ISSN:0975-9646

A Theoretical Approach for Augmenting Association Rule Mining to Predict Protein-Protein Interaction

Sony Snigdha Sahoo[1], Tripti Swarnkar[2]

[1]Department of Computer Science and Engg, Siksha 'O' Anusandhan University

[2]Department of Computer Applications, Siksha 'O' Anusandhan University
Jagamara, Bhubaneswar-751030, India

***Abstract*— Background:Every biological process occurring within the living body involves the formation of protein complexes. Interactions between proteins are an important protein feature. Therefore, determining protein interaction has become one of the most significant problems in the post genomic era.**

Methodology:For effectively determining the interactions occurring among proteins computational approaches like association rule mining could be used. But, only support and confidence measures used with association rule mining can be insufficient at filtering out interesting rules, because it fails in implying the kind of association between given datasets. Correlation measure when used along with association mining could augment the support-confidence framework by deciding whether the association is positive or negative.

Conclusion:In this study, we have presented a comparison between association rule mining and correlation in an attempt to indicate the ways in which correlation can augment the support-confidence framework.

***Keywords*— Protein interaction, computational approach, association rule mining, correlation, support-confidence framework.**

I. Introduction

Proteins also known as polypeptides are organic compounds made up of amino acids arranged in a linear chain. The amino acids in a protein polymer are joined together by the peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code. In general, the genetic code specifies twenty standard amino acids. Like other biological macromolecules such as polysaccharides and nucleic acids, proteins are essential parts of organisms and participate in virtually every process within cells. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signaling, immune responses, cell adhesion, and the cell cycle. Proteins have a typical feature that they often work together to achieve any particular function and associate to form stable complexes.

Such communications taking place between the proteins for processes occurring within the living body could be depicted in the form of an interaction network where proteins are represented as nodes and a connection between two proteins would exist only if they are interacting. The analysis of these interactions can prove significant for understanding the mechanisms of biological processes. These analyses can also help, in finding out rules that can predict the occurrence of proteins which are likely when some other proteins are present, clues regarding function of a protein by seeing whether it interacts with another protein of known function etc.

The different types of interactions among proteins are essential to various biological functions in a living cell. Information about these interactions provides a basis to construct protein interaction networks and improves our understanding of the general principles of the functioning of biological systems. Recent years have seen the development of various experimental techniques for systematic protein-protein interaction (PPI) analysis. At present, however, experimentally detected interactions represent only a small fraction of the real interaction network. Therefore, a number of computational approaches have been proposed to accelerate the PPI detection process based on only experimental techniques ([2]-[5], [8]).

But determining the protein-protein interaction network is not as easy as it seems to be. Because unlike genomics that remains fixed for a particular individual, proteomics for a single individual varies from cell to cell, process to process and from time to time. Moreover protein interaction are extremely transitory in nature i.e. they form complexes with other proteins for only a short span of time. For establishing this interaction, computational approaches like association rule mining can be used leading to determination of the protein-protein interaction network.

But, Gavin et al [7] have suggested that protein complexes consist of several versatile modules, with different functional modules that contribute to one superimposed biological function. They have validated this modularity of protein complexes in their recent experiments. These more complex relationships between proteins cannot be described by pair wise protein interactions represented in protein-protein interaction graphs. Also there remains the fact that biological network of proteins being enormously complex, the number of protein association would be complicated as well. The problems associated with protein-protein interaction can be reduced to a cluster of questions like: "Who's touching whom? How does it do it? Where does the touching take place? When does it happen? Why does it occur? And what are the consequences?"

This work provides an overview of, how to address few among such issues regarding protein-protein interaction networks and to augment support-confidence framework for predicting stronger association rules.

II. Protein-Protein interaction prediction as a frequent itemset problem

Finding sets of items in the data that frequently appear together is known as the frequent itemset problem. Protein-protein interaction can be sighted as a frequent itemset problem because proteins seldom act alone; they must interact with other biomolecular units to execute their function. Frequently associated items can be represented in the form of association rules, rule support and rule confidence being two measures of rule interestingness. These two measures respectively reflect the usefulness and certainty of discovered association rule. Typically association rules are considered to be interesting if they satisfy both a minimum support and a minimum confidence threshold.

A. Frequent Itemsets, Closed Itemsets and Association Rules

Let P = {i_1, i_2,i_3,... i_N} be a set of N distinct items. A *transaction* T is a set of items in P. A database D of size M is a set of M such transactions. A set, I $\subseteq$ P, of items is called an *itemset.* The number of items in an itemset is called the *length* of an itemset. Itemsets of some length k are referred to as k-itemsets.

An association rule is an implication of the form A$\Rightarrow$ B, where A $\subset$ P, B $\subset$ P, and A$\cap$B=$\varnothing$. The rule A $\Rightarrow$ B holds in the transaction set D with support S, where S is the percentage of transactions in D containing A that also contains B. this is taken to be the conditional probability, P (B|A). That is,

Support (A $\Rightarrow$ B) = P (A $\cup$ B) [1]

Confidence (A $\Rightarrow$ B) =P (B|A) [2]

Rules that satisfy minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called strong. By convention we write support and confidence values to occur between 0% and 100% or 0 and 1.0([1],[3]). The occurrence frequency of an itemset is the number of transaction that contains that itemset. This is also known as the frequency, support count or count of the itemset. The support defined in equation 1 is sometimes referred to as absolute support. If the relative support of an itemset I satisfies a prespecified minimum support threshold, then I is a frequent itemset. The set of frequent k-itemsets is commonly denoted by L_k. From equation 2, we have

Confidence (A$\Rightarrow$B) = P (B|A) = $\frac{Support\,(A\cup B)}{Support\,(A)}$

In general, association rule mining can be viewed as a two step process:

(i) Finding all frequent itemsets (ii) Generating strong association rules from the frequent itemsets.

This algorithm is based on the fact that any subset of a frequent itemset is also frequent. It also means that every frequent itemset with n items would result in n association rules having n single item on RHS. Therefore the first step incurs huge overhead in terms of memory usage, computation and I/O resources. The second step is quite straightforward, but can be expensive while dealing with large datasets like real world problems ([6]).

III. Different frequent itemset algorithms used

A. Apriori Algorithm

Several methods have been proposed for mining frequent patterns from given itemset. Apriori algorithm is the basic algorithm for generating frequent itemsets. It was proposed by R.Agrawal et al (1994) for mining frequent itemsets. It has been so named because it uses apriori property of the frequent itemsets.

Apriori Property: All nonempty subsets of a frequent itemset must also be frequent.

Apriori algorithm employs this property in two steps consisting of join and prune actions respectively.

i) Join step
ii) Prune step

Join step generates set of k itemsets by joining set of k-1 itemset with itself. Prune step, during each iteration eliminates those itemsets which don't satisfy minimum support and confidence.

Many variations of apriori algorithm have also been suggested which focus on improving its efficiency. Some of the variations are Hash based technique, transaction reduction, partitioning, sampling, dynamic itemset counting. All these variations aim at reducing the number of itemsets generated at each step for reducing the overhead.

But along with other overheads, association rule mining may lead to rules which are insufficient in certain respects. The confidence of an association rule such as A$\Rightarrow$ B can often be misleading as it is an estimate of the conditional probability of itemset B given itemset A. It doesn't directly imply the strength of correlation between A and B ([1],[9]).

For instance, we are here considering hypothetical regarding the number of protein interactions responsible for cell aging and apoptosis. Suppose we want to analyze these interactions among proteins responsible for cell growth and apoptosis. Let, among the 100 interactions analyzed, 60 of the interactions contain proteins responsible for cell growth, whereas 40 are responsible for apoptosis and 75 of the interactions include proteins accountable for both. Let, a data mining approach discovered association rules on the above data, using a minimum support of say 30% and a minimum confidence of 60%. Then the association rule can be represented as:

Interacts (P, “Growth”) ⇒Interacts (P, “apoptosis) [S=40%, C=60%]

Where ‘S’ and ‘C’ denote support and confidence values respectively.

The above association rule would be a strong rule as its support value is 40/100=40% and confidence value is 40/60=66%, which satisfy the minimum support and minimum confidence thresholds.

Though, the above rule is misleading because the probability of proteins concerned with apoptosis is 75/100=75% which is well above the minimum support value. But, in reality cell growth and apoptosis are negatively associated because cell growth can never be responsible for cell death or apoptosis.

This example instantiates that only confidence value can be illusory in that, it indicates only conditional probability of itemset B given itemset A. it can’t provide us with the real measure of the implication between A and B.

Therefore to measure the strength of association between given itemsets, another measure, correlation can be used.

B. Correlation

A correlation rule is measured not only by its support and confidence but also by the correlation between given itemsets. There are various correlation measures. Here we are taking into account, correlation using lift which is the simplest among all for evaluating against association rule mining.

Lift is a correlation measure given by

$$\text{Lift}(A, B) = P(A \cup B)/P(A)P(B) \qquad [3]$$

Where A and B are two given itemsets.

If the resulting value of eq. 3 is less than 1, then occurrence of A is negatively correlated with the occurrence of B. If the resulting value is greater than 1 then A and B are positively correlated. If the resulting value is equal to 1, then A and B are independent [1].

If we consider the same data as we have used for finding out association rules, we find that

Lift (growth, apoptosis)
=P (growth ∪ apoptosis)/P(growth) P(apoptosis)
=0.031
Since, lift < 1, we infer that cell growth and apoptosis are negatively correlated.
To help understand correlation analysis, we consider here the following contingency table.

TABLE I
A 2X2 CONTIGENCY TABLE SUMMARIZING THE HYPOTHETICAL PROTEIN INTERACTIONSWITH RESPECT TO CANCER AND AGING

	cancer	$\overline{\text{cancer}}$	Σ_{row}
aging	588	400	988
$\overline{\text{aging}}$	33	11	44
Σ_{col}	621	411	1032

From the table we have, P(cancer) = 621/1032 = 0.601, P(aging) = 988/1032=0.957, P(cancer,aging) = 588/1032 = 0.569, Lift = P(cancer,aging)/P(cancer)P(aging) = 0.989

Because the value is less than 1, there is negative correlation between aging and cancer.

From this we infer that correlation analysis can be used to calculate the degree of association among the items in a dataset.

IV. Discussion and future direction

In this paper, association rule mining and correlation have been used for evaluating protein interaction data. By using only support and confidence measures to mine association results in generation of large number of rules, many of which might be uninteresting. Instead, the support-confidence framework can be improved using correlation measure that results in stronger rules. It identifies negative association rules and the items that conflict each other. In addition to it, while searching for protein interaction, correlation rules can provide information regarding which proteins are acting as hubs in the interaction network in terms of rules that result in positive correlation. Thus, correlation adds to the support and confidence measure by identifying stronger positive and negative association rules.
Future studies: We intend to unveil valid relationships among proteins which are behind biological processes. We plan to mine strong protein-disease association rules by implementing support-confidence framework with correlation measure.

References

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, 2001.

[2] E. Chautard, T. N. Mieg, R. S. Blum, “Interaction networks as a tool to investigate the mechanism of aging,”Biogerontology , Vol. 11(4), ISSN 1389-5729,2010

[3] H. W. Chiu, F. H. Hung, “Association rule mining from yeast protein interaction to assist protein-protein interaction prediction,” Biomedical fuzzy and human sciences, Vol.13(1),pp.03-06,2008

[4] J. R. Managbanag, T. M. Witten, D. Bonchev, I. A. Fox, M. Tsuchiya, et al,” Shortest-path network analysis is a useful approach towards identifying genetic determinants of longevity,” PLoS ONE, Vol.3(11) ,p.e3802 , Nov 2008

[5] Kocatas A., Gursoy A. and Atalay R. (2003): “Application of Data Mining Techniques to Protein-Protein Interaction Prediction”, Springer Berlin / Heidelberg.

[6] M.R.Antonie, O.R.Zaiane,”Mining negative and positive association rules:an approach for confined rules”,International journal of Business Intelligence and Data Mining,vol.3(2),pp.158-176,2008

[7] Gavin et al,”Proteome survey reveals modularity of yeast cell”,nature,vol.440,No.7084,pp.631-636,2006.

[8] S. Lee, L. Yang, L. I. Jianrong, C. Friedman, Y. A. Lussier, “Discovery of protein interaction networks shared by diseases”. Pacific symposium on biocomputing , pp. 76-87,2007

[9] Z. Zheng, R. Kohavi, L. Mason,“Real World performance of association rule algorithm” ,Proceedings of seventh ACM SIGKDD international conference on knowledge discovery data mining,pp.401-406,2001