Guest Editorial
Data Mining and Knowledge Discovery With Evolutionary Algorithms

DATA mining (DM) consists of extracting interesting knowledge from real-world, large and complex data sets; and is the core step of a broader process, called knowledge discovery from databases (KDD). In addition to the DM step, which actually extracts knowledge from data, KDD process includes several preprocessing (data preparation) and post-processing (knowledge refinement) steps. The goal of data preprocessing methods is to transform the data to facilitate the application of a (or several) given DM algorithm(s), whereas the goal of knowledge refinement methods is to validate and refine discovered knowledge. Ideally, discovered knowledge should be not only accurate, but also comprehensible and interesting for the user. The total process is computation intensive.

The idea of automatically discovering knowledge from databases is a very attractive and challenging task, both for academia and for industry. Hence, there has been a growing interest in data mining in several machine learning related areas, including evolutionary algorithms (EAs). The main motivation for applying EAs to KDD tasks is that they are robust and adaptive search methods, which perform a global search in the space of candidate solutions. Intuitively, the global search performed by EAs can more effectively discover interesting patterns that would have been missed by the greedy search performed by many KDD methods.

At present, results on investigations integrating EAs and DM, both theory and applications, are being made available in different journals and conference proceedings mainly in the fields dedicated to knowledge discovery and data mining or evolutionary computing. 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 EAs*, thereby presenting a consolidated view to the interested researchers in the aforesaid fields, in general, and readers of the journal IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, in particular. The special issue emphasizes the utility of different evolutionary computing tools to various facets of KDD.

The issue has four papers. First two papers are on classification, while the third paper is on knowledge discovery from text. The fourth paper gives a comparison of various evolutionary and non-EAs for prototype selection and training set selection for data reduction in KDD. Experts of different active groups from the United States, Hong Kong, the United Kingdom, and Spain have written these articles; and 3–5 referees review each of them. These four papers were chosen from a set of 35 submissions for this issue. Let us scan these papers.

Digital Object Identifier 10.1109/TEVC.2003.819653

Classification is one of the fundamental tasks of data mining. In the first paper of this issue Zhou *et al.*proposed a new approach for discovering classification rules by using gene expression programming (GEP). The antecedent of discovered rules may involve many different combinations of attributes. To guide the search process, they suggested a fitness function considering both the rule consistency gain and completeness. A multiclass classification problem was formulated as a combination of multiple two-class problems by using the *one against all* learning method. Compact rule sets were subsequently evolved using a two-phase pruning method based on the minimum description length (MDL) principle. Their approach is claimed to be noise tolerant, and able to deal with both numeric and nominal attributes. Experiments with several benchmark data sets showed an improvement in classification accuracy compared with existing algorithms. Furthermore, the proposed GEP approach is claimed to be more efficient compared with canonical tree-based genetic programming classifiers.

Many algorithms have been developed to mine large data sets for classification and they have been shown to be very effective. However, when it comes to determining the likelihood of classification, they face problems. Thus, they are not readily applicable to such problems as churn prediction. For such an application, the goal is not only to predict whether a subscriber would switch from one carrier to another, rather it is important that the likelihood of the subscriber's doing so be predictable. The reason for this is that a carrier can then choose to provide personalized offer and services to those subscribers who are predicted with higher likelihood to churn. Au *et al.* proposed a new data mining algorithm, called data mining by evolutionary learning (DMEL), to handle classification problems of which the accuracy of each prediction needs to be estimated. DMEL searches through the possible rule space using an evolutionary approach that has the following characteristics: 1) the evolutionary process begins with the generation of an initial set of simple, one-condition rules; 2) interestingness measure is used for identifying interesting rules; 3) fitness of a chromosome is defined in terms of the probability that the attribute values of a record can be correctly determined using the rules it encodes; and 4) the likelihood of predictions made are estimated so that subscribers can be ranked according to their likelihood to churn. Experiments with different data sets showed that DMEL is able to effectively discover interesting classification rules.

Atkinson-Abutridy *et al.* presented a novel evolutionary model for knowledge discovery from texts (KDT), which deals with issues concerning shallow text representation and processing for mining purposes in an integrated way. The approach uses natural language technology and genetic algorithms to

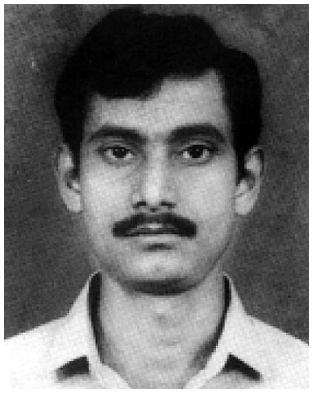
produce explanatory novel hypotheses. New kinds of genetic operations suitable for text mining were proposed in this regard; and they used multiobjective evaluations at the semantic level. Some promising results and their assessment by human experts are also discussed that indicate the plausibility of the model for effective KDT.

Cano *et al*. carried out an empirical study of the performance of four representative EAs in which they took into account two different instance selection perspectives, the prototype selection and the training set selection for data reduction in KDD. This study included a comparison between these algorithms and other nonevolutionary instance selection algorithms also. The results showed that the evolutionary instance selection algorithms consistently outperform the nonevolutionary ones, the main advantages being: better instance reduction rates, higher classification accuracy, and models that are easier to interpret.

Here, we take this opportunity to thank Prof. D. B. Fogel and Prof. X. Yao, Editors-in-Chief of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, 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 cooperation in making this issue a reality.

ASHISH GHOSH, *Guest Editor*
Indian Statistical Institute
Machine Intelligence Unit
Kolkata, 700108 India

ALEX A. FREITAS, *Guest Editor*
University of Kent
Computing Laboratory
Canterbury, Kent CT2 7NF U.K.



Ashish Ghosh received the B.E. degree in electronics and telecommunications from Jadavpur University, Calcutta, India, in 1987, and the M.Tech. and Ph.D. degrees in computer science from the Indian Statistical Institute, Calcutta, in 1989 and 1993, respectively.

He is an Associate Professor with the Machine Intelligence Unit, Indian Statistical Institute, Calcutta. He has been selected as an Associate of the Indian Academy of Sciences, Bangalore, in 1997. He visited the Osaka Prefecture University, Japan, with a Postdoctoral Fellowship from October 1995 to March 1997, and Hannan University, Japan, as a Visiting Scholar from September to October 1997. During May 1999, he was with the Institute of Automation, Chinese Academy of Sciences, Beijing, with a CIMPA (France) Fellowship. He was with the German National Research Center for Information Technology, Germany, with a German Government (DFG) Fellowship from January to April 2000. From October to December 2003, he was a Visiting Professor at the University of California, Los Angeles. He also visited various universities, academic institutes, and delivered lectures in different countries including South Korea, Poland, and The Nederland. His research interests include evolutionary computation, neural networks, image processing, fuzzy sets and systems, pattern recognition, and data mining. He has published about 55 research papers in internationally reputed journals and referred conferences, has edited four books, and is a Guest Editor of various journals.

Dr. Ghosh received the prestigious and most coveted Young Scientists Award in Engineering Sciences from the Indian National Science Academy in 1995 and in Computer Science from the Indian Science Congress Association in 1992.

Alex A. Freitas (M'99) received the B.Sc. degree in computer science from the Faculdade de Tecnologia de Sao Paulo, Brazil, in 1989, the M.Sc. degree in computer science from the Universidade Federal de Sao Carlos, Brazil, in 1993, and the Ph.D. degree in computer science from the University of Essex, U.K., in 1997.

From 1997 to 1998, he was a Visiting Lecturer at the Centro Federal de Educacao Tecnologica, Curitiba, Brazil, and a Lecturer at the Pontificia Universidade Catolica, Curitiba, Brazil, from 1999 to 2002. Since 2002, he has been a Lecturer at the University of Kent, Canterbury, U.K. His publications include two books on data mining, several invited book chapters, and more than 60 refereed research papers published in journals and conferences. He has organized two international workshops on data mining with evolutionary algorithms and delivered tutorials on this theme in several international conferences. He is a Member of the Editorial Board of *Intelligent Data Analysis*—an international journal. He has also coordinated a research cluster in Swarm Intelligence involving more than 15 institutions in the U.K., from July 2003 to December 2003. At present his main research interests are data mining, bioinspired algorithms, and bioinformatics.