ASSOCIATION ANALYSIS TECHNIQUES FOR ANALYZING COMPLEX BIOLOGICAL DATA SETS

Gaurav Pandey, Gowtham Atluri, Gang Fang, Rohit Gupta, Michael Steinbach and Vipin Kumar

Department of Computer Science and Engineering, University of Minnesota
{gaurav,gowtham,gangfang,rohit,steinbac,kumar}@cs.umn.edu
`http://www.cs.umn.edu/~kumar/dmbio`

ABSTRACT

Association analysis is one of the most popular analysis paradigms in data mining. In this paper, we present different types of association patterns and discuss some of their applications in bioinformatics. We present a case study showing the usefulness of association analysis-based techniques for pre-processing protein interaction networks. Finally, we discuss some of the challenges that need to be addressed to make association analysis-based techniques more applicable for bioinformatics.

1. INTRODUCTION

Association analysis [1, 2][1] is one of the most popular analysis paradigms in data mining. The techniques in this field seek to find patterns that describe the relationships among the binary attributes (variables) used to characterize a set of objects. The iconic example of data sets analyzed by these techniques is market basket data, where the objects are transactions consisting of sets of items purchased by a customer, and the attributes are binary variables that indicate whether or not an item was purchased by a particular customer. The interesting patterns in these data sets are either sets of items that are frequently purchased together (frequent itemset patterns) or rules that capture the fact that the purchase of one set of items often implies the purchase of a second set of items (association rule patterns). Association patterns, whether rules or itemsets, are local patterns in that they hold only for a subset of transactions. The size of this set of supporting transactions, which is known as the support of the pattern, is one measure of the strength of a pattern. A key strength of association pattern mining is that the potentially exponential nature of the search can often be made tractable by using support based pruning of patterns [1], i.e., the elimination of patterns supported by too few transactions early on in the search process.

In this paper, we present some standard formulations of association patterns, discuss a novel application of association analysis to the pre-processing of protein interaction data, and cite examples of challenges to be addressed to make association analysis more widely applicable for addressing bioinformatics problems.

2. ASSOCIATION PATTERNS

This section introduces some commonly used association patterns that have been proposed in the literature.

2.1. Traditional Frequent Patterns

Traditional frequent pattern analysis [2] focuses on binary data sets, such as the market basket data discussed above. These data sets can be represented as a binary matrix containing n rows (transactions) and m columns (items), and ij^{th} entry is 1 if the i^{th} transaction contains the j^{th} item, and 0 otherwise. Given such a representation, a key task in association analysis is to find frequent itemsets in this matrix, which are sets of items that frequently occur together in a transaction. The strength of an itemset is measured by its *support*, which is the number (or fraction) of transactions in the data set in which all items of the itemset appear together. Interestingly, support is an anti-monotonic measure in that the support of an itemset in a given data set can not be less than any of its supersets. This anti-monotonicity property allows the design of several efficient algorithms, such as Apriori [1] and FPGrowth [4], for discovering frequent itemsets in a given binary data matrix. With judicious choices for the support threshold, the number of patterns discovered from a data set can be made manageable. Also, note that, in addition to support, a number of additional measures have been proposed to determine the interestingness of association patterns [5].

2.2. Hyperclique Patterns

A hyperclique pattern [6] is a type of frequent pattern that contains items that are strongly associated with each other over the supporting transaction, and are quite sparse (mostly 0) over the rest of the transactions. In traditional frequent pattern mining, choosing the right support threshold can be quite tricky. If support threshold is too high, we may miss many interesting patterns involving low support items.

[1] Not to be confused with statistical association analysis [3].

If support is too low, it becomes difficult to mine all the frequent patterns because the number of extracted patterns increases substantially, many of which may relate a high-frequency item to a low-frequency item. Such patterns, which are called *cross-support patterns*, are likely to be spurious. Hyperclique patterns avoid these cross-support patterns by defining an anti-monotonic association measure known as *h-confidence* that ensures a high affinity between the itemsets constituting a hyperclique pattern [6]. Formally, the h-confidence of an itemset $X = \{i_1, i_2, \ldots i_m\}$, denoted as $h - confidence(X)$, is defined as,

$$h - confidence(X) = \frac{s(i_1, i_2, \ldots, i_k)}{max[s(i_1), s(i_2), \ldots, s(i_k)]}$$

where $s(X)$ is the support of an itemset X. Those itemsets X that satisfy $h - confidence(X) \geq \alpha$, where α is a user-defined threshold, are known as hyperclique patterns. These patterns have been shown to be useful for various applications, including data cleaning [7], and finding functionally coherent sets of proteins [8]. In the next section, we discuss how the $h - confidence$ measure can be used to pre-process protein interaction networks effectively.

3. PRE-PROCESSING OF PROTEIN INTERACTION NETWORKS

One of the most promising forms of biological data that are used to study the functions and other properties of proteins at a genomic scale are protein interaction networks. These networks provide a global view of the interactions between various proteins that are essential for the accomplishment of most protein functions. Due to the importance of the knowledge of these interactions, several high-throughput methods have been proposed for discovering them [9].

A protein interaction network can be represented as an undirected graph, where proteins are represented by nodes and protein-protein interactions as edges. Due to this systematic representation, several computational approaches have been proposed for the prediction of protein function from these graphs [10, 11, 12]. Also, owing to the rich functional information in these networks, several of these approaches have produced very good results, particularly those that use the entire interaction graph simultaneously and use global optimization techniques to make predictions [12].

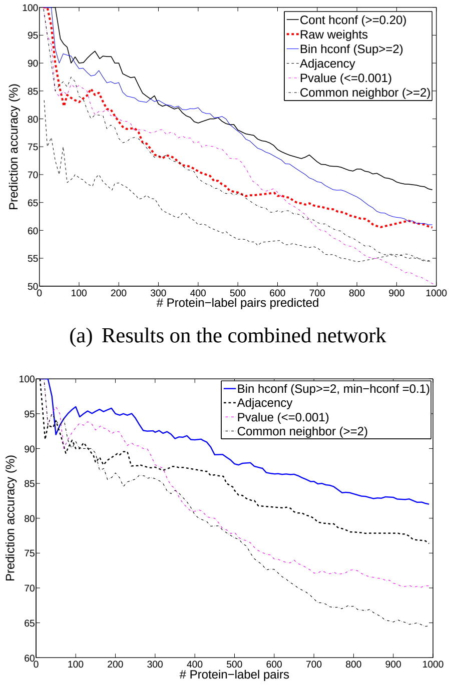
However, despite the advantages of protein interaction networks, they have several weaknesses which affect the quality of the results obtained from their analysis. The most prominent of these problems is that of noise in the data, which manifests itself primarily in the form of spurious or false positive edges [13]. Studies have shown that the presence of noise has significant adverse effects on the performance of protein function prediction algorithms [14]. Another important problem facing the use of these networks is their incompleteness, i.e., the absence of biologically valid interactions even from large sets of interactions [13]. This absence of interactions from the network prevents even the global optimization-based approaches from making effective use of the network beyond what is available, thus leading to a loss of potentially valid predictions.

A possible approach to address these problems is to transform the original interaction graph into a new weighted graph such that the weights assigned to the edges in the new graph more accurately indicate their reliability. The utility of hypercliques in noise removal from binary data [7], coupled with the representation of protein interaction graphs as a binary adjacency matrix to which association analysis techniques can be applied, motivated Pandey et al. [15] to address the graph transformation problem using an approach based on $h - confidence$ measure discussed earlier. This measure is used to estimate the common neighborhood similarity of two proteins P_1 and P_2 as

$$h - confidence(P_1, P_2) = \min\left(\frac{|N_{P_1} \cap N_{P_2}|}{|N_{P_1}|}, \frac{|N_{P_1} \cap N_{P_2}|}{|N_{P_2}|}\right)$$

where N_{P_1} and N_{P_2} denote the sets of neighbors of P_1 and P_2 respectively. As discussed earlier, this definition of $h - confidence$ is only applicable to binary data or, in the context of protein interaction graphs, unweighted graphs. However, the notion of $h - confidence$ can be readily generalized to networks where the edges carry real-valued weights indicating their reliability. In this case, the above equation can be conveniently modified to calculate $h - confidence(P_1, P_2)$ by making the following substitutions: (1) $|N_{P_1}| \rightarrow$ sum of weights of edges incident on P_1 (similarly for P_2) and (2) $|N_{P_1} \cap N_{P_2}| \rightarrow$ sum of minimum of weights of each pair of edges that are incident on a protein P from both P_1 and P_2. In both these cases, the $h - confidence$ measure is guaranteed to be bounded within the $[0, 1]$ interval.

Now, with this definition, it is hypothesized that protein pairs having a high $h - confidence$ score are expected to have a valid interaction between them, since a high value of the score indicates a high common neighborhood similarity, which in turn reflects greater confidence in the network structure for that interaction. For the same reason, interactions between protein pairs having a low $h - confidence$ score are expected to noisy or spurious. Accordingly, Pandey *et al* [15] proposed the following graph transformation approach for pre-processing available interaction data sets. First, using the input interaction network $G = (V, E)$, the $h - confidence$ measure is computed between each pair of constituent proteins, whether connected or unconnected by an edge in the input network. Next, a threshold is applied to drop the protein pairs with a low $h - confidence$ to remove spurious interactions and control the density of the network. The resultant graph $G' = (V, E')$ is hypothesized to be the less noisy and more complete version of G, since it is expected to contain fewer noisy edges, some biologically viable edges that were not present in the original graph, and more accurate weights on the remaining edges.



(a) Results on the combined network

(b) Results on the DIPCore network

Fig. 1. Comparison of performance of various transformed networks and the input networks (Best viewed in color).

In order to evaluate the efficacy of the resultant networks for protein function prediction, the original and the transformed graphs was provided as input to the FunctionalFlow protein function prediction algorithm [12]. The performance was also compared with transformed versions generated using other common neighborhood similarity measures for such networks, such as Samanta et al [11]'s p-value measure. Figure 1 shows the performance of this algorithm on these transformed versions of two standard interaction networks, namely the *combined* data set constructed by combining several popular yeast interaction data sets (combined) and weighted using the EPR Index tool [16], and the other being a confident subset of the DIP database [16] (DIPCore). The performance is evaluated using the accuracy of the top scoring 1000 predictions of the functions of the constituent proteins generated by a five-fold cross-validation procedure, where the functional annotations are obtained from the classes at depth two of the FunCat functional hierarchy [17].

The results in Figure 1 show that for both the data sets, the $h-confidence$-based transformed version(s) substantially outperform the original network and the other measures for this task. The margin of improvement on the highly reliable DIPCore data set is almost consistently 5% or above, which is quite significant. Similar results are observed using the complete precision-recall curves. The interested reader is referred to [15] for more details of the study.

4. CONCLUDING REMARKS

Association analysis has proved to be a powerful approach for analyzing traditional market basket data, and has even been found useful for some problems in bioinformatics [18, 8, 15]. However, there are a number of other important problems in this field, such as finding biomarkers using dense data like SNP data and real-valued data like gene-expression data, where such techniques could prove to be very useful, but cannot currently be easily and effectively applied.

An important example of patterns that are not effectively captured by the traditional association analysis framework and its current extensions, is a group of genes that are co-expressed together across a subset of conditions in a gene expression data set, which is real-valued. Such patterns have often been referred to as *biclusters*. Methods for transforming these data sets into binary form (for example, via discretization [18, 19, 20]) often suffer from loss of critical information about the actual values. Hence, a variety of biclustering algorithms have been developed for finding such patterns from gene expression data, such as ISA [21], Cheng and Church's algorithm [22] and SAMBA [23]. Although these algorithms are often able to find useful patterns, they suffer from a number of limitations. The most important one is an inability to efficiently explore the entire search space for such patterns without resorting to heuristic approaches that compromise the completeness of the search. Pandey et al. [24] have presented one of the first methods for directly mining association patterns from real-valued data, particularly gene expression data, that does not involve a transformation of the data. These techniques are able to discover all patterns satisfying the given constraints, unlike the biclustering algorithms that may only be able to discover a subset of these patterns. There are several open opportunities for designing better algorithms for addressing this problem.

Another challenge that has inhibited the use of association analysis in bioinformatics–even when the data is binary–is the density of several types of data sets. Algorithms for finding association patterns often break down when the data becomes dense because of the large number of patterns generated, unless a high support threshold is used. However, with a high threshold, many interesting, low-support patterns are missed. One particularly important category of applications with dense data are applications involving class labels, such as finding connections between genetic variations and disease. Consider the problem of finding connections between genetic variations and disease using binarized version of SNP-genotype data, which is 33% dense by design, since each subject must have one of the three variations of SNP pairs: *major-major*, *major-minor*, *minor-minor*. Most of the existing techniques for this problem only apply univariate analysis and rank individual SNPs using measures like p-value, odds ratio etc [3]. Some approaches like Multi-Dimensionality Reduction (MDR) [25] and Combinatorial Partitioning Methods (CPM) [26], which are designed to identify groups of SNPs, can only be applied to data sets with small number (few dozens) of SNPs. Also, existing discriminative pattern mining algorithms [27, 28] are only able to prune infrequent non-discriminative patterns, not the frequent non-discriminative patterns, which is the biggest challenge for dense data sets like SNP data and gene expression data. New association analysis approaches should be designed to enable efficient discriminative pattern mining on dense and high dimensional data, where effectively making use of class label information for pruning the large search space is crucial.

In conclusion, significant scope exists for future research on designing novel association analysis techniques for complex biological data sets and their associated problems. Such techniques will significantly aid in realizing the potential of association analysis for discovering novel knowledge from these data sets and solve important bioinformatics problems.

5. REFERENCES

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. VLDB*, 1994, pp. 487–499.

[2] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining," *Addison-Wesley*, 2005.

[3] D.J. Balding, "A tutorial on statistical methods for population association studies," *Nat Rev Genet.*, vol. 7, no. 10, pp. 781, 2006.

[4] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," *DMKD*, vol. 8, no. 1, pp. 53–87, 2004.

[5] P.N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," *Information Systems*, vol. 29, no. 4, pp. 293–313, 2004.

[6] Hui Xiong, Pang-Ning Tan, and Vipin Kumar, "Hyperclique pattern discovery," *Data Min. Knowl. Discov.*, vol. 13, no. 2, pp. 219–242, 2006.

[7] Hui Xiong, Gaurav Pandey, Michael Steinbach, and Vipin Kumar, "Enhancing data analysis with noise removal," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 3, pp. 304–319, 2006.

[8] H. Xiong, X. He, C. Ding, Y. Zhang, V. Kumar, and S. R. Holbrook, "Identification of functional modules in protein complexes via hyperclique pattern discovery," in *Proc. Pac. Symp. Biocomput. (PSB)*, 2005, pp. 221–232.

[9] Pierre Legrain, Jerome Wojcik, and Jean-Michel Gauthier, "Protein–protein interaction maps: a lead towards cellular functions," *Trends Genet.*, vol. 17, no. 6, pp. 346–352, 2001.

[10] G. Pandey, V. Kumar, and M. Steinbach, "Computational approaches for protein function prediction: A survey," Tech. Rep. 06-028, Deptt. of Comp. Sc. and Engg., Univ. of Minnesota, 2006.

[11] M. P. Samanta and S. Liang, "Predicting protein functions from redundancies in large-scale protein interaction networks," *PNAS*, vol. 100, no. 22, pp. 12579–12583, 2003.

[12] Elena Nabieva, Kam Jim, Amit Agarwal, Bernard Chazelle, and Mona Singh, "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps," *Bioinformatics*, vol. 21, no. Suppl. 1, pp. i1–i9, 2005.

[13] G. T. Hart, A. K. Ramani, and E. M. Marcotte, "How complete are current yeast and human protein-interaction networks?," *Genome Biol.*, vol. 7, no. 11, pp. 120, 2006.

[14] M. Deng, F. Sun, and T. Chen, "Assessment of the reliability of protein–protein interactions and protein function prediction," in *Pac Symp Biocomputing*, 2003, pp. 140–151.

[15] G. Pandey et al., "Association analysis-based transformations for protein interaction networks: a function prediction case study," in *Proc. 13th ACM SIGKDD International Conference*, 2007, pp. 540–549.

[16] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Mol Cell Proteomics*, vol. 1, no. 5, pp. 349–356, 2002.

[17] A. Ruepp et al., "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes," *NAR*, vol. 32, no. 18, pp. 5539–5545, 2004.

[18] Celine Becquet et al., "Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data," *Genome Biology*, vol. 3, 2002.

[19] C. Creighton and S. Hanash, "Mining gene expression databases for association rules.," *Bioinformatics*, vol. 19, no. 1, pp. 79–86, January 2003.

[20] Tara McIntosh and Sanjay Chawla, "High confidence rule mining for microarray analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 4, no. 4, pp. 611–623, 2007.

[21] Sven Bergmann, Jan Ihmels, and Naama Barkai, "Iterative signature algorithm for the analysis of large-scale gene expression data," *Physical Review*, vol. 67, 2003.

[22] Y. Cheng and G.M. Church, "Biclustering of Expression Data," in *Proc. Eighth ISMB Conference*, 2000, pp. 93–103.

[23] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, no. Suppl 1, pp. 136–144, 2002.

[24] G. Pandey, G. Atluri, M. Steinbach, Chad L. Myers, and V. Kumar, "An association analysis approach to biclustering," in *Proc. 15th ACM SIGKDD International Conference*, 2009, p. in press, Also TR 08-007, CS Department, UMN.

[25] M.D. Ritchie et al., "Multifactordimensionality reduction reveals high-order iteractions among estrogen- metabolism genes in sporadic breast cancer.," *Am J Hum Genet*, vol. 69(1), pp. 1245–1250, 2001.

[26] MR Nelson, SLR Kardia, RE Ferrell, and CF Sing, "A Combinatorial Partitioning Method to Identify Multilocus Genotypic Partitions That Predict Quantitative Trait Variation," *Genome Research*, vol. 11, no. 3, pp. 458–470, 2001.

[27] S. Bay and M. Pazzani, "Detecting group differences: Mining contrast sets," *DMKD*, vol. 5(3), pp. 213–246, 2001.

[28] Wei Fan, Kun Zhang, Hong Cheng, Jing Gao, Xifeng Yan, et al., "Direct discriminative pattern mining for effective classification," in *Proc. IEEE ICDE*, 2008, pp. 169–178.