P-tree Classification of Yeast Gene Deletion Data

Amal Perera, Anne Denton, Pratap Kotala, William Jockheck
North Dakota State University
C.S. Department, IACC 258

Willy Valdivia Granda
North Dakota State University
Plant Pathology Department
Fargo, ND, 58105-5012

William Perrizo
North Dakota State University
Computer Science Department
Fargo, ND, 58105-5164

{amal.perera, anne.denton, pratap.kotala, william.jockheck, willy.valdivia, william.perrizo}@ndsu.nodak.edu

ABSTRACT

Genomics data has many properties that make it different from "typical" relational data. The presence of multi-valued attributes as well as the large number of null values led us to a P-tree-based bit-vector representation in which matching 1-values were counted to evaluate similarity between genes. Quantitative information such as the number of interactions was also included in the classifier. Interaction information allowed us to extend the known properties of one protein with information on its interacting neighbors. Different feature attributes were weighted independently. Relevance of different attributes was systematically evaluated through optimization of weights using a genetic algorithm. The AROC value for the classified list was used as the fitness function for the genetic algorithm.

Keywords

P-tree, Data mining, Genetic Algorithm, Genomics, Bioinformatics.

1. INTRODUCTION

Genomics data poses many challenges and requires smart data mining techniques to unravel the secrets. The data mining competition held in conjunction with the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [1] was based on genomics data that had many interesting aspects such as multi-valued attributes, many null values, hierarchies within the attributes, a repository of unstructured information in the form of abstracts, an interaction table that introduces graph-like connectivity, and a very small probability of the class label "positive" (1.3% and 2.8% respectively). These properties are not commonly found in standard data sets such as the majority of those in the UCI Machine Learning Repository [8]. We accounted for multi-valued attributes and null values by using a P-tree-based [2,3,5,6] bit-vector approach, in which the presence of a particular attribute value was represented as a 1-bit. In this representation "unknown" is equivalent to the absence of a meaningful attribute value. Hierarchies were represented as separate attributes. Matches in different levels were thereby treated as multiple matches. Connectivity information provided by the interaction information was also used. The small probability of class label "positive" meant that accuracy was not a good measure of the quality of classification. Instead we followed the example of Dr. Mark Craven, using AROC characteristic to optimize the performance of our algorithm [7].

2. DATA PREPROCESSING

As the first step, we pre-processed the data to account for redundancies and to integrate information from the abstracts. An example of redundancy was that localization information occurred in both the sub-cellular localization (sub-hierarchy of the function-hierarchy) and in the localization attribute. We extracted sub-cellular localization from the function hierarchy. We found that treating it as a separate hierarchy, thus as an additional feature of the genes improved results. It is important to note that our algorithm does not require us to decide on the importance of the features. Other data cleaning efforts included recognizing "function unknown" entries as equivalent to missing information, and identifying duplicate names in lower levels of different hierarchies as distinct.

We generated additional features from abstracts based on observations made while classifying and also based on the nature of the experiments from which the data was derived. Relevance of these additional features was determined through the optimization step of our model. We not only included the properties of the genes that were to be classified in our study, but also investigated properties of genes to which an interaction exists. A successful example of such an indirect feature is the interactions with the essential (lethal) genes. Based on the nature of the task, essential genes could not, themselves, be change/control genes. Yet an interaction with these genes did indicate an increased likelihood of change/control.

3. PTREE DATA REPRESENTATION

The input data was converted to P-trees [2,3,5,6]. P-trees are a lossless, compressed, and data-mining-ready data structure. This data structure has been successfully applied in data mining applications ranging from Classification and Clustering with K-Nearest Neighbor, to Classification with Decision Tree Induction, to Association Rule Mining [2,3,5]. A basic P-tree represents one attribute bit that is reorganized into a tree structure by recursively sub-dividing, while recording the predicate truth value regarding purity for each division. Each level of the tree contains truth-bits that represent pure sub-trees and can then be used for fast computation of counts. This construction is continued recursively down each tree path until a pure sub-division is reached that is entirely pure (which may or may not be at the leaf level). The basic and complement P-trees are combined using boolean algebra operations to produce P-trees for values, entire tuples, value intervals, or any other attribute pattern. The root count of any pattern tree will indicate the occurrence count of that pattern. The P-tree data structure provides the perfect structure for counting patterns in an efficient manner.

The data representation can be conceptualized as a flat table in which each row is a bit vector containing a bit for each attribute of each gene. Representing each attribute bit as a basic P-tree generates a compressed form of this flat table. Hierarchical information is represented using a separate set of attribute bit columns for each level. For example to represent the information for protein-class we used 23 sets of bit columns at the highest level. “Molecular chaperone” at the highest level of the protein class hierarchy requires 15 bit columns to recursively identify all possible categorical attributes within 2 sub-levels. Experimental class labels "change," "control," and the other binary attributes such as "lethal gene" were each encoded in a single bit column. Protein-interaction was encoded using a bit column for each possible gene in the data set, where the existence of an interaction with that particular gene was indicated with a truth bit.

4. CLASSIFICATION

Similarity-based classification methods are a generalization of minimal distance methods, which form the basis of several machine learning and pattern recognition methods. Classification success depends on adaptive parameters and procedures used in the construction of the classification model. AROC evaluation requires genes to be ordered by their likelihood of being classified into a particular partition. Gene similarity and gene importance were quantified to rank-order the list of test genes. Gene similarity was quantified based on a weighted sum of matching features. We quantify similarity of a particular test gene with the collection of training genes in that particular partition (for example "change") without regard to similarity level to the rest of the partition ("control" or "no change"). The distribution of attribute values for change genes suggested that not only was similarity to a change gene an indication of change, but the fact that a gene had any listed function. This leads to the concept of importance as an attribute. Importance of a gene was calculated according to the count of the number of property attributes available for that particular gene, total number of interactions and the number of interactions with lethal genes.

The following equations summarize the computation, where **Rc**: root-count, **W**: weight, $\mathbf{P_x}$:P-tree for attribute x, **Atr**: attribute count, **ptn**: class-partition, **Im**: gene-importance, **Lth**: lethal gene, **g**: test gene to be classified, **f**: feature, **ClassEvl**: evaluated value for classification, $\wedge$: P-tree AND operator.

$$\mathrm{ClassEvl}_{ptn}(g) = \left(\sum_{f}^{f \in g} W_f \times Rc(P_{ptn} \wedge P_f) \right) + W_{Im} \times Im(g)$$

$$Im(g) = W_{Atr} \times Atr(g) + W_{Int} \times Rc(P_g) + W_{Lth} \times Rc(P_g \wedge P_{Lth})$$

For each feature attribute of each test gene, *g*, the count of matching features for the required partition was obtained from the root-count by ANDing the respective P-trees. We can obtain the number of lethal genes interacting with a particular gene, *g*, with one P-tree AND operation. It is possible to retrieve the required counts without a database scan due to vertical partitioning.

Due to the diversity of the experimental class label and the nature of the attribute data, we need a classifier that would not require a fixed importance measure on its features, i.e., we needed an adaptable mechanism to arrive at the best possible classifier. In our approach we optimize (column based scaling) the weight space, *W*, with a standard Genetic Algorithm (GA) [4]. The set of weights on the features represented the solution space that was encoded for the GA. The AROC value of the classified list was used as the GA fitness evaluator in the search for an optimal solution.

5. RESULTS AND LESSONS

This work resulted in both biological and data mining related insights. The systematic analysis of the relevance of different attributes is essential for a successful classification. We found that the function of a protein did not help to classify the hidden system. Sub-cellular localization, which is a sub-hierarchy of the function hierarchy, on the other hand, contributed significantly.

Furthermore, it was interesting to note that quantitative information, such as the number of interactions, played a significant role. The fact that a protein has many interactions may suggest that the deletion of the corresponding gene would cause changes to many biological processes. Alternatively it could be that a high number of listed interactions is an indication of the fact that previous researchers have considered the gene important and that it therefore is more likely to be involved in further experiments. For the purpose of classification we did not have to distinguish between these alternatives.

In summary we found that our systematic weighting approach and P-tree representation allowed us to evaluate the relevance of a rich variety of attributes. This included attributes we obtained in relational form as well as attributes we derived from abstracts, through counting, and as properties of attributes that could be reached through interactions.

We acknowledge Dr. Mark Craven for organizing this competition and Dr. Steven W. Meinhardt, Dept. of Biochemistry, NDSU, for his time and effort in helping us make biological sense of the data.

6. REFERENCES

[1] Craven, M., "KDD Cup 2002", http://www.biostat.wisc.edu/~craven/kddcup/.

[2] Ding, Q., Ding, Q., Perrizo, W., "ARM on RSI Using P-trees," Pacific-Asia KDD Conf., pp. 66-79, Taipei, May 2002.

[3] Ding, Q., Ding, Q., Perrizo, W., "Decision Tree Classification of Spatial Data Streams Using Peano Count Trees," ACM SAC, pp. 426-431, Madrid, Spain, March 2002

[4] Goldberg, D.E., Genetic Algorithms in Search Optimization, and Machine Learning, Addison Wesley, 1989.

[5] Khan, M., Ding, Q., Perrizo, W., "KNN on Data Stream Using P-trees," Pacific-Asia KDD, pp. 517-528, Taipei, May 2002.

[6] Perrizo, W., "Peano Count Tree Lab Notes," CSOR-TR-01-1, NDSU, Fargo, ND, 2001.

[7] Provost, F., Fawcett, T.; Kohavi, R., "The Case Against Accuracy Estimation for Comparing Induction Algorithms", 15th Int. Conf. on Machine Learning, pp 445-453, 1998.

[8] UCI, "UCI Machine Learning Repository", http://www.ics.uci.edu/~mlearn/MLSummary.html

About the authors:

The authors are members of the Data Systems Users and Research Group (DataSURG), Computer Science Dept., North Dakota State University, headed by Dr. William Perrizo. DataSURG has been supported by NSF, NASA, DARPA, GSA and is co-sponsor of the Virtual Conferences on Genomics and Bioinformatics series.
http://www.ndsu.nodak.edu/virtual-genomics/conference_2002.htm
http://web.cs.ndsu.nodak.edu/~perrizo/classes/03_cfp5.ppt
http://web.cs.ndsu.nodak.edu/~datasurg/