Chapter 1

Challenges in Scientific Data Mining: Heterogeneous, Biased, and Large Samples

Zoran Obradovic and Slobodan Vucetic
Center for Information Science and Technology
Temple University, 303 Wachman Hall (038-24)
1805 N. Broad St., Philadelphia, PA 19122, USA

Abstract:
With recent innovations in data collection technologies in many emerging scientific applications there is a need for a paradigm shift from a traditional hypothesize-and-test process to a partial automation of hypothesis generation, model construction and experimentation. To develop appropriate knowledge discovery tools in support of the paradigm shift in sciences numerous data mining challenges should be solved. The next generation data mining challenges considered in this article include (1) mining of heterogeneous data; (2) learning from biased labeled samples; (3) enhancing predictive modeling by exploiting large unlabeled databases; and (4) reducing large datasets with a controllable information loss for a more efficient mining in a distributed environment. Several effective methods for addressing these problems developed at our laboratory are described. Applications are illustrated on six scientific and engineering problems in precision agriculture, electric power systems and biochemistry studied in our collaboration with domain experts. It is demonstrated that some fairly general data mining algorithms with appropriate modifications can be adopted for a more successful mining of spatial observations, nonstationary time series and heterogeneous sequence databases in specific scientific applications.

Keywords: Temporal data mining, Spatial data mining, Data partitioning, Biased sampling, Outlier detection, Partially suppervised learning, Data reduction.

1.1 Introduction

Observing a phenomenon followed by creating and testing a hypothesis and developing and refining a model is a scientific paradigm shared across disciplines [49]. Temporal studies where observations of past outcomes of a phenomenon are used to anticipate its future behavior are very common in science [18]. Furthermore, sequence studies are often performed in biosciences where the objective is to explain certain property characterizing a given sequence or its specific position by observing and analyzing similar sequences or neighboring positions in a given sequence [70]. In a large number of scientific disciplines including geosciences, astronomy, and biology, spatial studies are carried out where observations at certain locations in space are used to understand the process at the same or other locations at different points in time [74]. In spatial-temporal scientific studies temporal and spatial aspects are considered together with an objective of understanding certain spatio-temporal patterns [60].

Data analysis techniques supporting traditional scientific processes were aimed at handling a fairly small amount of low dimensional data through a hypothesize-and-test paradigm. These extremely labor intensive techniques are becoming infeasible for analysis of enormous scientific datasets obtained at much higher speed and lower cost using improved or novel data collection technologies. In recent years astronomers, geoscientists, biochemists, high energy physicists and other scientists collect huge and high dimension datasets using advanced telescope technologies [9], multi-spectral remote sensors on satellites [53], integrating global positioning systems with high resolution sensors on ground [75], developing massively parallel instruments like microarrays that generate gene expressions for entire organisms at once [8], and employing other advanced technologies [28].

For example, in Earth sciences, in addition to a network of geostationary and polar orbiting weather and meteorological satellites, novel series of satellites have been recently introduced that provide steady data stream from multiple sensors to accomplish deeper understanding of climate and environmental changes [3, 19]. In particular, NASA Earth Observation System [53] consisting of several low-altitude satellites is the first observing system to offer integrated measurements of the Earth's processes. It supports a coordinated series of polar- orbiting and low-inclination satellites for long-term global observations of the land surface, biosphere, solid Earth, atmosphere, and oceans. Its Landsat 7 instrument has a data rate of 150Mbps [51], while Terra instrument produces data in order of 1 Terabyte daily [53].

Another source of large datasets in science is the result of using fast computational facilities in simulations of astrophysics, fluid dynamics, structural mechanics, chemical engineering, climate modeling and other fields. For example, the Reanalysis Project, jointly pursued by the National Center for Environmental Prediction and the National Center for Atmospheric Research, has a goal to produce new atmospheric analyses using historical data as well as to produce analyses of the current atmospheric state [37]. This effort results with 55GB/year of processed data, containing several temporal climate and weather attributes at a regular 3D spatial grid for 50+ years of atmospheric fields. The data has been employed in various domains including climatology, forestry and environmental sciences [56] as well to create snapshot annual CD- ROMs containing digests of raw reanalysis data.

In principle, data mining provides an opportunity for a shift from a traditional hypothesize-and-test scientific paradigm to a partial automation of the entire process including hypothesis generation, model construction and experimentation [49]. Some data mining techniques developed initially for automation of a business decision making process were quickly adopted to large scale scientific domains. For example, a general-purpose text mining technique of searching for co-occurrences of predefined terms was successfully applied to problems from bioscience for extracting known protein-protein interactions or gene names from Medline abstracts [35,47] and for finding similarities among abstracts [29, 67].

Also, innovative data mining methods were developed to address certain aspects of specific scientific problems that are distinct from typical business applications and were employed at other scientific domains [38]. At the 1996 report of the Workshop on Scientific Data Management, Mining, Analysis and Assimilation it was emphasized that regardless of a particular domain, scientific data sets share a lot of common properties and need a unified approach to efficiently solve numerous common problems involving data storage, organization, access, and knowledge discovery [24]. Terrabyte scale problems were proposed for evaluating scientific data mining technologies at this workshop. For example, one of the reported applications was aimed at analysis of 3 terabytes of radio astronomy data for determining size and distribution of objects. To analyze this data in five minutes would require a data handling system with an access rate of 10 gigabytes per second to the stored data. In more recent scientific data mining workshops [10, 11, 24, 25, 43] and elsewhere [40, 41, 52] testbed problems were increased to Petabytes of data [23, 69] distributed among multiple locations [42, 44].

In spite of significant advances in data mining and in related fields of remote sensing, databases, machine learning, temporal, spatial and spatial-temporal statistics [62], there is an urgent need for additional scientific data mining activities in order to make a genuine paradigm shift in science a reality [28]. Challenges that need more attention are numerous. Some of important scientific data mining issues that will not be considered at this article include:

- learning with prior knowledge [1, 55];
- incremental learning [17];
- handling short observation history [60];
- integrating information from many sources [30, 42, 44];
- performing effective data registration to relate information from various subjects [45].

Our discussion will focus on the following scientific data mining challenges:

- understaning heterogeneous phenomena with local relationships in observed attributes;
- handling spatial, temporal and sequence correlation in data;
- learning from small and biased labeled samples;

- exploiting large unlabeled datasets;
- reducing data with controllable information loss for efficient analysis.

Some of our techniques proposed for handling these problems will be described and their application will be illustrated on several scientific data mining problems currently studied at our laboratory.

1.2 Mining of Heterogeneous Data

1.2.1 Heterogeneous spatial data and nonstationary time series

Real-life spatial data are often heterogeneous, such that at different spatial locations the dependencies between observed attributes and the response are different. Heterogeneity could be caused by presence of unobserved or very noisy spatial attributes. If it is possible to partition heterogeneous spatial data into spatial regions with homogeneous distributions, specialized models could be learned on each homogeneous region separately. This would result in better overall prediction accuracy as compared to constructing a single model on all data.

Similarly, real-life time series are often nonstationary with data distribution changing over time. Piece-wise stationary time series consisting of time intervals that could be considered stationary are common in practice. Union of all intervals of a piece-wise stationary time series with approximately the same stationary distribution is called a regime. In general, the number of regimes and their switching dynamics is not known in advance. Fitting a global regression model to such data would fail to capture the underlying relationships and would result in learning an averaged behavior of its regimes.

Let us denote each data point as a pair $(\mathbf{x}_s, y_s)$, where s represents its spatial or time location, $\mathbf{x}_s$ is a vector of attributes, and y_s is a real-valued response variable. Let us also denote each homogeneous region/regime as S_j, $j = 1, \ldots, J$, where J is a number of homogeneous regions/regimes, and $s \in S_j$ means that data point s belongs to region/regime S_j. The relationship between attributes and response in region/regime S_j can be expressed as $y_s = f_j(\mathbf{x}_s) + \epsilon_s$, where f_j is a regression function, and ϵ_s is the error term representing the unexplained variance. The task of mining of heterogeneous spatial and time data can now be stated as:

1. discovering homogeneous regions/regimes S_j, their number J, and their location in space/time, and
2. analyzing each of the regions/regimes separately by estimating regression function f_j on each of them.

Both space and time domains are characterized by a significant degree of correlation where data points close in space/time are more likely to have similar properties and, therefore, to belong to the same homogeneous region/regime. This fact can be used to improve the mining of heterogeneous data. In our previous work [72, 75] we proposed an algorithm for mining of heterogeneous spatial and time data that:

(i) determines whether the data set is heterogeneous;

(ii) discovers an appropriate partitioning into more homogeneous subsets; and

(iii) constructs specialized regression model on each identified homogeneous subset.

This algorithm is based on partitioning of heterogeneous data through a competition of regression models for data points using information about spatial/temporal correlation to stabilize the partitioning process.

1.2.2 Partitioning Algorithm

The proposed partitioning algorithm relies on the competition among regression models for each part of a given data set S_0. The algorithm starts by learning a global regression model on the complete data set. Next, the dataset is split randomly into two equal sized disjoint subsets and separate regression models are trained on each of them. The two models then compete for each data point such that a given point is assigned to the model achieving higher accuracy. This competition procedure iterates until a stable partitioning of data into two regions/regimes is obtained.

Subsequently, the accuracy obtained by such a partition is compared to the accuracy of the global model. If the improvement is not sufficiently large, the algorithm is stopped and it is concluded that the data set is homogeneous. If using two regression models is better than one global model, the algorithm proceeds by partitioning data into three regions/regimes in an attempt to further increase the prediction accuracy. The procedure continues by adding new models into the competition until it is concluded that further partitioning does not improve prediction accuracy. Partitioning algorithm for heterogeneous spatial and temporal data is described next.

Partitioning Algorithm:

- Learn a single model on complete data set S_0 and calculate its accuracy.
- Split S_0 into disjoint equal size sets $S^0_{1,2}$ and $S^0_{2,2}$. Denote $S^0_2 = \{S^0_{1,2}, S^0_{2,2}\}$.
- Modify S^0_2 through **the competition procedure** and calculate the overall accuracy achieved by the obtained partitioning $S_2 = \{S_{1,2}, S_{2,2}\}$.
- Terminate the algorithm if the accuracy is not improved - data is homogeneous.
- L=2.

  repeat

  for $i = 1$ to L

  – Split $S_{i,L}$ into two disjoint subsets of the same sizes, $S^{'}_{i,L}$ and $S^{''}_{i,L}$, to obtain the initial partitioning $S^0_{L+1,i} = \{S_L/S_{i,L}, S^{'}_{i,L}, S^{''}_{i,L}\}$.

  – Modify $S^0_{L+1,i}$ through the competition procedure and calculate the accuracy achieved by the obtained partitioning $S_{L+1,i}$.

  end

  – Out of L partitions, $S_{L+1,i}, i = 1, \ldots, L$, choose the partitioning achieving the best accuracy as the new partitioning, $S_{L+1} = \{S_{1,L+1}, \ldots, S_{L+1,L+1}\}$.

  – $L \leftarrow L + 1$

  until there is improvement in accuracy

- Output: $L-1$ homogeneous subsets defined by partitioning S_{L-1}.

The competition procedure used in partitioning consists of the following steps.

Competition Procedure:

- Start from a partitioning into L subsets S_L^0 and set $n=0$.
  repeat
  - Learn L models using L subsets of S_L^n, $S_i, i=1,\ldots,L$.
  - Filter errors of all models over space/time.
  - If model M_j gives the best accuracy on data point s assign it to subset S_j.
  - Form a new partitioning, S_L^{n+1}, from reassignments.
  - $n \leftarrow n+1$

  until convergence to a stable solution

One of the most important mechanisms for efficient partitioning is competition of regression models with assigning a data point to the model achieving the highest accuracy. If data is characterized by a significant level of noise, the probability of incorrectly assigning a data point could become significant. The convergence of the competition and accuracy of partition can be aided by an error-filtering step that tends to assign data points close in space/time to the same homogeneous region. Following the construction of predictors in the competition step absolute prediction errors are filtered using a corresponding spatial/temporal filter. The coverage of spatial/temporal filter should correspond to the range over which spatial/temporal correlation of absolute prediction errors exists, and it could be easily determined experimentally.

In [75, 77] we showed that, when two homogeneous regions/regimes characterize the data heterogeneity, the partitioning algorithm converges and the quality of partitioning increases with the appropriate use of error filtering.

1.2.3 Application 1: Mining of heterogeneous spatial data

We performed an extensive study on synthetic spatial data to characterize the partitioning algorithm [75]. The results confirmed that convergence is very fast and that in the presence of unexplained variance in spatial data, incorporating error filtering to the competition procedure can significantly improve the partitioning accuracy.

We also performed an experiment on a real-life scientific database containing a grid of 8 topographic attributes and winter wheat yield from a 220 ha field. The goal was to predict wheat yield as a function of terrain attributes derived from USGS 30 m DEM data mapped to a 10 x 10 m grid, where the majority of the terrain attributes were continuous and non-Gaussian. The crop yield values were collected with a combine mounted yield monitor and a global positioning system. There were 24,592 data points in the entire data set.

From the previous results [78], it was known that topographical attributes alone could explain just a small part of crop yield variance. A global neural network with 5 hidden nodes achieved R^2 of only 0.11 on test data. The partitioning algorithm was applied in order to examine if the field was heterogeneous such that possible homogeneous regions could be treated differently. Using the advances in precision agriculture, this would allow farmers to apply different practices on each of the homogeneous regions to obtain larger profits as compared to treating the whole field uniformly.

From the correlogram of prediction errors of a global neural network achieving the range at about 150 meters, the error filter size was set to 150 meters. The partitioning algorithm discovered 3 homogeneous regions such that the achieved accuracy was improved to $R^2 = 0.31$. A brief analysis of discovered regions showed that the largest homogeneous region corresponded to the high-yielding low-slope part of the field, while the smaller two regions corresponded to the low-yielding higher-slope parts oriented toward south and north, respectively. This illustrated the benefits of careful mining of heterogeneous spatial scientific databases.

1.2.4 Application 2: Mining of nonstationary time series

We applied the partitioning algorithm to characterize recent price behavior in the California electricity market [72]. Publicly available data on day-ahead forecasted load l_t was known before trading and market clearing prices p_t of day-ahead market in the period from April 1, 1998 until September 30, 1999, were used in the experiments (first two plots from Figure 1). A total of 12,952 data points were constructed from hourly data covering this period. Since the relationship between price and generated power for thermal power generators is usually expressed as quadratic function, we have decided to model the relationship between $\{p_t\}$ and $\{l_t\}$ as $p_t = y(l_t, \beta) = \beta_0 + \beta_1 l_t + \beta_2 l_t^2 + \beta_3 l_t^3$, and used this functional form to construct competing price predictors in the partitioning algorithm. The length of error filter was set to 168 (representing a period of two weeks).

The partitioning algorithm discovered 4 specific regimes in the data (third plot from Figure 1). The mean squared error of 4 specialized price predictors was 61.6 representing a significant improvement over 93.9 achieved with a single price predictor. We observed that the 4 regimes differ significantly in price-load relationships (different price for the same load) as well as in price volatility. Regimes with relatively high prices indicated periods of exercising the market power with large suppliers possibly withholding a portion of generation and causing a price increase. The regime appearing during the nine weeks of May and June of 1998 was characterized by extremely low price and low volatility apparently caused by high hydro generation and relatively low load. Extremely high prices and high price volatility with a relatively low load characterize regime which appeared during the five weeks of October and November in 1998, one week at the end of December of 1998, and the last week of September of 1999. We speculated that this was caused by scheduled maintenance and lower hydro availability but a definitive answer depends on data not publicly available. As a conclusion, the partitioning algorithm proved as a potentially very valuable tool in analyzing price behavior and inefficiencies in electricity markets.

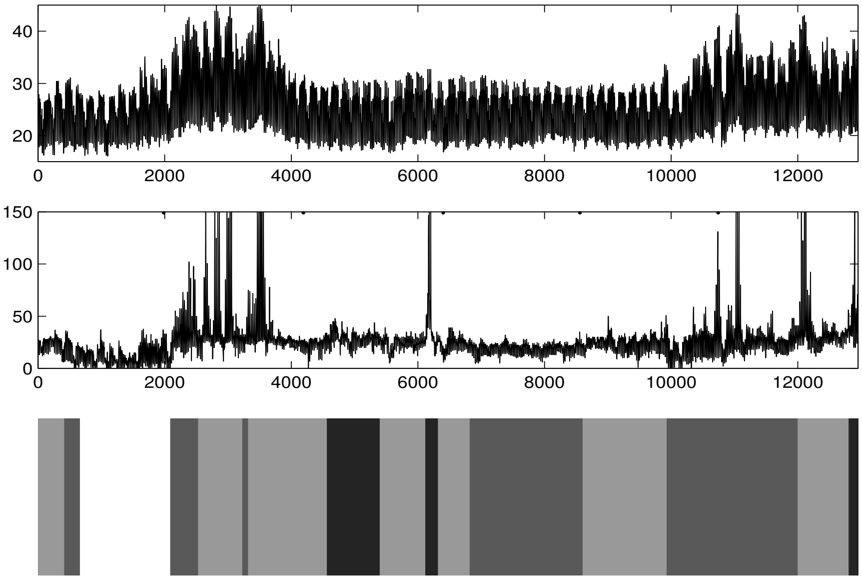

Figure 1.1: Plot of hourly load, hourly MCP between April 1, 1998 and September 30, 1999 as well as the four regimes discovered by the partitioning algorithm [72]. Permission for including this figure will be requested from IEEE Press.

1.2.5 Application 3: Mining of heterogeneous sequence data

We also modified our partitioning algorithm for mining of heterogeneous sequence data [70]. Our interest in this problem originated from the study of protein disorder property. Briefly, protein disorder is a biological concept that refers to regions in protein sequence that do not have unique 3D structure under physiological conditions [20]. Our previous studies suggested that predictors of protein disorder trained on disorder from one type of protein achieve poor accuracy on other disordered proteins, indicating the existence of different types, or flavors, of disorder. Therefore, our objective was to determine if disordered proteins could be partitioned in a systematic manner into a number of groups with distinct statistical and biological properties.

To a certain extent, the partitioning problem could be formulated as clustering. However, with a large number of relevant attributes and strong influence of unobserved effects, applying of-the-shelf clustering software results in clusters with low biological significance. So, we modified the described partitioning algorithm such that S_0 corresponds to the set of all disordered proteins and each model is learned to classify between a given subset of S_0 versus examples of protein sequences that are ordered.

The proposed algorithm [70] starts by fitting a global classifier based on all available disordered proteins and an equal number of examples of protein order. Next, disordered proteins are divided randomly into two disjoint subsets of equal size and separate classifiers are trained using each of them. The two classifiers then compete for each disordered protein such that a given protein is assigned to the predictor achieving the highest accuracy. This step is related with the error filtering since each protein is represented as a set of data points whose number equals protein length and each protein is assigned exclusively to a single flavor. The competition iterates until a stable partitioning is obtained. Subsequently, the overall accuracy obtained by such a partitioning is compared to the accuracy of the global classifier. The algorithm proceeds by gradually partitioning disordered proteins into more subsets in an attempt to further improve the accuracy.

Data sets with 145 nonredundant disordered protein regions and about equal number of ordered proteins of similar length were used in the experiment. For each position in a protein sequence we extracted 18 amino acid frequencies, flexibility, and sequence complexity calculated over a sliding window of length 41. By employing the proposed partitioning algorithm we concluded that 3 disorder flavors exist where accuracy of the global classifier of 71% was improved to 84% when 3 specialized classifiers were used. This difference in the accuracy of 3 flavor-specific predictors vs. a global predictor was significant enough to validate the claim that disorder flavors indeed exist. The 3 flavors, denoted as V, C and S, contained 52, 39 and 54 disordered proteins, respectively. A comparison of amino acid compositions revealed that the flavors have statistically significant differences where flavor C was very different from both the other two flavors and ordered proteins, while flavors V and S were more similar.

Experiments on 28 complete genomes revealed that 5 of the 6 archaea genomes were biased towards Flavor V and 12 of 18 eubacteria genomes were biased towards Flavor S (4 eubacteria have affinity towards Flavor C and 2 toward Flavor V), while all 4 eukaryote genomes were biased towards Flavor S. We also used flavor-specific predictions on SwissProt data bank to assess flavor-function relationships. The flavor V predictor gave the strongest prediction of disorder for helical regions. We have shown previously that disordered domains that become ordered helices upon binding to other proteins are compositionally distinct from other types of disorder. The flavor C predictor gave strong predictions of disorder for poly-and oligosaccharide binding domains. The flavor S predictor strongly predicted leucine rich regions such as leucine zippers to be disordered.

The partitioning was based exclusively on sequence statistics, whereas, post hoc, we found relationships between discovered flavors and function. This gives a strong indication that the proposed modification of the partitioning algorithm was very successful in discovering a biologically valid partition of disordered proteins and thus giving a novel insight into properties of protein disorder.

1.3 Learning from Biased Data

1.3.1 Biased labeled data

A common assumption made in machine learning is that labeled data used for training a classifier and unlabeled new data can be considered as samples from the same underlying distribution. In such a case one could apply a standard machine learning procedure to learn a classifier from labeled data (e.g., logistic regression, decision trees, neural networks), estimate its accuracy (e.g., directly from training set, using crossvalidation), and apply it on unlabeled examples in a straightforward manner. However, this assumption is often violated with labeled and/or unlabeled data obtained by biased sampling from an underlying distribution. If class distribution on labeled data is different from that of unlabeled data, a classifier trained on labeled data can cause incorrect inference and produce suboptimal classification on unlabeled data.

Large quantities of unlabeled data are often available in data-rich scientific domains such as remote sensing or bioscience. Predictive data mining, with a few exceptions, typically relies on labeled data without exploiting a much larger source of available unlabeled data. In this section we illustrate two approaches that exploit unlabeled data for more effective mining on biased data.

1.3.2 Reestimating the class distribution

The problem of learning from data with biased class distribution can be defined as the construction of a classifier using labeled data S_L with class probabilities $P_L[y = j]$, $j = 1, \ldots, c$, where c is the number of classes, for accurate prediction on unlabeled data S_U with unknown class probabilities $P_U[y = j]$. This problem can be addressed by (*step 1*) estimating the class distribution $P_U[y = j]$, followed by (*step 2*) using this estimate to construct a desired classifier. We proposed a bootstrap methodology to estimate distribution of $P_U[y = j]$ based on (*step 1.a*) an estimate of a classifier's accuracy obtained on labeled data, and (*step 1.b*) an estimate of a classifier's class predictions on unlabeled data [73].

Given a predictor, by $p_j = P_U[y = j]$ we denote the class probability on S_U, by $p(k|j) = P[\hat{y} = k|y = j]$ the probability of predicting class k if the true class is j, and by $q_k = P_U[\hat{y} = k]$ the probability of predicting class k on S_U. Using the law of total probability, q_k can be calculated as $q_k = \sum p(k|j) \times p_j$, or in the matrix form as $\mathbf{q} = \mathbf{P} \times \mathbf{p}$, where $\mathbf{q} = \{q_k\}$, $\mathbf{P} = \{p(k|j)\}$, $\mathbf{p} = \{p_j\}$. Therefore, the true class probability p_j can be calculated as $\mathbf{p} = \mathbf{P^{-1}} \times \mathbf{q}$ in the matrix form, under the assumption that values of q_j and $p(k|j)$ are known with certainty. This can occur only if S_L and S_U are very large. However, the size of available data sets is always limited and so q_k and $p(k|j)$ can be considered as random variables whose properties should be estimated first.

For statistical inference suitable for estimating the distribution of p_j from estimates of q_k and $p(k|j)$, we adopted a powerful bootstrap simulation methodology [21]. Given an original sample X with n examples the bootstrap sample X^* is obtained by randomly sampling n examples from X with replacement. Briefly, bootstrap samples S_L^* and S_U^* are repeatedly taken and used to estimate $p^*(k|j)$ from S_L^* and q_j^* from S_U^*. ¿From these two estimates p_j are be calculated. Bootstrap estimate of p_j can be obtained as an average p_j^* over B bootstrap rounds. It should be noted that that estimates $\mathbf{p}^* = \{p_i^*\}$ can be unfeasible if $\mathbf{P}^* = \{p^*(j|i)\}$ is noninvertible or if elements of $\mathbf{p}^* = (\mathbf{P}^*)^{-1}\mathbf{q}^*$ are negative. Such bootstrap estimates are rejected.

1.3.3 Application 4: Classification with biased class distribution

If a classifier could estimate posterior class probabilities $P(j|\mathbf{x})$ when presented with a new example $\mathbf{x}$, then it can be directly adjusted without the need for retraining according to estimated class probabilities p_j on S_U. For example, a neural network with a large hidden layer and c outputs (representing each of the c classes) trained by minimizing the mean squared error is known to be able to approximate posterior class probabilities. Adjusted posterior class probabilities $P_{adjust}(j|\mathbf{x})$ can be calculated as $P_{adjust}[j|x] = CP[j|x]P_U[y = j]/P_L[y = j]$, where $P_L[y = j]$ is class j probability on labeled data S_L and C is a normalization constant.

Retraining a predictor based on an estimate of class probability p_i on unlabeled data S_U is another choice for improving classification. In this case, labeled data S_L is resampled according to estimated p_j and a predictor is learned on the resampled data set. This simple procedure can be iterated several times by estimating p_j based on the improved predictor, resampling S_L, and learning a new predictor.

Our preliminary evaluation of the proposed bootstrap algorithm on synthetic data of various statistical properties showed that it accurately estimates the class probabilities on unlabeled data [73]. In another experiment on a 3-class benchmark dataset called Waveform [7] the S_L was constructed with balanced classes, while we experimented with different class distributions p_j on S_U. It was shown that for balanced classes $\mathbf{p} = [1/3, 1/3, 1/3]$ the original predictor had slightly higher accuracy then adjusted or retrained predictor for small data sets ($n_L = n_U = 500$), while all 3 strategies resulted in similar accuracy for larger data sets ($n_L = n_U > 1000$). For unbalanced classes, $\mathbf{p} \neq [1/3, 1/3, 1/3]$, our methodology with predictor adjusting or retraining resulted in significant accuracy improvement over original predictors. Moreover, retrained predictors were superior over adjusted predictors for large data. Bootstrap estimates of p_j were sufficiently accurate even for small data sets while this accuracy increased with the size of data sets.

1.3.4 Contrast classifiers for learning from biased data

Recently, we proposed use of contrast classifiers trained to discriminate between labeled and unlabeled examples for mining of biased data [59]. The name contrast classifier comes from the meaning of their outputs - they represent difference, or contrast, of a given example from the underlying distribution of labeled data. Given this explanation, it is apparent that contrast classifiers could be used in a number of important data mining applications including outlier detection and learning on biased data.

Given an unlabeled dataset representative of the underlying distribution and a K-class labeled sample that might be biased, our approach is to learn K contrast classifiers $cc_k(\mathbf{x})$, $k = 1, \ldots, K$, each trained to discriminate a certain class of labeled data from the unlabeled population. If an employed contrast classifier learned on a balanced set of labeled and unlabeled data points is able to approximate posterior class probabilities it can be shown that contrast classifiers can be used to construct the maximum a posteriori classifier for the K-classification problem with the optimal decision obtained as

$$\hat{c} = \arg\max_j \frac{1 - cc_j(\mathbf{x})}{cc_j(\mathbf{x})} \times p_j,$$

where p_j is prior probability for unlabeled data with class j.

Moreover, for a test data point $\mathbf{x}$ the ratio $(1 - cc_j(\mathbf{x}))/cc_j(\mathbf{x})$ measures difference from class j of labeled data, thus providing a mechanism for detection of unlabeled points under-represented in labeled data. Prediction on such uncharacteristic points can be expected to be less accurate since it corresponds to the extrapolation to less explored parts of the attribute space. Therefore, contrast classifiers are characterized by the ability to provide optimal classification on the explored portion of the attribute space, while detecting data points from the less-explored regions. This allows achieving an appropriate tradeoff between classification accuracy and prediction coverage.

Experiments on the Waveform data confirmed that contrast classifiers achieve accuracy comparable to standard classifiers in the case of unbiased labeled data, while they are significantly more efficient when the labeled data is biased. This approach is also shown to be more accurate than alternatives such as kernel density estimation that rely solely on a distribution of a positive class. Contrast classifiers are very robust to a small labeled data size and a high dimensional space with many irrelevant attributes.

1.3.5 Application 5: Using contrast classifiers in bioinformatics

In a recent work [59] we applied contrast classifiers to identify proteins in large sequence repositories that are statistically underrepresented in our labeled database of known ordered and disordered proteins. Understanding biological properties of these outliers could lead to more effective studies of protein disorder.

We were given a non-redundant labeled data (sequence identity less than 25%) consisting of 152 proteins with disordered regions longer than 30 consecutive residues, and 290 completely ordered proteins. In addition, we had access to 17,676 non-redundant unlabeled protein sequences derived from the SwissProt database [6] by grouping homologues (similar sequences) to the ProtoMap clusters [81].

Two class-specific contrast classifiers were constructed, one for contrasting disordered class from unlabeled data and another for contrasting order class from unlabeled data. A disordered predictor constructed from two contrast classifiers achieved accuracy of 84% that is practically identical to our currently most accurate disorder predictor [57] thus confirming the effectiveness of contrast classifiers.

The two contrast classifiers were used for selection and evaluation of underrepresented SwissProt proteins. From the unlabeled dataset we first selected a set we call $SWISS$ consisting of 6,964 sequences similar in lenght to our labeled sequences (200 to 500 amino-acids long). To detect proteins that are overall the most different from ordered and disordered proteins, we summarized each protein in $SWISS$ according to the average predictions of the two contrast classifiers over its sequence. Sequences with the largest average contrast with respect to both classes of labeled data (ordered and disordered proteins) were denoted as $OUTLIERS$ consisting of 1,259 sequences. For comparison we also derived from the SwissProt database 539 homologues to ordered proteins from the training data and 356 homologues of confirmed disordered proteins. For each of the 4 identified data sets we calculated the frequency of each of the 840 protein function keywords listed in SwissProt database. By comparing these frequencies we were able to observe significant differences between biological functions of proteins in the 4 grous.

The most interesting result was that frequency of Glycoprotein, Inner Membrane, Membrane, Transmembrane and Transport proteins in $OUTLIERS$ was extremely high as compared to the other 3 data sets. They are all members of a large family of membrane proteins known to be structurally and functionally diverse and are highly underrepresented among our labeled order and disorder sequences. This is also consistent with the results of our control study where outliers were identified by learning class-conditional distribution models of ordered and disordered protein sequences [71]. In fact, we found that the contrast for membrane proteins was so high that it is likely that most of the bias in our labeled data comes from this group. This interesting result is very important for the more efficient study of protein disorder and it also illustrates the promise of using contrast classifiers in scientific data mining on biased data.

1.4 Performance-Controlled Data Reduction

Much of machine learning research has concentrated on induction and model validation from limited data sets. However, in a number of scientific domains the data sets are extremely large, distributed over a number of locations, and dynamically increasing with time. While large data sets are desirable regarding the resolution of potentially extractable knowledge, they can easily cause computational and storage problems. Most approaches that can accomplish efficient transfer and storage include lossless attribute compression [48] or lossy attribute compression using vector quantization [64].

An important knowledge discovery problem is to establish a reasonable bound on the size of data sets needed for their accurate and efficient analysis. For example, for many applications a 10-fold increment of the data set size for a potential accuracy gain of 1% cannot be justified due to huge additional computational costs. Reducing large data sets into more compact representative subsets while retaining essentially the same extractable knowledge could speed up learning and reduce storage requirements.

1.4.1 Downsampling and quantization methods

We proposed an effective data reduction procedure [76] that, for a user-specified allowed accuracy loss, (1) applies down-sampling to identify a minimal representative random sample size, and follows it by (2) determining an appropriate quantization for each continuous attribute and performing compression of attributes and the target.

Assume a data set is given with N examples, each represented by a set of K attributes and the corresponding target variable. The task is to obtain a reduced data set with a minimal size that, using the same learning algorithm, allows knowledge extraction with accuracy reduced for at most α [%]. To achieve this goal we proposed the following two-phase method: (1) reduce the sample size from N to n_{min} while allowing accuracy loss α_D to achieve a data compression ratio $C_D = N/n_{min}$, and (2) perform proper quantization of continuous attributes in such down-sampled data set allowing accuracy loss α_Q, followed by Huffman coding [33] of discretized attributes to achieve compression ratio C_Q. Assuming that total accuracy loss α is close to zero, it follows that $\alpha \approx \alpha_D + \alpha_C$ with the achieved total compression ratio $C = C_D \times C_Q$.

A learning curve for least squares predictors depicts the dependence of the predictor accuracy measured as Mean Squared Error (MSE) on the training set size n. This curve consists of two regions: (1) the initial region characterized by fast drop of the Mean Squared Error (MSE) with the increasing sample size, and (2) the convergence region where addition of new samples into the training set does not significantly improve prediction accuracy. In our studies we used the idea of progressive sampling [61] to efficiently span the available range of sampling sizes and estimate the parameters of learning curve in search for the n_{min}.

Uniform quantization, where the range of a continuous attribute is split into equal subintervals (bins), was applied with subsequent Huffman coding of discretized values. Attributes on which the response has less sensitivity can be allowed higher distortion (the coarser quantization) and vice versa. To estimate the influence of attributes quantization on model prediction quality, we proposed the sensitivity analysis of a prediction model obtained on the down-sampled data set. This analysis allowed developing heuristics for proper quantization levels of all attributes in an efficient manner.

1.4.2 Application 6: Reduction of spatially correlated data

We performed a number of down-sampling and quantization experiments [76] on several large data sets including Covertype Data Set. This set is currently one of the largest databases in the UCI Database Repository [50] containing 581,012 examples with 54 attributes and 7 target classes and representing the forest cover type at 30 $\times$ 30 meter grid, obtained from US Forest Service Region 2 Resource Information System [4].

The data was transformed into 12 attributes, 10 of which were continuous. In its raw form the data set has 60MB and can be compressed 6 times by gzip software. We examined data reduction when neural networks are used as nonlinear least square estimators. Our experiments showed that the convergence region of the learning curve corresponded to sample sizes larger than 1,000, that asymptotic accuracy of a neural network is 70.1%, and that for allowed accuracy loss of $\alpha_C = 1\%$ the size of reduced dataset can be 3,600. To properly estimate the learning curve it was needed to train less than 20 neural networks on increasing sample sizes where the total number of data points used was around 30,000. Therefore, this learning effort was order of magnitude smaller than training a single neural network on the complete Covertype data. Quantization of the 10 continuous attributes with accuracy loss $\alpha_Q = 1\%$ allowed their representation with around 20 bits that compares favorably with 320 bits when standard double precision representation is used. To perform the quantization step around 20 neural networks were trained on the reduced data set with 3,600 data points. By properly combining the down-sampling and quantization steps we were able to reduce the size of Covertype data to 25kB, 12kB, and 5kB for user-specified accuracy loss of α = 1, 2, and 5%, respectively. Considering the original data size of 60 MB it is evident that reduction of more than 3 orders of magnitude are possible for a small acceptable prediction accuracy loss. This result indicates potentially huge benefits of the performance controlled data reduction in data-rich scientific data mining.

1.5 Discussion

Mining of heterogeneous data

Detection, understanding, and managing heterogeneity in large spatial/temporal scientific datasets with attributes that are noisy, collected with different resolutions, or even missing is a critical task of scientific data mining. Its success will depend on integration and extension of techniques from statistics [14, 27] and data mining [63]. The partitioning algorithm described in Section 2.1 and successfully applied to spatial and temporal domains as reported in Sections 2.2 and 2.3 resembles a large class of expectation-maximization (EM) algorithms [15], and, more particular, the mixture-of-experts approach [34, 79], annealed competition [58], and Markov switching models [26]. It is based on iterative improvements of the partitions and their corresponding specialized predictors where the overall goal is maximizing the accuracy of the specialized predictors.

The success of our algorithm could be attributed to error filtering that stabilizes partitioning by utilizing the spatial/temporal correlations existing in the data. This property is especially useful when unobserved or very noisy attributes are present, i.e. when attribute spaces of the corresponding homogeneous regions/regimes overlap. Additional advantage of our algorithm on scientific data mining domains is “hard“ assignment of a data point to a single region/regime. This improves the robustness of the partitioning algorithm to the deviations from the assumed distribution, such as existence of non-Gaussian error terms.

Sequence mining is an active research area with its most prominent application in bioinformatics. A large body of research in computational biology concentrates on effective algorithms for similarity search among biological sequences. The basic thrust behind this line of research is the fact that biological molecules with similar sequences tend to share structural and functional properties. Most available algorithms [2, 32, 68] are rather conservative: given a query sequence they select only a subset of sequences within a limited region of sequence space that guarantees their relatedness in a statistical sense. However, different biological properties tend to exist over a large number of families with unrelated sequences. Hence, important open problems of bioinformatics are inferring properties of sequences without well-studied homologues as well as increasing knowledge about certain insufficiently studied properties of the molecules. Thus, one would like to use more aggressive algorithms that generalize over a larger region of sequence space while handling an increased level of prediction error. The representatives of these algorithms are various predictors of structure [36] or function [39] from sequence data.

Partitioning of heterogeneous sequence data by our approach (Section 2.4) is based on sequence representation through a set of representative attributes. Examples from the negative class were used to provide partitioning based on the most relevant attributes, arguably, the ones that are helpful in classification between data points from the positive and the negative class. Based on this notion, we successfully modified our original partition algorithm to incorporate competition of specialized classifiers learned to discriminate between negative data and homogeneous subsets of positive data.

Learning from biased data

The biased labeled data problem discussed in Section 3 is common in science. It has been known in statistics as sample selection bias, and a partial solution for its correction even led to a Nobel prize in economy [31]. It is also a topic of active research in machine learning and data mining. The EM-based algorithms, used to iteratively estimate the model parameters and assign soft labels to unlabeled examples by treating the unknown labels as missing data and assuming generative model such as mixture of Gaussians, have been widely used in areas such as text document classification [54], image retrieval [80], and multispectral data classification [66]. Co-training [5] provides another popular strategy of incorporating the unlabeled data if the data can be described in two different sufficient views, or sets of attributes. However, it is often observed that these approaches could also degrade performance due to violated model assumption or convergence to local maxima [65]. Partially supervised classification [46] provides a notable solution to one-class classification problems that utilizes unlabeled data; it initially assumes that all unlabeled examples are from negative class, and then applies the EM algorithm to refine the assumption.

Although very effective, it should be noted that the methodology for reestimating the class distribution described in Section 3.2 is applicable only to the sampling bias in class distribution (as was the case in Section 3.3). Therefore, some theoretical or empirical evidence about sampling bias in labeled and unlabeled data should be provided to validate the algorithm application. The goal of our work in progress is to derive statistical tests that could determine: (i) if class distribution is biased, and (ii) if there are other types of sampling bias in data.

Use of contrast classifiers for mining on biased data, described in Section 3.4, is related to partially supervised classification [46]. Contrast classifiers provide a framework for utilizing unlabeled data on a larger class of problems than the previously proposed solutions for the partially supervised classification. The positive results obtained in a bioinformatics problem (Section 3.5) indicate that unlabeled data, if available in large amount, should be considered as an integral part of data mining process and, therefore, should not be ignored. This is particularly the case in biased data scenarios. An appropriate use of unlabeled data could be greatly beneficial to improvement of scientific data mining quality.

Performance-controlled data reduction

Data reduction is an active area of research in data mining. Common approaches include mining on down-sampled data or building predictors on progressively increasing samples until no significant improvement in accuracy is achieved [61]. One could choose to take a random sample from available large dataset or perform weighted sampling to increase the utility of reduced data. Examples of such weighted sampling include editing and condensing [16] or boosting-like methods [22] that concentrate on difficult-to-learn regions of attribute space. Our approach was based on random sampling since the goal was obtaining a representative sample of available data that could be reused in different incremental or distributed learning scenarios.

It is likely that modification of our methodology providing an appropriate weighted sampling could allow more efficient data reduction. It is worth noting that data reduction is similar to active sampling [12] where the goal is to maximize learning accuracy with the minimum amount of labeled data. It would be interesting to explore if the ideas of active sampling could be used in providing more efficient data reduction. Correlations existing in large spatial and temporal scientific datasets provide another dimension for data reduction. Starting from the preliminary results described in Section 4, further research is likely to result in efficient procedures specialized to performance controlled data reduction in spatial- temporal domains.

Overall summary

Applications briefly described at this article were drawn from important domains of precision agriculture, electric power systems and biochemistry. In many ways these areas are very different. However, they all have a common need for a partial automation of a scientific/engineering process due to a huge amount of collected data that prohibits a low capacity approach of a traditional hypothesize-and-test paradigm.

Common challenges for data mining in these areas considered in this article include efficient handling of heterogeneous data distributions, biased labeled samples and large unlabeled databases. The proposed techniques for addressing these challenges provide preliminary evidence in support of a claim that data mining methods can facilitate a shift in scientific methodology towards a more automatic hypothesis generation, model construction and experimentation in science. The current techniques for addressing the scientific data mining problems are far from comprehensive solutions. Therefore, a lot of work in this area is still needed.

Acknowledgements

We thank C.J. Brown, A.K. Dunker, T. Fiez, B. Han, K. Peng, D. Pokrajac, P. Radivojac, K. Tomsovic and H. Xie for collaboration on scientific studies used to illustrate benefits of data mining methods described at this article. These projects were supported in part by NIH grant 1 R01 LM06916-01 to A.K. Dunker and Z. Obradovic, NSF grant IIS-9711532 to Z. Obradovic and A.K. Dunker, NSF grant ECS-9988626 to K. Tomsovic and Z. Obradovic and NSF grant IIS-0219736 to Z. Obradovic and S. Vucetic.

Bibliography

[1] Abu-Mostafa, Y.S. 1995. Hints. Neural Computation, 7:639-671.

[2] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Res. 25: 3389-3402.

[3] Barrett, E.C., and Curtis, F.L. 1999. Introduction to Environmental Remote Sensing. Stanley Thornes Pub. Ltd, Cheltenham,UK.

[4] Blackard J. 1998. Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types, Ph.D. dissertation, Colorado State University, Fort Collins.

[5] Blum , A. and Mitchell, T. 1998. Combining Labeled and Unlabeled Data with Cotraining. Proc. 11th Annual Conf. Computational Learning Theory, 92-100.

[6] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider M. 2003. The SWISS-PROT Protein Knowledgebase and its Supplement TrEMBL. Nucleic Acids Res. 31:365-370.

[7] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. Classification and Regression Trees. Wadsworth International Group: Belmont, CA, 43-49.

[8] Brown, M. P. S., Grundy, W. N. , Lin, D., Cristianini, N. , Sugnet, C. W. , Furey, T. S., Ares Jr. M., and Haussler, D. 2000. Knowledge-based Analysis of Microarray Gene Expression Data by Using Suport Vector Machines. Proc. Natl. Acad. Sci. 97:262-267.

[9] Brunner, R.J. Djorgovski, S.G., Prince, T.A. and Szalay, A.S. 2002. Massive Datasets in Astronomy. In Handbook of Massive Datasets (eds: J. Abello, P. Pardalos, M. Resende). Kluwer Academic Publishers. 931.

[10] Burl, M., Kamatch, C., Kumar, V. and Namburu, R. (chairs) 2001. Third Workshop on Mining Scientific Datasets. First SIAM Int'l Conf. Data Mining. Chicago. IL.

[11] Burl, M., Kamatch, C., Kumar, V. and Namburu, R. (chairs) 2001. Fourth Workshop on Mining Scientific Datasets. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining. San Francisco. CA.

[12] Cohn, D., Ghahramani, Z., and Jordan, M. 1996. Active Learning with Statistical Models, Journal of Artificial Intelligence Research, 4: 129-145.

[13] Collier, N., Nobata, C. and Tsujii, J. 2000. Extracting the Names of Genes and Gene Products with a Hidden Markov Model. Proc. 18th Int'l Conf. Computational Linguistics. Saarbrucken. Germany. 201-207.

[14] Cressie, N.A.C., 1993. Statistics for Spatial Data, John Wiley & Sons, Inc., New York.

[15] Dempster, A.P., Laird, N.M. and Rubin,D.B. 1977. Maximum Likelihood from Data via the EM Algorithm. J. R. Stat. Soc. 39: 1-38.

[16] Devijver, P. and Kittler, J. 1982. Pattern Recognition: A Statistical Approach, Prentice Hall.

[17] Domingos, P. and Hulten, G. 2000. Mining High-Speed Data Streams. Knowledge Discovery and Data Mining, 71-80.

[18] Drossu, R. and Obradovic, Z. 2000. Data Mining Techniques for Designing Efficient Neural Network Time Series Predictors. chapter no. 10 in Cloete, I. and Zurada, J. Knowledge-Based Neurocomputing. MIT Press. 325-368.

[19] Drury, S.A. 1998. Images of the Earth, a Guide to Remote Sensing. Oxford Science Publications, Oxford, UK,

[20] Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M. and Obradovic, Z. 2002. Intrinsic Disorder and Protein Function. Biochemistry. May 28th, 41 (21) 6573-6582.

[21] Efron B. and Tibshirani, R. J. 1993. An Intrrduction to the Bootstrap. New York. Chimpan and Hall.

[22] Freund, Y., and Schapire, R.E. 1996. Experiments with a new boosting algorithm. Proc. Thirteenth Int'l Conf. Machine Learning, 148-156.

[23] Grossman, R.L., Creel, E., Harinath, S., Mazzucco, M., Reinhart, G. and Turinskiy, A. 2000. Terabyte Challenge 2000: Project DataSpace. Workshop on Mining Scientific Datasets, University of Minnesota, Minneapolis, MN.

[24] Grossman, R. and Moore, R. (chairs) 1996. Workshop on Scientific Data Management, Mining, Analysis, and Assimilation, San Diego Supercomputer Center.

[25] Grossman, R., Kamatch, C., Kumar, V. and Namburu, R. (chairs) 2000. Workshop on Mining Scientific Datasets, University of Minnesota, Minneapolis, MN.

[26] Hamilton, J. 1990. Analysis of Time Series Subject to Changes in Regime. Journal of Econometrics, 45: 39-70.

[27] Hamilton, J., 1994. Time Series Analysis, Princeton Univ. Press.

[28] Han, J., Altman, R.B., Kumar, V., Mannila, H. and Pregibon, D. 2002. Emerging Scientific Applications in Data Mining. Communications of the ACM. 45:(8) 54-58.

[29] Han, B., Vucetic, S. and Obradovic, Z. 2003. Re-Ranking Medline Citations by Relevance to a Difficult Biological Query. Proc. IAESTED Int'l Conf. Neural Networks and Computational Intelligence. Cancun, Mexico, 38-43.

[30] Hall, D. Llinas, J. 2001. Handbook of Multisensor Data Fusion, CRC Press, Boca Raton.

[31] Heckman J., 1979. Sample Selection Bias as a Specification Error, Econometrica, 47:(1) 153-161.

[32] Hofmann, K, Bucher, P, Falquet, L, Bairoch, A. 1999. The PROSITE Database, its Status in 1999. Nucleic Acids Res., 27, 215-219.

[33] Huffman D. 1952. A Method for the Construction of Minimum Redundancy Codes. Proc. IRE, 40: 1098-1101.

[34] Jacobs R.A., Jordan M.I., Nowlan S.J., and Hinton G.E. 1991. Adaptive Mixtures of Local Experts. Neural Computation. 3: 79-87.

[35] Jenssen, T.K., Laegreid, A., Komorowski, J., and Hovig, E., 2001. A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression. Nat. Genet. 28(1): 21-28.

[36] Jones, D.T. 1999. Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. J. Mol. Biol. 292: 195-202.

[37] Kalnay, E., et al., 1996. The NCEP/NCAR 40-Year reanalysis project. Bull. Amer. Meteor. Soc., 76: 437-471,

[38] Kamath, C. 2002. Introduction to scientific data mining. Presented at Mathematical Challenges in Scientific Data Mining, short program at Institute for Pure and Applied Marthematics, Univ. of California Los Angeles

[39] King, R.D., Karwath, A., Clare, A., Dehaspe, L. 2000. Genome Scale Prediction of Protein Functional Class from Sequence Using Data Mining. Proc. Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 384-389.

[40] Kamath, C. (chair) 2002. Mathematical Challenges in Scientific Data Mining, short program at Institute for Pure and Applied Marthematics, Univ. of California Los Angeles.

[41] Kargupta, H. and Joshi, A. (chairs) 2002. National Science Foundation Workshop on Next Generation Data Mining, Baltimore, MD.

[42] Kargupta, H. 2000. Distributed Data Mining: Towards the Next Generation of Knowledge Discovery Systems. Workshop on Mining Scientific Datasets. University of Minnesota, Minneapolis, MN.

[43] Kumar, V., Burl, M., Kamatch, C., and Namburu, R. (chairs) 2002. Fifth Workshop on Mining Scientific Datasets, Second SIAM Int'l Conf. Data Mining, Arlignton, VA.

[44] Lazarevic, A. and Obradovic, Z. 2002. Knowledge Discovery in Multiple Spatial Databases. Neural Computing and Applications, 10. no. 4, 339-350.

[45] Lester, H. and Arridge, S.R. 1999. A Survey of Hierarchical Non-Linear Medical Image Registration. Pattern Recognition, 32: 129-149.

[46] Liu, B., Lee, W.S., Yu, P.S. and Li, X. 2002. Partially Supervised Classification of Text Documents. Proc. 19th Intl. Conf. on Machine Learning, Sydney, Australia, 387-394.

[47] Marcotte EM, Xenarios I, Eisenberg D. 2001. Mining Literature for Protein-Protein Interactions. Bioinformatics. Apr; 17(4), 359-63

[48] Memon, N.D., Sayood, K., and Magliveras, S.S., 1994. Lossless Compression of Multispectral Image Data. IEEE Trans. Geosci. Rem. Sens, 32: 282-289

[49] Mjoslness, E. and DeCoste, D. 2001. Machine Learning for Science: State of the Art and Future Prospects. Science. 293: 14 Sept. 2051-2055.

[50] Murphy, P., Aha, D., and Robinson, G. 1992. University of California at Irvine Repository of Machine Learning Databases: Machine-Readable Data Repository.

[51] NASA. 2002. Science Data Users Handbook. Landsat Project Science Office. Goddard Space Flight Center.

[52] NASA. 1999. Workshop on Issues in the Application of Data Mining to Scientific Data, Univ. of Alabama, Huntsville, AL.

[53] NASA. 1999. EOS Reference Handbook, A Guide to NASA's Earth Science Enterprise and the Earth Observing System.

[54] Nigam, K., McCallum, A., Thrun, S. and Mitchell, T. 2000. Text Classification from Labeled and Unlabeled Documents using EM. Mach. Learning, 39(2/3):103-134.

[55] Niyogi, P. Girosi, F. and Poggio, T. 1998. Incorporating Prior Information in Machine Learning by Creating Virtual Examples. Proceedings of the IEEE., 86(11):2196–2209.

[56] Nogues-Paegle J., Mo, K., and Paegle, J. 1998. Predictability of the NCEP-NCAR Reanalysis Model During Austral Summer. Mon. Wea. Rev., 126: 3135-3152.

[57] Obradovic, Z, Peng, K, Vucetic, S., Radivojac, P., Brown, C., and Dunker, A.K. In press. Predicting Intrinsic Disorder from Amino Acid Sequence. Proteins: Structure, Function, and Genetics.

[58] Pawelzik, K., Kohlmorgen, J., and Muller, K.R. 1996. Annealed Competition of Experts for a Segmentation and Classification of Switching Dynamics, Neural Computation, 8: 340-356.

[59] Peng, K., Vucetic, S., Han, B., Xie H. and Obradovic, Z. In review. Exploiting Unlabeled Data for Improving Accuracy of Predictive Data Mining.

[60] Pokrajac, D., Hoskinson, R.L. and Obradovic, Z. In press. Modeling Spatial-Temporal Data with a Short Observation History. Knowledge and Information Systems.

[61] Provost F., Jensen D., Oates T. 1999. Efficient Progressive Sampling. Proc. Fifth Int'l Conf. on Knowledge Discovery and Data Mining.

[62] Roddick, J.F., and Spiliopoulou, M. 1999. A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research. SIGKDD Explorations, 1: 34-38.

[63] Roddick, J.F. and Spiliopoulou, M., 2002. A Survey of Temporal Knowledge Discovery Paradigms and Methods. IEEE Transactions on Knowledge and Data Engineering 14(4): 750-767.

[64] Ryan M.J., and Arnold, J.F. 1997. Lossy Compression of Hyperspectral Data using Vector Quantization. Remote Sensing of Environment. 61: 419-436.

[65] Seeger, M. 2001. Learning with Labeled and Unlabeled Data. Tech. Report, Institute for Adaptive and Neural Computation, University of Edinburgh.

[66] Shahshahani, B., and Landgrebe, D. 1994. The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon. IEEE Trans. Geoscience and Remote Sensing, 32:(5) 1087–1095.

[67] Shatkay, H., Edwards, S., Wilbur, W.J., and Boguski, M., Genes, 2000. Themes and Microarrays: Using Information Retrieval for Large- Scale Gene Analysis. Proc Int Conf Intell Syst Mol Biol, 8: 317-28.

[68] Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R, 1998. Pfam: Multiple Sequence Alignments and HMM-Profiles of Protein Domains. Nucleic Acids Res., 26, 320-322.

[69] Szalay,A.S., Gray,J., Vandenberg,J. 2002. Petabyte Scale Data Mining: Dream or Reality?, Proc. SPIE Conf. Advanced Telescope Technologies.

[70] Vucetic, S., Brown C., Dunker A.K and Obradovic, Z. In press. Supervised Partitioning of Disordered Proteins. Proteins: Structure, Function and Genetics.

[71] Vucetic, S., Pokrajac, D., Xie, H. and Obradovic, Z. 2003. Detection of Underrepresented Biological Sequences using Class-Conditional Distribution Models. Proc. Third SIAM Int'l Conf. on Data Mining, San Francisco, CA, 279-283.

[72] Vucetic, S., Obradovic, Z. and Tomsovic, K. 2001. Price-Load Relationships in California's Electricity Market. IEEE Trans. on Power Systems, 16:(2) 280-286.

[73] Vucetic, S. and Obradovic, Z. 2001. Classification on Data with Biased Class Distribution. Proc. 12th European Conf. on Machine Learning, Freiburg, Germany, 527-538.

[74] Vucetic, S., Fiez, T. and Obradovic, Z. 2000. Analyzing the Influence of Data Aggregation and Sampling Density on Spatial Estimation. Water Resources Research. 36:(12) 3721-3731.

[75] Vucetic S. and Obradovic Z. 2000. Discovering Homogeneous Regions in Spatial Data through Competition. Machine Learning: Proc. 17th Int'l. Conf., Stanford, CA, 1095-1102.

[76] Vucetic, S. and Obradovic, Z. 2000. Performance Controlled Data Reduction for Knowledge Discovery in Distributed Databases. Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, Kyoto, Japan. Computer Science Editorial 3, Springer-Verlag, 29-39.

[77] Vucetic, S. and Obradovic, Z. 2000. A Constructive Competitive Regression Method for Analysis and Modeling of Non-stationary Time Series. Proc. First Int'l Workshop on Computational Intelligence in Economics and Finance at the Fifth Int'l Conf. On Information Science, Atlantic City, N.Y., USA. 2: 978-981.

[78] Vucetic, S., Fiez, T. and Obradovic, Z. 1999. A Data Partitioning Scheme for Spatial Regression. Proc. IEEE/INNS Int'l Joint Conf. on Neural Networks, IEEE Press, Washington D.C., Session 8.1A.

[79] A.S. Weigend, A.S., Mangeas, M. and Srivastava, A.N. 1995. Nonlinear Gated Experts for Time Series: Discovering Regimes and Avoiding Overfitting. Int. Journal of Neural Systems, 6: 373-399.

[80] Wu, Y. Tian, Q. and Huang, T.S. 2000. Integrating Unlabeled Images for Image Retrieval Based on Relevance Feedback. Proc. 15th Int'l Conf. on Pattern Recognition, 1: 21-24, Barcelona, Spain.

[81] Yona, G., Linial, N., Linial, M. 1999. ProtoMap: Automatic Classification of Protein Sequences, A Hierarchy of Protein Families, and Local Maps of the Protein Space. Proteins, 37: 360-378.