

HICCUP: Hierarchical Clustering Based Value Imputation using Heterogeneous Gene Expression Microarray Datasets

Qiankun Zhao¹, **Prasenjit Mitra**², Dongwon Lee², and
Jaewoo Kang³
¹AOL Labs, China,
²The Pennsylvania State University, U.S.A.
³Korea University, Korea

11/6/2007

Prasenjit Mitra

1

Outline

- Motivation
- Algorithm
- Experiments
- Conclusion

11/6/2007

Prasenjit Mitra

2

The Problem

- Given a gene expression data matrix
- Some values are missing
- Impute the missing values

11/6/2007

Prasenjit Mitra

3

Motivation

- Inferring gene function
- Gene network discovery
- Drug discovery
- Patent diagnosis

11/6/2007

Prasenjit Mitra

4

Missing Values

- Some missing values may be crucial
 - Other values may imply whatever is being inferred from the missing gene expression value
 - For simplicity, rules may not keep the association between the other genes
 - Rules cannot account to keep redundancy for all missing data
 - Number of rules will be very large & redundant
 - Other data from which a missing value may be imputed may not be available or private

11/6/2007

Prasenjit Mitra

5

Related Work

- Local
 - KNN
 - [Troyanskaya, et al., Bioinformatics,'01]
 - Using similar genes
- Global
 - Basis eigen-genes using SVD
 - Regression models on the eigen-genes
 - [Bo, Dysvik, Jonassen, Nuc. Acids Res.,'04]
- Dependent upon datasets

11/6/2007

Prasenjit Mitra

6

Limitations

- Use single dataset
 - Small samples, large number of genes
 - Hard to build general imputation model
 - Covering biological properties
 - Sample space, gene space
 - Focus on gene space, ignore sample space
 - Unselectively use all samples of different types
 - Cancer + non-cancer samples

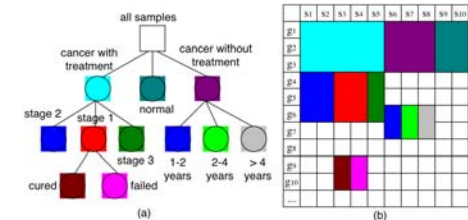
Our Model

- Aggregate samples from multiple microarray datasets
 - Careful
- Hierarchical clustering
 - reflect correlations in the sample space among the heterogeneous datasets
- Correlations in the genespace
 - Association rules within individual clusters

Value Imputation

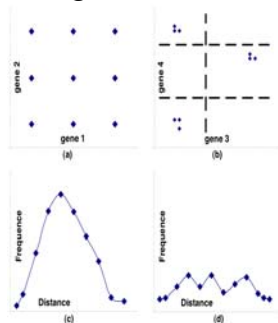
- Find the most similar cluster
- Use association rules within the cluster to perform imputation
- Similarity
 - Entropy-based metric for each level of cluster hierarchy

Hierarchy & Microarray Data



Clustering

- Clustering algorithm from Tang, Jhang, Pei, 2003
 - Extended it to perform hierarchical clustering
- Gene selection ~ Feature selection
 - Filter approach
 - Different point to point distance histograms than features (gene3, gene4) that do not form obvious clusters (gene1, gene2)
- Choose the set of genes with the lowest entropy measure (see paper)
 - These genes appear as the children of the root
 - Repeat



Discretization of Gene Expression Values

- Different discretization scheme for each siblings in the cluster
 - Effectively reflect the difference between different siblings

Discriminative Gene-values Sets

- Gene-values Set: a set of genes and their value ranges
- value ranges occurred only in this sample cluster but not in other clusters, especially, the sibling clusters
 - No overlap with its parent sample cluster
- Frequent: at least t % of the samples have values in the range for the genes in the set
- Different gene-values sets have different discriminative power
 - Similar to tf/idf
 - Importance to number of times it appears in a cluster
 - Appearing in multiple clusters => have less discriminative power

11/6/2007

Prasenjit Mitra

13

Top-k Gene Values Set Selection

- Any of the samples in the cluster can be modeled by at least one of the gene-value sets
- Top-k gene values sets that cover all samples in the cluster as the discriminative and covering gene-values sets
- Allow some redundancy to allow for missing values
- Frequent itemsets mining
 - Frequent gene-values sets

11/6/2007

Prasenjit Mitra

14

Cluster Selection

- Each sample cluster and target sample represented as a vector of discriminative and covering gene-value sets
 - Matching score: use cosine similarity between the two vectors
- Hierarchical matching
 - Choose cluster with the largest matching score at this level

11/6/2007

Prasenjit Mitra

15

Association Rule Mining

- Number of association rules for a single cluster in the hierarchy: very large
- Rhs: target gene
- Lhs: genes with known values that appear in the discriminative frequent gene-values sets
- Ranked according to confidence value
- Highest ranked rule used to estimate missing value
- If no discriminative frequent gene-value sets exist, use KNNImpute or row average imputation within this specific cluster of samples

11/6/2007

Prasenjit Mitra

16

Experiments

- PROSTATE CANCER DATASETS.
- Name Probe# #Normal #Tumor
#Total
- Singh 12600 50 52 102
- Welsh 12626 9 24 33
- LaTulippe 12626 3 23 26
- Normalized values

11/6/2007

Prasenjit Mitra

17

Error Computation

- Randomly pick and drop values
- Normalized mean-square error between the imputed value and the original value

$$y_{\text{impute}} - y = \begin{cases} 0, & \text{if } y \text{ is within the interval } y_{\text{impute}} \\ \text{Avg}[y_{\text{impute}}] - y, & \text{otherwise} \end{cases}$$

11/6/2007

Prasenjit Mitra

18

Legend

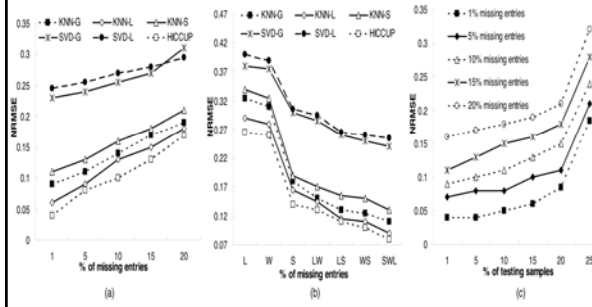
- *KNN-G* denotes the global KNNImpute that uses the whole set of integrated microarray dataset and the gene similarities are calculated using the whole set of genes.
- *KNN-L* denotes the local KNNImpute approach that uses only relevant clusters of samples, and gene similarities are calculated only using the discriminative gene-values sets of these clusters.
- *KNN-S* is similar to *KNN-G* but uses only single dataset for samples and the gene similarities are calculated using the whole set of genes.
- *SVD-G* represents the global SVD approach that uses the eigen-genes extracted from integrated microarray dataset.
- *SVD-L* is the local SVD approach that uses eigen-genes extracted from the corresponding single dataset.
- *HICcup* refers to our proposed imputation method, where only samples in the relevant cluster are used and the imputations are obtained by association rule, KNNImpute, and Row Average.

11/6/2007

Prasenjit Mitra

19

Results(1)

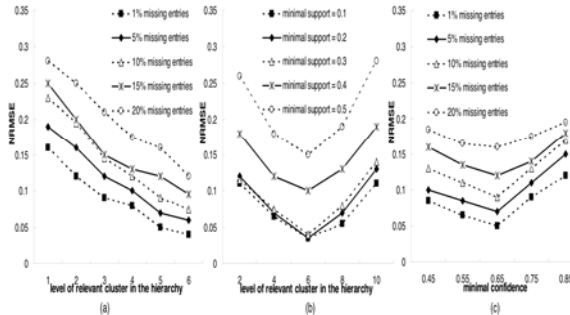


11/6/2007

Prasenjit Mitra

20

Results (2)



11/6/2007

Prasenjit Mitra

21

Conclusions

- Hierarchical clustering improves quality
- Integration of multiple datasets improves quality
- Association rule-based imputation in hierarchical setup better than local and global approaches

11/6/2007

Prasenjit Mitra

22