Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Sciences 163 (2004) 1–3
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Guest editorial

Soft computing data mining

Soft computing is a consortium of methodologies, (like fuzzy logic, neural networks, genetic algorithms, rough sets), that works synergistically and provides, in one form or another, flexible information processing capabilities for handling real life problems. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning and partial truth in order to achieve tractability, robustness, low solution cost, and close resemblance with human like decision-making. The process of knowledge discovery from data bases (KDD), on the other hand, is a real life problem solving paradigm and is defined as the non-trivial process of identifying valid, novel, potentially useful and understandable patterns from large data bases, where the data is frequently ambiguous, incomplete, noisy, redundant and changes with time. Data mining is one of the fundamental steps in the KDD process and is concerned with the algorithmic means by which patterns or structures are enumerated from the data under acceptable computational efficiency. Soft computing tools, individually or in integrated manner, are turning out to be strong candidates for performing data mining tasks efficiently. At present, the results on these investigations, integrating soft computing and data mining, both theory and applications, are being available in different journals and conference proceedings mainly in the fields of computer science, information technology, engineering and mathematics.

The objective of this issue is to assemble a set of high-quality original contributions that reflect the advances and the state-of-the-art in the area of *Data Mining and Knowledge Discovery with Soft Computing Methodologies*; thereby presenting a consolidated view to the interested researchers in the aforesaid fields, in general, and readers of the journal *Information Sciences*, in particular. It has ten articles. The first one is a title article. While the next four articles deal with classificatory rule analysis, the sixth one concerns with association rule mining. Utility of Self Organizing Map (SOM) in the context of text mining and clustering are discussed in seventh and eighth articles. The next contribution describes a novel multi-source data fusion strategy. The last article demonstrates an application of data mining in biological data analysis.

0020-0255/$ - see front matter
doi:10.1016/j.ins.2003.03.012

Experts from different active groups from the U.S.A, U.K, Brazil, Hong Kong, Finland and India author these articles; and each of them is reviewed by two to three referees.

In the leading article by Pal on soft data mining, computational theory of perceptions (CTP), and rough-fuzzy approach, key features of CTP and its significance in pattern recognition and data mining are explained. Merits of fuzzy granulation and the role of rough fuzzy approach are described with an example applicable to large-scale case based reasoning Systems. In the second article, Carvalho and Freitas propose a hybrid decision tree/genetic algorithm method to discover classification rules. In this approach, they develop two genetic algorithms specifically designed for discovering rules covering examples belonging to small disjuncts (i.e., a rule covering a small number of examples), whereas a conventional decision tree algorithm is used to produce rules covering examples belonging to large disjuncts. Some results evaluating the performance of the hybrid method using real-world data sets are presented. The next article by Cios and Kurgan describes a hybrid inductive machine-learning algorithm, called CLIP4. The algorithm first partitions data into subsets using a tree structure and then generates inequality rules only from subsets stored at the leaf nodes. The algorithm works with data that have large number of examples and attributes, noisy data, and may use numerical, nominal, continuous, and missing-value attributes. The flexibility and efficiency of the algorithms are demonstrated on many well-known benchmarking datasets, and the results are compared with several other machine learning algorithms. The study of Yifeng and Bhattacharyya demonstrates the potential of genetic programming (GP) as a base classifier algorithm in building ensembles in the context of large-scale data classification. Such an ensemble was found to significantly outperform its counterparts built upon base classifiers that were trained with decision tree and logistic regression.

The superiority of GP ensembles is attributed to the higher diversity, both in terms of the functional form and variables defining the models, among the base classifiers. In his paper, Hsu addresses the automated tuning of input specification for supervised inductive learning to reduce generalization error in classification by designing generic fitness functions. This fitness function is then used to develop two genetic algorithm wrappers: one for the variable selection and one for the variable ordering problem; and these wrappers are evaluated using real and synthetic data. Association rule mining problems are considered as a multi-objective problem rather than as a single objective one by Ghosh and Nath. Measures like *support count*, *comprehensibility* and *interestingness*, used for evaluating a rule can be thought of as different objectives of association rule mining problem. Using these three measures as the objectives of rule mining problem, this article uses a Pareto based genetic algorithm to extract some useful and interesting rules from any market-basket type database. Based on experimentation, the algorithm has been found suitable for large databases.

WEBSOM, a software System based on the SOM principle, orders a collection of textual items, according to their contents, and maps them onto a regular two-dimensional array of map units; similar texts are mapped to the same or neighboring map units, and at each unit there exist links to the document database. Thus, locating those documents that match best with the search expression can start text searching, further relevant search results can be found on the basis of the pointers stored at the same or neighboring map units. The work by Lagus, Kaski and Kohonen contains an overview to the WEBSOM method. Although SOM is a powerful tool for projecting high-dimensional data onto a regular, two-dimensional grid of neurons; due to dimensional conflict, the neighborhood preservation cannot always lead to perfect topology preservation. In the article of Jin, Shum, Leung, and Wong, the authors propose an Expanding SOM (ESOM) to preserve better topology between two spaces. Their experimental results on clustering demonstrate that the ESOM constructs better mappings than the classic SOM in terms of both the topological error and the quantization error.

Yager presents a general view of the multi-source data fusion process, based on a voting like process that tries to adjudicate conflict among the data. Situations in which the sources have different credibility weights, and fused values are granular objects together with the means of including any information available other than that provided by the sources are also considered.

In the last article, Yin and Wong present a System, called GeneScout, for predicting gene structures in vertebrate genomic DNA using hidden Markov models. The main hypothesis is that, given a genomic DNA sequence S, it is always possible to construct a directed acyclic graph G such that the path for the actual coding region of S is in the set of all paths on G; and thus, the gene detection problem is reduced to that of analyzing the paths in the graph G. A dynamic programming algorithm is used to find the optimal path in G, while the System is trained using expectation-maximization.

Finally, we take this opportunity to thank Prof. Paul P. Wang, Chief Editor, *Information Sciences*, for giving us an opportunity to act as the Guest Editors for this special issue. We believe, the issue is very timely. We are thankful to all the contributors and reviewers for their co-operation in making this issue a reality. Secretarial assistance provided by Shri Sanjoy Kumar Das is acknowledged.

Sankar K. Pal
Ashish Ghosh
Indian Statistical Institute
Machine Intelligence Unit
203 Barrackpore Trunk Road
Calcutta 700 108, India
Tel.: +91-33-25778085; fax: +91-33-25783357
E-mail address: sankar@isical.ac.in