

Additional paper for “Biological Data Mining”

Paper Reference:

Sebban, N et al. (2002) A data mining approach to spacer oligonucleotide typing of mycobacterium tuberculosis. *Bioinformatics*, 18, 235-243

Abstract:

In this paper the authors applied data mining and machine learning techniques to classify spacers into specific classes and compare it to that of the classification by human experts. In other words, the authors state that this was the first attempt to automatically discover simple knowledge rules from *spacer oligonucleotide* typing (a DNA analysis technique) or *spoligotyping* data. More technically, the algorithm used was C4.5 induction algorithm. Decision trees were generated and corresponding rules of inference established. Besides, prototype selection methods were later incorporated to remove irrelevant strains and a report regarding the usefulness of such a process was generated. The final section of the paper also talks about the contribution of additional 25 new spacers. An analysis of the correlation between different spacers was also presented.

Discussion:

The results from the experiments are very interesting and insightful. Let us briefly try to comprehend the problem definition before we try to discuss the results. In this paper the authors used spacers in each spoligotype of the databases as a feature input to the data mining algorithm. The total numbers of training sets were total number of spoligotypes. (n= 7352 for DB1 and n=323 for DB2). The class labels were the class of each spoligotype. Thus given the input features, the data mining algorithm builds a decision tree to predict the classes of the input spacer. One important aspect to note here is that such decision trees are not only useful for providing very efficient classification models but also for describing the process from the input vector to the output class labels. This is of great value in medical diagnosis systems.

The accuracy of each such model was then evaluated via well known statistical methods namely cross-validation techniques. The accuracy of each rule is described via the confidence interval of the proportion of correctly classified strains. The learning accuracy of one of the databases DB1 suggests that it is possible to build a predictive model, which would automatically classify new profile. A 5 fold cross validation resulted in a success rate of 98% and standard deviation of .8%. The authors implemented the PSRCG (PS Relative Certainty Gain) algorithm for deleting irrelevant strains. It globally improved information representation space. Moreover results proved the latter does not depend on a learning algorithm and does not generate an inductive bias. It resulted in direct decrease in number of rules and an increase in accuracy.

The study also resulted in simpler rules as compared to the ones generated by experts in previous studies. A possible explanation can be that decision trees use pruning to avoid overfitting whereas human experts take into account all data as signals. Thus data mining techniques allow distinguishing useful information from noise. This leads to downweighting or excluding uninformative spacers while dealing with phylogeny reconstruction using parsimony principles. Thus this a novel approach and definitely has distinct edge over traditional classification approaches.