Volume 2, No. 1, Jan-Feb 2011
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Granular Approach of Knowledge Discovery in Databases

P. R. Pal*
P G Deptt. Of Computer Science & Applications
Shri Vaishnav Institute of Management
Indore, M.P. India
prpal@rediffmail.com

R.C. Jain
Department of Computer Applications
Samrat Ashok Technological Institute (Degree)
Vidisha, M.P. India
dr.jain.rc@gmail.com

Abstract: The main objective of this paper is to examine the basic principles of granular computing and apply them for knowledge discovery in databases. Granular computing is an emerging field of research that provides a way of studying many issues and solving problems related to them. The paper consist introduction and overview of granular computing. It also examines some of those issues, including data and knowledge representation and processing. It is demonstrated that one of fundamental task of data mining is finding for knowledge up to right level of granularity in data and knowledge representation. Finally it concludes on that granular computing is powerful view that can be used to model many problems.

Keywords: KDD, Granular Computing, Data Mining.

I. INTRODUCTION

Granular Computing is a general computation theory for effectively using granules such as classes, clusters, subsets, groups and intervals to build an efficient computational model for complex applications with huge amounts of data, information and knowledge. Though the label is relatively recent, the basic notions and principles of granular computing, though under different names, have appeared in many related fields, such as information hiding in programming, granularity in artificial intelligence, divide and conquer in theoretical computer science, interval computing, cluster analysis, fuzzy and rough set theories, neutrosophic computing, quotient space theory, belief functions, machine learning, databases, and many others [6, 7, 8]. In the past few years, we have witnessed a renewed and fast growing interest in Granular Computing [6, 9, 10].

A subset of universe is called granule in granular computing. Granular computing is collection of theories, methodologies, techniques and tools that makes use of granules in problem solving [11, 12]. The way by which granules are used to solve problems is called granular approach. Ideas and principles of granular computing have been studied under various names in different fields. Many models have also been proposed for the same. The results enhance our understanding of granular computing

Even various models for granular computing have been proposed, but still there is lack of well-accepted framework. The results from the recent studies have shown that granular computing provides a common and conceptual framework for modeling human thinking and problem solving [13]. Granular computing framework deals with structured thinking at philosophical level and structured problem solving at application level. Structured thinking provides guidelines to structured problem solving while structured problem solving implements the philosophy of structured thinking.

The rest of this paper is organized as follows: section 2 presents the overview of granular computing. Section 3 presents overview of knowledge discovery in databases. Section 4 describes the role of granular computing in knowledge discovery and at last some remarks are given in section 5.

II. OVERVIEW OF GRANULAR COMPUTING

A subset of universe is called granule. Its gradients are subsets, classes, and clusters of the universe. Granular computing is new powerful philosophical view and a general problem solving theory. Granular computing can be studied in two steps: construction or representation of granules that concerns about the organization of granules in terms of levels, networks and hierarchies. Second is computation with granules or process. It deals with methods that manipulate granules and granular structure. Here is a list of some research areas in computer science where granular computing can play wide role.

A. Bio-informatics
B. e-Business
C. Security
D. Machine learning
E. Knowledge discovery in databases / Data mining
F. High-performance computing and
G. Wireless mobile computing in terms of efficiency, effectiveness, robustness and uncertainty.

III. OVERVIEW OF KNOWLEDGE DISCOVERY

Knowledge is a physical, mental or electronic record of relationships believed to exist between real or imaginary entities, forces and phenomena. Knowledge is gained either by experience, learning and perception or through association and reasoning.

Knowledge discovery is the process of identifying a valid, potentially useful and ultimately understandable structure in data. This process involves selecting or sampling data from data warehouses, cleaning or preprocessing it, transforming or reducing (if needed), applying data mining components to produce structure and then evaluating the derived structures [1, 2, 3].

The structures that are outcomes of the data mining process must be valid, understandable and of user interest so that these can be considered as knowledge. The process or knowledge discovery in databases consists, sequence of following steps:

A. Data cleaning - to remove noise and inconsistent data.
B. Data integration - where multiple data sources may be combined.
C. Data selection - where data, relevant to analysis task are retrieved from the databases.
D. Data transformation - where data are transformed or reduced into appropriate forms for mining.
E. Data mining - an essential process that applies intelligent methods to extract data patterns.
F. Pattern evaluation - to identify interesting patterns representing knowledge based on interestingness measures.
G. Knowledge representation - visualization and knowledge representation techniques are used to present the mined knowledge to the user.

IV. KNOWLEDGE DISCOVERY USING GRANULAR COMPUTING

Granular computing tries to mimic the perception and the societal instinct of humans when grouping similar items together. Data granulation [4], [5] is achieved by a simple two step iterative process that involves the following steps:

A. Find the two most 'compatible' information granules and merge them together as a new information granule containing both original granules.
B. Repeat the process of finding the two most compatible granules until a satisfactory data abstraction level is achieved.

The most important concept of the above process is the definition of the compatibility measure. This can be purely geometrical (distance between granules, size of granules, volume of granules), density driven (ratio of cardinality versus granule volume) or similarity driven. In this case the compatibility measure is a function of the distance between the granules and a function of the information density of the newly formed granule. A mathematical representation of the compatibility criterion is given in following equation.

$$\text{Compatibility} = f\,(w_1.\text{Distance}, w_2.\text{Density})$$

$$\text{Distance} = \sum_{i=1}^{k} (\text{position of granule}_b - \text{position of granule}_a)_i$$

$$\text{Density} = \frac{\text{Granule cardinality}}{\text{Granule volume}} = \frac{\text{no. of sub-granules}}{\prod_{i=1}^{k} \text{granule length}_i}$$

Where w_1, w_2 are weights for balancing the distance/density requirements and k the dimensionality of the data space.

Even though this process can be accidentally identified as hierarchical clustering there is a major difference; each granule consists of the same objects (sub-granules). In hierarchical clustering new objects are created and the boundaries of the new clusters can be in an area where no data are present. The growth of clusters allows strong linkage between the original data set (transparency) and it allows visual monitoring methods for terminating granulation. By considering the merging of each set of granules as some information condensation or information loss, it is possible to link information loss to the 'merged distance' and therefore plot an information loss graph. This graph can be used online or off-line as a criterion for terminating the granulation process.

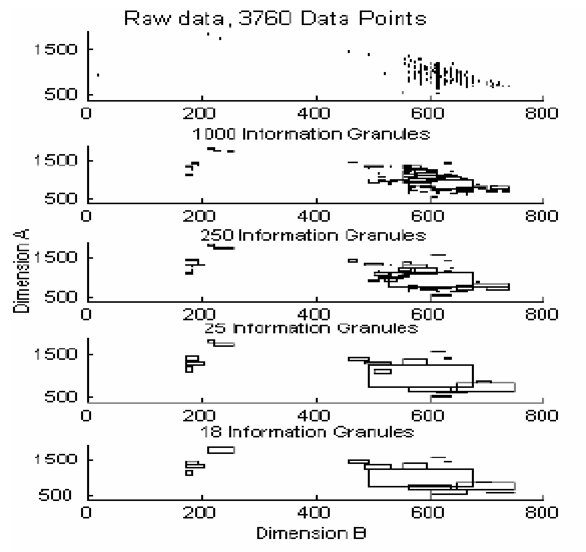

Figure 1: Iterative Information Granulation

Figure 1 shows snapshots of a 2-dimensional data granulation. Dimension A has units between 0-2000 and Dimension B between 0-800. The first snapshot, shown at the top of the figure, is the representation of the raw, pre-granulated, data, consisting of 3760 data points. As the iterative granulation algorithm progresses, snapshots of the granules are shown, consisting of 1000 granules (2nd snapshot of data), 250 granules, 25 and finally 18 granules. The added information collected during the granulation process is stored, and it consists of the cardinality and the multidimensional length of each granule.

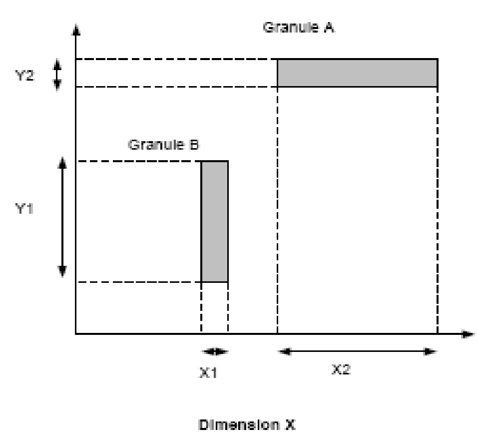

Figure 2: Orientation of granules

The following granulation example shows how the transparency and the additional information of the Granular Computing process can assist the modeling process. Granules A and B (figure 2) have same size, cardinality and density, but are of different orientation (90deg difference). Considering the fuzzy modeling Granular Computing structure, every granule consists of a linguistic rule; in this case:

GrA: if X = X2 then Y = Y2

GrB: if X = X2 then Y = Y1

Rule GrA has more output sensitivity (narrow space) and less input sensitivity (wide space). On the other hand, GrB has more input sensitivity rather than output one. By directing the algorithm towards solutions of type A or type B granules it is possible to enhance the model's performance, i.e. by increasing the sensitivity of an important variable.

V. CONCLUSION

Knowledge discovery in databases using Granular Computing is discussed in this paper. By taking advantage of the ability of Granular Computing to capture knowledge in a transparent and linguistic way it is possible to build very transparent systems capable of modeling various complex processes. The transparency of Granular Computing will always help to develop in the area of incremental learning, and system adaptation. By granulating and capturing information of new data sets using Granular Computing it is possible to explore the technique of combining new knowledge with the existing rule-base without any significant loss of performance.

VI. REFERENCES

[1] David Hand, H Mannila and P Smith, "Principles of Data Mining" Prentice Hall of India, 2001.

[2] J. Han and M. Kamber, "Data Mining: Concepts and techniques ", Academic Press, 2001.

[3] Arun Kumar Pujari, "Data Mining Techniques ", Universities Press (India), by 2001.

[4] Bargiela, A and W Pedrycz , 2003," Granular Computing An Introduction", Kluwer Academic Publishers, 2003

[5] Panoutsos, George and Mahdi Mahfouf, 2005, "Granular Computing and Evolutionary Fuzzy Modelling for Mechanical Properties of Alloy Steels", IFAC 2005, Proceedings of, 16th IFAC World Congress, July 4-8 2005, Prague Czech Republic, International Federation of Automatic Control (IFAC).

[6] Bargiela, A. and PedryczW. Granular Computing: an Introduction, Kluwer Academic Publishers, Boston, 2002.

[7] Hirota, K. and Pedrycz,W. Fuzzy computing for data mining, Proceedings of the IEEE, 87, 1575-1600, 1999.

[8] Lin, T.Y. Granular computing, Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Proceedings of the 9th International Conference (RSFDGrC 2003), LNCS 2639, 16-24, 2003.

[9] Inuiguchi,M., Hirano, S. and Tsumoto, S. (Eds.) Rough Set Theory and Granular Computing, Springer, Berlin, 2003.

[10] Liu, Q. and Jiang, S. L. Reasoning about information granules based on rough logic, International Conference on Rough Sets and Current Trends in Computing, 139-143, 2002.

[11] Yao, Y.Y. Granular computing: basic issues and possible solutions, Proceedings of 5th joint conference on Information Science, 186-189, 2000.

[12] Yao, Y.Y. and Zhong, N. Potential application of granular computing in knowledge discovery and data mining, proceedings of World Multiconference on Systemics, Cybernatics and Informatics, 19, 111-127, 1997.

[13] Yao, Y.Y. Perspectives of granular computing, Proceedings of 2005 IEEE International Conference on granular computing, Vol. 1, 85-90, 2005.

