



The *Institute for Operations Research and Management Sciences*
(INFORMS) presents an

Artificial Intelligence and Data Mining Workshop

in conjunction with the



With support from

INFORMS College of Artificial Intelligence (AI)

INFORMS Section on Data Mining (DM)

Industrial & Manufacturing Systems Engineering, Iowa State University

Industrial & Manufacturing Systems Engineering, University of Texas at Arlington

Decision & Information Sciences, University of Florida

Workshop Organizers

Program Chairs: Victoria (Tory) Chen (vchen@uta.edu)
Andrew Kusiak (andrew-kusiak@uiowa.edu)

Workshop Committee: Haldun Aytug (aytugh@ufl.edu)
Wei Jiang (wjiang@stevens-tech.edu)
Sigurdur (Siggi) Olafsson (olafsson@iastate.edu)
George Runger (George.Runger@asu.edu)
Riyaz Sikora (rsikora@uta.edu)
Janet Twomey (janet.twomey@wichita.edu)

Guest editors for *Annals of Operations Research* special volume: Siggi Olafsson, Tory Chen

INFORMS College of Artificial Intelligence

Website: <http://www.informs.org/Subdiv/Section/ai.html>

Chair: Gary Koehler
Vice Chair: Yuehwern Yih
Secretary-Treasurer: Selwyn Piramuthu
Newsletter Editor: Haldun Aytug

INFORMS Section on Data Mining

Website: <http://dm.section.informs.org/>

Chair: Tory Chen
Vice Chair: Andrew Kusiak
Secretary-Treasurer: Julia Tsai
Webmaster and Newsletter Editor: Siggi Olafsson
Council Members: Wei Jiang, T. Warren Liao, Shane Pederson, Marietta Tretter

INFORMS

The AI/DM Workshop Organizers would like to thank Terry Cryan, Christy Kline, Barry List, and Sandy Owens for their assistance; Melinda Brown and Shirley Mohr for producing the Workshop CD Proceedings; and Cheryl Clark for organizing facilities and catering.

Program Schedule

(All sessions will be held at the Westin Convention Center Hotel)

8:00 AM – 8:30 AM	Breakfast and Welcome
8:30 AM – 10:00 AM	Technical Sessions A1, A2
10:00 AM – 10:15 AM	Break
10:15 AM – 11:45 PM	Technical Sessions B1, B2
11:45 AM – 1:45 PM	Lunch
12:15 PM – 1:45 PM	Panel on Research and Funding Opportunities
1:45 PM – 3:15 PM	Tutorial
3:15 PM – 3:30 PM	Break
3:30 PM – 5:00 PM	Technical Sessions C1, C2

Track 1: Novel Methods in Learning and Data Mining (A1, B1, C1)

Track 2: Applications and Applied Methods (A2, B2, C2)

Panel on Research and Funding Opportunities

Session Chair: Andrew Kusiak, *University of Iowa* (andrew-kusiak@uiowa.edu)

Panelists:

Abhijit Deshmukh, *National Science Foundation* (adeshmuk@nsf.gov)

Wendy Martinez, *Office of Naval Research* (martinwe@onr.navy.mil)

Michael Vannier, *University of Chicago Medical Center* (mvannier@uchicago.edu)

Maria Zemankova, *National Science Foundation* (mzemanko@nsf.gov)

Tutorial

Session Chair: Janet Twomey, *Wichita State University* (janet.twomey@wichita.edu)

Presenters:

George Runger, *Arizona State University* (George.Runger@asu.edu)

Wei Jiang, *Stevens Institute of Technology* (wjiang@stevens-tech.edu)

Kwok Tsui, *Georgia Institute of Technology* (ktsui@isye.gatech.edu)

Andrew Kusiak, *University of Iowa* (andrew-kusiak@uiowa.edu)

Track 1: Novel Methods in Learning and Data Mining

Technical Session A1: Advances in Learning

Session Chair: Riyaz Sikora, *University of Texas at Arlington* (rsikora@uta.edu)

1. *Genetic Algorithm Based Learning Using Feature Construction*
Selwyn Piramuthu (selwyn@ufl.edu), Riyaz Sikora
2. *Learning Optimal Parameter Values in Dynamic Environment: An Experiment with Softmax Reinforcement Learning Algorithms*
Riyaz Sikora (rsikora@uta.edu)
3. *Using Genetic Algorithms to Solve the Strategic Learning Problem*
Fidan Boylu (fidan.boylu@business.uconn.edu), Haldun Aytug, Gary Koehler
4. *Some Recent Results on the Performance and Implementation of Manifold Learning Algorithms*
Xiaoming Huo (xiaoming@isye.gatech.edu)

Technical Session B1: Unsupervised Methods

Session Chair: Tom Au, *at&t Research Labs* (sau@att.com)

5. *Irregularity Analysis in Time Series Data*
Tom Au, Winnie Duan (rongduan@att.com), Wei Jiang
6. *Using Clustering to Improve Sales Forecasts in Retail Merchandizing*
Mahesh Kumar (maheshk@rutgers.edu)
7. *A Novel Approach to Classification in Financial Applications*
Marco Better (Marco.Better@colorado.edu), Fred Glover, Gary Kochenberger, Haibo Wang
8. *Entropy Maximizing Density Estimation Using a Genetic Algorithm*
Parag Pendharkar (pxp19@psu.edu), Jim Rodger

Technical Session C1: Support Vector Machines

Session Chair: Wei Jiang, *Stevens Institute of Technology* (wjiang@stevens-tech.edu)

9. *Solving Discrete Support Vector Machines with Tabu Search*
Stefan Lessmann (lessmann@econ.uni-hamburg.de), Stefan Voß
10. *Adjusted Support Vector Machines Based on a New Loss Function*
Shuchun Wang (swang1@isye.gatech.edu), Kwok Tsui, Wei Jiang
11. *Time Series Classification by Discrete Support Vector Machines*
Carlotta Orsenigo, Carlo Vercellis (carlo.vercellis@polimi.it)
12. *Hierarchical Local Clustering for Constraint Reduction in Rank-Optimizing Linear Programs*
Kaan Ataman (kaan-ataman@uiowa.edu), Nick Street

Track 2: Applications and Applied Methods

Technical Session A2: Bioinformatics and Methods for Biomedical Applications

Session Chair: Seoung Kim, *University of Texas at Arlington* (sbkim@uta.edu)

13. *Image Denoising via Solution Paths*
Li Wang, Ji Zhu (jizhu@umich.edu)
14. *A Bayesian Approach for the Alignment of High-Resolution NMR Spectra*
Seoung Kim (sbkim@uta.edu), Zhou Wang, Carlos Duran
15. *Disparate Data Fusion for Protein Phosphorylation Prediction*
Genetha Gray (gagray@sandia.gov), Pam Williams, Ken Sale
16. *Solving a Mixed-Integer Programming Formulation of a Multi-Category Constrained Discrimination Model*
Paul Brooks (jpbrooks@vcu.edu), Eva Lee

Technical Session B2: Interfacing Learning and Operations Research for Business and Industry

Session Chair: Tory Chen, *University of Texas at Arlington* (vchen@uta.edu)

17. *A Method for Reconciling Values of Parameters*
Shinya Kikuchi, Manoj Jha (mkjha@eng.morgan.edu)
18. *Improving the Estimation of Random Coefficient Logit Models of Demand*
Marietta Tretter (MTretter@mays.tamu.edu)
19. *Efficient Computer Experiment Based Optimization through Variable Selection*
Thomas Shih (dshih@uta.edu), Venkata Pilla, Seoung Kim, Jay Rosenberger, Tory Chen
20. *Modern Machine Learning for Automatic Optimization Algorithm Selection*
Patty Hough (pdhough@sandia.gov), Pam Williams

Technical Session C2: Advances in Data Mining for Manufacturing

Session Chairs: George Runger, *Arizona State University* (George.Runger@asu.edu)

21. *Time-Based Detection of Changes to Multivariate Patterns*
Jing Hu (Jing.Hu.1@asu.edu), George Runger
22. *Knowledge Discovery to Support Product Family Design*
Seung Ki Moon (moonsky@psu.edu), Timothy Simpson, Soundar Kumara
23. *Improving Productivity in Manufacturing Environments Using Data Mining*
Pam Ajoku (pnel@pitt.edu), Bart Nnaji
24. *Discovering Service Inventory Demand Patterns from Archetypal Demand Training Data*
Gene Beardslee (Eugene.A.Beardslee@saic.com), Ted Trafalis

Abstract 1

Genetic Algorithm Based Learning Using Feature Construction

Selwyn Piramuthu
University of Florida

Riyaz T. Sikora
University of Texas at Arlington

Genetic algorithms (GAs) are excellent for learning concepts that span complex space, especially those with a large number of local optima. Learning algorithms, in general, perform well on data that has been pre-processed to reduce complexity. Several studies have documented their effectiveness on raw as well as pre-processed data using feature selection, etc. Unlike other learning algorithms (e.g., those in feed-forward neural networks), GAs are not particularly effective in reducing data complexity while learning difficult concepts. Feature construction has been shown to reduce complexity of space spanned by input data. In this paper, we present an algorithm for enhancing the learning process of a GA through the use of feature construction as a pre-processing step. We also apply the same procedure on two other learning methods, namely C4.5 and Lazy Learner, and show the improvement in performance.

Abstract 2

Learning Optimal Parameter Values in Dynamic Environment: An Experiment with Softmax Reinforcement Learning Algorithms

Riyaz T. Sikora
University of Texas at Arlington

Many learning and heuristic search algorithms require tuning of parameters to achieve optimum performance. In stationary and deterministic problem domains this is usually achieved through off-line sensitivity analysis. However, this method breaks down in nonstationary and non-deterministic environments, where the optimal set of values for the parameters keep changing over time.

In this paper we present a simple meta-learning algorithm to learn the temperature parameter of the Softmax reinforcement-learning (RL) algorithm. We test the effectiveness of this meta-learning algorithm in two domains. The first is the classic reinforcement-learning problem known as the k -armed bandit problem. The second domain involves a context of strategic interaction consisting of homogeneous sellers of a single raw material or component vying for business from a single buyer. The sellers are modeled as artificial agents that learn increasingly effective bidding strategies.

We model non-stationarity in the first domain of k -armed bandit problem by periodically switching the reward distributions. The second domain is in effect a non-stationary and non-deterministic learning problem since it is a game of strategic interaction involving multiple agents. In both domains we show the improvement in performance brought on by the use of meta-learning.

Abstract 3

Using Genetic Algorithms to Solve the Strategic Learning Problem

Fidan Boylu
University of Connecticut

Haldun Aytug and Gary J. Koehler
University of Florida

None of the existing data mining algorithms take into account the possibility that future observed attributes might have been deliberately modified by their source when the source is a human or collection of humans. They fail to anticipate that people (and collections of people) might “game the system” and alter their attributes to attain a positive classification. For example, there are many websites that show how to increase one's credit score. Typically, attributes might be altered to help achieve a positive classification. We investigate this potential strategic gaming and have developed inference methods to determine better discriminant functions in the presence of such strategic behavior and show that this strategic behavior results in an alteration of the usual learning rules. We call this problem the Strategic Learning problem. In this paper, we provide a Genetic Algorithm for solving the Strategic Learning problem. We start by reducing the Strategic Learning problem to an unconstrained search over the space of linear functions. Once we have accomplished this, we develop a Genetic Algorithm (GA) to perform this search.

Abstract 4

Some Recent Results on the Performance and Implementation of Manifold Learning Algorithms

Xiaoming Huo
Georgia Institute of Technology

Manifold learning is becoming an important research topic in statistical and machine learning. Many works have appeared. A series of newly proposed algorithms have been proven to be effective in a wide range of applications. In many cases, the theoretical properties are not completely known. It is interesting to see that many algorithms depend on specific linear invariant subspaces. We recently derived performance bounds on these algorithms. Moreover, detailed analysis on these algorithms, especially the considerations of local utilized dimensions and the property of the associated algorithms, demonstrates that there are common principles in designing an efficient manifold learning algorithm. We reveal these principles and utilize them to analyze existing manifold learning algorithms. Our products include interpretation of some reported numerical examples, as well as predicted advantages and disadvantages of existing methods.

Abstract 5

Irregularity Analysis in Time Series Data

Siu-Tong Au and Rong Duan
at&t Research Labs

Wei Jiang
Stevens Institute of Technology

Government and corporations nowadays collect time series data at lowest possible details such as by locations, parts, products, or even individuals. Most of data cleaning methods assume one known type of irregularity. This paper provide a framework for the situation that there are multiple irregularities hiding in large volumes of cross sectional time series and develops a data mining platform to capture these key irregularities one by one based on their importance. It attempts to automate how a data analyst looking at time series graphs when cleaning the data (but there are too many to look at). Clustering is applied to group time series with similar pattern, and the principal irregular component of the dominated time series group is extracted and adjusted. The platform continues to cluster, extract and adjust the next significant irregular components iteratively. Finally all these significant irregular components are summarized in graphic forms to help analysts to know the data better and faster before any analysis and modeling.

Abstract 6

Using Clustering to Improve Sales Forecasts in Retail Merchandizing

Mahesh Kumar
Rutgers University

Given sales forecasts for a set of items along with the standard deviation associated with each forecast, we propose a new method of combining forecasts using the concepts of clustering. Clusters of items are identified based on similarity in their sales forecasts and then a common forecast (or combined forecast) is computed for each cluster of items. The objective of clustering is to minimize the mean square error (MSE), which is the sum of the variance and squared bias of the combined forecasts. It is easy to show that combining forecasts from a group of items generally decreases the variance but increases the bias of the combined forecast. A new clustering method is proposed based on this tradeoff between the decreased variance and increased bias. A useful property of the proposed clustering method is that it automatically finds the right number of clusters. On a real dataset from a national retail chain we have found that the proposed method of combining forecasts produces significantly better sales forecasts than either the individual forecasts (forecasts without combining) or an alternate method of using a single combined forecast for all items in a product line sold by this retailer.

Abstract 7

A Novel Approach to Classification in Financial Applications

Marco Better and Fred Glover
University of Colorado at Boulder

Gary Kochenberger
University of Colorado at Denver

Haibo Wang
Texas A&M International University

Modern methods for classification analysis involve processes for “learning” to correctly assign elements of a data set to certain classes. In many settings, the learning processes are supervised; i.e. the classes that the training data belong to are known in advance. In many other settings, however, the classes are not known a priori, and a process utilizing unsupervised learning is necessary.

We present a novel, two-stage unsupervised learning methodology for the classification problem. Stage one consists of a special clustering method based on a quadratic, unconstrained optimization model that finds optimal classes for the data. Stage two makes use of enhanced mathematical programming models for classifying the data into the optimal classes found during stage one.

A significant advantage of our approach, as demonstrated by computational testing, is the ability to yield more meaningful classifications than previously achieved in a variety of settings. We report the outcome of training and testing our method on various data sets from the data mining literature, with specific applications in finance. The comparative results disclose the effectiveness and versatility of the approach, and its merit as a tool for modeling and solving practical problems.

Abstract 8

Entropy Maximizing Density Estimation Using a Genetic Algorithm

Parag C. Pendharkar
Pennsylvania State Harrisburg

James A. Rodger
Indiana University of Pennsylvania

Density estimation is a commonly studied problem in data mining literature. Several unsupervised learning algorithms, neural networks, and support vector machine based clustering and classification approaches are kernel-based; and require sophisticated algorithms for density estimation. It is a well known fact that density estimation problem is a nontrivial optimization problem and most of the existing density estimation algorithms provide locally optimal solutions. In our paper we will provide a new entropy maximizing approach that uses global search genetic algorithm to estimate densities for a given data set. Unlike the traditional local search approaches, our approach uses global search and is more likely to provide solutions that are close to global optimum. Using a simulated dataset, we compare the results of our approach with the maximum likelihood approach.

Abstract 9

Solving Discrete Support Vector Machines with Tabu Search

Stefan Lessmann and Stefan Voß
University of Hamburg, Germany

We consider the case of classification with discrete support vector machines. While standard support vector machines use a continuous approximation to measure classification errors discrete support vector machines incorporate a step-function minimizing errors directly. We argue that this modification facilitates more accurate class predictions. In particular, applications that involve cost sensitive learning with asymmetric costs per error type should benefit from a discrete error measure. However, while support vector machines are trained by minimizing a convex quadratic program discrete support vector machines require solving a more complex mixed integer program. Therefore, we develop a tabu search heuristic to train the respective classifier. Considering the structure of the underlying optimization problem its optimal solution has to be contained in the set of extreme points of a relaxed problem that can be solved by fast linear programming methods. Our tabu search exploits this observation utilizing the respective extreme points as the neighborhood between candidate solutions. Using this extreme point tabu search we compare discrete support vector machines with standard support vector machines on well known benchmark data sets. Experiments include standard as well as cost sensitive classification settings. The discrete support vector machine is found to deliver superior results when cost-distributions are uneven while it performs competitive in standard settings.

Abstract 10

Adjusted Support Vector Machines Based on a New Loss Function

Shuchun Wang and Kwok-Leung Tsui
Georgia Institute of Technology

Wei Jiang
Stevens Institute of Technology

Support vector machine (SVM) since its first introduction has attracted many attentions from researchers and have been very successful in practical applications. The basic idea of SVM finds an optimal separating hyperplane by maximizing the margin of separation between two classes. However, there are some concerns associated with SVM. First, the SVM solution depends critically on only a few support vectors, which are often located near the boundary, and consequently is very sensitive to outliers. Second, the optimal separating hyperplane is equidistant to the two classes without considering the number of training samples and corresponding dispersions in each class. In this paper, we develop a new SVM algorithm, adjusted support vector machine (ASVM), based on a new loss function and adjust the SVM solution according to the sample sizes and dispersions of the two classes. Numerical experiments show that the proposed ASVM outperforms SVM, especially when the two classes have large differences in sample size and dispersion.

Abstract 11

Time Series Classification by Discrete Support Vector Machines

Carlotta Orsenigo
University of Milan, Italy

Carlo Verzellis
Polytechnic Institute of Milan, Italy

Time series classification is a supervised learning problem aimed at labeling temporally structured multivariate sequences of variable length. The most common approach reduces time series classification to a static problem by suitably transforming the set of multivariate input sequences into a rectangular table made by a fixed number of attributes. Then, any of the existing efficient methods for classification can be applied for learning and predicting the class of future temporal sequences.

In this paper we propose an extension of discrete support vector machines, that have been shown to outperform other competing classification methods on benchmark datasets, for time series classification. In order to transform a temporal dataset into the rectangular shape we also develop a constrained variant of dynamic time warping. Preliminary computational results on marketing datasets indicate the effectiveness of the proposed method in comparison to other techniques.

Abstract 12

Hierarchical Local Clustering for Constraint Reduction in Rank-Optimizing Linear Programs

Kaan Ataman and W. Nick Street
University of Iowa

Many real-world problems, such as lead scoring in marketing and treatment planning in medicine, require predictive models that successfully order cases relative to each other. We developed a linear-programming-based learning method, similar to SVMs, that optimizes ranking problems with binary output by maximizing an approximation to area under the ROC curve (AUC). This method consistently outperforms SVMs and other classification methods in terms of ranking. However, our formulation requires a quadratic number of constraints, limiting its application to moderate and large problems. In this paper, we present a localized hierarchical clustering algorithm that reduces the size of the problem by clustering points based on both geometric similarity and class labels. This method dramatically reduces the number of constraints while maintaining high-quality ranking ability.

Abstract 13

Image Denoising via Solution Paths

Li Wang and Ji Zhu
University of Michigan

Image denoising is a problem that arises in many engineering fields, because in practice images can be easily contaminated with noise when they are captured or transmitted. Many image denoising methods can be characterized as minimizing “loss + penalty,” where the “loss” measures the fidelity of the denoised image to the data, and the “penalty” measures the smoothness of the denoising function. In this paper, we consider a family of models that use the L1-norm of the pixel updates as the penalty.

The L1-norm penalty has the advantage of changing only the noisy pixels, while leaving the non-noisy pixels untouched. We derive efficient algorithms that compute entire solution paths of these L1-norm penalized models, which facilitate the selection of a balance between the “loss” and the “penalty.”

Abstract 14

A Bayesian Approach for the Alignment of High-Resolution NMR Spectra

Seoung Bum Kim, Zhou Wang, and Carlos M. Duran
University of Texas at Arlington

The rapid progresses in human genome project and biotechnologies result in the sheer volume of datasets associated with in-depth scientific knowledge. Metabolomics is defined as the study for understanding metabolic process in living systems. Metabolomics approaches that used high-resolution nuclear magnetic resonance (NMR) spectroscopy have been used to characterize metabolic variations in response to physiological alternation, disease states, genetic modification, and nutrition intake. An NMR spectrum usually involves tens of thousands of variables and the comparison of multiple spectra lead to huge number of data points and a situation that poses a great challenge to analytical and computational capabilities. When considering multiple spectra, small variations due to concentration, pH, and temperature, influence the spectral alignment and thus can interfere with direct comparisons between samples. Thus, it is crucial to align spectra before applying any subsequent statistical analyses such as clustering and classification. In this study, we propose a novel algorithm for the NMR spectra alignment within the Bayesian framework, which allows estimating the vertical and horizontal shifts simultaneously in the existence of noise. Effectiveness of our algorithm is demonstrated through the comparison of existing algorithms and experiments with real high-resolution NMR data.

Abstract 15

Disparate Data Fusion for Protein Phosphorylation Prediction

Genetha A. Gray, Pamela J. Williams, and Kenneth L. Sale
Sandia National Laboratories

Our work is motivated by the growing trends of collecting and interpreting large volumes of data from areas such as genomics, proteomics, chemistry, and medicine. Specifically, we want to make meaningful interpretations of data types that provide different views of the same situation, give complimentary information despite appearing dissimilar, and are collected and stored in a variety of formats. We are investigating the applicability of ensemble classification to disparate data sources. Traditionally, these techniques have been used to combine the predictions of different classifiers of the same data set into a single classification.

We will discuss an algorithm for fusing classifications of disparate data and its applicability to the phosphorylation prediction problem. Prediction is important in immune studies since phosphorylation is a key trigger of response pathways. Computational prediction of phosphorylation sites is critical to both uncovering the immune response pathway and understanding where in the pathway pathogens may be circumventing the immune response. Although most of the existing base classifiers perform well with ten-fold cross-validation on training sets, they tend in practice to produce many false positives. Thus, one goal of fusion is to increase the accuracy of phosphorylation prediction for proteins outside the training set.

Abstract 16

Solving a Mixed-Integer Programming Formulation of a Multi-Category Constrained Discrimination Model

J. Paul Brooks
Virginia Commonwealth University

Eva K. Lee
Georgia Institute of Technology

In classification, even if a Bayes-optimal rule has been developed, intragroup misclassification rates may be higher than desirable. We consider a two-stage model for multi-category constrained discrimination in which limits on misclassification rates of training observations may be pre-specified. The mechanism by which the misclassification limits are satisfied is a rejection option, also known as a reserved judgment group, for observations not demonstrating properties of membership to any of the groups.

The first stage of the model involves estimating conditional group density function values for training observations, and the second stage requires the solution of a mixed-integer program (MIP). The MIP is used to estimate the parameters that characterize an optimal classification rule.

The MIP is difficult to solve due to the formulation of constraints wherein certain variables are equal to the maximum of linear functions. Solution methods for the MIP are presented, including techniques for generating and exploiting edges of the conflict graph. Improvement in computation times over industry-standard software is demonstrated. Classification performance on real-world data is presented, including a demonstration of the trade-off between misclassification limits and rejection rates.

Abstract 17

A Method for Reconciling Values of Parameters

Shinya Kikuchi

Virginia Polytechnic Institute and State University

Manoj K. Jha

Morgan State University

Adjusting a set of values in order to conform to a set of requirements or relations usually involve subjective judgment and reconciliation. Consider the case of budget preparation, in which the total available fund is set. Most often, the sum of the requested amounts from different concerned parties exceeds the total available budget. Adjusting the individual requests requires negotiation and political maneuvering, and often the “losers” and the “winners” emerge. This paper presents a method that allocates the resource among the concerned parties in a rational and agreeable manner reflecting the degree of desires of the individual parties. A fuzzy set is used to represent the notion of desire, and the fuzzy optimization approach is used to find the most “acceptable” set of values. Mathematically, the approach reduces to a linear programming if the requirements are in the linear form. This process can be made iterative so that the elasticity of individual “desire” can be re-negotiated; in other words, change of mind is accommodated. The proposed approach can be used for various adjustment situations, both subjective resource allocation (money, space, manpower, and time) and also adjustment of physical quantities, e.g. measurement of distances, and weights, to achieve consistency.

Abstract 18

Improving the Estimation of Random Coefficient Logit Models of Demand

Marietta J. Tretter

Texas A&M University

Ordinary Logistic Regression is well used in Data Mining when the target variable is binary. However it is not possible to use such existing Data mining tools when one is interested in modeling demand over time from very large data bases with many different products, retailers, and prices. To mine such data it is necessary to fit random coefficient logit models of demand. Fitting these models is complex and often not very precise. Model fitting involves optimization and simulation. This research will present an approach using inexact arithmetic, also known as interval arithmetic, to make the estimation process more efficient and to produce interval error bound on the estimates. Applications of the technique will also be presented.

Abstract 19

Efficient Computer Experiment Based Optimization through Variable Selection

Dachuan T. Shih, Venkata L. Pilla, Seoung Bum Kim, Jay M. Rosenberger, Victoria C. P. Chen
University of Texas at Arlington

Variable selection has been widely used in regression data mining not only to select informative variables, but also to simplify the statistical model. A computer experiment based optimization approach employs design of experiments and statistical modeling to represent a complex objective function that can only be evaluated pointwise by solving an optimization subproblem. In large-scale applications, the number of variables is huge, and direct use of computer experiments would require an exceedingly large experimental design and, consequently, significant computational effort. Typically, a large portion of the variables have little impact on the objective; thus, there is a need to eliminate these before performing the complete set of optimization subproblem computer experiments.

Ideally, variable selection would be conducted after a small number of computer experiment runs, likely fewer runs (n) than the number of variables (p). Conventional variable selection techniques cannot be applied in this “large p and small n ” problem. We explore the use of regression trees and a multiple testing procedure based on false discovery rate. Performance of the selected variables is measured using the coefficient of determination (R^2). Two real world applications are studied, an air quality stochastic dynamic program and an airline fleet assignment problem.

Abstract 20

Modern Machine Learning for Automatic Optimization Algorithm Selection

Patricia D. Hough and Pamela J. Williams
Sandia National Laboratories

Optimization software is commonly used to solve simulation-based problems such as optimal design and control, model parameter estimation, best/worst-case scenario identification, etc. While the value of such software is widely recognized, user feedback indicates that these tools are difficult for non-experts to use. In particular, users experience difficulties in choosing an appropriate algorithm for the problem at hand. While the user has a great deal of intuition regarding the underlying problem, he/she usually must consult an optimization expert to determine the best algorithm or make a guess based on documentation written in the language of optimization developers for other optimization developers. In this talk, we will present a modern machine learning approach based on ensemble classifiers for automatically selecting an optimization algorithm to solve a given problem. We will discuss the feature sets used to characterize the optimization problems and algorithms. We will also review the metrics used to evaluate the performance of optimization algorithms and the challenges of managing competing metrics. Finally, we will present the results from our initial studies with the CUTer test set of optimization problems and discuss the generalization of the methodology to simulation-based optimization problems.

Abstract 21

Time-Based Detection of Changes to Multivariate Patterns

Jing Hu and George C. Runger
Arizona State University

Detection of changes to multivariate patterns is an important topic in a number of different domains. Modern data sets often include categorical and numerical data and potentially complex in-control regions. Given a flexible, robust decision rule for this environment that signals based on an individual observation vector, an important issue is how to extend the rule to incorporate time-based information. A decision rule can be learned to detect shifts through artificial data that transforms the problem to one of supervised learning. Then class probabilities ratios are derived from a relationship to likelihood ratios to form the basis for time-weighted updates of the monitoring scheme.

Abstract 22

Knowledge Discovery to Support Product Family Design

Seung Ki Moon, Timothy W. Simpson, and Soundar R. T. Kumara
Pennsylvania State University

Sharing and reusing product design knowledge can help reduce cost and time when developing new products and facilitate product family design. Knowledge associated with product design can be represented by combining constraints, functions, rules, and facts. An appropriate representation scheme for products and their design is important to share and reuse the knowledge effectively. The objective in this research is to develop a methodology for knowledge discovery related to product design using an ontology and data mining techniques. An ontology consists of a set of concepts or terms and their relationships that describe some area of knowledge or build a representation of it. An ontology can also be used to build a taxonomy representing products in a repository such as design depositories. Data mining can be used in the process of extracting valid, previously unknown, and easily interpretable information from large databases. Fuzzy clustering is employed to determine initial clusters based on the similarity among the functional features of the products. Based on the results of clustering, knowledge related to product family design is identified through association rules. We apply the proposed methodology to develop design knowledge for a family of power tools.

Abstract 23

Improving Productivity in Manufacturing Environments Using Data Mining

Pamela N. Ajoku and Bart Nnaji
University of Pittsburgh

Current production and manufacturing environments are drowning in data and starving for information. As a result, data mining is an area of interest to many manufacturing companies. Tons of data can be collected with relative ease, but the core issue is how to obtain timely critical information for decision-making and eventual organizational profitability. To compound the problem, today's customer can be described as both finicky and sophisticated. Customer demands are continually changing and the plant floor, associated supply chains and even the entire enterprise must keep up or close shop. In this paper, enhanced techniques will be used to explore solutions to data analysis problems in manufacturing environments. Integrating a pragmatic data mining framework within the manufacturing information infrastructure will provide access to minimally sufficient critical information and improved manufacturing productivity.

Abstract 24

Discovering Service Inventory Demand Patterns from Archetypal Demand Training Data

Eugene A. Beardslee and Theodore B. Trafalis
University of Oklahoma

A critical element of establishing inventory control parameters and consumption forecasts is determining the nature of demand generated by the processes supported by the inventory. Often these processes are stochastic and generate demand for inventory items that cannot easily be characterized using stationary probability distributions. The research presented here describes a method of classifying demand patterns using several data mining algorithms including support vector machine, C4.5 decision tree, Bayesian networks and Naïve Bayes classification methods. Using simulation to model primary demand source processes, archetypal demand time series were generated for use as training data input for the data mining analysis. Once the data mining models were trained, actual inventory transaction sets were classified into four types of demand: periodic, seasonal, level shift and sparse. These classified demand series, were then validated against a set of hand classified time series and through the application of the Box-Jenkins method for characterizing time series information. Results show that, using this method, managers of service part inventories could identify demand patterns with sufficient reliability, and flexibility to improve inventory management.