The workshop proceedings will be published as part of the ICCIT conference proceedings and will appear in IEEE Xplore® Digital Library.

Selected best papers, after extension, will be published in the following Book and the journal:

Book

Dynamic and Advanced Data Mining for Progressing Technological Development: Innovations and Systemic Approaches, IGI Global, USA.

Journal

Journal of Computers (Special issue on Artificial Intelligence) **,** Academiy Publisher, Finland.



International Workshop on Data Mining and Artificial Intelligence (DMAI 2008)

Accepted Paper List

Paper ID: 9
Title: A New Approach of Modified Transaction Reduction Algorithm for Mining Frequent Itemset
Author: Ramaraj Eswara Thevar and Rameshkumar Krishnamoorthy

Abstract: *Association rule mining is to extract the interesting correlation and relation between the large volumes of transactions. This process is divided into two sub problem: first problem is to find the frequent itemsets from the transaction and second problem is to construct the rule from the mined frequent itemset. Frequent itemsets generation is the requirement and most time vast process for association rule mining. Nowadays, most efficient Apriori-like algorithms rely heavily on the minimum support constraints to prune the vast amount of non-candidate itemsets. These algorithms store many unwanted itemsets and transactions. In this paper propose a novel frequency itemsets generation algorithm called MTR-FMA (Modified Transaction Reduction based Frequent Itemset Mining algorithm) that maintains its performance even at relative low supports. The experimental reports also show that proposed MTR-FMA algorithm on an outset is faster than High Efficient AprioriTid and other some algorithms.*

Paper ID: 12
Title: A Study of White Matter and Skull Inhomogeneous Anisotropic Tissue Conductivities on EEG Forward Head Modeling
Author: Md. Rezaul Bashar, Yan Li, and Peng Wen

Abstract: *The aim of this study is to investigate the effects of white matter (WM) and skull inhomogeneous anisotropic tissue conductivities on human head modeling. The non-homogeneity of WM and skull is included using fractional anisotropy (FA) method and the anisotropy is included according to Volume constraint in the head model construction. A five-layered spherical head model implemented using finite element method (FEM) is used as a volume conductor with a known current source to measure the electroencephalogram (EEG) on the head surface. Statistical measurement techniques are applied to analyze the EEGs obtained from inhomogeneous anisotropic head models and a homogeneous isotropic model. This study finds that the effects of WM and skull inhomogeneous anisotropy on EEG are significant.*

Paper ID: 13
Title: Selecting Signature Optical Emission Spectroscopy Variables Using Sparse Principal Component Analysis
Author: Beibei Ma, Sean McLoone, John Ringwood, and Niall Macgearailt

Abstract: *Principal component analysis (PCA) is a widely used technique in optical emission spectroscopy (OES) sensor data analysis for the low dimension representation of high dimensional datasets. While PCA produces a linear combination of all the variables in each loading, sparse principal component analysis (SPCA) focuses on using a subset of variables in each loading. Therefore, SPCA can be used as a key variable selection technique. This paper shows that using SPCA to analyze 2046 variable OES data sets, the number of selected variables can be traded off against variance explained to identifying a subset of key wavelengths, with an acceptable level of variance explained. SPCA-related issues such as selection of the tuning parameter and the grouping effect are discussed.*

Paper ID: 26
Title: A Data Mining Approach for Finding Optimal Discount of Retail Assortments

Author: Maryam Nafari and Jamal Shahrabi

Abstract: *Due to the recent competition in the retailing industry, retailers are striving to improve their operations in order to run their stores more efficiently. One of the most important factors that encourages customers to buy products is discount. The effects of discount on sales have rarely been dealt with academically. Moreover, in few previous researches in this case, the temporal characteristics of product discount have not been noticed. The problem addressed in this paper is the consideration of products' discounts in discovering association rules in different time intervals that a specific discount appears on a specific product and finding optimal discount for each product with the aim of maximizing total profits. Additionally, experiments on real world data demonstrate the effectiveness of the proposed approach.*

Paper ID: 29
Title: Identifying Stock Similarity Based on Episode Distances
Author: Abhi Dattasharma, Praveen Kumar Tripathi, and Sridhar Gangadharpalli

Abstract: *Predicting stock market movements is always difficult. Investors try to guess a stock's behavior, but it often backfires. Thumb rules and intuition seems to be the major tools. One approach suggested that instead of trying to predict one particular stock's movement with respect to the whole market, it may be easier to predict a stock* A0s *movement based on another stock* B0s *movement, because* A *may get affected by* B *after* B0s *movement. This may provide the investor invaluable time advantage. It would be very useful if a general framework can be introduced that can predict such dependence between stocks based on any user defined criterion. This article attempts to lay down one such framework, where the stock time series is encoded as a binary string. This binary representation depends on the user defined criterion. The string distances between two such encoded time series has been used as a measure of dependence. Further, this technique has been used in the '*Pairs Trading strategy*'; in fact, it is more powerful as varied user defined criterion can be handled while detecting similarity. The presented technique has been demonstrated with one typical user defined criterion.*

Paper ID: 43
Title: Behavior-based Robotics And The Reactive Paradigm A Survey
Author: L. De Silva and H. Ekanayake

Abstract: *Behavior-based robots have come under the spotlight in recent years by making their presence known in diverse fields of human interest. Applications in behavior-based robotics have continued to grow in areas such as demining, search and rescue, office automation, health care, etc and continue to replace human beings in risky and menial tasks. With this inspiration, this paper investigates a variety of aspects in behavior-based robotics and the reactive paradigm in the context of their origins, concepts, applications and current research and is intended to provide a comprehensive overview about this area in robotics. The paper will review several central issues including a brief history of robotics, transition of robotic systems from hierarchical paradigm to reactive paradigm, biological/ethological inspiration for behavior-based robots, fundamentals of the reactive paradigm, architectures for controlling robotic behavior and the hybrid deliberative/reactive paradigm.*

Paper ID: 45
Title: Reversible Data Hiding using Increased Peak Histogram
Author: T.C Thanuja, R Nagaraj, and M.Uttara kumari

Abstract: *In this paper, we present a reversible watermarking algorithm which provides higher embedding capacity and higher PSNR value compared to the existing algorithms. In the proposed algorithm, we utilize a peak point of image histogram and increase the height of the peak by transferring the pixels from the neighbouring gray values to achieve the higher embedding capacity.*

Paper ID: 47
Title: Web Document Clustering Approach using WordNet Lexical Categories and Fuzzy

Clustering
Author: Tarek F. Gharib, Mohammed M. Fouad, and Mostafa M. Aref

Abstract: *Web mining is defined as applying data mining techniques to the content, structure, and usage of Web resources. The three areas of Web mining are commonly distinguished: content mining, structure mining, and usage mining. In all these areas, a wide range of general data mining techniques, in particular association rule discovery, clustering, classification, and sequence mining, are employed and developed further to reflect the specific structures of Web resources and the specific questions posed in Web mining. In this paper, we introduced a web document clustering approach that uses WordNet lexical categories and fuzzy c-means algorithm to improve the performance of clustering problem for web document. Experiments show that Fuzzy c-means algorithm achieves great performance optimization with comparison with the recent algorithms for document clustering.*

Paper ID: 48
Title: A Unifying Viewpoint of some Clustering Techniques Using Bregman Divergences and Extensions to Mixed Data Sets
Author: Cecile Levasseur, Brandon Burdge, Ken Kreutz-Delgado, and Uwe F. Mayer

Abstract: *We present a general viewpoint using Bregman divergences and exponential family properties that contains as special cases the three following algorithms: 1) exponential family Principal Component Analysis (exponential PCA), 2) Semi-Parametric exponential family Principal Component Analysis (SP-PCA) and 3) Bregman soft clustering. This framework is equivalent to a mixed data-type hierarchical Bayes graphical model assumption with latent variables constrained to a low-dimensional parameter subspace. We show that within this framework exponential PCA and SPPCA are similar to the Bregman soft clustering technique with the addition of a linear constraint in the parameter space. We implement the resulting modifications to SP-PCA and Bregman soft clustering for mixed (continuous and/or discrete) data sets, and add a nonparametric estimation of the point-mass probabilities to exponential PCA. Finally, we compare the relative performances of the three algorithms in a clustering setting for mixed data sets.*

Paper ID: 49
Title: Making Good Choices of Non-Redundant N-gramWords
Author: Maria Fernanda Moura, Bruno Magalhaes Nogueira, Merley da Silva Conrado, Fabiano Fernandes dos Santos, and Solange Oliveira Rezende

Abstract: *A new complete proposal to solve the problem of automatically selecting good and non redundant n-gram words as attributes for textual data is proposed. Generally, the use of n-gram words is required to improve the subjective interpretability of a text mining task, with n _ 2. In these cases, the n-gram words are statistically generated and selected, which always implies in redundancy. The proposed method eliminates only the redundancies. This can be observed by the results of classifiers over the original and the non redundant data sets, because, there is not a decrease in the categorization effectiveness. Additionally, the method is useful for any kind of machine learning process applied to a text mining task.*

Paper ID: 54
Title: gSVMT: Aggregating SVMs over a Dynamic Grid Learned from Data
Author: Shaoning Pang, Tao Ban, Youki Kadobayashi, and Nik Kasabov

Abstract: *Addressing the problem of adaptively modeling a classifier as a modular system, a new type of SVM aggregating method termed gridding SVM Tree (gSVMT) is proposed in this paper. The proposed gSVMT achieves to discover data subregions with principal discriminant knowledge through a recursive SVM-supervised data partitioning procedure. For each subregion, an individual SVMis allocated to extract the subregion knowledge. A set of such SVMs are aggregated in a specific order, resulting in a globally reliable decision rule to predict new coming samples. Experiments on a synthetic Gaussian data*

set and 13 benchmark machine learning data sets, have highlighted the usability of the gSVMT on its competitive classification capability. In particular, the proposed gSVMT is found to have better generalization performance than SVM classifiers for data sets with high sparseness and/or class-imbalance. Its performance has been further demonstrated with the successful real application on a face membership authentication system.

Paper ID: 64
Title: Runtime Thread Rescheduling: An Extended Scheduling Algorithm to Enhance the Performance of the Gridbus Broker
Author: Altaf Hussain, Abu Awal Md. Shoeb, Md. Abu Naser Bikas and Mohammad Khalad Hasan

Abstract: *Grid computing is becoming a requirement for the processing of large amount of data now-a-days. The Gridbus broker schedules jobs depending on data and compute resources. Current scheduling process does not reassign a job from lower compute resource to higher compute resource if higher compute resource is available. In this paper, we have proposed a technique to reassign a thread to higher grade executor by preempting the thread in lower grade executor by using the data restoration technique which track the information of the thread so far ran on a lower rate compute resource. It is done only if there is an idle higher computer resource is available. The performance as well as the reliability of the Grid has been improved by this approach in a considerable extent.*

Paper ID: 65
Title: An Efficient Clustering based Texture Feature Extraction for Medical Image
Author: Marghny H. Mohamed, and M. M. AbdelSamea

Abstract: *In some medical applications where a tissue of interest covers a large fraction of the image or a prior knowledge on the region of interest is available, extracting features by fixed blocs in the image is sufficient. However in the general case, one would like to identify features for each tissue in the image. This would require prior image segmentation. Medical image segmentation is one of the most challenging problems in medical image analysis and a very active research topic. Therefore, there is no algorithm available in the general case for isolating medical image regions [1].*
This paper presents an accurate method for extracting texture features from medical image for classification. It is based on bloc wise clustering of medical images. The proposed technique extracts accurate and general set of texural features. Experimental result showed the high accuracy of the extracted textural features. Experiments held on Mammographic Image Analysis Society M(*IAS*) *dataset.*

Paper ID: 67
Title: Ant Colony Clustering by Expert Ants
Author: Zahra Sadeghi and Mohammad Teshnehlab

Abstract: *In this article a new ant clustering algorithm based on case based reasoning (CBR) is presented. Every ant has a case base which is updated iteratively by the process of CBR. The ant which is successful in dropping an item becomes an expert and can use its knowledge for future picked up items. Also expert ants are capable of cooperating to share their knowledge for even better clustering. Our simulation results demonstrated better performance than previous approaches.*

Paper ID: 70
Title: Detection of faulty products using data mining
Author: M. A. Karim, G. Russ, and A. Islam

Abstract: *The manufacturing process is complex due to the large number of processes, diverse equipment set and nonlinear process flows. Manufacturers constantly face yield and quality problems as they constantly redesign their processes for the rapid introduction of new products and adoption of new process technologies. Solving product yield and quality problems in a manufacturing process is becoming increasingly difficult. There are various types of failures and their causes have complex multi-factor*

interrelationships. High innovation speed forced today's manufacturers to find failure causes quickly by examining the historical manufacturing data. Data mining offers tools for quick discovery of relationships, patterns, and knowledge in large databases. This has been applied to many fields such as biological technology, financial analysis, medical information, etc. Application of data mining to manufacturing is relatively limited mainly because of complexity of manufacturing data. Growing self-organizing map (GSOM) algorithm has been proven to be an efficient algorithm to analyze unsupervised DNA data. However, it produced unsatisfactory clustering when used on some manufacturing data. Moreover, there was no benchmark to monitor improvement in clustering. In this study a method has been proposed to evaluate quality of the clusters produced by GSOM and to remove insignificant variables from the dataset. With the proposed modifications, significant improvement in unsupervised clustering was achieved with complex manufacturing data. Results show that the proposed method is able to effectively differentiate good and faulty products.

Paper ID: 71
Title: Significant Cancer Risk Factor Extraction: An Association Rule Discovery Approach
Author: Jesmin Nahar and Kevin S Tickle

Abstract: *Cancer is the top most death threat for human life all over the world. The research in the cancer area is still struggling to provide better support to a cancer patient. In this research our aim to discover the significant risk factors for a particular cancer. First, we construct a risk factor data set by an extensive literature review of bladder, breast, cervical, lung, prostate and skin cancer. Then we employ association rule mining algorithms, apriori, predictive apriori and tertius algorithm to discover most significant risk factor for a particular cancer. The discovery risk factor shows highest confidence values. Finally, we suggest apriori is the best association rule mining algorithm for significant risk factor discovery.*

International Workshop on
Data Mining and Artificial Intelligence (DMAI 2008)

Workshop Program

24 December 2008

8.30 – 9.00		**Workshop Kit Distribution**
9.00 – 9.15		**Inauguration** Vice-Chancellor, KUET, Bangladesh
9.15 – 9.25	**Welcome Talk**	A.B.M. Shawkat Ali and Md Rafiul Hassan
9.25 – 10.15	**Keynote Speech 1**	***Robust and Efficient Intrusion Detection Systems based on Conditional Random Fields*** Professor Ramamohanarao Kotagiri
10.15 – 10.30		**Morning Tea**
10.30 – 11.40	**Session 1**	**Data Mining 1** Session Chair: TBA
		Detection of faulty products using data mining M. A. Karim, G. Russ, and A. Islam
		A Data Mining Approach for Finding Optimal Discount of Retail Assortments Maryam Nafari and Jamal Shahrabi
		Significant Cancer Risk Factor Extraction: An Association Rule Discovery Approach Jesmin Nahar and Kevin S Tickle
		Reversible Data Hiding using Increased Peak Histogram T.C Thanuja, R Nagaraj, and M.Uttara kumara
11.40 – 13.00	**Session 2**	**Data Mining 2** Session Chair: TBA
		A New Approach of Modified Transaction Reduction Algorithm for Mining Frequent Itemset Ramaraj Eswara Thevar and Rameshkumar Krishnamoorthy
		A Unifying Viewpoint of some Clustering Techniques Using Bregman Divergences and Extensions to Mixed Data Sets Cecile Levasseur, Brandon Burdge, Ken Kreutz-Delgado, and Uwe F. Mayer
		Identifying Stock Similarity Based on Episode Distances Abhi Dattasharma, Praveen Kumar Tripathi, and Sridhar Gangadharpalli
		Selecting Signature Optical Emission Spectroscopy Variables Using Sparse Principal Component Analysis Beibei Ma, Sean McLoone, John Ringwood, and Niall Macgearailt
		gSVMT: Aggregating SVMs over a Dynamic Grid Learned from Data Shaoning Pang, Tao Ban, Youki Kadobayashi, and Nik Kasabov

13.00 – 14.10		**Lunch**
14.10 – 15.00	**Keynote Speech 2**	***Identifying Protein Complexes from Protein Interactome Maps*** Professor Limsoon Wong
15.00 – 15.15		**Afternoon Tea**
15.15 – 16.05	**Session 3**	**Artificial Intelligence 1** Session Chair: TBA
		Runtime Thread Rescheduling: An Extended Scheduling Algorithm to Enhance the Performance of the Gridbus Broker Altaf Hussain, Abu Awal Md. Shoeb, Md. Abu Naser Bikas and Mohammad Khalad Hasan
		Ant Colony Clustering by Expert Ants Zahra Sadeghi and Mohammad Teshnehlab
		Making Good Choices of Non-Redundant N-gramWords Maria Fernanda Moura, Bruno Magalhaes Nogueira, Merley da Silva Conrado, Fabiano
16.05 - 17.15	**Session 4**	**Artificial Intelligence 2** Session Chair: TBA
		A Study of White Matter and Skull Inhomogeneous Anisotropic Tissue Conductivities on EEG Forward Head Modeling Md. Rezaul Bashar, Yan Li, and Peng Wen
		An Efficient Clustering based Texture Feature Extraction for Medical Image Marghny H. Mohamed, and M. M. AbdelSamea
		Behavior-based Robotics And The Reactive Paradigm A Survey L. De Silva and H. Ekanayake
		Web Document Clustering Approach using WordNet Lexical Categories and Fuzzy Clustering Tarek F. Gharib, Mohammed M. Fouad, and Mostafa M. Aref
17.15 – 17.30		**Evening Tea**
17.30 – 18.10	**Tutorial**	***Secure Multi-party Computation: Problems, Techniques and Applications for Privacy-Preserving Data Mining*** Professor Durgesh Mishra
18.10 – 18.55	**Session 5**	**Doctoral Forum**
		Hybrid methods for the detection of regulatory signals in genomic sequences Md. Abdul Baten
		Pattern Discovery from Biological Data Jesmin Nahar
		Influence of Inhomogeneous and Anisotropic Tissue Conductivity on Human Head Modelling for EEG Md. Rezaul Bashar
18.55 – 19.10		**Workshop Closing**