IP1
Data Mining Biological Network Models

In this talk we survey work being conducted at the Centre for Integrative Systems Biology at Imperial College on the use of machine learning to build models of biochemical pathways. Within the area of Systems Biology these models provide graph-based descriptions of bio-molecular interactions which describe cellular activities such as gene regulation, metabolism and transcription. One of the key advantages of the approach taken, Inductive Logic Programming, is the availability of background knowledge on existing known biochemical networks from publicly available resources such as KEGG and Biocyc. The topic has clear societal impact owing to its application in Biology and Medicine. Moreover, object descriptions in this domain have an inherently relational structure in the form of spatial and temporal interactions of the molecules involved. The relationships include biochemical reactions in which one set of metabolites is transformed to another mediated by the involvement of an enzyme. Existing genomic information is very incomplete concerning the functions and even the existence of genes and metabolites, leading to the necessity of techniques such as logical abduction to introduce novel functions and invent new objects. Moreover, the development of active learning algorithms has allowed automatic suggestion of new experiments to test novel hypotheses. The approach thus provides support for the overall scientific cycle of hypothesis generation and experimental testing.

Stephen H. Muggleton
Imperial College London
s.muggleton@imperial.ac.uk

IP2
Examining the Relative Influence of Familial, Genetic and Covariate Information in Flexible Risk Models

Spline ANOVA (SS-ANOVA) models are a well known approach to penalized likelihood regression given heterogenous attribute variables, with the ability to model their various interactions. In many circumstances, one may observe attributes, along with some relationships between objects in the training set. We describe a new approach to incorporating relationship or (dis)similarity information in an SS-ANOVA model. For the objects under study, we have attributes along with relationship information between (some) pairs of objects in the study. As an example we consider a demographic study with the response a particular disease that is known to run in families. The data includes environmental/clinical observations, genetic data and pedigree information in a study where a large fraction of the population have relatives in the study. One issues is to evaluate the relative influence of the three distinct sources of information.

Grace Wahba
Department of Statistics
University of Wisconsin-Madison
wahba@stat.wisc.edu

IP3
Mining Scientific Data: Past, Present, and Future

The field of data mining grew out of the need of analyzing very large and complex data sets being generated in the scientific and commercial arena. This talk will focus on the advances made by our community in the context of mining scientific and engineering data sets. It will review the progress made so far, state of the art, as well as challenges for the future. A special focus will be on two domains of urgent societal interest: health and climate. We will discuss a number of applications that exemplify some of the most important questions faced by the scientists in these domains today and the role the data mining community can play in answering them.

Vipin Kumar
University of Minnesota
kumar@cs.umn.edu

IP4
Trumping the Multicore Memory Hierarchy with Hi-Spade

Date-intensive applications demand effective use of the cache/memory/storage hierarchy of the target computing platform(s) in order to achieve high performance. Algorithm designers and application/system developers, however, tend towards one of two extremes: (i) they ignore the hierarchy, programming to the API view of "memory + I/O" and often ignoring parallelism; or (ii) they are (pain)fully aware of all the details of the hierarchy, and hand-tune to a given platform. The former often results in very poor performance, while the latter demands high programmer effort for code that requires dedicated use of the platform and is not portable across platforms. Moreover, two recent trends—pervasive multi-cores and pervasive flash—provide both new challenges and new opportunities for maximizing performance. In the Hi-Spade (Hierarchy-Savvy parallel algorithm design) project, we are developing a hierarchy-savvy approach to algorithm design and systems for these emerging parallel hierarchies. The project seeks to create abstractions, tools and techniques that (i) assist programmers and algorithm designers in achieving effective use of emerging hierarchies, and (ii) leads to systems that better leverage the new capabilities these hierarchies provide. Our abstractions seek a sweet spot between ignoring and (pain)fully aware, exposing only what must be exposed for high performance, while our techniques aim to deliver that high performance across a variety of platforms and platform-sharing scenarios. Key enablers of our approach include novel thread schedulers and novel uses of available flash devices. This talk summarizes our progress to date towards achieving our goals and the many challenges that remain.

Phillip B. Gibbons
Intel Labs Pittsburgh
phillip.b.gibbons@intel.com

CP1
Text Categorization Using Word Similarities Based on Higher Order Co-Occurrences

This paper proposes an extension of the χ-*Sim* co-clustering algorithm to deal with the text categorization task. The χ-*Sim* method iteratively learns the similarity between documents using similarity between words and vice-versa using higher order co-occurrences.
We provide a graph-theoretical explanation of the χ-*Sim* algorithm and propose two ways to incorporate training data labels into the algorithm for supervised learning. We test the proposed algorithm on different text data sets and

provide the results.

Fawad Hussain
TIMC-IMAG
fawad.hussain@imag.fr

Gilles Bisson
TIMC-IMAG, CNRS
gilles.bisson@imag.fr

CP1

Improving Accessibility of Transaction-Centric Web Objects

This paper addresses the problem of making clickable web-objects required for doing online-transactions readily accessible to blind users. An Information-Retrieval based technique is proposed that uses the context of clickable objects to identify and classify them even when their captions are missing and a reinforcement mechanism based on user feedback, to accommodate previously unseen captions of objects, as well as new categories of objects. Experimental evidence of the techniques effectiveness and end-user experience is presented.

Muhammad A. Islam, FAISAL Ahmed, YEVGEN Borodin
Department of Computer Science
Stony Brook University
maislam@cs.sunysb.edu, faiahmed@cs.sunysb.edu, borodin@cs.sunysb.edu

JALAL Mahmud
IBM Almaden Research Center
jumahmud@us.ibm.com

I.V. Ramakrishnan
Stony Brook University
ram@cs.sunysb.edu

CP1

Semi-Supervised Bio-Named Entity Recognition with Word-Codebook Learning

We describe a novel semi-supervised learning method called Word-Codebook Learning (WCL), and apply it to the task of bio-named entity recognition (bioNER). Typical bioNER systems can be seen as tasks of assigning labels to words in bio-literature text. To improve supervised tagging, WCL learns a class of word-level feature embeddings to capture word semantic meanings (general WCL) or word label patterns (task-oriented WCL) from a large unlabeled corpus. A word codebook is then learned using the obtained embedding vectors. Finally codewords are treated as new word attributes and are added to the system for entity labeling. Without the need for complex linguistic features, WCL yields state-of-the-art performance and shows great improvements over supervised baselines and semi-supervised competitors as demonstrated on BioCreativeII gene name recognition competition data.

Pavel P. Kuksa
Rutgers University
pkuksa@cs.rutgers.edu

Yanjun Qi
NEC Laboratories America, Inc.
yanjun@nec-labs.com

CP1

Exploiting Associations Between Word Clusters and Document Classes for Cross-Domain Text Categorization

Cross-domain text categorization targets on adapting the knowledge learnt from a labeled source-domain to an unlabeled target-domain, where the documents from the source and target domains are drawn from different distributions. However, in spite of the different distributions in raw word features, the associations between word clusters (conceptual features) and document classes may remain stable across different domains. In this paper, we exploit these unchanged associations as the bridge of knowledge transformation from the source domain to the target domain by the nonnegative matrix tri-factorization. Specifically, we formulate a joint optimization framework of the two matrix tri-factorizations for the source and target domain data respectively, in which the associations between word clusters and document classes are shared between them. Then, we give an iterative algorithm for this optimization and theoretically show its convergence. The comprehensive experiments show the effectiveness of this method. In particular, we show that the proposed method can deal with some difficult scenarios where baseline methods usually do not perform well.

Fuzhen Zhuang
Room 506, Kexueyuan Nanlu #6, Zhongguan Cun, Haidian District, Beijing, China
zfz20081983@gmail.com

Ping Luo
Hewlett Packard Labs China
ping.luo@hp.com

Xiong Hui
MSIS Department, Rutgers University
hxiong@rutgers.edu

Qing He
Room 506, Kexueyuan Nanlu #6, Zhongguan Cun, Haidian District, Beijing, China
heq@ics.ict.ac.cn

Yuhong Xiong
Hewlett Packard Labs China
yuhong.xiong@hp.com

Zhongzhi Shi
Room 506, Kexueyuan Nanlu #6, Zhongguan Cun, Haidian District, Beijing, China
shizz@ics.ict.ac.cn

CP2

Do You Trust to Get Trust? A Study of Trust Reciprocity Behaviors and Reciprocal Trust Prediction

Trust reciprocity, a special form of link reciprocity, exists in many networks of trust among users. In this paper, we seek to determine the extent to which reciprocity exists in a trust network and develop quantitative models for measuring reciprocity and reciprocity related behaviors. We identify several reciprocity behaviors and their respective measures. These behavior measures can be employed for predicting if a trustee will return trust to her trustor given that the latter initiates a trust link earlier. We develop for this reciprocal trust prediction task a number of ranking method and classification methods, and evaluated them on

an Epinions trust network data. Our results show that reciprocity related behaviors provide good features for both ranking and classification based methods under different parameter settings.

Ee-Peng Lim, Viet-An Nguyen
Singapore Management University
eplim@smu.edu.sg, vanguyen@smu.edu.sg

Aixin Sun
Nanyang Technological University
axsun@ntu.edu.sg

Hwee-Hoon Tan, Jing Jiang
Singapore Management University
hhtan@smu.edu.sg, jingjiang@smu.edu.sg

CP2

Reconstructing Randomized Social Networks

Many times in social networks the nodes are associated with features. Noise, missing values or efforts to preserve privacy in the network may transform the original network and its feature vectors. We address the problem of reconstructing the original network and features given their randomized counterparts and knowledge of the transformation. We identify cases in which the original network and feature vectors can be reconstructed in polynomial time and illustrate the efficacy of our methods.

Niko Vuokko
Helsinki University of Technology
Department of Information and Computer Science
niko.vuokko@tkk.fi

Evimaria Terzi
Boston University
evimaria@cs.bu.edu

CP2

Publishing Skewed Sensitive Microdata

A highly skewed microdata contains some sensitive attribute values that occur far more frequently than others. Such data violates the eligibility condition assumed by existing works for limiting the probability of linking an individual to a specific sensitive attribute value. Specifically, if the frequency of some sensitive attribute value is too high, publishing the sensitive attribute alone would lead to linking attacks. In many practical scenarios, however, this eligibility condition is violated. In this paper, we consider how to publish microdata under this case. A natural solution is minimally suppressing dominating records to restore the eligibility condition. We show that the minimality of suppression may lead to linking attacks. To limit the inference probability, we propose a randomized suppression solution. We show that this approach has the least expected suppression in a large family of randomized solutions, for a given privacy requirement. Experiments show that this solution approaches the lower bound on the suppression required for this problem.

Yabo Xu
Sun Yat-sen University, China
xuyabo@mail.sysu.edu.cn

Ke Wang
Simon Fraser University, Canada
wangk@cs.sfu.ca

Ada Fu
Chinese University of Hong Kong
adafu@cse.cuhk.edu.hk

Raymond Wong
Hong Kong University of Science and Technology
raywong@cse.ust.hk

CP2

Reconstruction from Randomized Graph Via Low Rank Approximation

The privacy concerns associated with data analysis over social networks have spurred recent research on privacy-preserving publishing of social network data. This paper investigates whether we can reconstruct a graph from the edge randomized graph such that accurate feature values can be recovered. We exploit spectral graph properties and use low rank approximation to filter noise. Our results show key differences from previous findings of point-wise reconstruction methods on numerical data.

Leting Wu, Xiaowei Ying, Xintao Wu
University of North Carolina at Charlotte
lwu8@uncc.edu, xying@uncc.edu, xwu@uncc.edu

CP3

Generation of Alternative Clusterings Using the Cami Approach

Exploratory data analysis aims to discover and generate multiple views of the structure within a dataset. Conventional clustering techniques, however, are designed to only provide a single grouping or clustering of a dataset. In this paper, we introduce a novel algorithm called CAMI, that can uncover alternative clusterings from a dataset. CAMI takes a mathematically appealing approach, combining the use of mutual information to distinguish between alternative clusterings, coupled with an expectation maximization framework to ensure clustering quality. We experimentally test CAMI on both synthetic and real-world datasets, comparing it against a variety of state-of-the-art algorithms. We demonstrate that CAMI's performance is high and that its formulation provides a number of advantages compared to existing techniques.

Xuan Hong Dang, James Bailey
Department of Computer Science and Software Engineering
The University of Melbourne, Australia
xdang@csse.unimelb.edu.au, jbailey@csse.unimelb.edu.au

CP3

Making k-Means Even Faster

The k-means algorithm is widely used for clustering, compressing, and summarizing vector data. We propose a new acceleration for exact k-means that gives the same answer, but is much faster in practice. Like Elkan's accelerated algorithm, our algorithm avoids distance computations using distance bounds and the triangle inequality. Our algorithm uses one novel lower bound for point-center distances, which allows it to eliminate the innermost k-means loop 80% of the time or more in our experiments. On datasets of low and medium dimension (e.g. up to 50 dimensions), our algorithm is much faster than other methods, including methods based on low-dimensional indexes, such as k-d trees. Other advantages are that it is very sim-

ple to implement and it has a very small memory overhead, much smaller than other accelerated algorithms.

Greg Hamerly
Baylor University
hamerly@cs.baylor.edu

CP3

Spectral and Semidefinite Relaxations of the Cluhsic Algorithm

CLUHSIC is a recent clustering framework that unifies the geometric, spectral and statistical views of clustering. In this paper, we show that the recently proposed discriminative view of clustering, which includes the DIFFRAC and DisKmeans algorithms, can also be unified under the CLUHSIC framework. Moreover, CLUHSIC involves integer programming and one has to resort to heuristics such as iterative local optimization. In this paper, we propose two relaxations that are much more disciplined. The first one uses spectral techniques while the second one is based on semidefinite programming (SDP). Experimental results on a number of structured clustering tasks show that the proposed method significantly outperforms existing optimization methods for CLUHSIC. Moreover, it can also be used in semi-supervised classification. Experiments on real-world protein subcellular localization data sets clearly demonstrate the ability of CLUHSIC in incorporating structural and evolutionary information.

Wen-Yun Yang
University of California, Los Angeles
wenyun.yang@gmail.com

James Kwok
Dept. Computer Science and Engineering
Hong Kong Univ. of Science and Technology
jamesk@cse.ust.hk

Bao-Liang Lu
Shanghai Jiao Tong University
blu@cs.sjtu.edu.cn

CP4

Formal Concept Sampling for Counting and Threshold-Free Local Pattern Mining

We describe a Metropolis-Hastings algorithm for sampling closed sets according to any desired strictly positive distribution. Applications are (a) estimating the number of all formal concepts or (b) discovering any number of interesting, non-redundant, and representative local patterns. Setting (a) is useful for estimating the runtime of algorithms examining all formal concepts, setting (b) can be used for the construction of data mining systems not requiring any user-specified thresholds like minimum frequency.

Mario Boley, Henrik Grosskreutz
Fraunhofer IAIS
mario.boley@iais.fraunhofer.de,
henrik.grosskreutz@iais.fraunhofer.de

Thomas Gaertner
University of Bonn
and Fraunhofer IAIS
thomas.gaertner@iais.fraunhofer.de

CP4

An Information-Theoretic Approach to Finding Informative Noisy Tiles in Binary Databases

The task of finding informative recurring patterns in data has been central to data mining research since the introduction of the task of frequent itemset mining in [**?**, **?**, **?**]. In these seminal papers, the informativeness of a recurring itemset in a binary database was formalized by its support in the database. However, it is now widely recognized that an itemset's support is not the best measure of its informativeness. Furthermore, recent work has highlighted that the support of an itemset is highly susceptible to noise, such that it may be more appropriate to search for itemsets that recur only approximately. In this paper, we present a new measure of informativeness for noisy itemsets in binary databases within the formalism of tiles [**?**]. We demonstrate the benefits of our new measure by means of experiments on artificial and real-life data, allowing for objective and subjective evaluation.

Kleanthis Kontonasios
University of Bristol
Department of Computer Science
kk8232@bristol.ac.uk

Tijl De Bie
University of Bristol
tijl.debie@gmail.com

CP4

Mining Top-K Patterns from Binary Datasets in Presence of Noise

The discovery of patterns in binary dataset has many applications, e.g. in electronic commerce, TCP/IP networking, Web usage logging, etc. Still, this is a very challenging task in many respects: overlapping vs. non overlapping patterns, presence of noise, extraction of the most important patterns only. In this paper we formalize the problem of discovering the Top-K patterns from binary datasets in presence of noise, as the minimization of a novel cost function. According to the Minimum Description Length principle, the proposed cost function favors succinct pattern sets that may approximately describe the input data. We propose a greedy algorithm for the discovery of Patterns in Noisy Datasets, named PaNDa, and show that it outperforms related techniques on both synthetic and real- world data.

Claudio Lucchese
ISTI-CNR
claudio.lucchese@isti.cnr.it

Salvatore Orlando
UniVe
orlando@unive.it

Raffaele Perego
ISTI-CNR
perego@isti.cnr.it

CP4

On Mining Statistically Significant Attribute Association Information

Knowledge of the association information between the attributes in a data set provides insight into the underlying structure of the data and explains the relationships (in-

dependence, synergy, redundancy) between the attributes. Complex models learnt computationally from the data are more interpretable to a human analyst when such interdependencies are known. In this paper, we focus on mining two types of association information among the attributes - correlation information and interaction information which capture multivariate dependencies between the data attributes. Identifying the statistically significant attribute associations is a computationally challenging task - the number of possible associations increases exponentially and many associations contain redundant information when a number of correlated attributes are present. In this paper, we explore efficient data mining methods to discover non-redundant attribute sets that contain significant association information indicating the presence of informative patterns in the data.

Pritam Chanda
State University of New York at Buffalo
pchanda@buffalo.edu

Jianmei Yang, Aidong Zhang
Department of Computer Science
State University of New York at Buffalo
jy38@buffalo.edu, azhang@buffalo.edu

Murali Ramanathan
Department of Pharmaceutical Sciences
State University of New York at Buffalo
murali@buffalo.edu

CP5

Collaborative Filtering: Weighted Nonnegative Matrix Factorization Incorporating User and Item Graphs

In this paper, we proposed a unified model for collaborative filtering based on graph regularized weighted nonnegative matrix tri-factorization. In our model, two graphs are constructed on users and items, which exploit the internal information (e.g. neighborhood information in the user-item rating matrix) and external information (e.g. content information such as user's occupation and item's genre, or other kind of knowledge such as social trust network).

Quanquan Gu, Jie Zhou
Department of Automation, Tsinghua University
gqq03@mails.thu.edu.cn, jzhou@tsinghua.edu.cn

Chris Ding
University of Texas at Arlington
chqding@uta.edu

CP5

Alleviating the Sparsity Problem in Collaborative Filtering by Using An Adapted Distance and a Graph-Based Method

Collaborative Filtering (CF) is the process of predicting a user's interest in various items, such as books or movies, based on taste information. One of the key issues in CF is how to deal with data sparsity. We propose two ways to alleviate this problem. The first is a probability-based distance measure, adapted for use with sparse data. The second is a probabilistic graph-based CF algorithm. Experiments show that both approaches lead to more accurate predictions.

Beau Piccart, Jan Struyf, Hendrik Blockeel
K.U.Leuven
beau.piccart@cs.kuleuven.be, jan.struyf@cs.kuleuven.be, hendrik.blockeel@cs.kuleuven.be

CP5

Residual Bayesian Co-Clustering for Matrix Approximation

Matrix approximation for missing value prediction has emerged as an important problem in various domains. We propose residual Bayesian co-clustering (RBC), which learns a generative model corresponding to the matrix from the non-missing entries, and uses the model to predict the missing entries. The model allows mixed memberships of rows and columns to multiple clusters, and can naturally handle the prediction on new rows and columns which are not available in the training process.

Hanhuai Shan
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
shan@cs.umn.edu

Arindam Banerjee
University of Minnesota
banerjee@cs.umn.edu

CP5

Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization

Real-world relational data are seldom stationary, yet traditional collaborative filtering algorithms generally assume this. Motivated by our relational sales prediction problem, we propose a factor-based algorithm that is able to take time into account. By introducing additional factors for time, we formalize this problem as a probabilistic tensor factorization with a special constraint on the time dimension. Further, we provide a fully Bayesian treatment to avoid tuning parameters and achieve automatic model complexity control. To learn the model we develop an efficient sampling procedure that is capable of analyzing large-scale data sets. This new algorithm, called Bayesian Probabilistic Tensor Factorization (BPTF), is evaluated on several real-world problems including sales prediction and movie recommendation. Empirical results demonstrate the superiority of our temporal model.

Liang Xiong
Carnegie Mellon University
lxiong@cs.cmu.edu

CP6

Nonnegative Principal Component Analysis for Proteomic Tumor Profiles

Identifying cancer molecular patterns with high accuracy from proteomic profiles presents a challenge for statistical learning. In this study, we develop a nonnegative principal component analysis and propose a nonnegative principal component analysis based support vector machine with sparse coding to conduct effective feature selection and high-performance proteomic pattern classification. We rigorously show that the over-fitting problem associated with SVMs can be overcome by nonnegative principal component analysis with exceptional sensitivities and specifici-

ties.

Xiaoxu Han
Eastern Michigan University
xiaoxu.han@gmail.com

CP6

Two-View Transductive Support Vector Machines

We propose a novel two-view transductive SVM that takes advantage of both the abundant amount of unlabeled data and their multiple representations to improve the performance of classifiers. The idea is fairly simple: train a classifier on each of the two views of both labeled and unlabeled data, and impose a global constraint that each classifier assigns the same class label to each labeled and unlabeled data. Experimental results show the utility of our method.

Guangxia Li, Steven C. H. Hoi, Kuiyu Chang
Nanyang Technological University
Singapore
ligu0005@ntu.edu.sg, chhoi@ntu.edu.sg, askychang@ntu.edu.sg

CP6

Fast Stochastic Frank-Wolfe Algorithms for Nonlinear Svms

The high computational cost of nonlinear support vector machines has limited their usability for large-scale problems. We propose two novel stochastic algorithms to tackle this problem. These algorithms are based on a simple and classic optimization method: the Frank-Wolfe method, which is known to be fast for problems with a large number of linear constraints. Formulating the nonlinear SVM problem to take advantage of this method, we achieve a provable time complexity of $O(dQ^2/\epsilon^2)$. The proposed algorithms achieve comparable or even better accuracies than the state-of-the-art methods, and are significantly faster.

Hua Ouyang, Alexander Gray
Georgia Institute of Technology
houyang@cc.gatech.edu, gray@cc.gatech.edu

CP6

Single-Pass Distributed Learning of Support Vector Models Using Core-Sets

We propose a method to learn Support Vector Models (SVMs) when the training data is partitioned among several data sources. The basic idea is to consider SVMs which can be reduced to Minimal Enclosing Ball (MEB) problems in an dot-product space. The algorithm requires a single pass through each source of data in order to compute local core-sets: summaries of the data available at each one which are enough to recover locally optimal SVMs. Our main result is that the union of such core-sets provides a global core-set from which the optimal global model can be obtained. Relaxation of the enclosing-ball property using a user-defined tolerance allows to control both the local convergence times and the network load of the distributed solution. Experiments in small and large scale datasets shows that the method is effective in terms of prediction accuracy and scales well in the network size.

Ricardo anculef, Stefano Lodi, Claudio Sartori
University of Bologna
calcetin@gmail.com, stefano.lodi@unibo.it, claudio.sartori@unibo.it

CP7

Bridging Domains with Words: Opinion Analysis with Matrix Tri-Factorizations

With the explosion of user-generated web2.0 content in the form of blogs, wikis and discussion forums, the Internet has rapidly become a massive dynamic repository of public opinion on an unbounded range of topics. A key enabler of opinion extraction and summarization is sentiment classification: the task of automatically identifying whether a given piece of text expresses positive or negative opinion towards a topic of interest. Building high-quality sentiment classifiers using standard text categorization methods is challenging due to the lack of labeled data in a target domain. In this paper, we consider the problem of cross-domain sentiment analysis: can one, for instance, download rated movie reviews from rottentomatoes.com or IMBD discussion forums, learn linguistic expressions and sentiment-laden terms that generally characterize opinionated commentary and then successfully transfer this knowledge to the target domain, thereby building high-quality sentiment models without manual effort? We outline a novel sentiment transfer mechanism based on constrained non-negative matrix tri-factorizations of term-document matrices in the source and target domains. The constrained matrix factorization framework naturally incorporates document labels via a least squares penalty incurred by a certain linear model and enables direct and explicit knowledge transfer across different domains. We obtain promising empirical results with this approach.

Tao Li
Florida International University
taoli@cs.fiu.edu

Vikas Sindhwani
IBM Research
vsindhw@us.ibm.com

Chris Ding
University of Texas at Arlington
chqding@uta.edu

Yi Zhang
Florida International University
yzhan004@cs.fiu.edu

CP7

Exact Passive-Aggressive Algorithm for Multiclass Classification Using Support Class

The Passive Aggressive framework is a principled approach to various online tasks. While the PA framework allows integration with a loss function for multiclass classification, it is yet to have an exact solution without resorting to a numerical approach. We obtain the solution analytically and present an efficient algorithm for updating the weight vectors. We call the method the Support Class Passive Aggressive Algorithm. Experiments demonstrated that our method improves the traditional PA algorithms.

Shin Matsushima, Nobuyuki Shimizu, Kazutohiro Yoshida, Takashi Ninomiya, Hiroshi Nakagawa
The University of Tokyo
masin@r.dl.itc.u-tokyo.ac.jp, shimizu@r.dl.itc.u-tokyo.ac.jp, kyoshiddha@gmail.com, ninomi@r.dl.itc.u-tokyo.ac.jp, n3@dl.itc.u-tokyo.ac.jp

CP7

Efficient Nonnegative Matrix Factorization with Random Projections

The recent years have witnessed a surge of interests in Nonnegative Matrix Factorization (NMF) in data mining and machine learning fields. Despite its elegant theory and empirical success, one of the limitations of NMF based algorithms is that it needs to store the whole data matrix in the entire process, which requires expensive storage and computation costs when the data set is large and high-dimensional. In this paper, we propose to apply the random projection techniques to accelerate the NMF process. Both theoretical analysis and experimental validations will be presented to demonstrate the effectiveness of the proposed strategy.

Fei Wang
Department of Statistical Science, Cornell University
feiwang03@gmail.com

CP8

Frequentness-Transition Queries for Distinctive Pattern Mining from Time-Segmented Databases

We propose a new data mining method called frequentness-transitional pattern mining for finding patterns with interesting sequential behavior, which is specified in a regular expression as a user's query. To cope with the unavoidably large number of candidate patterns, we use Zero-suppressed BDDs to store and operate a large number of candidate itemsets in a short time. Our method detects distinctive itemsets of user-specific models of sequential behaviors.

Shin-Ichi Minato
Hokkaido University
minato@ist.hokudai.ac.jp

Takeaki Uno
National Institute of Informatics
uno@nii.ac.jp

CP8

Robust Mining of Time Intervals with Semi-Interval Partial Order Patterns

We present a new approach to mining patterns from symbolic interval data that extends previous approaches by allowing semi-intervals and partially ordered patterns. The mining algorithm combines and adapts efficient algorithms from sequential pattern and itemset mining for discovery of the new semi-interval patterns. An empirical evaluation with seven real life interval databases demonstrates the flexibility and usefulness of the patterns for sequence classification in comparison with patterns over full intervals.

Fabian Moerchen, Dmitriy Fradkin
Siemens Corporation, Corporate Research
fabian.moerchen@siemens.com,
dmitriy.fradkin@siemens.com

CP8

Cascading Spatio-Temporal Pattern Discovery: A Summary of Results

Given a collecton of boolean spatio-temporal (ST) event types, the cascading spatio-temporal pattern (CSTP) discovery process finds partially ordered subsets of event types whose instances are located together and occur in stages. For example, analysis of crime datasets may reveal frequent occurrence of misdemeanors and drunk driving after bar closings on weekends and after large gatherings such as football games. CSTP discovery is challenging due to the conflicting requirements of extracting statistically meaningful patterns while maintaining computational efficiency in the face of a candidate space that is exponential in the number of event types. This paper proposes a novel interest measure that is statistically meaningful. A novel algorithm that prunes out uninteresting candidates quickly by using the ST nature of the datasets is also proposed.

Pradeep Mohan
Computer Science and Engineering
University of Minnesota
mohan@cs.umn.edu

Shashi Shekhar
University of Minnesota, Twin-Cities
shekhar@cs.umn.edu

James Shine, James Rogers
Engineering Research and Development Corporation, US Army
james.a.shine@usace.army.mil,
james.p.rogers.ii@usace.army.mil

CP8

Consecutive Ones Property and Spectral Ordering

Binary matrix with all 1s consecutive in each column has the consecutive ones property. Modifying data to reach this property exactly is rarely possible and so methods giving good approximate solutions with minimal number of 0s between 1s are needed. Spectral ordering solves the problem approximately and works well empirically. We give theoretical basis for the connection between the property and spectral ordering. We also prove spectral ordering's optimality within a class of algorithms.

Niko Vuokko
Helsinki University of Technology
Department of Information and Computer Science
niko.vuokko@tkk.fi

CP9

On Multidimensional Sharpening of Uncertain Data.

In this paper, we will propose a technique for multi-dimensional enhancement of uncertain data. In many applications, the uncertainty in the different dimensions is caused by independent factors, especially if the different dimensions have been collected from independent sources. In such cases, it is possible to enhance the quality of the data and reduce the underlying uncertainty by using multidimensional uncertainty analysis. In this paper, we will discuss techniques for uncertainty reduction of multidimensional uncertain data. We will examine the effectiveness of the approach over a variety of real and synthetic data sets.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

CP9

On the Use of Combining Rules in Relational Probability Trees

A relational probability tree is a type of decision tree that can be used for probabilistic classification of instances with a relational structure. We show how to integrate probability models based on combining rules in the leaves of such trees, introduce two corresponding learning algorithms and experimentally compare these algorithms to the standard algorithm. The results show that combining rules are useful but do not have an added value when aggregates tests are used.

Daan Fierens
Katholieke Universiteit Leuven
daan.fierens@cs.kuleuven.be

CP9

Subspace Clustering for Uncertain Data

We propose a method for subspace clustering of high-dimensional uncertain data. For this data, deciding whether dimensions are relevant for a subspace cluster is challenging. In uncertain scenarios a strict assignment of objects to single clusters is not appropriate; therefore we enrich our model with the concept of membership degree. Subspace clustering for uncertain data is computationally expensive; thus, we propose an efficient algorithm. In thorough experiments we show the effectiveness of our novel method.

Stephan Günnemann, Hardy Kremer, Thomas Seidl
RWTH Aachen University
guennemann@cs.rwth-aachen.de,
kremer@cs.rwth-aachen.de, seidl@cs.rwth-aachen.de

CP9

Naive Bayes Classifier for Positive Unlabeled Learning with Uncertainty

Existing algorithms for positive unlabeled learning (PU learning) only work with certain data. However, data uncertainty is prevalent in many real-world applications. Based on positive naive Bayes (PNB), a PU learning algorithm for certain data, we propose an algorithm to handle uncertain data and further improve it to avoid user-specified parameter. The conducted experiments show that the proposed algorithm yields good performance without user-specified parameter and has satisfactory performance even on highly uncertain data.

Jiazhen He
College of Information Engineering, Northwest A&F University
Yangling, Shaanxi Province, P.R. China, 712100
hejiazhen@nwsuaf.edu.cn

Yang Zhang
College of Information Engineering, Northwest A&F University
zhangyang@nwsuaf.edu.cn

Xue Li
School of Information Technology and Electrical Engineering,
The University of Queensland, Brisbane, Queensland 4072
xueli@itee.uq.edu.au

Yong Wang
School of Computer, Northwestern Polytechnical University
wangyong@nwpu.edu.cn

CP10

The Application of Statistical Relational Learning to a Database of Criminal and Terrorist Activity

We apply statistical relational learning to a database of criminal and terrorist activity to predict attributes and event outcomes. The database stems from a collection of news articles and court records which are carefully annotated with a variety of variables. We use this data to build relational models from historical data to predict attributes of groups, individuals, and events, such as social network roles. Collective classification is used to boost the accuracy under data poor conditions.

Brian Delaney
MIT Lincoln Laboratory
bdelaney@ll.mit.edu

Andrew Fast
Elder Research, Inc.
fast@datamininglab.com

William Campbell, Clifford Weinstein
MIT Lincoln Laboratory
wcampbell@ll.mit.edu, cjw@ll.mit.edu

David Jensen
UMass Amherst
jensen@cs.umass.edu

CP10

ContexTour: Contextual Contour Visual Analysis on Dynamic Multi-Relational Clustering

Rich context social network data are generated everyday from various platforms such as. The dynamic, multi-relational data pose tremendous challenges on the users who try to understand the underlying patterns in the social media. We introduce ContexTour that generates visual representations for exploring multiple dimensions of community activities and their evolutions. ContexTour consists of (1) Dynamic Relational Clustering and and (2) Dynamic Network Contour-map visualization. The effectiveness of ContexTour is demonstrated on the DBLP dataset.

Yu-Ru Lin
Arizona State University
yu-ru.lin@asu.edu

Jimeng Sun
IBM Research
jimeng@us.ibm.com

Nan Cao
Hong Kong University of Science and Technology
IBM Reserach
nancao@cse.ust.hk

Shixia Liu
IBM Reserach
liusx@cn.ibm.com

CP10
Mining Actionable Subspace Clusters in Sequential Data

Extraction of knowledge from data and using it for decision making is vital in various real-world problems, particularly in the financial domain. We identify several financial problems, which require the mining of *actionable* subspaces defined by objects and attributes over a sequence of time. These subspaces are actionable in the sense that they have the ability to suggest profitable action for the decision-makers. We propose to mine *actionable subspace clusters* from sequential data, which are subspaces with high and correlated *utilities*. To efficiently mine them, we propose a framework MASC (Mining Actionable Subspace Clusters), which is a hybrid of numerical optimization, principal component analysis and frequent itemset mining. We conduct a wide range of experiments to demonstrate the actionability of the clusters and the robustness of our framework MASC. We show that our clustering results are not sensitive to the framework parameters and full recovery of embedded clusters in synthetic data is possible. In our case-study, we show that clusters with higher utilities correspond to higher actionability, and we are able to use our clusters to perform better than one of the most famous value investment strategies.

Kelvin Sim
Institute for Infocomm Research
shsim@i2r.a-star.edu.sg

Ardian K. Poernomo, Vivekanand Gopalkrishnan
Nanyang Technological University, Singapore
ardi0002@ntu.edu.sg, asvivek@ntu.edu.sg

CP10
Identifying Multi-Instance Outliers

We have studied a new data mining problem called multi-instance outlier identification. We have defined the multi-instance outliers (MIO) and analyzed the basic types of MIO. Two general identification approaches are proposed. Based on the two approaches, four concrete multi-instance outlier detectors are then introduced. We conduct experiments over four synthetic data collections and three real-world data collections and the results show the initial success of our methods.

Ou Wu
NLPR, Institute of Automation, Chinese Academy of Sciences
wuou@nlpr.ia.ac.cn

CP11
On Clustering Graph Streams.

In this paper, we will examine the problem of clustering massive graph streams. Graph clustering poses significant challenges because of the complex structures which may be present in the underlying data. The massive size of the underlying graph makes explicit structural enumeration very difficult. Consequently, most techniques for clustering multi-dimensional data are difficult to generalize to the case of massive graphs. Recently, methods have been proposed for clustering graph data, though these methods are designed for static data, and are not applicable to the case of graph streams. Furthermore, these techniques are especially not effective for the case of massive graphs, since a huge number of distinct edges may need to be tracked simultaneously. This results in storage and computational challenges during the clustering process. In order to deal with the natural problems arising from the use of massive disk-resident graphs, we will propose a technique for creating *hash-compressed micro-clusters* from graph streams. The compressed micro-clusters are designed by using a hash-based compression of the edges onto a smaller domain space. We will provide theoretical results which show that the hash-based compression continues to maintain bounded accuracy in terms of distance computations. We will provide experimental results which illustrate the accuracy and efficiency of the underlying method.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

Yuchen Zhao, Philip Yu
University of Illinois at Chicago
yzhao@cs.uic.edu, psyu@cs.uic.edu

CP11
Mining Frequent Graph Sequence Patterns Induced by Vertices

The mining of a complete set of frequent subgraphs from labeled graph data has been studied extensively. Furthermore, much attention has recently been paid to frequent pattern mining from graph sequences (dynamic graphs or evolving graphs). In this paper, we define a novel class of subgraph subsequence called an "induced subgraph subsequence' to enable efficient mining of a complete set of frequent patterns from graph sequences containing large graphs and long sequences. We also propose an efficient method to mine frequent patterns, called "FRISSs (Frequent Relevant, and Induced Subgraph Subsequences)', from graph sequences. The fundamental performance of the method has been evaluated using artificial datasets, and its practicality has been confirmed through experiments using a real-world dataset.

Akihiro Inokuchi
Osaka University
inokuchi@ar.sanken.osaka-u.ac.jp

Takashi Washio
ISIR, Osaka University
washio@ar.sanken.osaka-u.ac.jp

CP11
Grass: Graph Structure Summarization

It is increasingly to replace large graph databases with summaries, either for space efficiency or for privacy protection (e.g., in the case of social network graphs). We propose a formal semantics for answering queries on graph summaries, and we show that important graph- structure queries can be answered efficiently using these semantics. Further, based on this approach to query answering, we formulate three novel graph partitioning/compression problems and algorithms to solve these problems.

Kristen Lefevre
University of Michigan
klefevre@umich.edu

Evimaria Terzi
Boston University
evimaria@cs.bu.edu

CP11

Inferring Probability Distributions of Graph Size and Node Degree from Stochastic Graph Grammars

Stochastic graph grammars are useful models for mining graph databases. In this paper, we extend the utility of such grammars by presenting techniques for learning the probability mass functions of the number of nodes, the number of edges, and the degree of a randomly selected node from graphs in the distribution. Empirical results using both synthetic grammars and a grammar from the domain of AIDS research demonstrate the accuracy of our methods.

Sourav Mukherjee, Tim Oates
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County, Baltimore, USA
sourav1@umbc.edu, oates@cs.umbc.edu

CP12

P-Isomap: Efficient Parametric Update for Isomap with Applications to Visualization

One of the most widely-used nonlinear data embedding methods is ISOMAP. Based on manifold learning, ISOMAP has a parameter k or o that controls how many edges a neighborhood graph has. However, a suitable parameter value is often difficult to determine due to a time-consuming optimization process based on certain criteria, which may not be justified clearly. In addition, when ISOMAP is considered to visualize the data, users might want to test different parameter values to see if they can obtain various insights about the data, but interactions between humans and such visualizations require reasonably efficient updating, even for large-scale data. To tackle these problems, we propose what we call p-ISOMAP, an efficient updating algorithm for ISOMAP when a parameter changes. We present not only its complexity analysis but also its empirical running time comparison, which shows its advantage over ISOMAP. Furthermore, we show interesting visualization applications of p-ISOMAP and how to discover various characteristics of the data through visualization using different parameter values.

Jaegul Choo
College of Computing
Georgia Institute of Technology
joyfull@cc.gatech.edu

Chandan Reddy
Wayne State University
reddy@cs.wayne.edu

Hanseung Lee, Haesun Park
Georgia Institute of Technology
hanseung.lee@gatech.edu, hpark@cc.gtech.edu

CP12

Co-Selection of Features and Instances for Unsupervised Rare Category Analysis

Previous research in rare category analysis focuses on the supervised settings. In this paper, we address the challenge of unsupervised rare category analysis, including feature selection and rare category selection. We propose to jointly deal with them so that they benefit from each other. Therefore, we design an optimization framework to co-select the relevant features and the examples from the rare category. Furthermore, we develop the Partial Augmented Lagrangian Method to solve the optimization problem.

Jingrui He
Machine Learning Department
Carnegie Mellon University
jingruih@cs.cmu.edu

Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
jgc@cs.cmu.edu

CP12

Active Ordering of Interactive Prediction Tasks

Many applications involve a set of prediction (classification, regression, and retrieval) tasks that must be accomplished sequentially through user interaction. If the tasks are interdependent, the order in which they are posed may have a significant impact on their overall performance. We present a novel approach for dynamically ordering such prediction tasks by taking into account the effect of user feedback on the performance of multiple prediction systems.

Abhimanyu Lad
Language Technologies Institute
Carnegie Mellon University
alad@cs.cmu.edu

Yiming Yang
Carnegie Mellon University
yiming@cs.cmu.edu

CP12

Confidence-Based Feature Acquisition to Minimize Training and Test Costs

We present Confidence-based Feature Acquisition (CFA), a novel supervised learning method for acquiring missing feature values when there is missing data at both training and test time. Previous work has considered the cases of missing data at training time (e.g., Active Feature Acquisition, AFA), or at test time (e.g., Cost-Sensitive Naive Bayes, CSNB), but not both. At training time, CFA constructs a cascaded ensemble of classifiers, starting with the zero-cost features and adding a single feature for each successive model. For each model, CFA selects a subset of training instances for which the added feature should be acquired. At test time, the set of models is applied sequentially (as a cascade), stopping when a user-supplied confidence threshold is met. We compare CFA to AFA, CSNB, and several other baselines, and find that CFAs accuracy is at least as high as the other methods, while incurring significantly lower feature acquisition costs.

Marie desJardins, James MacGlashan
University of Maryland, Baltimore County
mariedj@cs.umbc.edu, jmac1@cs.umbc.edu

Kiri Wagstaff
Jet Propulsion Laboratory
kiri.wagstaff@jpl.nasa.gov

CP13
Label Propagation on Heterogeneous Networks

Label propagation is an effective and efficient technique to utilize local and global features in a network for semi-supervised learning. In the literature, no general learning framework or algorithm is available for label propagation in heterogeneous networks comprising several subnetworks, each of which has its own cluster structures that need to be explored independently. In this paper, we introduce an efficient algorithm MINProp (Mutual Interaction-based Network Propagation) and a general regularization framework for propagating information between subnetworks in a heterogeneous network. MINProp sequentially performs label propagation on each individual subnetwork with the current label information derived from the other subnetworks and repeats this step until convergence. The independent label propagation on each subnetwork explores the cluster structure in the subnetwork. The label information from the other subnetworks is used to capture mutual interactions (bicluster structures) between the vertices in each pair of the subnetworks. The iterative propagation algorithm finally converges on each individual subnetwork to the global optimal solution to the convex objective function in the regularization framework. MINProp algorithm was evaluated in simulations and application to disease gene prioritization. MINProp significantly outputperformed the original label propagation algorithm on a single network and the-state-of-the-art methods for discovering disease genes. The experiments also suggest that MINProp is more effective in utilizing the modular structures in a heterogenous network. Finally, MINProp discovered new disease-gene associations that are only reported recently.

Taehyun Hwang
University of Minnesota
Dept. of Computer Science and Engineering
thwang@cs.umn.edu

Rui Kuang
Dept Computer Science
University of Minnesota
kuang@cs.umn.edu

CP13
Radius Plots for Mining Tera-Byte Scale Graphs: Algorithms, Patterns, and Observations

Given large, multi-million node graphs (e.g., FaceBook, web-crawls, etc.), how do they evolve over time? How are they connected? What are the central nodes and outliers of the graphs? We show that the Radius Plot (pdf of node radii) can answer these questions. However, radii and graph diameter estimation are prohibitively expensive for graphs that reach planetary scale. There are two major contributions in this paper: (a) We propose HADI (HAdoop DIameter and radii estimator), a carefully designed and fine-tuned algorithm to compute the diameter of massive graphs, that runs on the top of the hadoop/MapReduce system, with excellent scale-up on the number of available machines (b) We run HADI on several real world dataset including YahooWeb (6B edges, 1/8 of a Terabyte), one of the largest public graphs ever analyzed. Thanks to HADI, we report fascinating patterns on large networks, like the surprisingly small effective diameter, the multi-modal/bi-modal shape of the Radius Plot, and its palindrome motion over time.

U Kang
Carnegie Mellon University
Computer Science Department
ukang@cs.dot.cmu.dot.edu

CP13
Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization

We study the application of spectral clustering, prediction and visualization methods to graphs with negatively weighted edges. We show that several characteristic matrices of graphs can be extended to graphs with positively and negatively weighted edges, giving signed spectral clustering methods, signed graph kernels and network visualization methods that apply to signed graphs. In particular, we review a signed variant of the graph Laplacian. We derive our results by considering random walks, graph clustering, graph drawing and electrical networks, showing that they all result in the same formalism for handling negatively weighted edges. We illustrate our methods using examples from social networks with negative edges and bipartite rating graphs.

Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch
Technische Universität Berlin
kunegis@dai-lab.de, stephan.schmidt@dai-labor.de,
andreas@dai-labor.de

Jürgen Lerner
Universität Konstanz
lerner@inf.uni-konstanz.de

Ernesto De Luca, Sahin Albayrak
Technische Universität Berlin
ernesto.deluca@dai-labor.de, sahin.albayrak@dai-labor.de

CP13
Fast Single-Pair SimRank Computation

SimRank is an intuitive and effective measure for link-based similarity that scores similarity between two nodes as the first-meeting probability of two random surfers, based on the random surfer model. However, when a user queries the similarity of a given node-pair based on SimRank, the existing approaches need to compute the similarities of other node-pairs beforehand, which we call an all-pair style. In this paper, we propose a Single-Pair SimRank approach. Without accuracy loss, this approach performs an iterative computation to obtain the similarity of a single node-pair. The time cost of our Single-Pair SimRank is always less than All-Pair SimRank and obviously efficient when we only need to assess similarity of one or a few node-pairs. We confirm the accuracy and efficiency of our approach in extensive experimental studies over synthetic and real datasets.

Pei Li
School of Information, Renmin Univ of China
lp@ruc.edu.cn

Hongyan Liu
Tsinghua University
hyliu@tsinghua.edu.cn

Jeffrey Xu Yu
The Chinese University of Hong Kong
yu@se.cuhk.edu.hk

Jun He, Xiaoyong Du
Renmin Univ of China
hejun@ruc.edu.cn, duyong@ruc.edu.cn

CP14

The Generalized Dimensionality Reduction Problem

The dimensionality reduction problem has been widely studied in the database literature because of its application for concise data representation in a variety of database applications. The main focus in dimensionality reduction is to represent the data in a smaller number of dimensions that the least amount of information is lost. In this paper, we study the dimensionality reduction problem from an entirely different perspective. We discuss methods to find a representation of the data so that a user-defined objective function is optimized. For example, we may desire to find a reduction of the data in which a particular kind of classifier works effectively. Another example (relevant to the similarity search domain) would be a reduction in which the cluster of k closest points provides the best distance based separation from the remaining data set. We discuss a general abstraction for the problem and provide the broad framework of an evolutionary algorithm which solves this abstraction. We test our framework on two separate instantiations of this framework and provide results illustrating the effectiveness and efficiency of our method.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

CP14

Generalized and Heuristic-Free Feature Construction

State-of-the-art learning algorithms take data in feature vector format as input. Examples belonging to different classes may not always be easy to separate in the original feature space. One may ask: can transformation of existing features into new space reveal significant discriminative information not obvious in the original space? Since there can be infinite number of ways to extend features, it is impractical to first enumerate and then perform feature selection. Second, evaluation of discriminative power on the complete dataset is not always optimal. This is because features highly discriminative on subset of examples may not necessarily be significant when evaluated on the entire dataset. Third, feature construction ought to be automated and general, such that, it doesn't require domain knowledge and its improved accuracy maintains over a large number of classification algorithms. In this paper, we propose a framework to address these problems through the following steps: (1) divide-conquer to avoid exhaustive enumeration; (2) local feature construction and evaluation within subspaces of examples where local error is still high and constructed features thus far still do not predict well; (3) weighting rules based search that is domain knowledge free and has provable performance guarantee. Empirical studies indicate that significant improvement (as much as 9% in accuracy and 28% in AUC) is achieved using the newly constructed features over a variety of inductive learners evaluated against a number of balanced, skewed and high-dimensional datasets.

Wei Fan
IBM T.J.Watson Research,
weifan@us.ibm.com

Erheng Zhong
Sun Yat-Sen University
sysu.zeh@gmail.com

Jing Peng
Montclair State University
pengj@mail.montclaire.edu

Olivier Verscheure
IBM T.J.Watson Research
ov1@us.ibm.com

Kun Zhang
Xavier University of Louisiana
kzhang@xula.edu

CP14

Convex Principal Feature Selection

A popular approach for dimensionality reduction and data analysis is principal component analysis (PCA). A limiting factor with PCA is that it does not inform us on which of the original features are important. There is a recent interest in sparse PCA (SPCA). By apply- ing an L1 regularizer to PCA, a sparse transformation is achieved. However, true feature selection may not be achieved as non-sparse coecients may be distributed over several features. Feature selection is an NP-hard combinatorial optimization problem. This paper re- laxes and re-formulates the feature selection problem as a convex continuous optimization problem that min- imizes a mean-squared-reconstruction error (a criterion optimized by PCA) and considers feature redundancy into account (an important property in PCA and fea- ture selection). We call this new method Convex Principal Feature Selection (CPFS). Experiments show that CPFS performed better than SPCA in selecting features that maximize variance or minimize the mean-squared- reconstruction error.

Mahdokht Masaeli, Yan Yan, Ying Cui
Northeastern University
masaeli.m@neu.edu, yan.y@neu.edu, cui.yi@neu.edu

Glenn M. Fung
Siemens Medical Solutions USA
glenn.fung@siemens.com

Jennifer Dy
Northeastern University
jdy@ece.neu.edu

CP14

Direct Density Ratio Estimation with Dimensionality Reduction

Methods for directly estimating the ratio of two probability density functions without going through density estimation have been actively explored recently since they can be used for various data processing tasks such as nonstationarity adaptation, outlier detection, conditional density estimation, feature selection, and independent component analysis. However, even the state-of-the-art density ratio estimation methods still perform rather poorly in high-dimensional problems. In this paper, we propose a new density ratio estimation method which incorporates dimensionality reduction into a density ratio estimation procedure. Our key idea is to identify a low-dimensional subspace in which the two densities corresponding to the

denominator and the numerator in the density ratio are significantly different. Then the density ratio is estimated only within this low-dimensional subspace. Through numerical examples, we illustrate the effectiveness of the proposed method.

Masashi Sugiyama
Tokyo Institute of Technology
sugi@cs.titech.ac.jp

CP15

An Integrated Framework for Simultaneous Classification and Regression of Time-Series Data.

Zero-inflated time series data often leads to poor model fitting using standard regression methods because they tend to underestimate the frequency of zeros and the magnitude of non-zero values. This paper presents an integrated framework that simultaneously performs classification and regression to accurately predict future values of a zero-inflated time series. We demonstrate the effectiveness of our framework in the context of its application to a precipitation downscaling problem for climate impact assessment studies.

Zubin Abraham, Pang-Ning Tan
Michigan State University
abraha84@msu.edu, ptan@cse.msu.edu

CP15

Multiresolution Motif Discovery in Time Series

Time series motif discovery is an important problem. Available algorithms are not scalable and only consider motifs at a single resolution. Our approach is time and space efficient. We tackle the motif discovery problem as an approximate Top-K frequent subsequence discovery problem. We use the iSAX representation multiresolution capability to obtain motifs at different resolutions. This property allows the user to navigate along the Top-K motifs structure, permitting deeper understanding of the time series database.

Nuno C. Castro, Paulo Azevedo
CCTC - Department of Informatics
University of Minho
castro@di.uminho.pt, pja@di.uminho.pt

CP15

Time-Series Classification in Many Intrinsic Dimensions

We investigate the impact of hubness on time-series classification. Hubness refers to the tendency of some data instances to be included in unexpectedly many k-NN lists of other instances. We show the cause of hubness is high *intrinsic* dimensionality of a time-series data set. We describe the mechanism through which hubs emerge, focusing on DTW distance, and demonstrate how hubness information can be used to improve k-NN classification performance.

Milos Radovanovic
Department of Mathematics and Informatics
University of Novi Sad
radacha@dmi.uns.ac.rs

Alexandros Nanopoulos
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim
nanopoulos@ismll.de

Mirjana Ivanovic
Department of Mathematics and Informatics
University of Novi Sad
mira@dmi.uns.ac.rs

CP15

Unsupervised Discovery of Abnormal Activity Occurrences in Multi-Dimensional Time Series, with Applications in Wearable Systems

We present a method for unsupervised discovery of abnormal occurrences of activities in multi-dimensional time series data. Unsupervised activity discovery approaches differ from traditional supervised methods in that there is no requirement for manually labeled training datasets. In addition, they minimize the need for field experts' knowledge during the setup phases, which makes the deployment phase faster and simpler. We focus our attention on wearable computing systems and their applications in human activity monitoring for health care and medicine. The developed method constructs activity models in multi-dimensional time series based on the frequency and coincidence of activity perceptual primitives in single-dimensional time series data. We study the frequent variations exposed in human activity time series data and leverage physical attributes of the data to classify the activity primitives. A graph clustering approach is used to construct the frequent activity structures. Such structures are used to locate normal and abnormal occurrences of activities in time series. A method is presented to distinguish the abnormal activity occurrences from the normal occurrences. Two state-of-the-art wearable embedded systems (Smartcane and Smartshoe) are used to perform empirical evaluation of the developed methods.

Alireza Vahdatour
University of California, Los Angeles
Computer Science Department
alireza@cs.ucla.edu

Majid Sarrafzadeh
University of California, Los Angeles
majid@cs.ucla.edu

CP16

Scalable Tensor Factorizations with Missing Data

The problem of missing data is ubiquitous in multi-way data. Therefore, we need a robust and scalable approach for factorizing tensors in the presence of missing data. We focus on one of the most well-known tensor factorizations, CANDECOMP/PARAFAC (CP), and formulate the CP model as a weighted least squares problem that models *only* the known entries. We develop an algorithm called CP-WOPT (CP Weighted OPTimization) using a first-order optimization approach to solve the weighted least squares problem.

Tamara G. Kolda
Sandia National Laboratories
tgkolda@sandia.gov

CP16

On Low-Rank Updates to the Singular Value and

Tucker Decompositions

The problem of computing low-rank updates to the thin singular value decomposition and Tucker decomposition is important in data mining applications where data arrives in a stream. We examine the technique by Brand and provide new justification for it as well as modify it to trade off accuracy for speed. We extend the technique to the Tucker decomposition of multi-arrays, and validate the algorithms on datasets of network traffic.

Michael J. O'Hara
Lawrence Livermore National Laboratory
ohara7@llnl.gov

CP16
Mach: Fast Randomized Tensor Decompositions

We propose MACH, a randomized algorithm for computing accurately and efficiently low rank tensor decompositions. Our method is of significant practical value for tensor streams where large amounts of multi-aspect data are accumulated. We provide the theoretical analysis of our proposed method and verify its efficacy on synthetic data and a real world monitoring system application.

Charalampos Tsourakakis
CARNEGIE MELLON UNIVERSITY
ctsourak@cs.cmu.edu

CP17
HCDF: A Hybrid Community Discovery Algorithm

We introduce a novel Bayesian framework for hybrid community discovery in graphs. Our framework, HCDF (short for Hybrid Community Discovery Framework), can effectively incorporate hints from a number of other community detection algorithms and produce results that outperform the constituent parts. We describe two HCDF-based approaches which are: (1) effective, in terms of link prediction performance and robustness to small perturbations in network structure; (2) consistent, in terms of effectiveness across various application domains; (3) scalable to very large graphs; and (4) nonparametric. Our extensive evaluation on a collection of diverse and large real-world graphs, with millions of links, show that our HCDF-based approaches (a) achieve up to 0.22 improvement in link prediction performance as measured by area under ROC curve (AUC), (b) never have an AUC that drops below 0.91 in the worst case, and (c) find communities that are robust to small perturbations of the network structure as defined by Variation of Information (an entropy-based distance metric).

Keith Henderson
Lawrence Livermore National Laboratory
henderson43@llnl.gov

CP17
Toward Finding Valuable Topics

Enterprises depend on their information workers finding valuable information to be productive. However, existing search and recommendation systems can exploit few studies on the correlation between information content and information workers' productivity. In this paper, we combine content, social network and revenue analysis to identify computational metrics for finding valuable information content in people's electronic communications within a large-scale enterprise. Specifically, we focus on two questions: (1) how do the topics extracted from such content correlate with information workers' performance? and (2) how to find valuable topics with high impact on employee performance without having to access performance data? For the first question, we associate the topics with the corresponding workers' productivity measured by the revenue they generate. This allows us to evaluate the topics' influence on productivity. We further verify that the derived topic values are consistent with human assessor subjective evaluation. For the second question, we identify and evaluate a set of significant factors including both content and social network factors. In particular, the social network factors are better in filtering out low-value topics, while content factors are more effective in selecting a few top high-value topics. In addition, we demonstrate that a Support Vector regression model that combines the factors can already effectively find valuable topics. We believe that our results provide significant insights towards scientific advances to find valuable information, especially for scenarios where performance data may not be available.

Zhen Wen
IBM T. J. Watson Research Center
zhenwen@us.ibm.com

Ching-Yung Lin
IBM Reserach
chingyung@us.ibm.com

CP17
Directed Network Community Detection: A Popularity and Productivity Link Model

In this paper, we consider the problem of community detection in directed networks using probabilistic models. Most existing probabilistic models for community detection are either *symmetric* in which incoming links and outgoing links are treated equally or *conditional* in which only one type (i.e., either incoming or outgoing) of links is modeled. We present a probabilistic model for directed network community detection that aims to model both incoming links and outgoing links *simultaneously* and *differentially*.

Tianbao Yang
Michigan State University
yangtia1@msu.edu

Yun Chi, Shenghuo Zhu, Yihong Gong
NEC Laboratories America
ychi@sv.nec-labs.com, zsh@sv.nec-labs.com,
ygong@sv.nec-labs.com

Rong Jin
Michigan State University
rongjin@cse.msu.edu

CP17
Predicting Customer Churn in Mobile Networks Through Analysis of Social Groups

Churn prediction aims to identify subscribers who are about to transfer their business to a competitor. It has emerged as a crucial Business Intelligence application for modern telecommunication operators. We propose a novel churn prediction framework that exploits the structure of customer interactions to predict which groups of subscribers are most prone to churn, before even a single member in the group has churned. Our experimental re-

sults demonstrate the unique advantages of the proposed method.

Yossi Richter
IBM Haifa Lab
richter@il.ibm.com

Elad Yom-Tov, Noam Slonim
IBM Haifa Research Lab
yomtov@il.ibm.com, noams@il.ibm.com

CP18

On Classification of High-Cardinality Data Streams

The problem of massive-domain stream classification is one in which each attribute can take on one of a large number of possible values. Such streams often arise in applications such as IP monitoring, super-store transactions and financial data. In such cases, traditional models for stream classification cannot be used because the size of the storage required for intermediate storage of model statistics can increase rapidly with domain size. Furthermore, the one-pass constraint for data stream computation makes the problem even more challenging. For such cases, there are no known methods for data stream classification. In this paper, we propose the use of massive-domain counting methods for effective modeling and classification. We show that such an approach can yield accurate solutions while retaining space- and time-efficiency. We show the effectiveness and efficiency of the sketch-based approach.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

Philip Yu
University of Illinois at Chicago
psyu@cs.uic.edu

CP18

A Robust Decision Tree Algorithm for Imbalanced Data Sets

We propose a new decision tree algorithm which is robust and insensitive to size of classes. We introduce a new measure, Class Confidence Proportion (CCP), to form the basis of our algorithm. We use Fisher's exact tests to prune branches of the tree. Together these two changes yield a classifier that performs statistically better than not only traditional decision trees but also trees learned from data that has been balanced by well known sampling techniques.

Wei Liu, Sanjay Chawla
School of IT, the University of Sydney
weiliu.au@gmail.com, chawla@it.usyd.edu.au

David Cieslak, Nitesh Chawla
University of Notre Dame
dcieslak@cse.nd.edu, nchawla@nd.edu

CP18

Fast and Accurate Gene Prediction by Decision Tree Classification

We present a novel homology-based gene prediction method that integrates the principled entropy and decision tree concepts. Our goal is to identify "coding" and "non-coding" regions on genomic sequences. However, there is no training data with explicit class labels. We deduce class labels based on biologically-relevant homology measures and use decision tree construction techniques in gene prediction. This method was shown to have comparable accuracy as state-of-the-art methods, while being several orders of magnitude faster.

Rong She, Jeffffrey Chu
Simon Fraser University
rshe@cs.sfu.ca, jeff.sc.chu@gmail.com

Ke Wang
Simon Fraser University, Canada
wangk@cs.sfu.ca

Nansheng Chen
Simon Fraser University
chenn@sfu.ca

CP18

Multi-Label Classification Without Multi-Label Cost

Existing methods of multi-label classification have limited performance in both efficiency and accuracy. We propose an extension over decision tree ensembles that can handle both challenges. Our method is almost without computational cost on handling multipl labels. The experiments shows it is robust on accracy and runs 1-3 orders of magnitude faster than selected algorihms on several different datasets.

Xiatian Zhang
IBM Research - China
Beijing, China
xitianz@cn.ibm.om

Quan Yuan, Shiwan Zhao
IBM Research - China
Beijing, 100193, China
quanyuan@cn.ibm.com, zhaosw@cn.ibm.com

Wei Fan
IBM T.J.Watson Research,
weifan@us.ibm.com

Wentao Zheng
IBM Research - China
Beijing, 100193, China
zhengwt@cn.ibm.com

Zhong Wang
Northeast University
Shenyang, 110819, China
wangzhong.neu@gmail.com

CP19

A Compression Based Distance Measure for Texture

The analysis of texture is an important subroutine in diverse application areas. Almost all existing texture similarity measures require the careful setting of many parameters which make them exceptionally difficult to avoid over fitting. In this work, we introduce a compression based method for texture measures and construct an efficient and robust parameter-free texture similarity measure, CK-1. We demonstrate the utility of our measure with an exten-

sive empirical evaluation on real-world case studies.

Bilson J. Campana, Eamonn Keogh
University of California, Riverside
bcampana@cs.ucr.edu, eamonn@cs.ucr.edu

CP19

A Probabilistic Framework to Learn from Multiple Annotators with Time-Varying Accuracy

This talk will address the challenging problem of learning from multiple annotators whose labeling accuracy differ and vary over time. I will present a framework based on Sequential Bayesian Estimation to learn the expected accuracy at each time step while simultaneously deciding which annotators to query for a label in an incremental learning framework. I will demonstrate a thorough empirical evaluation, showing the strength of the proposed method in time-varying accuracy estimation and label prediction.

Pinar Donmez
Carnegie Mellon University
pinard@cs.cmu.edu

Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
jgc@cs.cmu.edu

Jeff Schneider
Carnegie Mellon University
schneide@cs.cmu.edu

CP19

Predictive Modeling with Heterogeneous Sources

Lack of labeled training examples is a common problem for many applications. At the same time, there is often an abundance of labeled data from related tasks, although they have different distributions and outputs. In this paper, we propose a method to utilize these labeled examples by first performing a sample selection to draw source examples similar to the target data; and then "re-scaled" and assigned new output values to the source examples.

Xiaoxiao Shi
Computer Department, University of Illinois at Chicago
xiao.x.shi@gmail.com

Qi Liu
College of Life Science and Biotechnology, Tongji University
qiliu@tongji.edu.cn

Wei Fan
IBM T.J.Watson Research,
weifan@us.ibm.com

Qiang Yang
Department of Computer Science,
Hong Kong University of Science
qyang@cse.ust.hk

Philip Yu
Computer Science Department,
University of Illinois at Chicago
psyu@uic.edu

CP19

An Integrative Approach to Identifying Biologically Relevant Genes

We propose a novel approach to integrate different types of knowledge for identifying biologically relevant genes. The approach converts different types of external knowledge to its internal knowledge, which can be used to rank genes. Upon obtaining the ranking lists, it aggregates them via a probabilistic model and generates a final list. Experimental results show that using different types of knowledge together can help detect biologically relevant genes.

Zheng Zhao, Jiangxin Wang, Shashvata Sharm
Arizona State University
zhaozheng@asu.edu, jiangxin.wang@asu.edu,
sssharma@asu.edu

Nitin Agarwal
University of Arkansas at Little Rock
nxagarwal@ualr.edu

Huan Liu, Yung Chang
Arizona State University
huan.liu@asu.edu, yung.chang@asu.edu

CP20

Fast Implementation of ℓ_1 Regularized Learning Algorithms Using Gradient Descent Methods

With the advent of high-throughput technologies, ℓ_1 regularized learning algorithms have attracted much attention recently. Dozens of algorithms have been proposed for fast implementation, using various advanced optimization techniques. In this presentation, we demonstrate that ℓ_1 regularized learning problems can be easily solved by using gradient-descent techniques. The basic idea is to transform a convex optimization problem with a non-differentiable objective function into an unconstrained non-convex problem, upon which, via gradient descent, reaching a globally optimum solution is guaranteed. We present detailed implementation of the algorithm using ℓ_1 regularized logistic regression as a particular application. We conduct large-scale experiments to compare the new approach with other state-of-the-art algorithms on eight medium and large-scale problems. We demonstrate that our algorithm, though simple, performs similarly or even better than other advanced algorithms in terms of computational efficiency and memory usage.

Yijun Sun, Yunpeng Cai, Yubo Cheng, Jian Li
University of Florida
sunyijun@biotech.ufl.edu, caiyp@ufl.edu, ————,
————

Steve Goodison
MD Anderson Cancer Center
————

CP20

Adaptive Informative Sampling for Active Learning

In this work we introduce a method that automatically find the most suitable ensemble for active learning for a given data set, which we call adaptive informative sampling (AIS). The algorithm periodically adds data points to the training set, adapts the ratio of classifier types in the ensemble, and optimizes the classifiers using stochas-

tic methods. Experimental results show that the proposed method performs consistently better than homogeneous ensembles and peer active learning methods.

Zhenyu Lu
the University of Vermont
zlu@uvm.edu

CP20

A Permutation Approach to Validation

We give a permutation approach to validation (estimation of out-sample error). One typical use of validation is model selection. We establish the legitimacy of the proposed permutation complexity by proving a uniform bound on the out-sample error, similar to a VC-style bound. We extensively demonstrate this approach experimentally on synthetic data, standard data sets from the UCI-repository, and a novel diffusion data set. The out-of-sample error estimates are comparable to cross validation (CV); yet, the method is more efficient and robust, being less susceptible to overfitting during model selection.

Malik Magdon-Ismail, Konstantin Mertsalov
Rensselaer Polytechnic Institute
magdon@cs.rpi.edu, kmertsalov@gmail.com

CP20

Learning Compressible Models

In this paper, we study the combination of compression and L1-norm regularization in a machine learning context: learning compressible models. By including a compression operation into the L1 regularization, the assumption on model sparsity is relaxed to compressibility: model coefficients are compressed before being penalized, and sparsity is achieved in a compressed domain rather than the original space. We focus on the design of different compression operations, by which we can encode various compressibility assumptions and inductive biases, e.g., piecewise local smoothness, compacted energy in the frequency domain, and semantic correlation. In this sense, use of a compression operation provides an opportunity to leverage auxiliary information from various sources, e.g., domain knowledge, coding theories, unlabeled data, etc. We conduct extensive experiments on brain-computer interfacing, handwritten character recognition and text classification. Empirical results show clear improvements in prediction performance by including compression in L1 regularization. We also analyze the learned model coefficients under appropriate compressibility assumptions, which further demonstrate the advantages of learning compressible models instead of sparse models.

Yi Zhang, Jeff Schneider, Artur Dubrawski
Carnegie Mellon University
yizhang1@cs.cmu.edu, schneide@cs.cmu.edu, awd@cs.cmu.edu

CP21

Mining Maximally Banded Matrices in Binary Data

Binary data occurs often in several real world applications ranging from social networks to bioinformatics. Extracting patterns from such data has been a focus of fundamental data mining tasks including association rule analysis, sequence mining and bi-clustering. Recently, the utility of banded structures in binary matrices has been pointed out with applications in paleontology, bioinformatics and social networking. A binary matrix has a banded structure if both the rows and columns can be permuted so that the 1's exhibit a staircase pattern down the rows, along the leading diagonal. In this paper we show the correspondence between bi-clustering and banded structures in matrices; and the MMBS (Mine Maximally Banded Sub-matrices) algorithm is presented as a direct result of this correspondence. The current state of the art algorithm, MBS, only allows for the discovery of a single band and assumes a fixed column permutation. On the other hand, MMBS facilitates the discovery of multiple bands that may possibly be overlapping or segmented. Our experimental results, presented here, clearly indicate the advantage of MMBS over MBS with both, synthetic and real data sets.

Faris Alqadah, Raj Bhatnagar
Universtiy of Cincinnati
alqadaf@email.uc.edu, raj.bhatnagar@uc.edu

Anil Jegga
Cincinnati Children's Hospital Medical Center
anil.jegga@cchmc.org

CP21

A Generalized Tree Matching Algorithm Considering Nested Lists for Web Data Extraction

This paper studies structured data extraction from Web pages. It presents a generalized tree matching algorithm for template patterns detection from web pages for data extraction. It is the first tree matching algorithm which can handles nested lists of patterns (through a novel grammar generation algorithm). It is the first algorithm which can solve both problems of extracting data from multiple pages with the same template or from a single page which has multiple lists.

Nitin Jindal, Bing Liu
University of Illinois at Chicago
nitin.jindal@gmail.com, liub@cs.uic.edu

CP21

Cross-Selling Optimization for Customized Promotion

The profit of a product depends on its influence on the sales of other products. How to promote the right products to the right customers becomes a key issue in marketing. In this presentation, we propose a new formulation of promotion value by considering cross-selling effects. We investigate the problem of customized promotion, which identifies products and customers to maximize promotion effect. We propose greedy and randomized algorithms to conduct promotion in an efficient manner.

Nan Li
Computer Science Department
University of California, Santa Barbara
nanli@cs.ucsb.edu

Yinghui Yang
Graduate School of Management
University of California, Davis
yiyang@ucdavis.edu

Xifeng Yan
Department of Computer Science
University of California at Santa Barbara

xyan@cs.ucsb.edu

CP21

Evaluating Query Result Significance in Databases Via Randomizations

Many sorts of structured data are commonly stored in a multi-relational format of interrelated tables. Under this relational model, exploratory data analysis can be done by using relational queries. We consider the problem of assessing whether the results returned by such a query are statistically significant or just a random artifact in the data. Our approach is based on randomizing the tables occurring in the queries and repeating the original query on the randomized tables.

Markus Ojala, Gemma Garriga
HIIT, Aalto University School of Science and Technology
Department of Information and Computer Science
Markus.Ojala@tkk.fi, gemma.garriga@hut.fi

Aristides Gionis
Yahoo! Research
Barcelona, Spain
gionis@yahoo-inc.com

Heikki Mannila
HIIT, Aalto University School of Science and Technology
Department of Information and Computer Science
heikki.mannila@aaltouniversity.fi

MS1

Estimation of Topic Cardinality in Document Collections

The exponential growth of the size and popularity of the world wide web has increased the interest in text analysis. One of the applications of text analysis consists in grouping (i.e. clustering) texts according to the main topic they deal with. This paper presents the first part of a fundamental new approach towards this problem. A competitive setting for production of documents by n data providers is introduced. It is reasoned that in this setting, production of documents about topics is not random, but obeys Zipfs law. Under this assumption, the number of topics can be estimated with fairly high accuracy. Three main advantages of this technique are noticed. Firstly, the estimated number is not a fixed constant in terms of the size of the text collection. Secondly, the estimation does not make assumptions on the clustering method that is used. Thirdly, this method provides a dynamical instrument to verify the recall of clustering.

Antoon Bronselaer
Ghent University
antoon.bronselaer@ugent.be

Saskia Debergh
University of Antwerp
saskia.debergh@ua.ac.be

Dirk Van Hyfte
Katholieke Universiteit
dirk.van.hyfte@skynet.be

Guy De Tre
Ghent University
guy.detre@ugent.be

MS1

Importance and Ontology-based Enhancement of Concepts in Structured Queries

Concept-based information retrieval systems treat each document as a set of concepts instead of terms, words or phrases. Definitions of concepts provide information which words are required to identify specific concepts. In this paper, we use a domain ontology as a source of concept definitions a single concept is defined not only with single terms/words but also with relations to other concepts. In order to represent true contributions of terms and word towards concept definitions we propose a novel schema for automatic assignment of term importance AATI. The main component of the AATI schema is an iterative unsupervised algorithm capable of determining importance values of terms/words constituting concept definitions. The merge of ontology providing words/terms defining concepts, and AATI that determines importance values of those terms/words allows for defining and identifying concepts in documents.

Zhan Li, Marek Reformat
University of Alberta
xxx@xxx.xxx, xxx@xxx.xxx

Ronald Yager
Iona College
yager@panix.com

MS1

Evaluation of Abstraction-Based Data Models for Text via Supervised Learning Methods

With an increase in computational power, demand for enhanced data models for massive collections of text documents is growing. In creating these data structures, the aim is to increase the accuracy of the text representation while maintaining a compact data format. In this paper, we present abstraction-based data models that use paths that contain keywords (words appearing in a document) and their abstracts for document representation rather than merely using the presence of the keywords themselves. We also experimentally evaluate the usefulness of the new data models, based on experiments that involve the classification of documents from the well-known 20 Newsgroups text benchmark.

Richard McAllister, Rafal Angryk
Montana State University
mcallis@montana.edu, angryk@cs.montana.edu

MS1

Some Features of Partitions Useful for Linguistic Data Mining

The semantics of linguistic terms used in linguistic data mining and summarization are generally based on a partition of attribute domains. An implication of this is the centrality of the process of partitioning in these linguistically focused tasks. Here we look to provide a deeper understanding of partitions by investigating a number of measures associated with partitions. We first discuss congruence measures, which are used to calculate the similarity between two partitions. We provide a number of examples of this type of measure. Another class of measures we investigate are prognostication measures. This measure, closely related to a concept of containment between partitions, are useful in indicating how well knowl-

edge of an objects class in one partition predicts its class in a second partitioning. Finally we introduce a measure of the non-specificity of a partition. This measures a feature of a partition related to the generality of the constituent classes of the partition. A common task in data mining is developing rules that allow us to predict the class of an object based upon the value of some features of the object. The more narrowly we partition the features in the rules the better we can predict an objects classification. However counterbalancing this is the fact that to many narrow feature categories are difficult for human experts to cognitively manage, this introduces a fundamental issue in data mining. We shown how the combined use of our measures prognostication and non-specificity allow us navigate this issue.

Ronald Yager
Iona College
yager@panix.com

MS1

A Perspective on the Role of Fuzzy Logic, Computational Linguistics and Natural Language Processing in Data Mining and Data Summarization

A perspective and novel approach to the use of fuzzy logic based tools and techniques for a human consistent data mining is presented adopting as a point of departure that for the human being the only fully natural means of articulation and communication is natural language. The use of fuzzy and possibilistic tools, and computing with words, for handling vagueness (imprecision) is shown, concentrating on linguistic data summaries, emphasizing a linguistic quantifier driven aggregation and the use of protoforms. A brief account of relations between this approach and natural language generation (NLG) and systemic functional linguistics (SFL) is given. Other papers from the Minisymposium are summarized emphasizing their conceptual closeness to the topic and scope of the Minisymposium and our arguments and proposals.

Slawomir Zadrozny, Janusz Kacprzyk
Polish Academy of Sciences
spzadrozny@gmail.com, kacprzyk@ibspan.waw.pl