Missing Data Estimation in Microarrays Using Multi-Organism Approach

Marcel Nassar and Hady Zeineddine

Progress Report: Data Mining Course Project, Spring 2008

Prof. Inderjit S. Dhillon

April 02, 2008

I. Introduction

Microarrays measure gene expression levels and provide data samples that are sparse compared to the number of the interacting genes that are to be modeled [1]. In addition to that, due to insufficient resolution, image corruption, robotic methods and dust on slides in the microarray some values in the gene expression data are missing [2]. Unfortunately, many algorithms (hierarchical clustering, K-means clustering,$\cdots$) for gene expression analysis require complete matrix of gene array values as inputs. The effectiveness of such algorithms degrade rapidly even with few missing values [2]. Due to the sparsity of data, the algorithms for estimating missing values in the DNA microarrays will not be, in many cases, reliable and will lead to poor estimation results.

Our main work in this project is to exploit the similarity between genes across different organisms to overcome the data sparsity problem and obtain more accurate missing value estimates. Therefore, we try to use mutual information between data sets, corresponding to different organisms, to enhance learning the structure and properties in each data set individually, and consequently reach a better missing data estimation. We will try to approach this problem in two directions: (1) using a Bayesian-based method and (2) enhancing existing estimation methods. In the first approach, we use Bayesian networks to model the interactions between genes, apply multi-task learning algorithms on similar gene data sets to learn the Bayesian structure accurately [3], and then make use of the learned relations to estimate the missing data. In the second approach, we use similarity between the data sets to enhance the existing single-set estimation algorithms by imposing new constraints and/or modifying the estimation process. The rest of the report is organized as follows: Section II gives the biological background for the similarity between genes of different organisms. Section III presents some known effective missing-value-estimation algorithms and discusses the options for exploiting multi-organism data-sets in the context of the K-nearest-neighbor based algorithms. Section IV discusses the Bayesian-based method, and Section V concludes the report.

II. Biology Background

Recent studies have revealed highly conserved proteins that exhibit similar expression patterns in different organisms, and play a role in gene regulation. Therefore, this evolutionary information can be used to refine

regulatory relationships among genes, which are estimated from gene expression data and bias the learning process. In this project, we will try to use this fact to improve missing value estimators for microarray data.

In genetics, sequence homology implies that the given sequences share a common ancestor. Orthology is a type of sequence homology that results from a speciation event; i.e. a parent species diverging into two new species. Orthologous genes usually preserve protein expression and regulation functions and can be used to quantify the evolutionary information between two organisms [4], [5]. The idea of utilizing evolutionary information to build gene networks using a multiorganism approach was used in [5], where two organisms A and B were considered, with D_A and D_B representing their gene expression data.

We have researched available datasets for orthologous genes and found two good candidates for this work at http://genome-www.stanford.edu/cellcycle/ [6] for yeast genes and genome-www.stanford.edu/Human-CellCycle/Hela/ [7] for human genes. It is interesting to note that even yeast and humans have some genetic pathways common. This highlights the importance of this approach in exploiting this type of dependencies between organisms to improve performance and accuracy.

III. Single-set Estimation: Overview and Possible Improvements

Several single set estimation have been suggested in the literature for missing value estimation in microarray data, including K-nearest-neighbor (KNN) imputation, Singular Value Decomposition Imputation, least square imputation, Bayesian PCA, and Collateral missing value imputation (CMVE). All these techniques try to exploit correlation between data to reduce prediction errors, and thus their efficiency is directly related to the underlying, highly unknown relations between genes. The way in which similarity between different data sets can be exploited is thus related to how every technique makes use of the data of the different genes to estimate a missing value.
We will work on *SVDimpute*, *KNNimpute*, and *CMVE* algorithms. In this report, we take $\mathbf{H}_{M\times N}$ to be the matrix containing the microarray data for M genes at N instances.

KNNimpute: If the gene A has a missing value in experiment 1, this method would find K-other genes, which have a value present in experiment 1, with expression most similar to A in experiments 2-N. A weighted average of values in experiment 1 from the K-closest genes is then used as an estimate for the missing value in gene A. The best candidate metric for gene similarity in this context was shown to be the Euclidean distance, obtained after log-transforming the data (and thus reducing the effect of the outliers). [2]

CMVE: Similar to the KNNimpute, the CMVE selects K-nearest-neighbors to each gene according to the covariance metric between two gene expressions $\sum_{k=1}^{N}(\mathbf{H}_{jk} - E(\mathbf{H}_j)(\mathbf{H}_{ik} - E(\mathbf{H}_i))$ where $E(\mathbf{H}_i)$ is the average of ith row. Then, three parallel estimations of gene missing expression $\mathbf{H}_{ij}$ are done by using the K-nearest-gene values as an input. The first estimate uses the least-square-error linear regression method, while the other two use the non-negative least-square (NNLS) algorithm. A weighted average of the three estimates is then taken to reach the final prediction. [8]

SVDimpute: In this method singular value decomposition is employed to obtain a set of mutually orthogonal expression patterns that can be linearly combined to approximate the expression of all genes in the data set. In the SVD $\mathbf{H} = \mathbf{U\Sigma V^T}$, the gene expression pattern of each gene (i.e. row of $\mathbf{H}$) is a linear combination of the columns of $\mathbf{V}$, called eigengenes, where the contribution of each eigengene to the expression pattern in $(i.e. row of \mathbf{H}$ is quantified by the corresponding singular value in $\mathbf{\Sigma}$ (being multiplied by it). Therefore, the k eigengenes with the highest singular values are selected, and each gene A is regressed against them and the coefficients are used to reconstruct missing values A from a linear combination of the k-eigengenes. [2]

A. Exploiting similarity across data sets

The similarity between different data sets can be used to overcome the data sparsity problem, therefore, the performance of multi-organism approach on each of the techniques mentioned above is upper bounded by the performance of the technique on a data set with a size that equals the sum of the different data-set sizes, i.e. 100% similarity. The techniques presented above have two major stages: first, the k nearest-neighbor genes or principal pattern components for a gene expression are found and then a regression is applied to predict the relation between the "k" components and the gene. Mutual information across data sets enhance the selection process performance, while more care is needed regarding the regression stage, since it is more sensitive to differences between organisms. These differences may turn to an additional source of noise in the regression stage. The details are shown below.

In the search for K-nearest-neighbors of gene A, we try to deduce a similarity measure $0 \leq f_{P,R} \leq 1$ between each pair P and R of datasets, and then do one (or more) of the following: Denote K_R as the set of genes that is K-nearest to the gene A in the data set R, and let D be the number of the different data sets.

1. Find K_R under the $\forall 1 \leqslant R \leqslant D$, such that $|K_P \cap K_R| \geqslant F(f_{P,R}, K).K \ \forall 1 \leqslant P \leqslant D$, where $F(f_{P,R}, K)$ is an increasing function of $f_{P,R}$. If the K-nearest-genes are weighted, the condition can be slightly modified according to the details of the algorithm, for example $W(K_P \cap K_R) \geqslant F(f_{P,R}, K)$ where $W(.)$ is the sum of the weights of the elements of the set and $W(K_R) = 1$.
2. Compute the final distance between 2 genes A and B in a data set R, as a weighted average of the distance $d_{AB,P} \ \forall 1 \leqslant P \leqslant D$.

In the case of *SVDimpute*, work on how to exploit mutual information between data sets is yet to be done. Using across-set information in the regression stage may cause noisy observations to happen since it is highly probable that the underlying functions have evolved with time and changed, quantitatively, their interaction style. It is left to simulations to show whether the learned functions will change drastically across different organisms. However, in the choice of computational model (e.g. basis functions, degree,$\cdots$)for the regression, a very intuitive requirement is to minimize the sum of error values across all considered organisms.

A major part any multi-organism approach is to estimate and quantify the similarity of data sets. The choice of how to quantify the similarity depends on the context in which it is used. The resulting estimate is a result of "lab"

verified information and also by using learning techniques (for example use SVD to test similarity and then apply KNN) and applying other measure (the Kullback-Leibler distance for example).

IV. Bayesian-Based Method

The methods described in the previous section assume no specific structure for genes relations, and thus, they make no use of the verified fact that genes interactions involve control/regulation mechanisms and that, consequently, some genes control the expression levels of others. To reflect these regulation mechanisms, gene interactions are modeled as Bayesian networks, described below. In the context of our problem, we use multiple data sets to learn the structure of the underlying network and then predict the gene missing values accordingly.

A. Bayesian networks

A Bayesian network is a graphical model for representing conditional independencies between a set of random variables. This representation consists of two components, G and $Theta$. G is a directed acyclic graph (DAG) whose n nodes (genes) correspond to the random variables $X_1, X_2, ..., X_n$. Θ, describes the conditional probability distribution of each node, given its parents [1] $P_a(X_i)$, i.e. $P(X_i|P_a(X_i)), X_i \in G$. A basic assumption in Bayesian networks is that it must respect the Markov assumption: Each variable X_i is independent of its non-descendants, given its parents in G. This property allows the joint probability distribution $P(X_1, X_2, ..., X_n)$ to be decomposed into the product form

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \tag{1}$$

Modeling a gene network as a Bayesian network is motivated by the fact that causal relations between genes can be described qualitatively through the presence of edges and quantitatively through the conditional probability distributions. Biological hidden factors and the imperfect modeling (i.e. gene interactions may be more complex than Bayesian networks) prohibit the prediction of deterministic relations between the genes and act as a noise source that results in the probabilistic nature of the relations between the genes. Bayesian gene networks can be extended to incorporate temporal order to account for a direct causal effect of one gene on anther through different time stages.

The algorithms developed to learn the Beyesian network structure from the data can be classified into two categories: search and score algorithms and information theoretic algorithms. Search and score algorithms are more robust under data sparsity conditions, and thus are used to learn gene networks (specifically the K2 algorithm). Several multi-organism approaches have been suggested to enhance the learning structure by making use of similar data sets. We try here to follow similar approach to enhance the missing values estimation.

[1] all nodes that have an edge directed towards X_i in G

B. Bayesian-based estimation

Based on the Bayesian model assumption, we will try the following estimation methodology:

1. Discretize the observed gene expression levels as a preprocessing step. The number of possible discrete levels is kept low due to the saprsity of data. A relatively high number of the discrete levels would lead to poor performance of the structure and parameter learning algorithms.
2. Apply multi-task/multi-organism learning algorithms to build a gene network of each organism.
3. For every gene network:

   For each gene A , Apply one of two approaches:

   a. Apply linear regression to find the expression of A as a function of the expression levels of its parents, using the microarray data as training data. A major subtlety in this approach is the probabilistic nature of the relation between the A and its parents, which makes same parent-value yield different child values according to some probability distribution distribution. This probabilistic relation can be attributed to hidden unknown factors in the controlling mechanism. To account for this hidden factor, we add a new node λ_A to the list of parents of A. Our regression problem is then slightly changed into: Find the coefficients of the basis functions $\mathbf{w}$ and the pattern expression vector (i.e. the row of values of λ_A that should have appeared in $\mathbf{H}$ had λ_A not been hidden) of λ_A that minimizes the error function(least square error for example). To predict a missing value of A, we first sample λ_A using the distribution of its estimated pattern expression vector.

   b. Apply Bayesian standard parameter learning using data from all gene networks that have (nearly)identical parent set of A. To predict the value of A, we sample from the learned probability distribution given the values of the parents of A. The predicted value v is discrete, and is replaced with the average of the expression levels of A that were discretized to v in step 1.

If the controlling function by which the parents regulate the child's expression is continuous in nature, the first approach would be more compatible. If the control mechanism is digital, i.e. changes in the input values produce little change in the output, then the problem is close to a classification problem, and the second approach is more useful. The latter assumption can be justified by the fact that digital controlling mechanism is immune to noise and is thus prevents the propagation of possible errors during the functioning of the genes in the cells.

V. Conclusion

In conclusion, the multiorganism approach shows promise in terms of improving the accuracy of the estimates for the missing data problem. We have explored possible approaches of taking advantage of the multiorganism similarity. In addition to that, we identified possible datasets to be used in our simulations. We will be using BNT Matlab toolbox developed by Murphy to perform our simulations. The remaining work to be done can be summarized by modifying the mentioned algorithms to take advantage of the similarity and perform various simulations on synthetic data and real datasets in order to verify the validity and usefulness of this approach.

REFERENCES

[1] N. Friedman, M. Linial, I. Nachman, D. Peer, "Using Bayesian Networks to Analyze Expression Data," *Journal of Computational Biology, Vol. 7, No. 3/4, pp. 601-620*, 2000.

[2] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, 2001.

[3] M. Nassar and R. Abdallah and H. A. Zeineddine and E. Yaacoub and Z. Daey, "A New Multitask Learning Algorithm Applied to Multiorganism Gene Network Estimation," *ISIT*, 2008.

[4] S. Bergmann, J. Ihmels, and N. Barkai, "Similarities and differences in genome-wide expression data of six organisms," *PLOS Biology*, vol. 2, no. 1, p. 0085, January 2004.

[5] Tamada et al., "Utilizing Evolutionary Information and Gene Expression Data for Estimating Gene Networks with Bayesian Network Models," *Journal of Bioinformatics and Computational Biology, Vol. 3, No. 6, pp. 1295-1313*, 2005.

[6] Spellman et al., "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell, Vol. 9, pp. 3273-3297*, December 1998.

[7] Whitfield et al., "Identification of Genes Periodically Expressed in the Human Cell Cycle and their Expression in Tumors," *Molecular Biology of the Cell, Vol. 13, pp. 1977-200*, June 2002.

[8] M. Sehgal, I. Gondal, and L. Dooley, "Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data," *Bioinformatics*, vol. 11, no. 10, 2005.