Introduction

Recent studies[1, 2] into the use of a selection of similarity coefficients, when applied to searches of chemical databases represented by binary fingerprints, have shown considerable variation in their retrieval performance and in the sets of compounds being retrieved. The main factor influencing performance is the density of the similarities for the class of the query compound, a feature which is closely related to the molecular size of the active class.

It was found that some coefficients, the Fortes and Simple Match for instance, are more efficient at retrieving classes of relatively small compounds, whereas others, like the Russell/Rao, are more useful for larger actives.

If this is the case when these coefficients are applied to similarity searches, then we would expect considerable variation in performance when applied to dissimilarity methods, namely clustering and compound selection.

Here we report on several studies which have been undertaken to investigate the relative performance of thirteen association and correlation coefficients (Table 1), which have been shown to exhibit complementary performance in similarity searches, when used to cluster a 20K subset of the MDL Drug Data Report database (MDRD) using hierarchical and non-hierarchical methods. In addition, the same coefficients have been applied to a compound selection routine to select a diverse selection of the 20K compounds. In all cases, the representation used was the 1025 standard 505 fingerprints from Digital Chemistry.

Hierarchical Clustering

The group-average agglomerative clustering algorithm was used to cluster the dataset thirteen times, using one of the coefficients of Table 1 as the similarity metric each time. During the clustering process, a measure of relative performance was calculated at every 100th iteration of the agglomeration process. The measure used was

$$\text{AC} = \frac{\text{Number of clusters retrieved}}{\text{Number of classes associated}}$$

wherein, for a given active class, AC is the total number of active compounds in the active clusters (an active cluster being any cluster containing at least one member of the active class) and n is the total number of compounds in the active classes.

In order to illustrate any size dependency which might exist between coefficients, eleven separate active classes were chosen to evaluate the performance measure. These are shown in Table 2, with the number of actives given for the 20K MDDR subset.

Figure 1 illustrates the results for the Angiotensin II AT1 antagonists and the Protein kinase C inhibitors (other classes are comparable to these). These results clearly illustrate the poor performance of the Russell/Rao, Fortes and Simple Match which give consistently low values for the measure. They also illustrate that the correlation coefficients, Yule, Dennis, Pearson and Sikes, are generally good performers, comparable, and often better, than the Tanimoto. The Baroni-Urbani/Buser is also a consistently good performer.

Similar results are seen for all active classes tested, indicating that it is unlikely that the size is a size class relationship between coefficient and performance.

Non-hierarchical Clustering

Non-hierarchical clustering was performed using the Jarvis-Patrick clustering algorithm with the nearest neighbor list length varying from 14 to 20 and with common nearest neighbors varying from 6 to 8. Table 4 shows the results, at 14 and 8, with singleton clusters excluded. For each class, the best performing coefficient is shown in red and those performing within 10% of this are shown in gold. The final row indicates the number of times a coefficient is a good performer (red or gold).

Tables 3a and 3b illustrate the relative performance at clustering levels of 2000 and 1000 clusters. For each class, the best performing coefficient is shown in red, those performing within 10% of this are shown in gold. The final row indicates the number of times a coefficient is a good performer (red or gold). Table 3b gives these values for the 500 cluster level. Notably, the Russell/Rao is always the poorest performer.

Conclusions

We have applied thirteen similarity coefficients to clustering and compound selection routines to assess their relative performance. The results appear to indicate that the correlation coefficients (Pearson, Yule, Sikes and Dennis), as well as the Baroni-Urbani/Buser coefficient, are the most consistently efficient when assessed using our performance measure. The standard measures used in the chemoinformatics field, the Tanimoto and the Euclidean Distance (equivalent to the Simple Match) have been found to be inferior choices when applied to these techniques. In particular, the Simple Match is one of the worst tested.

Comparison of Similarity Coefficients for Clustering and Compound Selection

Maciej Haranczyk1 and John Holliday2

References


1 Department of Chemistry, University of Gdansk, 80-952 Gdansk, Poland
2 Department of Information Studies, University of Sheffield, Sheffield S1 4DP, UK