Preface

Du Zhang and Shu-Ching Chen

As the age of digital information arrives, more and more information is being stored electronically in the form of files, databases and Web pages. Data can grow without limit at a high rate of millions of data items per day. The problem of being inundated with data looms ominously ahead. To address this issue, data mining has been increasingly recognised as a key technique to analyse and understand the huge amount of data. Many data mining techniques and applications, which are proving to be extremely useful to handle various problems in diverse domains, are developed. The main theme of this special issue of the *International Journal of Computer Applications in Technology* (IJCAT) is 'Data Mining Applications.' From all the submissions, eight papers were selected for inclusion in this special issue.

The first paper of this special issue is entitled 'An efficient intrusion detection system using boosting-based learning algorithm' by Zhenwei Yu and Jeffrey J.P. Tsai. In this paper, the authors developed a Multi-Class SLIPPER (MC-SLIPPER) system for intrusion detection, and proposed multiple prediction-confidence-based strategies to arbitrate the final prediction among predictions from all binary SLIPPER modules. The MC-SLIPPER system uses the boosting-based learning algorithm since boosting is effective to improve the accuracy of learners. Their proposed MC-SLIPPER system was evaluated using the KDDCUP'99 intrusion detection dataset, and the experimental results showed that it achieved the best performance using a BP neural network, and got better performance in terms of misclassification cost than that of the winner of the KDDCUP'99 contest when it used other prediction strategies.

In the second paper entitled 'Indirect classification approaches: a comparative study in network intrusion detection', a set of indirect classification techniques for addressing the multi-category classification problem in network intrusion detection was proposed by Taghi M. Khoshgoftaar, Kehan Gao and Hua Lin. The idea of an indirect classification technique is to decompose the original multi-category problem into multiple binary classification problems based on some criteria. In this study, the authors investigated the '*one vs. one*' and '*one vs. rest*' approaches for building the binary classifiers, and the results are then merged using a combining strategy. Three different combining strategies: Hamming decoding, loss-based decoding and soft-max function were considered, and consequently six different indirect classification techniques were evaluated in the context of the DARPA KDD-1999 offline intrusion detection project. Their study demonstrated the usefulness of the indirect classification approach for network intrusion detection.

Noise filtering techniques can improve the quality of training datasets by removing data points that are likely to be noisy. In the next paper 'Noise elimination with partitioning filter for software quality estimation', Taghi M. Khoshgoftaar and Pierre Rebours discussed the use of partitioning-based filtering techniques. That is, the training dataset is first split into subsets, and base learners are induced on each of these subsets. The predictions are then combined in such a way that an instance in the training data is identified as noisy if it is misclassified by a certain number of base learners. They proposed two new noise filtering techniques: one is a Multiple-Partitioning Filter and the other one is an Iterative-Partitioning Filter, and conducted several empirical studies using software measurement data from a high assurance software project to assess the efficiencies of these two noise filtering approaches. The empirical results suggested that using several base classifiers as well as performing several iterations with a conservative filtering scheme can improve the efficiency of the filtering technique.

In the fourth paper, See-Kiong Ng and Soon-Heng Tan addressed the challenges in mining biological literature for bio-molecular interaction pathways in their paper entitled 'Challenges in biological literature mining for online discovery of molecular interaction pathways'. Despite the previous accomplishments from the text mining community and the increasing research activities in biological text mining, biologists are still expending great efforts by laborious hand curation of the scientific literature to create quality online databases of bio-molecules and their interactions. The authors proposed a methodology for training and evaluating biological literature-based data mining applications with annotated biological review papers with the intention that a road-map can be furnished for the text-based data mining community to collectively solve this complex but increasingly important data mining task in bioinformatics.

The next paper was contributed by Renáta Iváncsy and István Vajk with the title 'A time- and memory-efficient frequent itemset discovering algorithm for association rule mining'. Association rule mining is to find hidden, previously unknown relationships in a large amount of data. The first phase of the association rule mining process is the step of discovering the frequent itemsets. The task of discovering the frequent patterns in large databases is mostly computational and I/O complex. To address this issue, a novel algorithm that is efficient both in time and memory was proposed in this study. Their new algorithm discovers the small frequent itemsets quickly by taking advantage of the easy indexing opportunity of the suggested

candidate storage structure. The main benefit of the novel algorithm is its advantageous time behaviour when using different types of datasets as well as its low I/O activity and moderate memory requirement.

Several types of data constitute what are commonly known as vital statistics data, including births, deaths, fetal deaths, marriages and divorces. Vital statistics data offer a fertile ground for data mining. In the sixth paper, Du Zhang, Quoc Luan Ha and Meiliu Lu discuss the results of a data mining project on the causes of death aspect of the vital statistics data in the state of California in their paper entitled 'Mining California vital statistics data'. In this study, a data mining tool called Cubist was used to build predictive models out of two million cases over a nine-year period with the objective of discovering knowledge (trends, correlations or patterns) that may not be gleaned through standard techniques. Their results showed that the generated predictive models allowed pertinent state agencies to gain insight into various aspects of death rates, to predict health issues related to the causes of death, to offer an aid to the decision-making or policy-making process and to provide useful information services to the customers in the state of California.

Choochart Haruechaiyasak, Mei-Ling Shyu and Shu-Ching Chen proposed a new recommender system framework based on data mining techniques and the Semantic Web concept with an attempt to intelligently generate a list of information that matches the users' preferences, in their paper 'A web-page recommender system via a data mining framework and the Semantic Web concept'. In their proposed recommendation system, two information filtering methods for providing the recommended information (i.e., content-based and collaborative filtering) were considered. Both filtering techniques are based on data mining algorithms, which provide efficiency in handling large data sets. In addition, the Semantic Web concept, in which the information is given well-defined meaning, was incorporated into the framework to provide the users with semantically enhanced information. The authors implemented a prototype of their proposed recommendation system to demonstrate the potential use of their proposed framework and the performance enhancement to the traditional query-based information retrieval approach provided on the website.

The last paper in this special issue is entitled 'A multimodal data mining framework for soccer goal detection based on decision tree logic' by Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang and Min Chen. In this paper, a new multimedia data mining framework for the extraction of soccer goal events in soccer videos by utilising both multi-modal analysis and decision tree logic was proposed. This framework first adopted an advanced video shot detection method to produce shot boundaries and some important visual features, and then the visual/audio features were extracted for each shot at different granularities. The rich multi-modal feature set was then filtered by a pre-filtering step to clean the noise as well as to reduce the irrelevant data. A decision tree model was built upon the cleaned data set to classify the goal shots. Experimental results over diverse video data from different sources demonstrating the robustness of their framework for soccer goal extraction in terms of recall and precision were also reported.

This special issue covers a wide range of research studies that apply data mining techniques to various applications. We hope that you will enjoy reading this special issue. We are grateful to Dr. M.A. Dorgham and to Inderscience Publishers for giving us the opportunity to organise this special issue in the *International Journal of Computer Applications in Technology*. We would like to thank the authors and referees for their contributions to this special issue.