Knowledge-based clustering of numerical data sets incorporating confidence values: applications to yeast genomic data sets

Bill Andreopoulos[1]

Department of Computer Science, York University, Toronto, Canada, M3J 1P3
billa@cs.yorku.ca

1 INTRODUCTION

Much of the existing knowledge on the functions of an organism's genes that is represented as *Gene Ontology (GO)* annotations has uncertain correctness. However, confidence values can be extracted on this knowledge as a quantitative measure of the certainty of correctness, by using GO Evidence Codes [1]. Ideally, clustering of numerical data sets from biological domains should consider the existing knowledge as *categorical annotations (CAs)* on genes and the uncertainties as *confidence values (CVs)* on the categorical annotations [3].

2 OVERVIEW OF THE KNOWLEDGE-BASED CLUSTERING ALGORITHM

We have designed, implemented and tested a clustering tool for numerical data sets that incorporates the existing uncertain knowledge and confidence values both during and after clustering, for predicting gene function. Our method provides a different type of insight for predicting the functional roles of genes, by giving the 'full picture' of the data set including existing knowledge and confidence on its correctness.

Our technique performs clustering of numerical data sets, but bases the process on existing knowledge in the form of CVs and CAs on the genes. The clustering process consists of two levels. The first clustering level involves only the subset of genes having CAs, or the genes the CAs of which have the highest CVs. The first level uses a variation of the k-Modes algorithm [3]. The results of the first level are given as input to the second clustering level. The second level involves all genes in the data set that have a numerical annotation. The second level uses a numerical variation of the ROCK algorithm [2]. The output of the second clustering level is the final result of the whole process.

3 CONFIDENCE VALUES AND CATEGORICAL ANNOTATIONS

We created six pools of CAs for each gene and each pool contained GO annotations of a specific type. Three pools contained GO annotations on the molecular function, biological process and cellular location of a gene. The other three pools contained GOSlim annotations for each GO annotation (GOSlim are high level GO annotations that represent high level knowledge on genes).

CVs are real numbers (0.0 to 1.0) assigned to the CAs of a gene. A CV indicates the confidence that a CA is correct; the CVs on a gene also specify how strongly the gene's CAs should influence the clustering process. We determined the CVs by using GO evidence codes: GO evidence codes indicating strong evidence supporting a GO annotation got mapped to a CV close to 1.0, while those indicating weak evidence got mapped to a CV close to 0.0.

4 FUNCTIONAL PREDICTION ON YEAST

Given a resulting cluster, we assigned a P-value to each CA in the cluster. A P-value is a statistical test that measures whether a cluster contains a specific CA more frequently than would be expected by chance [4]. A P-value close to 0.0 indicates a frequent occurrence of the CA in the cluster, while a P-value close to 1.0 indicates its seldom occurrence. We multiplied the resulting P-value with the reciprocal of the average of all CVs assigned to the CA in the cluster (1/avg(CV)), thus deriving what we call an *M-value*. We used this metric for functional prediction by applying certain CAs to genes with CAs of 'Unknown'. For this purpose, we located in each cluster the CAs with the lowest M-values for molecular function, biological process and cellular location.

We validated the accuracy of the functional predictions for 35 clusters, by focusing on a cluster having 224 genes. We set to 'Unknown' the annotations of some genes in which the CAs 'component:nucleus' and 'nucleus(a granular annotation)' occurred originally and then re-clustered the data set into 35 clusters. From the results we extracted the CAs with the lowest M-values and we were able to predict correctly that some of the genes should be annotated as either 'component:nucleus' or 'nucleus(a granular annotation)'.

References

1. The Gene Ontology Consortium. (2001). Creating the gene ontology resource: design and implementation. Genome Research 11: 1425-1433.
2. Guha S., Rastogi R., Shim K. (2000). ROCK: A Robust Clustering Algorithm for Categorical Attributes. Information Systems 25(5): 345-366.
3. Huang Z. (1998) Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2(3): 283-304.
4. Wu L.F., Hughes T.R., Davierwala A.P., Robinson M.D., Stoughton R. and Altschuler S.J. (2002). Large-scale Prediction of Saccharomyces Cerevisiae Gene Function Using Overlapping Transcriptional Clusters. Nature Genetics 31:255-265.