Sanjay Chawla

Statement on Research

The bulk of my research output has been in the area of data mining with a recent emphasis on outlier detection with applications to biological data. I also continue to work in spatial data mining - which was how I was introduced to the field.

Outlier Detection

Of all the data mining techniques that are in vogue, Outlier Detection comes closest to the metaphor of *mining* for nuggets of information in large databases. My working definition of outlier detection is slightly more broad than what is conventionally accepted. For example, I consider mining for rare association rules (low support and high confidence), classification for imbalanced data sets and of course the use of non-parametric techniques (e.g., distance and density based) to find isolated entities as examples of Outlier Detection.

Discovery of Rare but Confident Rules

From a Frequent Mining(FM) perspective, microarray expression data is an $N \times M$ boolean matrix where N is the number of experiments and M is the number of genes and a non-zero entry (i, j) indicates that gene j was expressed in experiment i. The objective of FM is to find rules of the form

$$\textsc{gene}1 \Rightarrow \textsc{gene}2 \text{ (support 10\%, confidence 90\%)}$$

which mean that when GENE1 is expressed, 90% of the time GENE2 is also expressed, and that GENE1 and GENE2 are expressed together in 10% of the microarray experiments.

The original Apriori and related algorithms (FP-Tree) were designed for data sets where effectively $N >> M$. In order to handle the ”almost transpose” nature of microarray data, a new family of algorithms, collectively referred to as *row-enumeration* algorithms have emerged in the literature. While an inordinate amount of work has gone into the design of new algorithms for frequent mining, my work has focused on an orthogonal issue: how to design and customize *measures* for specific application domains which retain the anti-monotonicity property. Once such a measure has been invented it can be plugged into an appropriate pattern mining algorithm and the output examined to determine if the *measure* captures the appropriate semantics of the domain.

Coming back to the mining in micoarray data, an important point to be aware of is that there are different types of microarray experiments and support-based FM is not necessarily meaningful for all types.

In our paper, *High-Confidence Rule Mining for Microarray Analysis* which has been accepted to appear in *IEEE/ACM Transactions on Computational Biology and Bioinformatics* , we exclusively focus on perturbation microarray experiments. Perturbation experiments are based on the rationale, that if a gene or cell is no longer able to function normally, the expression levels of other genes that are functionally related will be altered. In perturbation data, each row corresponds to a cell which may be genetically altered to prevent the expression of a selected gene, to infer its affect. We have two original contributions in the above mentioned paper.

1. We have shown that support-based mining is not appropriate for mining in perturbation data.

2. We designed a new monotonic measure, based solely on confidence, specifically for perturbation data.

3. We carried out an extensive domain-based validation to determine the usefulness of the rules discovered. The domain-based validation was carried out using the Biomolecular Interaction Network Database(BIND), which documents gene interactions in wet lab experiments and the Gene Ontology.

Validation of rules is absolutely essential for rare rule discovery because if the output is not properly validated we could be mining *meaningless noise* instead of *meaningful information*. Another important point to note is that randomization based validation cannot be used for perturbation data -because each row has been individually perturbed. Thus they (the rows) are not identical (not iid).

Sequential Outliers

Suppose we have a set of sequences (for example sequences of amino acids). For example, suppose we have three sequences:

1. SRHPAZBGKPBFLBCYVSGFHPXZIZIBLLKB

2. IXCXNCXKEGHSARQFRA

3. MGVRNSVLSGKKADELEKIRLRPGGKKKYMLKHIVWAANELDRFGLAESLLENKEGCQKI

We want to determine which one of the sequence is an outlier compared to the rest?
In our paper, *Mining for Outliers in Sequential Databases*, which won the best application paper award in the 2006 SIAM International Conference on Data Mining, we proposed a solution to the above problem. Our solution was based on constructing a Probabilistic Suffix Tree(PST) to encode a variable-length markov chain. This basically means that given a sequence $s_1s_2\ldots s_n$ we can use a PST to efficiently calculate

$$P(s_1s_2\ldots s_n) = P(s_1)P(s_2|s_1)P(s_3|s_1s_2)\ldots P(s_n|s_1\ldots s_{n-1})$$

Now each of the terms on the right hand side, except the first one, are conditional probabilites and a variable length markov chain model automatically determines the appropriate length of conditioning suffix. Furthemore, and this is an important observation, for outlier detection, all the outliers typically tend to be close to the root of the PST. Thus if the objective is to mine for outliers, a very small portion of the tree has to be retained. In our experiments we have shown that by retaining around 10% of the original PST we can mine more than 95% of the original outliers. To the best of our knowledge this was the first paper which illustrated the use of PSTs for mining outliers.

Imbalanced Classification

One of the great challenges in classification is to handle the imbalance class problem. This situation occurs when only a few training examples of one class (usually the class of interest) are given compared to the number of examples from the other classes. In some sense the minority class can be considered as documented examples of outliers (for example in the case of fraud detection these are the actual fraudulent cases). The usual way of handling the imbalanced class problem is to oversample the minority class or undersample the majority class.

In a recent paper, *CCCS: A Top-Down Associative Classifier for Imbalanced Data*, which appeared in the 2006 ACM SIGKDD proceedings we have proposed a new anti-monotonic measure which captures the relationship between the two classes. The new measure, called the Complement Class Support (CCS), effectively makes asssociation rule mining discriminative when the data set also contains class information. While there are several approaches which use association rule mining for classification, ours is the first approach which contrasts[1] the class information. Furthemore we have shown that the CCS measure is closely related to correlation. In particular we guarentee that the rules which survive the mining process will be positively correlated (between the antecedent and consequent). This is the first paper which shows how Associative Classifiers can be used to solve the classification problem for imbalanced data sets.

Going Forward..

My plans going forward are to continue work in data mining with a major focus in outlier detection. Specifically

1. Along with Prof. David Hand (Imperial College, UK) I am co-editing a book on outlier detection which will be published by Elsevier. Our plan is to have the book out by early 2008.

2. Our work on designing new computationally efficient measures (e.g., anti-monotonic) has a lot of potential in bioinformatics and other application verticals.

3. I am currently working with a Sydney-based company on a neuroinformatics project. In particular we are using data mining techniques to predict the onset of certain brain related diseases.

4. The application of outlier detection techniques to discover insider-trading patterns in the securities market is another area for which I have recently won industry funding.

[1] This is very different from contrast sets which are used for feature generation