Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Sciences 163 (2004) 5–12
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Soft data mining, computational theory of perceptions, and rough-fuzzy approach

Sankar K. Pal *

Machine Intelligence Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700 108, India

Received 8 May 2002; accepted 17 March 2003

Abstract

Data mining and knowledge discovery is described from pattern recognition point of view along with the relevance of soft computing. Key features of the computational theory of perceptions and its significance in pattern recognition and knowledge discovery problems are explained. Role of fuzzy-granulation (f-granulation) in machine and human intelligence, and its modeling through rough-fuzzy integration are discussed. Merits of fuzzy granular computation, in terms of performance and computation time, for the task of case generation in large scale case-based reasoning systems are illustrated through an example.

Keywords: Soft computing; Fuzzy granulation; Granular computation; Rough sets; Case-based reasoning

1. Introduction

In recent years, the rapid advances being made in computer technology have ensured that large sections of the world population have been able to gain easy access to computers on account of falling costs worldwide, and their use is now commonplace in all walks of life. Government agencies, scientific, business and

* Tel.: +91-33-25778085; fax: +91-33-25783357.
E-mail address: sankar@isical.ac.in (S.K. Pal).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2003.03.014

commercial organizations are routinely using computers not just for computational purposes but also for storage, in massive databases, of the immense volumes of data that they routinely generate, or require from other sources. Large-scale computer networking has ensured that such data has become accessible to more and more people. In other words, we are in the midst of an information explosion, and there is urgent need for methodologies that will help us bring some semblance of order into the phenomenal volumes of data that can readily be accessed by us with a few clicks of the keys of our computer keyboard. Traditional statistical data summarization and database management techniques are just not adequate for handling data on this scale, and for extracting intelligently, information or, rather, knowledge that may be useful for exploring the domain in question or the phenomena responsible for the data, and providing support to decision-making processes. This quest had thrown up some new phrases, for example, *data mining* [1,2] and *knowledge discovery in databases* (*KDD*) which are perhaps self-explanatory, but will be briefly discussed in the following few paragraphs. Their relationship with the discipline of pattern recognition (PR), certain challenging issues, and the role of soft computing will also be mentioned.

The massive databases that we are talking about are generally characterized by the presence of not just numeric, but also textual, symbolic, pictorial and aural data. They may contain redundancy, errors, imprecision, and so on. KDD is aimed at discovering natural structures within such massive and often heterogeneous data. Therefore PR plays a significant role in KDD process. However, KDD is being visualized as not just being capable of knowledge discovery using generalizations and magnifications of existing and new pattern recognition algorithms, but also the adaptation of these algorithms to enable them to process such data, the storage and accessing of the data, its preprocessing and cleaning, interpretation, visualization and application of the results, and the modeling and support of the overall human–machine interaction. What really makes KDD feasible today and in the future is the rapidly falling cost of computation, and the simultaneous increase in computational power, which together make possible the routine implementation of sophisticated, robust and efficient methodologies hitherto thought to be too computation-intensive to be useful. A block diagram of KDD is given in Fig. 1 [3].

Data mining is that part of knowledge discovery which deals with the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data, and excludes the knowledge interpretation part of KDD. Therefore, as it stands now, data mining can be viewed as applying PR and machine learning principles in the context of voluminous, possibly heterogeneous data sets. Furthermore, soft computing-based (involving fuzzy sets, neural networks, genetic algorithms and rough sets) PR methodologies and machine learning techniques hold great promise for data mining. The motivation for this is provided by their ability to handle imprecision, vagueness,

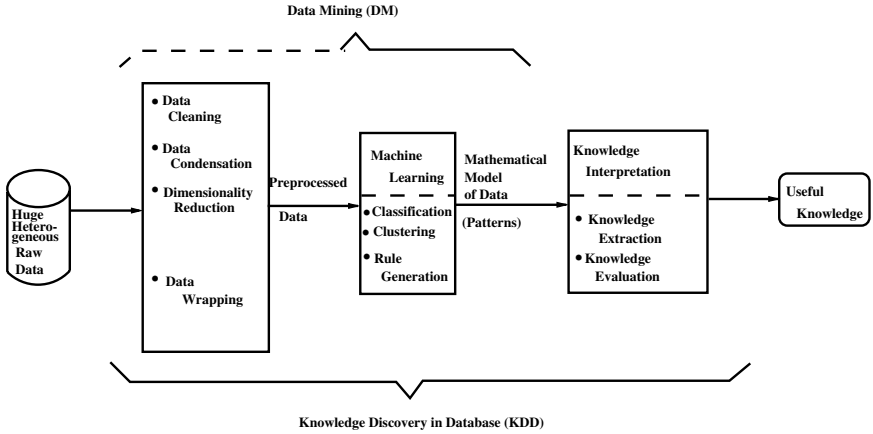

Fig. 1. Block diagram for knowledge discovery in databases [3].

uncertainty, approximate reasoning and partial truth and lead to tractability, robustness and low-cost solutions [4]. An excellent survey demonstrating the significance of soft computing tools in data mining problem is recently provided by Mitra et al. [5]. Some of the challenges arising out of those posed by massive data and high dimensionality, nonstandard and incomplete data, and over-fitting problems deal mostly with issues like user interaction, use of prior knowledge, assessment of statistical significance, learning from mixed media data, management of changing (dynamic) data and knowledge, integration of different classical and modern soft computing tools, and making knowledge discovery more understandable to humans by using linguistic rules, visualization, etc.

Web mining can be broadly defined as the discovery and analysis of useful information from the web or WWW which is a vast collection of completely uncontrolled heterogeneous documents. Since the web is huge, diverse and dynamic, it raises the issues of scalability, heterogeneity and dynamism, among others. Recently, a detailed review explaining the state of the art and the future directions for web mining research in soft computing framework is provided by Pal et al. [6]. One may note that web mining, although considered to be an application area of data mining on the WWW, demands a separate discipline of research. The reason is that web mining has its own characteristic problems (e.g., page ranking, personalization), because of the typical nature of the data, components involved and tasks to be performed, which can not be usually handled within the conventional framework of data mining and analysis. Moreover, being an interactive medium, human interface is a key component of most web applications. Some of the issues which have come to light, as a result, concern with—(a) need for handling context sensitive and imprecise

queries, (b) need for summarization and deduction, and (c) need for personalization and learning. Accordingly, *web intelligence* became an important and urgent research field that deals with a new direction for scientific research and development by exploring the fundamental roles and practical impacts of machine intelligence and information technology (IT) on the next generation of web-empowered products, systems, services and activities. It plays a key role in today's IT in the era of WWW and agent intelligence.

While talking about pattern recognition and data mining in the 21st century, it will remain incomplete without the mention of the *computational theory of perceptions* (*CTP*), recently explained by Zadeh [7,8], which has a significant role in the said tasks. In the following section we discuss its basic concepts and features, and relation with soft computing.

2. Computational theory of perceptions and F-granulation

CTP [7,8] is inspired by the remarkable human capability to perform a wide variety of physical and mental tasks, including recognition tasks, without any measurements and any computations. Typical everyday examples of such tasks are parking a car, driving in city traffic, cooking meal, understanding speech, and recognizing similarities. This capability is due to the crucial ability of human brain to manipulate perceptions of time, distance, force, direction, shape, color, taste, number, intent, likelihood, and truth, among others.

Recognition and perception are closely related. In a fundamental way, a recognition process may be viewed as a sequence of decisions. Decisions are based on information. In most realistic settings, decision-relevant information is a mixture of measurements and perceptions; e.g., the car is six-year old but looks almost new. An essential difference between measurement and perception is that in general, measurements are crisp, while perceptions are fuzzy. In existing theories, perceptions are converted into measurements, but such conversions in many cases, are infeasible, unrealistic or counterproductive. An alternative, suggested by the CTP, is to convert perceptions into propositions expressed in a natural language, e.g., it is a warm day, he is very honest, it is very unlikely that there will be a significant increase in the price of oil in the near future.

Perceptions are intrinsically imprecise. More specifically, perceptions are f-granular, that is, both fuzzy and granular, with a granule being a clump of elements of a class that are drawn together by indistinguishability, similarity, proximity or functionality. For example, a perception of height can be described as very tall, tall, middle, short, with very tall, tall, and so on constituting the granules of the variable 'height'. F-granularity of perceptions reflects the finite ability of sensory organs and, ultimately, the brain, to resolve detail and store information. In effect, f-granulation is a human way of achieving

data compression. It may be mentioned here that although information granulation in which the granules are crisp, i.e., c-granular, plays key roles in both human and machine intelligence, it fails to reflect the fact that, in much, perhaps most, of human reasoning and concept formation the granules are fuzzy (f-granular) rather than crisp. In this respect, generality increases as the information ranges from singular (age: 22 years), c-granular (age: 20–30 years) to f-granular (age: "young"). It means CTP has, in principle, higher degree of generality than qualitative reasoning and qualitative process theory in AI [9,10]. The types of problems that fall under the scope of CTP typically include: perception-based function modeling, perception-based system modeling, perception-based time series analysis, solution of perception-based equations, and computation with perception-based probabilities where perceptions are described as a collection of different linguistic *if-then* rules.

F-granularity of perceptions puts them well beyond the meaning representation capabilities of predicate logic and other available meaning representation methods. In CTP, meaning representation is based on the use of so called constraint-centered semantics, and reasoning with perceptions is carried out by goal-directed propagation of generalized constraints. In this way, the CTP adds to existing theories the capability to operate on and reason with perception-based information.

This capability is already provided, to an extent, by fuzzy logic and, in particular, by the concept of a linguistic variable and the calculus of fuzzy if-then rules. The CTP extends this capability much further and in new directions. In application to pattern recognition and data mining, the CTP opens the door to a much wider and more systematic use of natural languages in the description of patterns, classes, perceptions and methods of recognition, organization, and knowledge discovery. Upgrading a search engine to a question-answering system is another prospective candidate in web mining for CTP application. However, one may note that dealing with perception-based information is more complex and more effort-intensive than dealing with measurement-based information, and this complexity is the price that has to be paid to achieve superiority.

3. Granular computation and rough-fuzzy approach

Rough set theory [11] provides an effective means for analysis of data by synthesizing or constructing approximations (upper and lower) of set concepts from the acquired data. The key notions here are those of "information granule" and "reducts". Information granule formalizes the concept of finite precision representation of objects in real life situation, and reducts represent the core of an information system (both in terms of objects and features) in a granular universe. *Granular computing* refers to that where computation and

operations are performed on information granules (clump of similar objects or points). Therefore, it leads to have both data compression and gain in computation time, and finds wide applications. An important use of rough set theory and granular computing in data mining has been in generating logical rules for classification and association. These logical rules correspond to different important regions of the feature space, which represent data clusters.

For the past few years, rough set theory and granular computation has proven to be another soft computing tool which, in various synergistic combinations with fuzzy logic, artificial neural networks and genetic algorithms, provides a stronger framework to achieve tractability, robustness, low cost solution and close resembles with human like decision making. For example, rough-fuzzy integration can be considered as a way of emulating the basis of f-granulation in CTP, where perceptions have fuzzy boundaries and granular attribute values. Similarly, rough neural synergistic integration helps in extracting crude domain knowledge in the form of rules for describing different concepts/classes, and then encoding them as network parameters; thereby constituting the initial knowledge base network for efficient learning. Since in granular computing computations/operations are performed on granules (clump of similar objects or points), rather than on the individual data points, the computation time is greatly reduced. The results on these investigations, both theory and real life applications, are being available in different journals and conference proceedings. Some special issues and edited volumes have also come out [12–14].

4. Rough-fuzzy granulation and case-based reasoning

Case-based reasoning (CBR) [15], which is a novel Artificial Intelligence (AI) problem-solving paradigm, involves adaptation of old solutions to meet new demands, explanation of new situations using old instances (called cases), and performance of reasoning from precedence to interpret new problems. It has a significant role to play in today's pattern recognition and data mining applications involving CTP, particularly when the evidence is sparse. The significance of soft computing to CBR problems has been adequately explained in a recent book by Pal et al. [16]. In this section we demonstrate an example [17] of using the concept of f-granulation, through rough-fuzzy computing, for performing an important task, namely, *case generation*, in large scale CBR systems.

A case may be defined as a contextualized piece of knowledge representing an evidence that teaches a lesson fundamental to achieving goals of the system. While case selection deals with selecting informative prototypes from the data, case generation concerns with construction of 'cases' that need not necessarily include any of the given data points. For generating cases, linguistic repre-

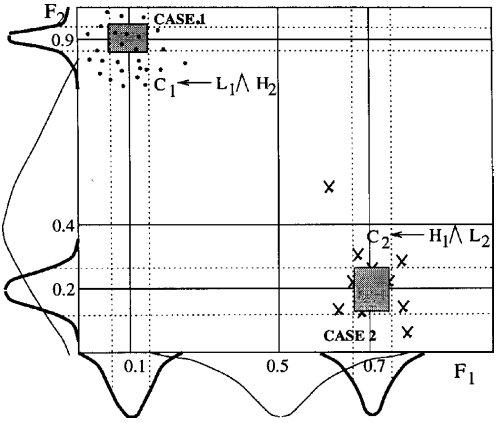

Fig. 2. Rough-fuzzy case generation for a two-dimensional data [15].

sentation of patterns is used to obtain a fuzzy granulation of the feature space. Rough set theory is used to generate dependency rules corresponding to informative regions in the granulated feature space. The fuzzy membership functions corresponding to the informative regions are stored as cases. Fig. 2 shows an example of such case generation for a two-dimensional data having two classes. The granulated feature space has $3^2 = 9$ granules. These granules of different sizes are characterized by three membership functions along each axis, and have ill-defined (overlapping) boundaries. Two dependency rules: $\text{class}_1 \leftarrow L_1 \wedge H_2$ and $\text{class}_2 \leftarrow H_1 \wedge L_2$ are obtained using rough set theory. The fuzzy membership functions, marked bold, corresponding to the attributes appearing in the rules for a class are stored as its case.

Unlike the existing methods, the cases here are cluster granules and not sample points. Also, each case involves a reduced number of relevant features. The methodology is suitable for mining data sets, large both in dimension and size, due to its low time requirement in case generation as well as retrieval.

5. Conclusions

Data mining and knowledge discovery in databases, which has recently drawn the attention of researchers significantly, have been explained from the view-point of pattern recognition. As it appears, soft computing methodologies, coupled with CTP, have great promise for efficient mining of large, heterogeneous data and solution of real-life recognition problems. Fuzzy granulation, through rough-fuzzy computing, and performing operations on fuzzy granules provide both information compression and gain in computation time; thereby making it suitable for data mining applications. We believe the

next decade will bear testimony to this in several fields including web intelligence/mining which is considered to be a forefront research area in today's era of IT.

References

[1] J.G. Shanahan, Soft Computing for Knowledge Discovery: Introducing Cartesian Granule Feature, Kluwer Academic, Boston, MA, 2000.
[2] S.K. Pal, A. Pal (Eds.), Pattern Recognition: From Classical to Modern Approaches, World Scientific, Singapore, 2002.
[3] A. Pal, S.K. Pal, Pattern recognition: evolution of methodologies and data mining, in: S.K. Pal, A. Pal (Eds.), Pattern Recognition: From Classical to Modern Approaches, World Scientific, Singapore, 2002, pp. 1–23.
[4] L.A. Zadeh, Fuzzy logic, neural networks and soft computing, Communications of the ACM 37 (1994) 77–84.
[5] S. Mitra, S.K. Pal, P. Mitra, Data mining in soft computing framework survey, IEEE Trans. Neural Networks 13 (1) (2002) 3–14.
[6] S.K. Pal, V. Talwar, P. Mitra, Web mining in soft computing framework: relevance, state of the art and future directions, IEEE Trans. Neural Networks 13 (5) (2002) 1163–1177.
[7] L.A. Zadeh, A new direction in AI: toward a computational theory of perceptions, AI Magazine 22 (2001) 73–84.
[8] L.A. Zadeh, Foreword, in: S.K. Pal, S. Mitra (Eds.), Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing, Wiley, New York, 1999.
[9] B.J. Kuipers, Qualitative Reasoning, MIT Press, Cambridge, 1984.
[10] R. Sun, Integrating Rules and Connectionism for Robust Commonsense Reasoning, Wiley, New York, 1994.
[11] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic, Dordrecht, 1991.
[12] S.K. Pal, A. Skowron (Eds.), Rough-Fuzzy Hybridization: A New Trend in Decision Making, Springer-Verlag, Singapore, 1999.
[13] S.K. Pal, W. Pedrycz, R. Swiniarski, A. Skowron (Eds.), Rough-Neuro Computing, Neurocomputing 36 (124) (2001) (special issue).
[14] S.K. Pal, A. Skowron (Eds.), Rough Sets, Pattern Recognition and Data Mining, Pattern Recognition Letters 24 (6) (2003) (special issue).
[15] J.L. Kolodner, Case-Based Reasoning, Morgan Kaufmann, San Mateo, CA, 1993.
[16] S.K. Pal, T.S. Dillon, D.S. Yeung (Eds.), Soft Computing in Case Based Reasoning, Springer, London, 2001.
[17] S.K. Pal, P. Mitra, Case generation using rough sets with fuzzy discretization, IEEE Trans. Knowledge and Data Engineering 16 (3) (2004).