Classification by Data Integration: From Clinical Records to Protein data

Xiaowei Zhu

The aim of functional genomics is the prediction of function for unknown proteins. This type of prediction is very complicated. A lot of properties, such as protein folds, protein sequences and gene expression data have some relations to the protein functions. However, no single relation is strong enough for correct protein function prediction. Recently more and more researchers are learning that good prediction results can only be achieved by integration of different kinds of data. The power of data integration is obvious, but it is even more complicated since data are from diverse natures. Classification by aggregating emerging patterns (CAEP), a classifier for disease diagnosis, is based on integration of different patient attributes. I propose that the same idea as CAEP can be used in protein function classification.

1. Background:

Datamining is a natural evolution of information technology along the path of data collection, database creation, database management, and data analysis and interpretation (1).

1.1 Datamining and Classification:

In general, datamining can be spilt into two categories: descriptive and predictive (1). Descriptive datamining emphasizes on discovering the internal patterns of the data. On the other hand, predictive datamining performs inference on the current data in order to make prediction.

Classification, which will be discussed in this paper, belongs to predictive datamining tasks. A training set with known classes is necessary for constructing a model to discriminate different classes. Then this model can be used for prediction. The model is based on the relations between targeted attribute and other related attributes. Integration of different attributes should be performed in order to make the model more accurate, since any single link is not strong enough to make the prediction.

1.2 The process of datamining:

Datamining can be divided into six stages (2): data selection, cleansing, enrichment, coding, datamining, and reporting.

The first step is data selection. For disease diagnosis classification, the clinical information from the normal and disease groups should be collected. For the protein function classification, a training set with known functions, structures and expression data etc. should be collected.

The second step is data cleansing. The aim of this step is to tidy up the data. For example, clinical records from different hospital should be combined into same terminologies and units. As for the proteins, redundancy in the database should be filtered before datamining.

The third step is data enrichment, which means to acquire other kinds of data that can be integrated into the exiting data. For instance, protein-protein interaction and protein localization are also useful for protein function prediction.

The fourth stage is coding, where data are transformed or consolidated into forms appropriate for datamining. The description in the clinical information should be transformed into numbers.

The fifth step, datamining, is an essential process where intelligent methods are applied in order to extract data patterns.

The last step, reporting, is to give the results to the users by some techniques such as graphs and visualization.

2. CAEP method for clinical data

2.1 Problem:

Given a training set of patients with known classification and other clinical information, we can use CAEP to predict whether new persons have a disease by their clinical information (3,4).

Diabetes is a good example for interpreting the problem. Different kinds of clinical records of patients and normal people were collected to make a training set. The problem is to construct a model from this dataset, and the model should reflect the relations between the disease and other features. Then this model can be used to analyze the clinical data of new people, and prediction of whether they have the disease can be made.

Clinical records are very different from other kinds of data. The clinical information related to diabetes is heterogeneous. Many attributes relate to the disease, such as the age of the patients, triceps skin fold thickness and body mass index etc. Furthermore, any data in a single attribute is not strong enough for a correct predictor. Hence combination of different attributes is necessary.

2.2 CAEP—the method for data integration

The CAEP method is based on the idea of *emerging patterns* (3). An emerging pattern is a pattern whose frequency increases significantly from one class to another. For example, the pattern {plasma glucose concentration > x%} is shown in 60% of diabetes patients versus 6% of normal people. The pattern frequency is 10 times in patient group than in normal group, so it is an emerging pattern. This difference can be used to distinguish different groups latterly in class prediction.

The main idea of CAEP is to describe all clinical attributes respectively by emerging patterns. Then the relations between the emerging patterns and the disease classification are calculated to construct the prediction model. The process is explained as follows.

In the training set, the percent of members that has pattern X is calculated in each group D_i. It is called the *support* of X in $D_{i.}$ Thus each emerging pattern has higher support in some groups than others.

$$\mathrm{Supp}_{Di}(X)=\frac{\text{number of } D_i \text{ members that have pattern X}}{\text{number of } D_i \text{ memebers}} \qquad (1)$$

Then we can use the *support* for disease prediction. Given a person with an emerging pattern X, the likelihood that he or she is in D_i can be calculated. Thus the person can be predicted to be in a specific group D if the $\mathrm{likelihood}_D(X)$ is the largest among all groups D_i.

$$\mathrm{likelihood}_{Di}(X)=\frac{\mathrm{Supp}_{Di}(X) * |D_i|}{\sum_j \mathrm{Supp}_{D_j}(X) * |D_j|} \qquad (2)$$

$|D_i|$ means the number of D_i memebers

This kind of prediction only uses one single emerging pattern X. CAEP method integrates different kinds of data by adding up the differentiating power of every emerging pattern linearly. In the equation below, t means a new person who needs disease prediction. Score (t, D_i) means the probability that t belongs to group D_i. It is calculated by adding all the emerging pattern information.

$$\mathrm{score}(t, D_i) = \sum_{X} \mathrm{likelihood}_{Di}(X) * \mathrm{Supp}_{Di}(X) \qquad (3)$$

In order to be more robust, the score should be normalized by dividing with a score at a fixed percentile of the members in each group. Now we can classify a new person to a group D_i if norm_score(t, D_i) is higher than any others. The base_score (D_i) is usually chosen as the median of score (t, D_i) over all members in D_i.

$$\mathrm{norm\_score}(t, D_i) = \frac{\mathrm{score}(t, D_i)}{\mathrm{base\ score}(D_i)} \qquad (4)$$

3. Protein function classification

3.1 Problem and difficulties

The aim of functional genomics is to predict functions for unknown proteins. The other attributes of proteins, such as protein sequences, protein structures and gene expression data, are much easier to be gathered than their functions. It was also proved that these attributes are related to protein functions, although the extent of these relations had not been identified clearly. Thus the challenge is to construct a model for connecting these attributes to protein function classification. Then we can use this model for function prediction.

This problem is similar to disease prediction, but in this case the prediction is much more complicated because of the distinct characteristics of protein functions (5).

First of all, there is currently no universal classification for protein functions. However, the CAEP method depends on the known classification in the training set. It cannot generate new classification.

Second, for disease prediction, one person belongs to and only belongs to one group, whereas for proteins, one protein can have more than one function. Furthermore, sometimes one function needs several proteins.

Third, classifications may mean different things when they refer to molecular action, cellular roles and phenotypic manifestation.

Fourth, the terminology for proteins is disordered.

Fifth, in the diabetes example, there are only two groups: patients with the disease and the healthy group. With protein functions, there are much more protein classes.

Finally, the relations among the protein structures, protein sequences and gene expression are not clear. Thus it may be dangerous if we simply integrate all these kinds of data by a linear addition. One modification can be applying one factor, or multiplier, to each kind of data, and the factors relate to the importance of that data for function prediction.

In general, I think we can use the same idea as CAEP to protein function prediction, but this case will definitely need some modifications to the method.

3.2 Method for protein function classification

3.2.1 Translating attributes data into emerging patterns.

In the training data set, each protein has known function, known expression pattern, known folds information and known sequence. We can then put that protein into specific function classes, structure classes, expression patterns and sequence similarity groups. Some of these groups such as folds already exit, and we can use that classification. While as for protein sequences, we can cluster protein according to their sequence similarity. Thus we can define an emerging pattern as “protein belongs to a specific fold group” or "protein belongs to a specific expression pattern", and then analyze its frequency in different function classes.

The parameters such as *support* and *likelihood* for each emerging pattern can be calculated similarly as the CAEP method.

3.2.2 Integrating the emerging patterns.

To integrate the different data is much more complicated. As I mentioned before, the data such as structures and expression data may not be independent to each other. Hence it will be misleading if they are integrated by a simple addition. To analyze the real relations among this data is too complicated to be practical. Thus I still choose a simplified linear method, but in compensation for this simplification, one multiple factor is applied to each kind of data. Different kinds of data obviously have different effects to the function classification. Furthermore, the effect also changes in different function class. For instance, folds information may be very useful to one function class, while expression pattern is more important for predicting another function class. The factor reflects this difference. Thus the equation for score (t, D_i) changes to:

$$score(t, D_j) = \sum_{X} likelihood_{Di}(X) * Supp_{Di}(X) * F_{ij} \qquad (5)$$

The factor F_{ij} depends on the nature of the emerging pattern j and the specific function class D_i. For instance, all emerging patterns in protein structures have the same F_{ij} for a specific function class $D_{i.}$

For each class D_i in the training set, we can choose F_{ij} that maximizes the difference between *norm_score*(t, D_i) (D_i is the right class) and the *norm_score*(t, D_i) (D_i is the wrong class).

3.2.3 Interpreting the results.

Changes should also be made to interpret the results. We may need to select more than one class in protein function prediction. In CAEP, we can compare the number *norm_score* (t, D_i) and then determine the class by selecting the largest number. However, this method cannot be used in protein function classification any more, since one protein may have several functions. Thus we must determine a threshold for each function class. If the score is larger than this threshold, we will say that this protein belongs to this function class, no matter how many classes the protein belongs to.

To determine this threshold for each function class, the scores of proteins that do not belong to this class are gathered and their distribution of them is calculated. Then for a special *norm_score*(t, D_i), the probability P that the protein “t” does not belong to class D_i can be calculated by the distribution curve. Thus a threshold can be selected to this probability. We can say that the protein belongs to the functional classes if the probability P is lower than the threshold.

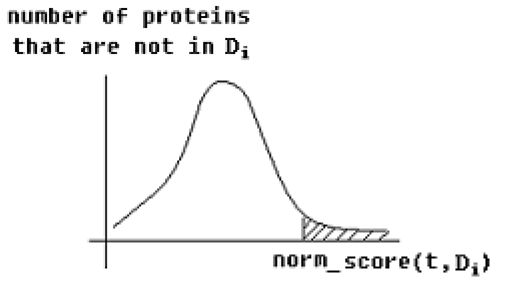

Fig.1: distribution of unrelated protein scores to a function class

The distribution curve is distinct for each function class. Unfortunately, to generate distribution curves for all function classes may require a large number of proteins in the training set.

4. Discussion

4.1 CAEP and Bayesian method

By simple calculation, I find that the CAEP method is similar to the Bayesian method. The *support* of X in D_i is the probability P (t has X| t $\in D_i$) and the likelihood$_D$ (X) is the probability P (t $\in D_i$| t has X). According to the Bayesian law:

$$P(t \in D_i \mid t \text{ has } X) = \frac{P(t \text{ has } X \mid t \in D_i) * P(t \in D_i)}{\sum_j P(t \text{ has } X \mid t \in D_j) * P(t \in D_j)} \qquad (6)$$

$$= \frac{Supp_{Di}(X) * |D_i|}{\sum_j Supp_{D_j}(X) * |D_j|} \qquad \left(P(t \in D_i) = \frac{|D_i|}{\sum_j |D_j|} \right)$$

4.2 The model needs more evaluation

The CAEP method has very good predictive accuracy on clinical data sets (3,4). It gives better accuracy than previous classification algorithms such as C4.5 (6) and CBA (7) in general. Furthermore, CAEP is equally accurate on all classes. This is very useful for many applications, where there are a dominant class and a minority class.

However, nobody has tried to use this method for protein function prediction. The difference and difficulties of this prediction has been discussed before. I made several modifications to the method. The result of these changes is unknown, for example, it may cause the model to over fit to the training set. Thus an evaluation on a protein data set by this method will be needed.

5.Reference

1. Han, J., and Kamber, M. (2000). *Data Mining: Concepts and Techniques.* San Francisco: Morgan Kaufmann.
2. Adriaans, P., and Zantinge, D. (1996). *Data Mining.* Harlow, UK: Addison Wesley Longman.
3. Dong, G., and Li, J. (1999). Efficient mining of emerging patterns: Discovery trends and difference. In *Proc. 5th ACM SIGDD Intl. Conf. on Knowledge Discovery & Data Mining,* 15-18. New York: ACM Press.
4. Dong, G., Zhang, X., Wong, L., and Li, J. (1999). CAEP: Classification by aggregating emerging patterns. In *LNCS 1721: Discovery Science*, Arikawa, S., and Furukawa, K., eds., 30-42. Berlin: Springer
5. Gerstein, M. (2000). Integrative database analysis in structural genomics. *Nat Struct Biol.* **Suppl**, 960-963.
6. Quinlan, J. R. (1992). *C4.5: Programs for Machine learning.* San Mateo, Calif.: Morgan Kaufmann.
7. Liu, B., Hsu, W., and Ma, Y. (1998). Integrating classification and association rule mining. In *Proc. 4th Intl. Confl. on Knowledge Discovery in Databases and Data Mining,* Agrawal, R., Stolorz, P.E., and Piatesky-Shapiro, G., eds., 80-86. Menlo Park, Calif.: AAAI Press.