Predicting Protein–Protein Interactions using Numerical Associational Features

Waleed Aljandal, William H. Hsu, and Jing Xia
Department of Computing and Information Sciences, Kansas State University
234 Nichols Hall, Manhattan, KS 66506-2302

Abstract **-** We investigate the problem of predicting protein-protein interaction (PPI) using numerical features constructed from parent-child relations within a partial network based on known protein interactions. For each pair of proteins, we use a validation-based approach to normalize quantitative features that represent interestingness measures for associations between parents and children. The primary contribution of this work is the parametric normalization formula we derive and calibrate using data for the PPI task. This formula improves basic interestingness measures by considering itemset sizes. Our derived itemset size-sensitive measures emphasize small sets of frequently co-interacting proteins, for which we hypothesize a greater conditional probability of direct interaction. We evaluate our work using *k*-nearest neighbor and rule-based classification approaches.

I. INTRODUCTION

We explore association rule-based approaches to the problem of predicting protein-protein interactions (PPI) and other related proprieties. Existing methods for PPI prediction include information extraction from literature – specifically, using dynamic programming-based alignment between biomedical texts and key verbs describing interactions to compute distinguishing patterns [1]. In interaction prediction, [2] considers two methods: the neighborhood-counting method and the chi-square method. They used protein-protein interaction network to predict protein function. Another approach using common-neighbor based model and a Bayesian framework to predict protein function introduced in [3]. The Mixture-of-Feature-Experts method has been used in [4] to predict protein-protein interaction where they combine a set of features as a weighted expert.

II. BACKGROUND

A. Link Prediction

Link prediction is discovering the presence of links or connections between two or more instances based on properties, such as shared relational features. Link prediction methods can be used to provide an expectation of unknown relations which come from the massive amount of data related to gene expression, known regulatory relationships, RNA, protein sequences, and protein interaction. Association rule mining has been used in this area for discovering associations between different concepts with different structures. In order to discover assocation rules, researchers have investigated some special algorithms to handle bioinformatics datasets for discovering frequent patterns [5]. [6] introduces different algorithms to mine frequent closed pattrens and propose new ones. *GenMiner* is an implementation of a special association rule generator from genomic data using an algorithm called NorDi which is more efficient than the Apriori approach, as shown in [7]. Other studies [8] propose new data structures such as the BSC-tree and FIS-tree, to prepare the gene expression data for the association mining step.

B. Association Rules

Most previous work on applying association rules techniques to PPI prediction has been devoted to building predictive rules of identifying function regions pairs engaged in protein-protein interactions [9][10] [11]. Recently, new frequent pattern identification techniques specific to protein networks have been proposed [12] that were used to find patterns for predicting protein-protein interaction, specifically recurring functional interaction patterns. In [13] they also integrated association mining approach to integrating several diverse types of evidence. Features of primary structure and associated physicochemical properties were used in [14] and gene expression profiles, as features, were considered in [11] with large number of protein network and difference among them. The concept of differential association rule mining was introduced in [15], the annotations of proteins in the context of one or more interaction networks. Furthermore, to evaluate the rules extracted by association rules more efficiently, specific scoring measures of the rules were introduced in [16].

In this paper, we consider the positive protein-protein interaction network and use numerical features to predict protein-protein interaction from only the parent-child relationships. However, using additional information such as gene expression and other protein features, we can improve the link prediction component of the system.

III. Computing numerical Features based using Association Rules Measures

An association rule has the form $u \rightarrow v$, where both u and v are subsets of an observed itemset $L = \{I_1, I_2, \ldots, I_k\}$. There are many approaches for association rule learning from itemset data and additional derived objective measures of interestingness. We refer the interested reader to a survey of association rule mining [17] that reviews the interestingness measures for rules and summaries, classifies them from several perspectives and compares their properties. The authors present 38 probability-based objective interestingness measures for association rules. In a previous survey, [18] discusses the properties of 21 objective interestingness measures and concludes that there is no measure that is consistently better than others in all application domains.

Interestingness measures are descriptive statistics computed over rules of the form $u \rightarrow v$, which in our application denotes that "when u has parent or child, then v also has this parent or child". This allows us to apply algorithms for association rule (AR) mining based on calculation of frequent itemsets, which by analogy with market basket analysis (grocery basket) denote sets of proteins which all share parenthood or childhood. Each interestingness measure captures one or more desiderata of a data mining system: novelty (surprisingness), validity (precision, recall, and accuracy), expected utility, and comprehensibility (semantic value).

We use the count of common interests, plus eight normalized AR interestingness measures over common interests, as numerical friendship prediction features. Each measure is a statistic over the set common interests of u and v, and expressed as a function of the rule $u \rightarrow v$.

1. The number of common interests:
   $|\ \text{Itemsets}(u) \cap \text{Itemsets}(v)\ |$
2. Support $(u \rightarrow v)$ = Support $(v \rightarrow u) = P(u, v)$
3. Confidence $(u \rightarrow v) = P(v|u)$
4. Confidence $(v \rightarrow u) = P(u|v)$
5. Lift $(u \rightarrow v) = \frac{P(v|u)}{P(v)}$
6. Conviction $(u \rightarrow v) = \frac{P(u)P(\neg v)}{P(u,\neg v)}$
7. Match $(u \rightarrow v) = \frac{P(u,v)-P(u)*P(v)}{P(u)*(1-P(u))}$
8. Accuracy $(u \rightarrow v) = P(u, v) + P(\neg u, \neg v)$
9. Leverage $(u \rightarrow v) = P(u, v) - P(u)P(v)$

We apply a normalization step to sensitize the AR mining algorithm to the popularity of parent or child proteins, which measured by the sizes of itemsets. Intuitively, it is more significant for two candidate proteins to share rare parents or children in the interaction network than popular ones, a property that gives itemset size a particular semantic significance in this application domain. For the derivation of a parametric normalization function, we refer the interested reader to [19].

IV. Experiment design

In this experiment, we use data set containing known protein–protein interactions (PPI) used in [20] (for both negative and positive examples). The goal of our experiments is to predict PPI using normalized and unnormalized numerical feature constructed from parent-child relationships. Our experiment design is modeled after that used by Taskar *et al.* [21] in the social network domain.

The data set of PPI consists of more than 10,000 positive protein pairs and around 10,000 known negative protein pairs. For preparing the datasets, both positive and negative sets are split into two parts for testing/training and all links (positive pairs) that connect the two sets are removed. The first step in training is to build a graph based on positive PPI and to represent the parent-child relations in the dataset like a market basket for similar analysis. The next step is to use 10,000 protein pairs – made of 50% positive and 50% negative examples – as the training set, and construct numerical features from the co-occurrence of proteins in the training parent-child dataset. For testing, the positive proteins pairs were used after eliding the number of existing links (positive PPI) (50%, 75%) and the rest are used to build an incomplete graph of positive pairs. Next, we construct four 5,000 protein-pair-datasets with 1%, 2%, 5%, 10% positive and 99%, 98%, 95%, 90% negative examples respectively as the test set. We do this because the real ratio of negative examples to positive examples is currently unknown. Finally, from only the known part of the graph we construct numerical features based on co-occurrence of proteins in the testing parent-child dataset. Therefore, the module will predict unknown links (from the hidden part) using numerical features of the known part.

This experiment uses only the connection structure of the positive PPI. We evaluated the normalized and unnormlized numerical features using two classifier models and inductive learning algorithms: the *k*-nearest neighbor approach IB1, and the rule based approach OneR.

V. Experiment Result

The performance expectation of the result shows how

models learned from the unnormalized numerical features are able to predict PPI relationships, and how performance further improves using normalized measures. In this paper, we have presented only the significant results for two of the numerical features, Accuracy and Leverage, where other features (interestingness measures) achieved a similar or a little less performance result. The results shown in Table 1 illustrate the classification performance measures in terms of precision, recall, F-measure, and area under curve (AUC) based on either Accuracy or Leverage features alone and 50% observed positive proteins pairs using the IB1 classification method. We see that the best AUC recorded was 0.854 in the dataset with 2% positive examples with normalized accuracy.

Using the second test sets with only 25% observed positive proteins (75% hidden), we see that the AUC is 0.781, representing a degradation of performance from the previous case. Table 2 shows complete results.

50%					
	method	Precision	Recall	F-Measure	AUC
1%	U- Accuracy	0.304	0.280	0.292	0.637
	N- Accuracy	0.174	0.660	0.275	0.814
	Different	-42.76%	135.71%	-5.82%	27.79%
	U- Leverage	0.319	0.300	0.309	0.647
	N- Leverage	0.288	0.340	0.312	0.666
	Different	-9.72%	13.33%	0.97%	2.94%
2%	U- Accuracy	0.458	0.270	0.340	0.632
	N- Accuracy	0.320	0.740	0.447	0.854
	Different	-30.13%	174.07%	31.47%	35.13%
	U- Leverage	0.492	0.310	0.380	0.652
	N- Leverage	0.468	0.370	0.413	0.681
	Different	-4.88%	19.35%	8.68%	4.45%
5%	U- Accuracy	0.624	0.212	0.316	0.603
	N- Accuracy	0.505	0.640	0.564	0.804
	Different	-19.07%	201.89%	78.48%	33.33%
	U- Leverage	0.648	0.236	0.346	0.615
	N- Leverage	0.622	0.276	0.382	0.634
	Different	-4.01%	16.95%	10.40%	3.09%
10%	U- Accuracy	0.761	0.204	0.322	0.599
	N- Accuracy	0.667	0.630	0.648	0.799
	Different	-12.35%	208.82%	101.24%	33.39%
	U- Leverage	0.778	0.224	0.348	0.609
	N- Leverage	0.753	0.256	0.382	0.624
	Different	-3.21%	14.29%	9.77%	2.46%

Table 1. IB1-Classification measures for 50% hiding
(U – Unnormalized, N – Normalized)

Figure 1 shows the comparison between normalized and unnormalized accuracy features based on AUC from the 50% observed data. The superiority of normalized features came from their ability to capture the rarity of childhood and parenthood of positive proteins.

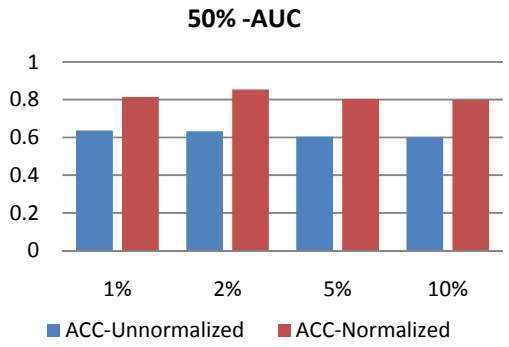

Fig-1: AUC result for 50% observed data for normalized and unnormalized Accuracy

75%					
	method	Precision	Recall	F-Measure	AUC
1%	U- Accuracy	0.118	0.480	0.189	0.722
	N- Accuracy	0.135	0.600	0.221	0.781
	Different	14.41%	25.00%	16.93%	8.17%
	U- Leverage	0.117	0.380	0.179	0.676
	N- Leverage	0.138	0.760	0.234	0.856
	Different	17.95%	100.00%	30.73%	26.63%
2%	U- Accuracy	0.211	0.480	0.293	0.722
	N- Accuracy	0.238	0.600	0.341	0.781
	Different	12.80%	25.00%	16.38%	8.17%
	U- Leverage	0.192	0.340	0.245	0.656
	N- Leverage	0.235	0.730	0.356	0.841
	Different	22.40%	114.71%	45.31%	28.20%
5%	U- Accuracy	0.381	0.444	0.410	0.704
	N- Accuracy	0.416	0.548	0.473	0.755
	Different	9.19%	23.42%	15.37%	7.24%
	U- Leverage	0.381	0.352	0.366	0.662
	N- Leverage	0.445	0.760	0.561	0.856
	Different	16.80%	115.91%	53.28%	29.31%
10%	U- Accuracy	0.554	0.448	0.496	0.706
	N- Accuracy	0.601	0.578	0.589	0.770
	Different	8.48%	29.02%	18.75%	9.07%
	U- Leverage	0.545	0.342	0.420	0.657
	N- Leverage	0.617	0.762	0.682	0.857
	Different	13.21%	122.81%	62.38%	30.44%

Table 2. IB1-Classification measures for 75% hiding
(U – Unnormalized, N – Normalized)

Figure 2 presents the AUC measure for 75% hidden. The result shows that the unnormalized measure is affected more when hiding more pairs that are positive.

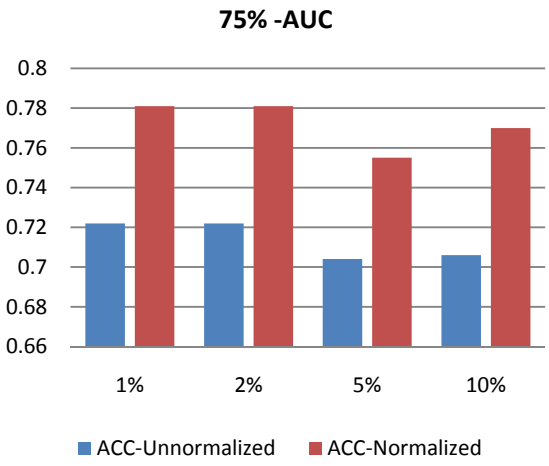

Fig-2: AUC result for 75% observed data for normalized and unnormalized Accuracy

Using all numerical features, in the case where we hide 50% we see the difference between the normalized and unnormalized measures as shown in Table 3. In the next case where we hide a 75%, there is no difference as shown in Table 4 because the skewness of itemsets size becomes insignificantly recognizable. In general, the numerical feature records a significant result where the AUC record between 0.973 and 0.98 in 50%-hidden datasets and between 0.873 and 0.89 in 75% hidden datasets.

The results further show the usefulness of using numerical features with proteins that share properties. The results obtained using normalized features are superior to those obtained using the original features.

50%					
%	method	Precision	Recall	F-Measure	AUC
1%	U- Accuracy	0.222	0.96	0.361	0.963
	N- Accuracy	0.221	0.980	0.360	0.973
	Different	-0.45%	2.08%	-	1.04%
2%	U- Accuracy	0.359	0.940	0.519	0.953
	N- Accuracy	0.364	0.990	0.532	0.978
	Different	1.39%	5.32%	2.50%	2.62%
5%	U- Accuracy	0.586	0.952	0.726	0.959
	N- Accuracy	0.588	0.988	0.737	0.977
	Different	0.34%	3.78%	1.52%	1.88%
10%	U- Accuracy	0.739	0.952	0.832	0.959
	N- Accuracy	0.742	0.994	0.850	0.980
	Different	0.41%	4.41%	2.16%	2.19%

Table -3 OneR-Classification measures for 50% hiding
(U- : Unnormalized, N- : Normalized)

75%					
%	method	Precision	Recall	F-Measure	AUC
1%	U- Accuracy	0.365	0.760	0.494	0.873
	N- Accuracy	0.365	0.760	0.494	0.873
	Different	0.00%	0.00%	0.00%	0.00%
2%	U- Accuracy	0.532	0.750	0.622	0.868
	N- Accuracy	0.532	0.750	0.622	0.868
	Different	0.00%	0.00%	0.00%	0.00%
5%	U- Accuracy	0.747	0.780	0.763	0.883
	N- Accuracy	0.747	0.780	0.763	0.883
	Different	0.00%	0.00%	0.00%	0.00%
10%	U- Accuracy	0.857	0.794	0.825	0.890
	N- Accuracy	0.857	0.794	0.825	0.890
	Different	0.00%	0.00%	0.00%	0.00%

Table -4 OneR Classification measures for 75% hiding
(U- : Unnormalized, N- : Normalized)

For our future work, we shall continue working in the domain of protein-protein interaction by adding more numerical features extracted from repositories of biological information. In addition, there is the possibility of using a genetic algorithm (GA) to selecting from among structural and biological features, which may lead to a further incremental boost in prediction quality.

ACKNOWLEDGMENT

We thank the Department of Defense for partial support of this research, Doina Caragea for helpful discussions, and Tim Weninger for his help in preparing this paper.

REFERENCES

[1] Y. Hao, X. Zhu, M. Huang, and M. Li, "Discovering patterns to extract protein--protein interactions from full texts," *Bioinformatics*, vol. 20, no. 18, pp. 3604-3612, 2004.

[2] M. Deng, F. Sun, and T. Chen, "Assessment of the Reliability of Protein-protein Interactions and Protein Function Prediction ," in *Pacific Symposium of Biocomputing (PSB2003)*, 2003.

[3] C. Lin, D. Jiang, and A. Zhang, "Prediction of Protein Function Using Common-Neighbors in Protein-Protein Interaction network," in *Proceedings of the Sixth IEEE Symposium on BionInformatics and BioEngineering* , Arlington, VA , 2006.

[4] Y. Qi, J. Klein-seetharaman, and Z. Bar-joseph, "A mixture of feature experts approach for protein-protein interaction prediction," *BMC Bioinformatics*, vol. 8(Suppl 10):S6, 2007.

[5] F. Pan, G. Cong, A. K. H. Tung, J. Yang, and M. J. ,. Zaki, "Carpenter: finding closed patterns in long

biological datasets," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* , Washington, D.C, 2003.

[6] G. Cong, K.-L. Tan, A. K. H. Tung, and F. Pan, "Mining Frequent Closed Patterns in Microarray Data," in *Proceedings of the Fourth IEEE International Conference on Data Mining*, Washington, DC, USA, 2004.

[7] R. Martinez, N. Pasquier, and C. Pasquier, "GenMiner: Mining non-redundant association rules from integrated gene expression data and annotations," *Bioinformatics*, 2008.

[8] X.-R. Jiang and L. Gruenwald, "Microarray gene expression data association rules mining based on BSC-tree and FIS-tree," *Data & Knowledge Engineering*, vol. 53, no. 1, 2005.

[9] F.-H. Hung and H.-W. Chiu, "Protein-Protein Interaction Prediction based on Association Rules of Protein Functional Regions," in *Proceedings of the Second International Conference on Innovative Computing, Informatio and Control*, 2007, p. 359.

[10] T. Oyama, K. Kitano, K. Satou, and T. Ito, "Mining Association Rules Related to Protein-Protein Interactions," *Genome Informatics*, vol. 11, p. 358–359, 2000.

[11] T. Oyama , K. Kitano , K. Satou and T. Ito "Extraction of knowledge on protein-protein interaction by assciation rule descovery," *Bioinformatics*, vol. 18, no. 5, pp. 705-714, 2002.

[12] M. E. Turanalp and T. Can, "Discovering functional interaction patterns in protein-protein interaction networks," *BMC Bioinformatics*, p. 9:276, 2008.

[13] M. Kotlyar and I. Jurisica, "Predicting protein-protein interactions by association mining," *Information Systems Frontiers*, pp. 37-47, 2006.

[14] T. Oyama, et al., "Automatic Extraction of Expression-Related Features Shared by a Given Group of Genes," *GENOME INFORMATICS SERIES*, pp. 312-313, 2003.

[15] C. Besemann, A. Denton, A. Yekkirala, R. Hutchison, and M. Anderson, "Differential Association Rule Mining for the Study of ProteinProtein Interaction Networks," in *BIOKDD*, Seattle, WA, USA, 2004, pp. 72-80.

[16] J. R. Bock and D. A. Gough, "Predicting protein–protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, pp. 455-460, 2001.

[17] L. Geng and H. J. Hamilton, " Interestingness measures for data mining: A survey," in *ACM Comput. Surv*, Sep.2006, p. 38,3.

[18] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in *Proceeding of the ACM SIGKDD international conference on knowledge discovery in databases (KDD'02)*, Edmonton, Canada, 2002, p. 32–41.

[19] W. Aljandal, W. H. Hsu, V. Bahirwani, and D. W. T. Caragea, "Validation-based normalization and selection of interestingness measures for association rules," in *In Proceedings of the 18th International Conference on Artificial Neural Networks in Engineering (ANNIE 2008)*, St. Louis, MO, 2008.

[20] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, pp. 38-46, 2005.

[21] B. Taskar, M. Wong, P. Abbeel, and D. Koller "Link Prediction in relation data" Advances in Neural Information Processing Systems (NIPS 2003)