Mining for Most Certain Predictions from Dyadic Data

Meghana Deodhar
Department of Electrical and Computer Engg.
University of Texas at Austin
Austin, TX, USA
deodhar@ece.utexas.edu

Joydeep Ghosh
Department of Electrical and Computer Engg.
University of Texas at Austin
Austin, TX, USA
ghosh@ece.utexas.edu

ABSTRACT

In several applications involving regression or classification, along with making predictions it is important to assess how accurate or reliable individual predictions are. This is particularly important in cases where due to finite resources or domain requirements, one wants to make decisions based only on the most reliable rather than on the entire set of predictions. This paper introduces novel and effective ways of ranking predictions by their accuracy for problems involving large-scale, heterogeneous data with a dyadic structure, i.e., where the independent variables can be naturally decomposed into three groups associated with two sets of elements and their combination. These approaches are based on modeling the data by a collection of localized models learnt while simultaneously partitioning (co-clustering) the data. For regression this leads to the concept of “certainty lift”. We also develop a robust predictive modeling technique that identifies and models only the most coherent regions of the data to give high predictive accuracy on the selected subset of response values. Extensive experimentation on a variety of real life datasets highlights the utility of our proposed approaches.

1. INTRODUCTION

In certain prediction problems, one is interested in only the most accurate or useful predictions rather than in the performance over the entire test/scoring data. Such scenarios typically arise in applications where acting on each prediction requires a certain quantity of a limited resource and/or the cost incurred depends on the inaccuracy of the prediction. Perhaps the most well known example comes from direct marketing where one is interested in identifying a subset of the target population most likely to respond to a solicitation (positive class) [18]. Since making irrelevant offers is not only unprofitable but also harmful to customer experience, the choice of the target sub-population is critical. In this context, the performance of a model is evaluated based on the response rate in the top x quintiles via a gain

(lift) chart, rather than on the overall classification accuracy. Determining the most relevant documents/pages for a search query has a similar flavor. Note that these are both classification examples in which there is a rare positive class, and one tries to identify a small subset of the records that most likely belong to this class.

A qualitatively different consideration is encountered in certain regression settings. Consider a stock market player in options, who wants to predict closing stock prices. The sign of price change is not critical as he can use both “puts” and “calls”, and he can bet on only a limited number of stocks because of a limited budget. In this situation, the accuracy of prediction (difference from the actual closing value) for a limited number of stocks is key to this person’s ability to maximize his expected returns. A similar situation is faced in predicting other futures (energy usage, currency rates, etc.) for arbitraging strategies. Note that there has been little study of selection/ranking of regression results based on accuracy in the data mining literature. In this paper, we focus on the problem of ranking predictions of numeric quantities by their reliability, so that one is able to identify a subset of predictions that are the most accurate. Though the emphasis is on regression settings, we will also show how the methodology can be applied to classification problems as well.

The problem of estimating the uncertainty associated with the prediction of a regression model has been examined in the statistics and neural networks literature [11, 3, 12]. The typical approach is to learn a linear/non-linear map from the independent variables to the response variable and then estimate the uncertainty of each prediction as a function of the uncertainty in the regression coefficients. In this paper we show that a radically different and more effective approach can be taken if the data is structured such that the independent variables can be naturally partitioned into two (or more) groups and a pivoting operation on the response variables can result in organizing them as entries in a matrix. Data with this structure is referred to as dyadic (bi-modal) data [2] and is now pervasive in a variety of real life applications. Dyadic data consists of measurements on dyads, which are pairs of elements from two different sets (modes). The measurements can be represented as the entries of a matrix, whose rows and columns are the two sets of elements. The independent variables, also referred to as covariates are associated with the entities along the two modes. Continuing with the stock market example, assume that we have the prices for a set of stocks predicted by a certain number of analysts. The independent variables are attributes asso-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WOODSTOCK '97 El Paso, Texas USA

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

ciated with each stock, characteristics of each analyst and attributes associated with a stock-analyst pair, e.g., how good the analyst’s prediction has been for the stock in the past. The data can now be thought of as a stocks by analysts matrix with the response variable, i.e., the predicted price, represented by each cell of the matrix. Note that the known entries in this matrix form the labeled dataset and the missing values are the ones that we desire to estimate. Datasets consisting of web page-ad clicks, search query-web page relevance, user-movie ratings or customer-product preferences are other examples of dyadic data. While this paper primarily considers dyadic data, the extension of our proposed approach to tensor (multi-modal) data is straightforward.

We begin by formulating a simple but effective ranking strategy using a single model learned on the data. Motivated by the observation that in a realistic setting the dataset is typically too large and heterogeneous to be adequately represented by a single global model, we propose ranking strategies based on the Simultaneous CO-clustering And Learning (SCOAL) framework [7]. SCOAL is a technique proposed for modeling large-scale, heterogeneous data by partitioning the data matrix into a grid of blocks (co-clusters) while simultaneously learning models in each block. We formulate two novel ranking schemes, (i) row-column ranking, and (ii) block ranking, and show that the use of multiple localized models that evolve on dyadic data leads to a better as well as more actionable ranking.

Another contribution of our work is Robust SCOAL, a novel predictive modeling technique that selectively models only the most coherent parts of the data. A different viewpoint of this technique is as an approach for automatic outlier detection and removal that is carried out simultaneously with the modeling process. Robust SCOAL identifies and retains only those parts of the data where accurate predictive models can be learnt. The predictions within the selected data subset can be subsequently ranked. Finally, we introduce a new certainty lift measure that evaluates the improvement in error for a selected subset of predictions.

What distinguishes our proposed approaches from the traditional techniques for error bar estimation [11, 3, 12] is that, our approaches can take advantage of the structure of dyadic data rather than treat it as a flat collection of points. Moreover, while most existing techniques are based on learning a single model using the available training data, we learn a collection of localized models that provide a better representation of the input space and finer resolution of prediction uncertainty. Comparisons of our ranking procedure with ranking based on standard methods of prediction error estimation on real life datasets highlight the benefits of our approach. Additionally, our ranking strategies with a collection of localized models are readily applicable to both regression and classification problems.

2. RELATED WORK

Determining the standard error of the predictions of a linear regression model is well established, text-book material [13]. This is not as direct for non-linear models, and there has been a considerable amount of work in the neural network community on determining the uncertainty of network predictions [16, 12]. There are several references in the literature on the application of resampling based techniques for prediction error estimation. Heskes [11] uses the variance of the outputs of an ensemble of neural networks,

trained on bootstrap replicates of the training dataset, for computing confidence intervals on the output for new data points. This technique is shown to do better than [16] on synthetic data. High computational effort however is one of the main drawbacks of the resampling approaches. While resampling techniques construct multiple global models, each of which represents the entire input space, our approach is based on a collection of local models (Section 5). To the best of our knowledge, no one has used an ensemble of localized models for estimating the accuracy of individual predictions.

Bayesian techniques provide another way of estimating prediction uncertainty based on the variance of the predictive distribution computed as a function of the input [4, 3]. We compare our proposed techniques with a Bayesian linear regression approach [4] in Section 10.

An example of a recent non-Bayesian approach is the work by Cawley et al. [6] where they assume heteroscedasticity as in [3] and use an iterative procedure based on kernel ridge regression to simultaneously model the mean and variance of the target distribution as functions of the input. A novel approach by Shrestha and Solomatine [19] involves clustering the residuals into homogeneous groups using fuzzy c-means and then constructing the prediction limits for each cluster, based on the distribution of the errors within the cluster. The prediction limits for each training point are then estimated according to its membership in each cluster. A linear/non-linear mapping learnt from each input point to its prediction intervals is used to make estimates for test points. While the learning of the model and clustering the model residues are sequential steps in this work, our approach discussed in Section 5 is based on performing these two tasks simultaneously.

The recently proposed SCOAL [7] and PDLF [2] approaches deal with predictive modeling of large, heterogeneous dyadic data. More details on the SCOAL approach are presented in Section 5. The Predictive Discrete Latent Factor (PDLF) model simultaneously incorporates the effect of the covariates, through a global model as well as any local structure that may be present in the data, through a block (co-cluster) specific constant. However, neither of these techniques consider the uncertainty of the predicted values.

Our proposed idea of ranking co-clusters and rows/columns of the data matrix based on their prediction accuracy is inspired by the recent work in co-clustering, which involves the mining of the dense, coherent co-clusters from large, noisy datasets, by pruning away irrelevant regions of the data [8]. Note that [8] is only focused on clustering; no prediction models are formed.

Notation: Lower case letters represent scalars, e.g., a, z , lower case, bold face letters represent vectors, e.g., $\mathbf{b}, \mathbf{c}, \boldsymbol{\beta}$, upper case letters like Z, W represent matrices and calligraphic upper case letters like $\mathcal{T}$ represent sets. Individual elements of a matrix, e.g., Z are represented as z_{ij} , where i and j are the row and column indices respectively.

3. PROBLEM DESCRIPTION

Let the set of training data points be represented by pairs $(\mathbf{x}_k, y_k)$, where $\mathbf{x}_k$ is the vector of independent variables (covariates) and y_k is the corresponding target. It is assumed that the response variable is generated by a certain model involving the covariates. The problem is to learn a model that maps $\mathbf{x}_k$ to y_k and outputs a ranking of the model predictions by an estimate of their accuracy.

In some cases, as discussed in Section 1, the independent variables can be naturally partitioned into 3 groups associated with 2 sets of entities and their combination. The response variable y can now be represented as the cells z_{ij} of an $m \times n$ matrix Z , and we now denote the corresponding covariates $\mathbf{x}$ by $\mathbf{x}_{ij}$, which represents the independent variables associated with z_{ij} ¹. We specifically adopt this notation to emphasize the dyadic nature of the data. The covariate vector can be partitioned into row attributes, $\mathbf{r}_i$, column attributes, $\mathbf{c}_j$ as well as annotations associated with a row-column pair, $\mathbf{a}_{ij}$. Note that up to two of these three vectors $(\mathbf{r}_i, \mathbf{c}_j, \mathbf{a}_{ij})$ could be empty. The matrix Z will have missing values, which form the set of responses to be predicted. The missing values are handled by associating a weight w_{ij} with each cell z_{ij} . The weights of the known (training) matrix cell values can be set to 1, while missing values can be ignored during the modeling process by setting their weights to 0. In general, the weights are not restricted to 0 or 1 and can take other non-negative values that reflect uncertainties or importances associated with the responses.

The problem can now be posed as predicting the missing entries in matrix Z , along with an estimate of the goodness of the prediction. This will enable the selection of a specified number of missing data entries on which the most accurate predictions are expected to be made.

4. RANKING WITH A SINGLE MODEL

A simple approach to the above problem is to train a single “global” model that learns a map from the covariates to the response variable. For ease of exposition, consider a linear model $y_k = \beta^T \mathbf{x}_k + u_k$, where $(\mathbf{x}_k, y_k)$ is the k^{th} data point. u_k is the unobserved error term, assumed to be a random variable with $E(u_k | \mathbf{x}_k) = 0$.

In the dyadic data notation, the response variable $z_{ij} \in \mathbb{R}$ is estimated by a linear combination of the corresponding covariates: $\hat{z}_{ij} = \beta^T \mathbf{x}_{ij}$, where $\mathbf{x}_{ij}^T = [1, \mathbf{r}_i^T, \mathbf{c}_j^T, \mathbf{a}_{ij}^T]$ is a vector consisting of the covariates augmented with $x_0 = 1$ for the intercept, and the least squares estimates of the model parameters are $\beta^T = [\beta_0, \beta_x^T]$. Let $\mathcal{T}$ represent the set of training data instances, of size s and let p be the number of independent variables.

Since β is estimated with imprecision, the predicted value $\hat{z}_{ij}$ is also subject to error. For the above multivariate linear regression model, the error variance is estimated by the following standard approach [13] we refer to as Global Var

$$E_{ij}^{\text{GlobalVar}} = \text{var}(z_{ij} - \hat{z}_{ij}) \quad (1)$$

$$= \sigma^2 [1 + \mathbf{x}_{ij}^T (Y^T Y)^{-1} \mathbf{x}_{ij}].$$

Y is the design matrix, with rows of the form $\mathbf{x}_{uv}^T$, where $(u, v) \in \mathcal{T}$, i.e., $w_{uv} = 1$ for a 0-1 weight matrix W . σ^2 can be estimated by the mean squared error, given by

$$\hat{\sigma}^2 = \frac{\sum_{uv \in \mathcal{T}} (z_{uv} - \hat{z}_{uv})^2}{s - p}.$$

The square root of (1) gives the standard error of the prediction $\hat{z}_{ij}$. In the experiments in Section 10, we use ridge regression with error variance estimated as

$$E_{ij}^{\text{GlobalVar}} = \sigma^2 [1 + \mathbf{x}_{ij}^T (Y^T Y + \lambda I)^{-1} (Y^T Y) (Y^T Y + \lambda I)^{-1} \mathbf{x}_{ij}], \quad (2)$$

¹Two subscripts are now used to denote row and column indices.

where $\lambda \geq 0$ controls the amount of regularization [9].

Other alternatives to estimate prediction uncertainty, as discussed in Section 2, are to employ a resampling based technique [11] or a Bayesian technique [4], where the uncertainty is estimated by the variance of the predictive distribution. However, all these procedures implicitly flatten the data matrix and hence fail to exploit the inherent dyadic structure of the data.

For dyadic data fitted with a single (linear or non-linear) model, we propose a third approach, referred to as Global Rank, to rank predictions by their estimated accuracy.

1. The average error for row i is its training mean squared error, computed as

$$E_i = \frac{\sum_{j=1}^n w_{ij} (z_{ij} - \hat{z}_{ij})^2}{\sum_{j=1}^n w_{ij}}.$$

2. Similarly, the average error for column j is given by

$$E_j = \frac{\sum_{i=1}^m w_{ij} (z_{ij} - \hat{z}_{ij})^2}{\sum_{i=1}^m w_{ij}}.$$

3. Assuming independence of row and column errors, the error for a missing entry z_{ij} is given by

$$E_{ij}^{\text{GlobalRank}} = E_i + E_j. \quad (3)$$

4. A ranking of the predicted entries is obtained based on their computed errors.

While a single learned model is adequate for simple prediction problems, it may not be sufficient to represent the heterogeneous population that very large classification or regression problems often involve. For instance, the problem of predicting customer preferences for products typically involves a diverse population of customers and a wide range of different products. It is unlikely that all the customer-product preferences can be well explained by a single model. Since it is more natural for a model to closely represent the preferences of only a subset of the customers for a subset of the products, the data might be better modeled by a collection of suitably trained localized models [7].

5. RANKING WITH A COLLECTION OF LOCAL MODELS

The recently proposed simultaneous co-clustering and learning (SCOAL) framework [7] aims to solve prediction problems in heterogeneous domains, where the independent variables can be naturally partitioned into two (or more) groups. The key idea is to co-cluster the entire data matrix into a grid of blocks (co-clusters) such that each block can be well characterized by a single predictive model that relates the independent variables to the response variables in the block.

For building the necessary background, we now provide a brief summary of SCOAL. The problem setting assumed is similar to that described in Section 3. Formally, let k, l represent the number of row and column clusters and let ρ, γ be mappings from the m rows to the k row clusters and from the n columns to the l column clusters respectively. Considering a regression problem, let us start with linear generative models for the data, for ease of exposition². The

²Generalization to other models is discussed in Section 5.1.

aim of SCOAL is to find a co-clustering defined by (ρ, γ) and the associated $k \times l$ regression models that minimize the following objective function

$$\sum_{g=1}^k \sum_{h=1}^l \sum_{u:\rho(u)=g} \sum_{v:\gamma(v)=h} w_{uv} (z_{uv} - \hat{z}_{uv})^2, \quad (4)$$

where

$$\hat{z}_{uv} = \beta^{gh^T} \mathbf{x}_{uv}. \quad (5)$$

A suitable co-clustering (ρ, γ) can be obtained by an algorithm that iterates over two steps; the model update step and the row and column cluster reassignment step. In the model update step, the β for each co-cluster is a solution to a least squares problem. The cluster reassignment step greedily assigns each row/column to the row/column cluster that minimizes the squared error for the row/column. Each step is guaranteed to reduce the objective function and the algorithm hence converges to a locally optimal solution. After the algorithm converges, a missing value z_{ij} is predicted as $\hat{z}_{ij} = \beta^{gh^T} \mathbf{x}_{ij}$, where $\rho(i) = g$ and $\gamma(j) = h$.

By generating multiple localized models, SCOAL provides novel ways to rank predictions by their expected quality. In this section we propose two such approaches: (i) row-col ranking, (ii) block ranking.

Row-Col Ranking. The row-col ranking technique is based on estimation of the prediction errors as given by Equation (3). The average error for row i , where $\rho(i) = g$ is given by

$$E_i = \frac{\sum_{h=1}^l \sum_{j:\gamma(j)=h} w_{ij} (z_{ij} - \beta^{gh^T} \mathbf{x}_{ij})^2}{\sum_{j=1}^n w_{ij}}.$$

The average error for column j is computed similarly. As before, the error for a missing entry z_{ij} is the sum of the corresponding average row and column errors. Intuitively, since the collection of local models characterize the data better than a single model, the resulting ranking would be expected to be better as well.

Block Ranking. The intuition here is that for a heterogeneous dataset the fit of the co-cluster models, quantified by the training reconstruction error (MSE), differs across co-clusters. Models in noisy and sparse regions of the input space will have poorer fit, while other models may capture more coherent relationships between the covariates and the response variables. Hence, the first step is to rank the predictions by the mean squared error of the co-clusters they are assigned to. The mean error for co-cluster (g, h) is computed as

$$\frac{1}{\sum_{i:\rho(i)=g} \sum_{j:\gamma(j)=h} w_{ij}} \sum_{i:\rho(i)=g} \sum_{j:\gamma(j)=h} w_{ij} (z_{ij} - \beta^{gh^T} \mathbf{x}_{ij})^2.$$

Ranking by co-cluster membership helps in identifying regions of the input space that are not adequately represented by the learnt models and gives the predictions in these regions a lower rank. Further, a second level of ranking can be done within each co-cluster (g, h) , in which the predictions are ranked by estimating their prediction errors as a sum of the average co-cluster row and co-cluster column errors, i.e., the error for a predicted value in row i , column j is $E_i + E_j$, where,

$$E_i = \frac{\sum_{k:\gamma(k)=h} w_{ik} (z_{ik} - \hat{z}_{ik})^2}{\sum_{k:\gamma(k)=h} w_{ik}},$$

$$E_j = \frac{\sum_{k:\rho(k)=g} w_{kj} (z_{kj} - \hat{z}_{kj})^2}{\sum_{k:\rho(k)=g} w_{kj}}.$$

5.1 Generalization to Other Models

SCOAL and the associated row-col and block ranking procedures are readily extensible to other predictive models as well as to classification problems by modifying (4) and/or (5) [7]. For example, the data can be modeled by a collection of neural network models or regression models with the L_1 norm (Lasso [10]). In our experiments described in Section 10 we use ridge regression, which reduces the possibility of overfitting while learning the larger number of model parameters ($k \times l$ β vectors) as compared to a single linear regression model.

6. MODELING A SELECTED DATA SUBSET

As discussed in Section 1, for several applications the objective is not to make predictions for every missing data entry, but to make very accurate predictions for a certain fraction of the missing data. Moreover, several domains involve datasets that are often large and noisy and the discriminative patterns are limited to only a small subset of the data. Also, the data could contain outliers that could largely skew the prediction models that include them. Simpler and more generalizable models can be learnt by ignoring the outliers and the non-informative parts of the data. Hence we propose Robust SCOAL, a technique for selecting and modeling a suitable subset of the data.

Robust SCOAL aims at simultaneously co-clustering the data and learning predictive models in each co-cluster, with the added constraint that only a predetermined number of rows and columns are assigned to row and column clusters. More concretely, with the same problem setting as before (Section 3), our aim here is to simultaneously select and cluster $s_r \leq m$ rows and $s_c \leq n$ columns of Z , into a grid of k row clusters and l column clusters, such that the response values in each co-cluster are predicted by a common regression model. The exact value of s_r and s_c is not critical in the setting of this paper, as this induces the first stage of selection only, and a second stage is needed in any case so long as s_r and s_c are not too small. Let $\mathcal{K}$ and $\mathcal{L}$ denote the sets consisting of the s_r clustered rows and the s_c clustered columns respectively. Let ρ be a mapping from the s_r rows $\in \mathcal{K}$ to the k row clusters and γ be a mapping from the s_c columns $\in \mathcal{L}$ to the l column clusters. We want to find a co-clustering defined by (ρ, γ) , sets $\mathcal{K}$ and $\mathcal{L}$ and the associated set of regression models $\{\beta_s\}$ for specified s_r and s_c that minimize the following objective function

$$\frac{1}{\sum_{u \in \mathcal{K}, v \in \mathcal{L}} w_{uv}} \sum_{g=1}^k \sum_{h=1}^l \sum_{u \in \mathcal{K}:\rho(u)=g} \sum_{v \in \mathcal{L}:\gamma(v)=h} w_{uv} (z_{uv} - \hat{z}_{uv})^2, \quad (6)$$

where $\hat{z}_{uv} = \beta^{gh^T} \mathbf{x}_{uv}$.

Note that the objective function (6) (i) involves the squared error summed only over $s_r \times s_c$ elements of Z , and that (ii) it is based on the mean error rather than the total error. In the presence of missing entries in the data matrix, the denominator in (6) will differ for different choices of sets $\mathcal{K}$

and $\mathcal{L}$ and hence needs to be considered in the cost function. This point is especially important for sparse data, where a substantial number of matrix entries are missing. In this scenario, if the total error was optimized rather than the mean error the solution would be a very sparse subset of the data with highly overfit models, since missing entries have $w_{uv} = 0$ and do not contribute to the error. In the most extreme case, the solution would be an empty $s_r \times s_c$ block of the matrix Z with zero error.

6.1 Robust SCOAL Algorithm

As in the case of SCOAL, a co-clustering (ρ, γ) , that decreases the cost function can be obtained by an iterative algorithm that alternately updates co-cluster models and row/column cluster assignments. The algorithm begins by initializing (ρ, γ) either randomly or by a suitable clustering procedure followed by a selection of s_r rows and s_c columns. Given the initial cluster assignments (ρ, γ) of the selected s_r rows and s_c columns, the co-cluster models are constructed, i.e., the coefficient vector β is learnt for each co-cluster. In case of 0/1 weights, this is done by least squares regression using only the non-missing (training) values within the co-cluster. In case of a general set of weights, the β is a solution to a weighted least squares problem.

For a fixed column cluster assignment and a set of co-cluster models, the row cluster update step involves assigning rows to row clusters and selecting s_r of m rows to participate in the current clustering. If row u is assigned to row cluster g , i.e., $\rho(u) = g$, the row error is the error summed over the appropriate s_c elements in the row given by

$$E_u(g) = \sum_{h=1}^l \sum_{v \in \mathcal{L}:\gamma(v)=h} w_{uv}(z_{uv} - \hat{z}_{uv})^2.$$

For a fixed $\mathcal{L}, \gamma$ and set of β s, the best choice of the row cluster assignment for row u is the g that minimizes the row error, i.e., $\rho^{new}(u) = \argmin_g E_u(g)$. Each of the m rows is hence assigned to the row cluster that minimizes the row error.

Since the objective function (6) includes the normalization term, it cannot be decomposed over the rows. A naive greedy solution of selecting the s_r rows with lowest mean row error is hence not optimal for the row selection problem. Instead, we select the s_r rows that minimize

$$\frac{\sum_{u \in \mathcal{K}} E_u(\rho^{new}(u))}{\sum_{u \in \mathcal{K}} \sum_{v \in \mathcal{L}} w_{uv}}.$$

An optimal selection of s_r rows can now be obtained by using a standard dynamic programming approach akin to the solution of the knapsack problem in [15].

A similar approach is used to update the column clusters. Note that the rows/columns that are not included in the current s_r/s_c rows/columns assigned to co-clusters are still retained since they could be included in the co-clusters in future iterations. The algorithm alternates between the row/column cluster update and model update steps until convergence. The overall algorithm is described in Figure 1. Step 1 reduces the objective function because of the optimality of the pseudo-inverse solution to linear regression, steps 2 and 3 are greedy updates and directly reduce the objective function. The objective function hence decreases in every iteration. Since this function is bounded from below by zero, the algorithm is guaranteed to converge to a local minimum.

Algorithm: Robust SCOAL

Input: $Z_{m \times n}$, $W_{m \times n}$, covariates, s_r , s_c, k, l
Output: (ρ, γ) , co-cluster models β s

1. Begin with an initial co-clustering (ρ, γ)
2. Repeat
 3. **Step 1: Update co-cluster models**
 4. $\forall [g]_1^k, [h]_1^l$,
 5. Train a linear regression model with
 6. all the training samples $(\mathbf{x}_{uv}, z_{uv})$ in
 7. co-cluster (g, h) , with weights w_{uv} ,
 8. to obtain an updated β^{gh} .
 9. **Step 2: Update ρ**
 10. **2a.** $\forall [u]_1^m$,
 11. $\rho(u) = \argmin_g \sum_{h=1}^l \sum_{v \in \mathcal{L}:\gamma(v)=h} w_{uv}(z_{uv} - \beta^{ghT} \mathbf{x}_{uv})^2$
 12. **2b.** $\mathcal{K}$ = the set of s_r rows selected
 13. by dynamic programming.
 14. **Step 3: Update γ**
 15. **3a.** $\forall [v]_1^n$,
 16. $\gamma(v) = \argmin_h \sum_{g=1}^k \sum_{u \in \mathcal{K}:\rho(u)=g} w_{uv}(z_{uv} - \beta^{ghT} \mathbf{x}_{uv})^2$
 17. **3b.** $\mathcal{L}$ = the set of s_c columns selected
 18. by dynamic programming.
 19. **20.**
 21. Until Convergence
 22. Return (ρ, γ) and β 's

Figure 1: Pseudo-code for Robust SCOAL

If the data is extremely sparse then, for given s_r and s_c , the row/column selection step could involve empty rows/columns with zero weight and hence zero cost. Although selecting such rows/columns will be optimal from the point of view of the cost function, it will not result in a well generalizable set of models for making accurate predictions on test data. We hence restrict the dynamic programming steps (2b and 3b) to select from the set of non-empty rows/columns. This avoids the modeled part of the data matrix from including rows and columns with all entries missing.

While steps 1, 2a and 3a of the Robust SCOAL algorithm are linear in the size of the data, the complexity of steps 2b and 3b is $O(ms_r)$ and $O(ns_c)$, i.e., quadratic in the worst case. If the number of missing entries in the data matrix is comparatively small and the entries are missing at random, then the $\sum_{u \in \mathcal{K}, v \in \mathcal{L}} w_{uv}$ term in the objective function (6) does not vary too much over different $s_r \times s_c$ blocks of the matrix and can be ignored. An efficient heuristic then is to replace the dynamic programming steps 2b and 3b by a greedy selection of the s_r rows and s_c columns with the lowest row and column errors respectively. This greedy selection can be performed in linear time by the efficient order statistics based algorithm proposed by Blum et al. [5]. We verified empirically on the ERIM dataset (Section 10) that there is no significant difference between the greedy strategy and dynamic programming in terms of prediction accuracy.

Similar to the row-column ranking procedure described in Section 5, the Robust SCOAL predictions can be ranked

by an estimation of the prediction error, which is the sum of the average row and column errors. Note that the average row/column error is computed over only the rows and columns that belong to sets $\mathcal{K}$ and $\mathcal{L}$ respectively.

Analogous to the discussion in Section 5.1, Robust SCOAL can also be generalized to other predictive models and/or loss functions. For instance, the experiments in Section 10 use Robust SCOAL with ridge regression.

7. EXTENSION: RANKING FOR CLASSIFICATION PROBLEMS

The focus so far has been on assessing the reliability of the output of a regression model. In classification problems as well, it may be of interest to select a specified number of the most accurate or reliable predictions. A motivating example is the ecological problem of predicting the absence/presence of a set of species in a set of locations. Extremely reliable predictions for a few location-species pairs lead to more fruitful ground research rather than the burdensome task of investigating a larger set of predictions with lower overall accuracy. In this section, we show that if a classification problem involves data with a dyadic structure as described in Section 1, we can adapt the proposed ranking technique to obtain a ranking of the classifier predictions by their reliability. Analogous to the regression setting, note that for a 2-class problem we treat the 2 classes symmetrically and aim to find the most accurate predictions, irrespective of the predicted class label. Our perception of prediction “accuracy” in this case is simply the probability of correct classification for $\mathbf{x}_{ij}$, and so a posteriori class probabilities very close to 0 or 1 are ranked higher as compared to probabilities close to the prior. Hence, if the data is modeled by a single, “global” probabilistic classifier, $\max_i(P(\text{class}_i|x_k))$ can be used to rank predicted values.

For data with a dyadic structure, the response variable is represented as the entries of a matrix Z as before, with covariates associated with the corresponding modes. The entries in matrix Z are now class labels, e.g., $+1, -1$ for 2-class problems. We use the SCOAL (meta)-algorithm to simultaneously partition the matrix into co-clusters and learn classification models in each co-cluster. One can use logistic regression classifiers, in which case the SCOAL objective function will be the total log loss. We then get a ranking of the missing class labels based on the class membership probabilities estimated by the co-cluster models. We illustrate in Section 10.3 that the ranking provided by this approach is better than that obtained by a global model. The Robust SCOAL algorithm described in Section 6.1 can also be easily extended to classification models by suitably changing the objective function (6) to the loss function of the classifier.

8. CONTRASTING DIFFERENT RANKING TECHNIQUES

The ranking techniques discussed so far in Sections 4, 5 and 6 result in 3 conceptually different approaches for selecting the subset of the missing entries to be predicted. Both Global Rank and row-col ranking select individual entries in the data matrix. In contrast, for block ranking the unit of selection is a co-cluster rather than individual matrix entries, since the first level of ranking is based on co-cluster membership. Finally, Robust SCOAL selects the most accurately predictable parts of the data matrix by throwing away

rows and columns with large errors. Its unit of selection is hence individual rows and columns. The selection technique that is the most relevant to a certain problem depends on the application.

In applications involving extremely large datasets, Robust SCOAL identifies a smaller subset of the data that can be focused on and analyzed in further detail. For example, in a direct marketing application, it identifies a group of customers and products for which predictions can be made with very high confidence. This result is possibly more actionable from a marketing viewpoint than selecting individual customer-product combinations in an unstructured manner from the entire customer-product base. Robust SCOAL is also well suited to datasets known to have substantial outliers since it performs simultaneous outlier detection and pruning during the modeling process.

Selection at the co-cluster level is the most appropriate when the subset of the predictions on which decisions/actions are taken is required to satisfy certain criteria imposed by the application. For example, a certain application might constrain the predictions to be localized to a certain geographical region or to users in a certain age or income group. A retail business may want to identify customers whose purchase decisions are heavily influenced by the price, and provide special discounts to them on products they are most likely to purchase. Ranking by co-clusters is also very interpretable since the co-cluster models can provide insights on the degree to which different factors influence customer purchases in various customer-product subgroups.

An interesting point is that the proposed ranking approaches can be combined to make selections at multiple resolutions. Robust SCOAL can first select a set of rows and columns, which can be followed by a ranking of the co-clusters, with a finer ranking by individual entries within each co-cluster.

9. CERTAINTY LIFT

Having discussed a number of approaches to select the most certain predictions, we now formulate a “certainty lift” measure that evaluates the improvement in the error of a selected subset of model predictions as compared to the error of a baseline model on the entire test set. With squared error as the loss function, the certainty lift is defined as

$$\text{Certainty Lift}(\text{model}, \mathcal{V}) = \frac{\frac{1}{s} \sum_{i \in \mathcal{U}} (y_i - \bar{y}_{\mathcal{T}})^2}{\frac{1}{s} \sum_{i \in \mathcal{V}} (y_i - y^{\text{model}})^2}, \quad (7)$$

where $\mathcal{U}$ represents the test set, $\mathcal{V} \subseteq \mathcal{U}$ and $s = |\mathcal{U}|$ and $\bar{s} = |\mathcal{V}|$. The baseline model is simply the mean of the response values in the training dataset, given by $\bar{y}_{\mathcal{T}}$ and y^{model} is the value output by a predictive model learnt on $\mathcal{T}$. For example, for a given model, a certainty lift of 5 on a subset $\mathcal{V}$ of $\mathcal{U}$, implies that the mean squared error on $\mathcal{V}$ is 5 times lower than the baseline error on $\mathcal{U}$.

10. EXPERIMENTAL RESULTS

In this section we evaluate the proposed ranking procedures on a number of real life datasets from a variety of applications. Regression problems are considered on the MovieLens and ERIM Marketing datasets in Sections 10.1 and 10.2 respectively. Section 10.3 explores the ranking mechanisms

in a classification setting on a dataset of student-course choices.

10.1 MovieLens Dataset

The MovieLens dataset consists of 100,000 ratings (1-5) from 943 users on 1682 movies made available by the GroupLens Research Project at the University of Minnesota³. We use a set of 23 attributes including user demographics like age, gender, employment status and movie release year and genre. The dataset is extremely sparse, with the proportion of known user-movie ratings as small as 0.7 % of the size of the data matrix. We assume ridge regression models to predict the ratings and evaluate the quality of the predictions using mean squared error.

We evaluate our ranking techniques by comparing with three traditional approaches for computing the uncertainty of predictions of a linear regression model. The first is Global Var, which computes the standard error of a predicted value using Eq. (2)⁴. The second approach, Global Resample, trains a set of ridge regression models on bootstrap samples of the training dataset and computes the variance of the values predicted by the ensemble of models. A ranking of the data points in the test set is then obtained by sorting them in ascending order of their estimated standard error/variance values. We also consider Bayesian linear regression (Bayesian LinReg), with a zero-mean, isotropic Gaussian prior [4]. The precision parameter of the Gaussian prior is selected by cross validation, while the variance of the additive Gaussian noise is set to the MSE obtained with a linear least-squares model. The ranking is based on the estimated variance of the predictive distribution for each test point. Figure 2 compares the Global Var, Global Resample (with 25 bootstrap samples) and Bayesian LinReg techniques with Global Rank, which uses the dyadic data structure to provide the ranking (Eq. (3)), the SCOAL based row-col ranking and block ranking and Robust SCOAL. $k = 4$ and $l = 4$ for the SCOAL techniques. Robust SCOAL is run to select $s_r = 890$ rows and $s_c = 1550$ columns, followed by row-col ranking to rank order the selected predictions. The SCOAL techniques use ridge regression models with $\lambda = 0.78$, selected by cross validation. The x-axis specifies the fraction of the points in the test set that are predicted, while the y-axis gives the cumulative mean squared error on the predicted set of points. At the last point on the x-axis all the entries in the test set are predicted and included in the MSE. The plotted values are averaged over 10 random 80 – 20 % splits of the data as the training and test set.

The plot shows a significant improvement in the ranking obtained by techniques that use the dyadic data structure as compared to Global Var, Global Resample and Bayesian LinReg (almost superimposed in Figure 2, with MSE around 1.2). Since a collection of co-cluster models characterizes the data better than a single global model [7], the MSE of SCOAL on the entire training dataset is significantly lower than the MSE of a single global model, illustrated by the last point on the x-axis. Row-col ranking results in a useful ranking of the predicted values, with 10% of the test en-

Figure 2: Comparison of ranking techniques on the MovieLens dataset.

tries very accurately predicted with a MSE of 0.52. Robust SCOAL performs as good as row-col ranking, while block ranking, which ranks predicted values by co-cluster MSE, results in a flatter plot, possibly due to its more constrained ranking mechanism.

Figure 3 illustrates the certainty lift, computed according to Eq. (7), for different prediction techniques. The SCOAL based approaches as well as Global Rank provide substantial improvement over the baseline and also do much better than Global Var.

Figure 3: Certainty Lift of different prediction approaches on the MovieLens dataset.

10.2 ERIM Marketing Dataset

In this section we use the publicly available ERIM dataset⁵ consisting of household panel data collected by A.C. Nielsen, which is well known in the marketing community and has been used by several researchers [14, 17]. This dataset has purchase information for six product categories over a period of 3 years from 1985-1988 for households in Sioux Falls, South Dakota and includes household demographics and product characteristics. We preprocess the data as discussed in [7], resulting in a data matrix of 1714 households and 121 products from 6 product categories (ketchup, tuna, sugar, tissue, margarine and peanut butter). The covari-

³The data can be downloaded from http://www.grouplens.org/system/files/ml-data.tar__0.gz

⁴The value of the regularization parameter (λ), selected by cross validation, came out to be 0 for a global ridge regression model. Hence, the standard error is equivalent to that given by Eq. (1).

⁵URL: <http://www.gsb.uchicago.edu/kilts/research/db/erim/>

ates include 6 household attributes - income, number of residents, male head employed, female head employed, total visits and total expense and 3 product attributes - market share, price and the number of times the product was advertised. The matrix cell values are the number of units of a product purchased by a household, aggregated over the time the household was tracked.

The data matrix is extremely sparse, with 74.86% of the values being 0. The distribution of the response values is also very skewed. 99.12% of the values are below 20, while the remaining values are very large and range upto around 200. We consider these few, large values (> 20 # of units purchased) as outliers with respect to the rest of the data.

Since the products are from 6 different product categories, the product attributes and the number of units of the product purchased vary widely from one category to another. We hence standardize the product attributes and the matrix cell values as done in [7] to make them comparable across categories.

Figure 4 compares various ranking techniques on the ERIM dataset. The SCOAL techniques are run with $k = 4$ and $l = 4$ and Robust SCOAL is run with $s_r = 1700$, $s_c = 150$. The ridge regression parameter λ for the SCOAL techniques is set to 0.98 and is selected by cross validation. The λ selected for a global ridge regression model is 0.84. Note that the MSE is computed on the original data, by back transforming the standardized values. The trend of the results is similar to that in Figure 2. Global Rank and the SCOAL (row-col and block ranking) approaches that use the dyadic data structure do significantly better than the resampling, the Bayesian linear regression and the standard error estimation approaches. Although the average prediction error of row-col ranking and Global Rank on the entire test set is almost the same, one can observe that over different fractions of the test entries predicted, row-col ranking is slightly better than Global Rank. This is possibly due to better estimation of prediction errors caused by an improvement in the overall fit of the co-cluster models as compared to a single global model. Figure 5 compares the certainty lift for different prediction techniques. Observe that row-col ranking gives an improvement of almost 25 times on 10% of the test set.

Figure 4: Comparison of ranking techniques on the ERIM dataset.

Linear regression based on the squared error cost function is not very robust to outliers and a few outliers could

Figure 5: Certainty Lift of different prediction approaches on the ERIM dataset.

completely skew the model results. While the use of ridge regression models mitigates the effect of outliers to some extent, there is still some sensitivity to outliers, resulting in a large MSE on the entire test set. We now evaluate the ability of Robust SCOAL to automatically identify and prune away outliers, where an outlier is defined as a response value (# of units purchased) > 20 . Table 1 displays the percentage of the total number of outliers detected and discarded by Robust SCOAL for different sizes of the data modeled, specified by the s_r and s_c values. k and l are set to 4 for Robust SCOAL. The results are averaged over 10 random 80 – 20 % train-test splits of the data. The last row of the table is for the case where $s_r = m$ and $s_c = n$, i.e., no rows and columns are discarded. One can observe that the % of outliers discarded is consistently much larger relative to the % of data discarded. The ratio of the % of outliers discarded to the % of data discarded (last column) is hence high, indicating that Robust SCOAL is able to selectively prune away outliers. It is interesting that discarding even a few rows and columns results in throwing away a large fraction of the outliers. This results in a dramatic reduction in the prediction error on the selected subset of test points, which is still a substantial fraction of the missing values, e.g., the MSE on 85.9 % of the test points is 8.225 when $s_r = 1573$ and $s_c = 115$, as compared to a MSE of 15.787 on the entire test set.

s_r	s_c	% outliers discarded	% data discarded	Ratio
900	90	98.414	61.549	1.599
1068	96	96.128	51.330	1.873
1236	102	91.359	40.154	2.275
1405	109	83.889	27.303	3.073
1573	115	70.770	14.130	4.949
1741	121	0	0	-

Table 1: Outlier pruning by Robust SCOAL for varying s_r and s_c values on the ERIM dataset.

10.3 MBA Student Course Dataset

This dataset arises from a collaborative filtering application that deals with predicting course choices made by masters students at a large Midwestern University. The aim is to use the information of known course choices of students, and additional student and course information to predict unknown choices. These predictions can be used to recommend

the right courses for students to take in the future. The data includes a matrix of 326 students vs. 32 courses with class labels = 1 if the student took the course and -1 otherwise. Each student has attributes including the student's career aspiration and undergraduate degree. The course attributes include the department offering the course, the course evaluation score and a binary variable indicating whether the course is quantitative.

Since this is a classification problem, we use logistic regression models for prediction. Figure 6 compares the ranking of the predicted values obtained by a collection of local models, learnt using SCOAL with that obtained by a single global model. For a missing entry z_{ij} in the student-course matrix Z , both approaches estimate $P(z_{ij} = 1|\mathbf{x}_{ij})$. The missing entries are ranked in descending order of $\max(P(z_{ij} = 1|\mathbf{x}_{ij}), P(z_{ij} = -1|\mathbf{x}_{ij}))$. The y-axis in Figure 6 is the log loss averaged over the predicted entries. SCOAL is run with $k = 2, l = 2$. As in the regression case, since the local models represent the data better as compared to a single global model, they result in a better ranking.

Figure 6: Comparison of ranking using SCOAL vs. a single classification model on the MBA dataset.

11. CONCLUDING REMARKS

The task of selecting the top n predictions based on their certainty has not been adequately explored, although it is very naturally applicable in a wide range of practical scenarios. In this paper we propose novel approaches for mining the most certain predictions in regression and classification problems involving large, heterogeneous data with a dyadic structure. Our approaches are based on the key idea of modeling the data by a collection of multiple, localized models learnt by exploiting the dyadic structure. As illustrated in Section 10, the proposed ranking techniques give superior results on a variety of real life datasets as compared to classical uncertainty estimation approaches based on global models. Of the proposed ranking methods, Robust SCOAL and block ranking are more useful than row-col ranking from a practical viewpoint since they enable selection of structured subspaces of the data, which is more actionable. Another advantage of our multiple models based approach is that the selection of “certain” predictions can be done at varying resolutions, providing application specific flexibility as well as interpretability.

The idea of “certainty lift” can be useful in several applications. In addition it can be used as a basis for developing active learning techniques that are radically different from traditional uncertainty sampling based approaches [1]. Another aspect worth exploring is the use of robust error func-

tions, which suitably weigh errors from different entries in the data matrix differently. This will provide a “soft” selection of the subset of the data to be modeled, which may be more flexible than pruning entire matrix rows/columns as done by Robust SCOAL.

12. REFERENCES

- [1] N. Abe and H. Mamitsuka. Query learning strategies using boosting and bagging. In *ICML '98*, pages 1–9, 1998.
- [2] D. Agarwal and S. Merugu. Predictive discrete latent factor models for large scale dyadic data. In *KDD '07*, pages 26–35, 2007.
- [3] C. Bishop and C. Quazaz. Bayesian inference of noise levels in regression. In *ICANN '96*, pages 59–64, 1996.
- [4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [5] M. Blum, R. Floyd, V. Pratt, R. Rivest, and R. Tarjan. Time bounds for selection. *J. Comput. Syst. Sci.*, 7(4):448–461, 1973.
- [6] G. Cawley, N. Talbot, and O. Chapelle. Estimating predictive variances with kernel ridge regression. In *Machine Learning Challenges '06*, pages 56–77, 2006.
- [7] M. Deodhar and J. Ghosh. A framework for simultaneous co-clustering and learning from complex data. In *KDD '07*, pages 250–259, 2007.
- [8] M. Deodhar, G. Gupta, J. Ghosh, H. Cho, and I. Dhillon. Hunting for coherent co-clusters in high dimensional and noisy datasets. In *Foundations of Data Mining, ICDM '08 Workshop*, 2008.
- [9] J. Fox. *Applied Regression Analysis, Linear Models, and Related Methods*. Sage Publications, 1997.
- [10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.
- [11] T. Heskes. Practical confidence and prediction intervals. In *NIPS '97*, pages 176–82, 1997.
- [12] J. Leonard, M. Kramer, and L. Ungar. Using radial basis functions to approximate a function and its error bounds. *IEEE Transactions on Neural Networks*, 3:4:624–627, July 1992.
- [13] J. Johnston. *Econometric Methods*. New York: McGraw-Hill, 1963.
- [14] B. Kim and M. Sullivan. The effect of parent brand experience on line extension trial and repeat purchase. *Marketing Letters*, pages 181 – 193, 1998.
- [15] J. Kleinberg and E. Tardos. *Algorithm Design*. Pearson Education, 2006.
- [16] A. Nix and A. Weigend. Estimating the mean and variance of the target probability distribution. In *Neural Networks '94*, pages 55–60, 1994.
- [17] P. Seetharaman, A. Ainslie, and P. Chintagunta. Investigating household state dependence effects across categories. *JMR*, pages 488 – 500, 1999.
- [18] D. Shepard. *The New Direct Marketing: How to Implement A Profit-Driven Database Marketing Strategy*. McGraw-Hill, 1999.
- [19] D. Shrestha and D. Solomatine. Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks*, 19(2):225–235, 2006.