3DM)

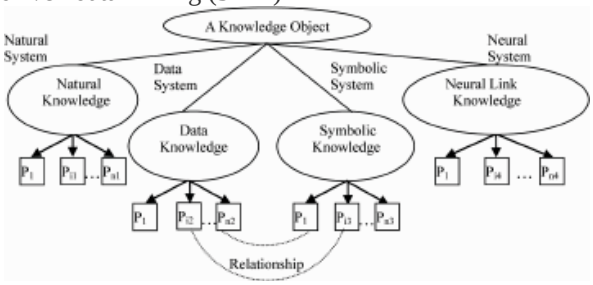

Figure 1. Knowledge transformation framework for data mining

Granular Computing

Human problem solving involves the perception, abstraction, representation and understanding of real world problems, as well as their solutions. Granular computing (GrC) is an emerging conceptual and computing paradigm of information processing at different levels of granularity. It has been motivated by the increasingly urgent need for intelligent processing for large quantities heterogeneous data. By taking human knowledge generation as a basic reference, granular computing offers a landmark change from the current machine-centric to human-centric approach for information and knowledge.

In our opinion, granular computing is a conceptual framework for data mining. The process of data mining is a transformation of knowledge in different granularities. In general, the original data is not understandable for human. However, human is often sensitive with knowledge in a coarser granularity. So, the process of data mining is to transform the knowledge from a finer granularity to a coarser granularity

Rough Set Theory

The theory of rough set is motivated by practical needs to interpret, characterize, represent, and process indiscernibility of individuals.

A decision information system is defined as $S =< U, A, V, f >$, where U is a non-empty finite set of objects, called universe, A is a non-empty finite set of attributes, $A = C \cup D$, where C is the set of condition attributes and D is the set of decision attributes. With every attribute $a \in A, V_a$ denotes the domain of attribute a. Each attribute has a determine function $f: U \times A \rightarrow V$.

Given a decision information system $S =< U, A, V, f >$, each subset of attribute $B \subseteq A$ determines an indiscernibility relation $IND(B) = \{(x, y) \mid (x, y) \in U \times U, \forall_{b \in B}(b(x) = b(y))\}$. Obviously, the indiscernibility relation *IND (B)* is an equivalence relation on U (reflexive, symmetric and transitive). The quotient set of equivalence classes induced by *IND (B)*, denoted by *U/IND (B)*, forms a partition of U, and each equivalence class of the quotient set is called an elementary set.

Let $S =< U, A, V, f >$ be a decision information system, for any subset X C U and indiscernibility relation *IND* (B), the B lower and upper approximation of X is defined as:

$$B_{-}(X) = \bigcup_{Y_i \in U/IND(B) \wedge Y_i \subseteq X} Y_i$$

$$B^{-}(X) = \bigcup_{Y_i \in U/IND(B) \wedge Y_i \cap X \neq \varnothing} Y_i .$$

If $B_{-}(X) = B^{-}(X) X$ is definable with respect to *IND(B)*. Otherwise, X is rough with respect to *IND(B)*. The lower approximation is the union of elementary $B_{-}(X)$ sets which are subsets of X, and the upper approximation $B^{-}(X)$ is the union of elementary $B_{-}(X)$ sets which have a non-empty intersection with X. That is, is the greatest definable set contained by X, while $B^{-}(X)$ is the least definable set containing X.

The lower approximation $B_{-}(X)$ is also called the positive region, the complement of the upper approximation $B^{-}(X)$ is called the negative region, and the difference of the upper approximation $B^{-}(X)$ with the lower approximation B (X) is called boundary region. The relationship among them is as Figure 2.

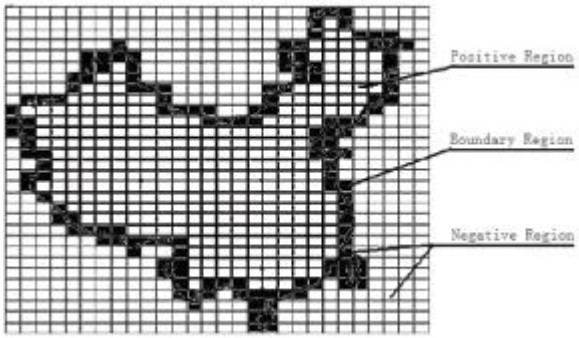

Figure 2. The positive region, negative region, and boundary region of a China map

Rough set theory provides a systematic method for representing and processing vague concepts caused by indiscernibility in situations with incomplete information or a lack of knowledge. In the past years, many successful applications can be found in various fields, such as process control, economics, medical diagnosis, biochemistry, environmental science, biology, chemistry psychology, conflict analysis, emotion recognition, video retrieval, and so on.

Fuzzy Set Theory

The notion of fuzzy set theory provides a convenient tool for representing vague concepts by allowing partial memberships.

In classical set theory, an object either belongs to a set or does not. That is:

$$A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$, where $A(x)$ is a characteristic function of set A.

An object can partially belong to a set in fuzzy set. Formally, consider a fuzzy set A, its domain D, and an object x. The membership function / specifies the degree of membership of x in A, such that

$\mu_A(X): D \rightarrow [0, 1]$.

For example, given a real domain $\mathbb{R}$, let A be a fuzzy set of numbers which are much greater than 0, the membership function with regard to A is

$\mu_A(X): \mathbb{R} \rightarrow [0, 1]$:

$$\mu_A(x) = \begin{cases} 0 & , x \leq 0 \\ \dfrac{1}{1+\dfrac{100}{x^2}} & , x > 0 \end{cases} .$$

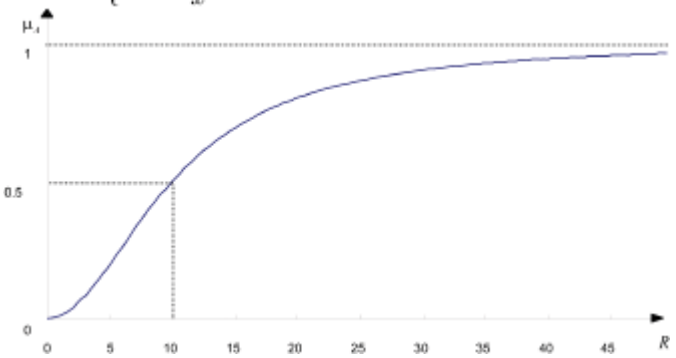

Figure 3. The membership function $\mu_A(x)$ of x in fuzzy set A

In Figure 3, it means x does not belong to A when $\mu_A(x) = 0$, whereas it means x completely belongs to $\mu_A(x)$ when $\mu_A(x) = 1$. Intermediate values represent varying degree of membership, for example, $\mu_A(x)(10) = 0.5$.

3. GRANULAR COMPUTING BASED DATA MINING

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market

related - also known as "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.

In granular computing, there are two important issues. One is the construction of granules, and the other is the computation with granules. Rough set uses equivalence relation to partition the universe into a granular space. However, it is difficult to define an equivalence relation in an incomplete information system. In addition, developing high efficient algorithms and increasing the generalization ability of knowledge are another two important issues in data mining.

Some Extended Models of Classical Rough Set Theory:

For an object *x* of an information system, let a(x) be its value on attribute *a*, we can define the following rough logic formulas:

1. *(a, v)* is an atomic formula. All atomic formulas are formula.
2. if *Ø* and *x* are both atomic formulas, then ¬*Ø, Ø Λ x, Ø ∨ x, Ø → X, Ø ↔ X* are also formulas.
3. Formulas resulted by using logic function ¬, Λ, ∨, →, ↔ on formulas defined in (1) and (2) in finite steps are also formulas.

According to the above definition, in the extension of rough set theory based on tolerance relation, objects with same value or "*" on an attribute can constitute a set. Therefore, we can define granules with this set.

Let $f^1(a, v)$ be a set of objects, $\forall_{x \in f^{-1}(a,v)} (a(x) = v \vee a(x) = *)$. Then, we can define a granule: $Gr^* = ((a, v), f^1(a, v))$, where *(a, v)* is called as the syntax of granular Gr^* and Gr^* an atomic granule. Suppose φ be a logic combination of atomic formulas, $F^{-1}(\varphi)$ be an object set, and all the objects in $F^{-1}(\varphi)$ satisfy logic combination φ. Granule $Gr^* = (\varphi, f^1(\varphi))$ is called a combination granular.

$gs(Gr^*)$ is a mapping function from granules to object sets,

$$\forall_{Gr^*} (Gr^* = (\varphi, f^{-1}(\varphi)) \Rightarrow gs(Gr^*) = f^{-1}(\varphi))$$

Rules for Granular Computing

Suppose $Gr_1^* = (\varphi, f^1(\varphi))$ and $Gr^2 * = (\varphi, f^1(\varphi))$ be two granules, according to the five logic conjunctions in classic logic theory, the rules of granular computing can be defined as follows:

1. $\neg Gr_1^* = \neg(\varphi, f^{-1}(\varphi)) = (\neg\varphi, f^{-1}(\neg\varphi))$
2. $Gr_1^* \wedge Gr_2^* = (\varphi \wedge \psi, f^{-1}(\varphi \wedge \psi))$
3. $Gr_1^* \vee Gr_2^* = (\varphi \vee \psi, f^{-1}(\varphi \vee \psi))$
4. $Gr_1^* \rightarrow Gr_2^* = (\varphi \rightarrow \psi, f^{-1}(\varphi) \subseteq f^{-1}(\psi))$
5. $Gr_1^* \leftrightarrow Gr_2^* = (\varphi \leftrightarrow \psi, f^{-1}(\varphi) = f^{-1}(\psi))$

Here, the logic conjunction → is not an implication relation between granules, but an inclusion relation between granules.

In order to generate knowledge with granular computing theory, we should have enough granules from an incomplete information system. These granules should contain enough information about the original incomplete information system. That is, we need to decompose the original incomplete information system into granules.

Given an incomplete information system $S =< U, A, V, f >$, suppose *GrS* be a set of granules. $\forall Gr^* \in GrS,\ Gr^* = (\varphi, f^{-1}(\varphi))$, where $\varphi = (a_1, v_1) \wedge (a_2, v_2) \wedge \ldots \wedge (a_m, v_m), (m = |C|, v_i \neq *)$, and $\forall_{x \in U} \exists_{Gr^*} (x \in gs(Gr^*))$. We call this kind of granule sets as a granule space of the original information system *S*.

A Rule Generation Algorithm Based on Granular Computing

An information table consists of some instances. If it represents the whole granule space, it can be divided into some small spaces. Each small space would be taken as a basic granule. These basic granules are further composed or decomposed into new granules so that new granules could describe the whole problem space or solve the problem at different hierarchies. During the process of problem solving, how to compose or decompose basic granules into new rule granules, and adjust the solution and granule space to improve the algorithm's efficiency, are two key aspects of the algorithm.

In order to improve the performance of rule granule generation, RGAGC adjusts the solution space with the "false preserving" property of quotient space theory. That is, if RGAGC cannot generate rule granules from the granule space, it should enlarge the current solution space, so that it can generate rule granules from the granule space, otherwise it generates rule granules from the current solution space directly. The above steps are repeated until rule granules contain all instances of an information table. Finally, rules are generated from these rule granules.

4. CONCLUSION

In this chapter, a new understanding of data mining, domain-oriented data-driven data mining (3DM), is introduced. Moreover, its relationship with granular computing is analyzed. From the view of granular computing, data mining could be considered as a process of transforming the knowledge from a finer granularity to a coarser granularity. Rough set and fuzzy set are two important computing paradigms of granular computing. For their ability in processing vague information, they were often used in data mining. Some applications of granular computing in data mining are introduced in the views of rough set. Although several problems in data mining have been partial solved by rough set theory, there are some problems needed to be further studied.

Uncertainty measure is a key problem for computing significant of attribute, attribute core, and attribute reduction in rough set theory. The uncertainty measure can be classified into uncertainty of knowledge, uncertainty of rough set, and uncertainty of decision. The uncertainty measures in Pawlak approximation space have been studied deeply, while the uncertainty measures in covering approximation space have not obtain adequate attention. In addition, what are the axioms of uncertainty measures? This is still an open problem.

Bitmap technology is a method computing with bit string. Because it is machine oriented, it can improve the computation efficiency evidently. In this chapter, we encoded granules with bit strings, and then developed a high efficient attribute reduction algorithm by translating set operation of granules into logical operation of bit strings. That is to say, if a problem

could be converted into a bit computable problem, then the efficiency of this problem can be improved by bitmap technology.

5. REFERENCES:

1.http://202.154.59.182/mfile/files/Information%20System/Informatics%20Engineering%20and%20Information%20Science%3B%20ICIEIS%202011%20PART%20II/Chapter%2052%20Granular%20Computing%20Based%20Data%20Mining%20in%20the%20Views%20of%20Rough%20Set%20and%20Fuzzy%20Set.pdf

2.www.isical.ac.in/~sankar/paper/iib_vol13no1_article1.pdf

3. http://www.igi-global.com/chapter/granular-computing-based-data-mining/44703

4. http://link.springer.com/chapter/10.1007/978-3-642-25453-6_52

5. http://en.wikipedia.org/wiki/Granular_computing

6. Granularity of knowledge, indiscernibility and rough sets,by pawlak.z

7. Rough Sets and Data Mining Analysis of Imprecise Data by T.Y.Lin. and N.Cercone

AUTHORS BIOGRAPHY

D. Fernandez raj was born in Sirkali, Tamilnadu, India, in 1973. He received the Bachelor degree in Mathematics (1995) , the Batchler degree in Education (1996), and the Master degree in English (2006) from the Annamalai University, Chidambaram, Tamilnadu. He received professional Master degree in Computer Science Applications (2002) from the Bharadhidasan University,Tiruchirapalli, Tamilnadu and the Master degree in Business Administration in Computer systems (2010) from the University of Madras , Chennai, Tamilnadu . He is currently pursuing the Ph.D degree with the Department of Computer Science and Engineering, St. Peter's University, Chennai. His research interests include Data Mining concepts, Big Data analytics and web data mining.

Dr. G. GUNASEKARAN, Principal, Meenakshi College of Engineering was born in Tamilnadu, India in 1965. He received the Bachelor degree in Computer Science and Engineering from the Madurai Kamaraj University, in 1989 and the Master degree in Computer Science and Engineering from Jadavpur University, Kolkata, in 2001. He got his Ph.D. degree from the Department of Computer Science and Engineering, Jadavpur University, Kolkata in 2009. His research interests include Data Mining, Bio informatics, Software Engineering, Graphics and Multimedia.