

Additional Paper for “Biological Data Mining”

Paper reference

Kretschmann, E., Fleischmann, W., and Apweiler, R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on Swiss-Prot. *Bioinformatics*, **17**, 920-926.

Abstract

In this age of high-throughput sequencing, a huge amount of newly available protein data is being continuously submitted to protein databases. This data needs to be annotated in order to be made more useful and valuable for the users. Previously human experts manually handled this annotation task by a literature curation process mainly concerning protein descriptions such as names and synonyms, comments, keywords, and sequence features. However, this manual annotation process has been unable to deal with the ever increasing volume of the data being submitted. Hence, greatly needed were some automated methods to tackle this task.

In this paper, the authors present such a method utilizing a Java-based implementation of the renowned C4.5 data mining algorithm to generate rules, based on taxonomy and sequence similarity and signature searches, to automate the Keyword annotation for SWISS-PROT.

In order to obtain a reasonable speed in producing the decision trees (of C4.5) while maintaining a reasonable confidence in the result, they divide the proteins in the SWISS-PROT to a number of groups each of which ideally contains similar proteins. To do so groups common to InterPro entries were utilized.

They have also devised a criterion using which to select only the best rules with a reasonably high confidence. To do so, they use the number of TP (True Positive) and FP (False Positive) examples to calculate which rules lie above a given threshold (in terms of their confidence) in a reasonably high percentage of all cases.

After selecting the satisfactorily high-confidence rules, the authors also evaluated the reliability of the obtained rules using the popular tenfold cross-validation procedure run over the SWISS-PROT data. Observed in this evaluation process was the fact that error-rates obtained in the cross-validation stage were better than what was suggested by the confidence calculated for the rules. To sum up, the method worked quite satisfactorily.

Discussion

Noticing the increasing need for automated annotation of protein databases, the problem considered in this paper is of real practical value. The utilization of the C4.5 algorithm to solve this problem seems favorable, as it frequently produces human readable rules lending themselves to intuitive interpretations.

Also the idea of dividing the entire database into a number of groups, adds to the practical value of the presented method, by making it fast while maintaining the quality of the result at a reasonable level.

Calculating the confidence of the resulted rules and rejecting those rules with confidence levels below a threshold brings up a trade off between coverage and confidence; that is, the method is able to produce either more annotations with less confidence or vice versa.

Despite all these, the method bears some limitations and disadvantages as well. One of the main limitations of the method is that it only produces Keyword annotations; that is, it does nothing for other annotation items such as Description, Comment, or Feature Lines.

Another drawback of the presented method is the way it calculates confidence. The proposed routine does not take into account the number of TN (True Negative) or FN (False Negative) instances. Hence, it prefers the generation of frequent keywords (i.e. general Keywords) rather than rare ones (i.e. specific ones) while it is clear that the latter is of much more value.