Comparison of Similarity Coefficients for Clustering and Compound Selection

Maciej Haranczyk[1] and John Holliday[2]

The University Of Sheffield.

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 bitstrings 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 Forbes 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 (MDDR) 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 BCI standard 1052 fingerprints from Digital Chemistry.

Similarity Coefficients

$\frac{a}{a+b+c}$ Tanimoto (Tan)	$\frac{a}{n}$ Russell/Rao (Rus)	$\frac{a+d}{n}$ Simple Match (SM)
$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$ Baroni-Urbani/Buser (Bar)	$\frac{a}{\sqrt{(a+b)(a+c)}}$ Ochiai/Cosine (Cos)	$\frac{na}{(a+b)(a+c)}$ Forbes (For)
$\frac{a}{2}\frac{(2a+b+c)}{(a+b)(a+c)}$ Kulczynski(2) (Ku2)	$\frac{n\left(a-\frac{1}{2}\right)^2}{(a+b)(a+c)}$ Fossum (Fos)	$\frac{a}{\min(a+b,a+c)}$ Simpson (Sim)
$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$ Pearson (Pea)	$\frac{ad-bc}{ad+bc}$ Yule (Yul)	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$ Dennis (Den)
$\log_{10}\frac{n\left(\lvert ad-bc\rvert-\frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$ Stiles (Sti)	In which *a*: number of bits in common *b*: bits set uniquely in first bitstring *c*: bits set uniquely in second bitstring *d*: bits set in neither bitstring *n*: size of bitstring	Association Coefficients Correlation Coefficients

Table 1

Activity Classes

Active Class	Class ID	Number of Actives
5HT3 Antagonist	06233	154
5HT1A Agonist	06235	160
5HT Reuptake Inhibitor	06245	68
D2 Antagonist	07701	75
Renin Inhibitor	31420	230
Angiotensin II AT1 Antagonist	31432	183
Thrombin Inhibitor	37110	168
Substance P Antagonist	42731	231
HIV-1 Protesae Inhibitor	71523	147
Cyclooxygenase Inhibitor	78331	130
Protein Kinase C Inhibitor	78374	83

Table 2

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

$$\frac{nA}{nC}$$

wherein, for a given active class, *nA* 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 *nC* is the total number of compounds in the active clusters.

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, Forbes and Simple Match which give consistently low values for the measure. They also illustrate that the correlation coefficients, Yule, Dennis, Pearson and Stiles, 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 is a size or class relationship between coefficient and performance.

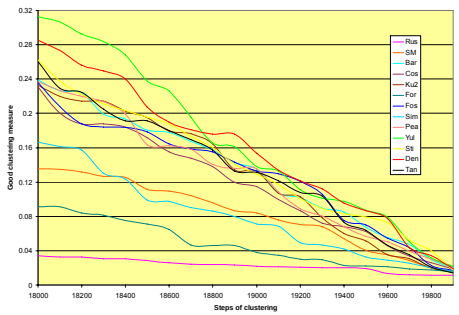

Figure 1

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 3c gives these values for the 500 cluster level. Notably, the Russell/Rao is always the poorest performer.

a: 2000 Clusters:

ID	Tan	Rus	SM	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
06233	0.1409	0.0256	0.0791	0.1542	0.1522	0.1570	0.0440	0.1588	0.1260	0.1591	0.1628	0.1557	0.1528
06235	0.1124	0.0253	0.0910	0.1174	0.1063	0.1121	0.0401	0.1154	0.0969	0.1331	0.1143	0.1269	0.1292
06245	0.0798	0.0141	0.0402	0.0640	0.0708	0.0775	0.0148	0.0740	0.0652	0.0743	0.0808	0.0783	0.0706
07701	0.0896	0.0120	0.0554	0.0859	0.0810	0.0879	0.0178	0.0828	0.0690	0.0939	0.0748	0.0866	0.0727
31420	0.1351	0.0416	0.1580	0.1459	0.1241	0.1472	0.1051	0.1267	0.1237	0.1285	0.1570	0.1335	0.1388
31432	0.2603	0.0343	0.1355	0.2392	0.2308	0.2331	0.0913	0.2358	0.1664	0.2370	0.3128	0.2662	0.2850
37110	0.1086	0.0272	0.1107	0.1296	0.1137	0.0996	0.0421	0.1033	0.0929	0.1597	0.1327	0.1466	0.1296
42731	0.1151	0.0302	0.1088	0.1486	0.1196	0.1286	0.0480	0.1147	0.0925	0.1359	0.1314	0.1314	0.1197
71523	0.0938	0.0212	0.0832	0.1102	0.0866	0.0963	0.0309	0.0894	0.0788	0.0877	0.0971	0.0831	0.0983
78331	0.1285	0.0188	0.0573	0.1064	0.1186	0.1432	0.0236	0.1212	0.1059	0.1162	0.1335	0.1201	0.1105
78374	0.0621	0.0134	0.0357	0.0652	0.0732	0.0534	0.0149	0.0716	0.0547	0.0724	0.0744	0.0747	0.0628
>90%	2	0	1	5	2	5	0	3	0	7	6	6	3

b: 1000 Clusters:

ID	Tan	Rus	SM	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
06233	0.0798	0.0162	0.0374	0.0795	0.0802	0.0711	0.0281	0.0809	0.0693	0.0783	0.0773	0.0918	0.0758
06235	0.0513	0.0132	0.0421	0.0654	0.0470	0.0586	0.0265	0.0540	0.0480	0.0535	0.0604	0.0592	0.0568
06245	0.0348	0.0081	0.0164	0.0288	0.0349	0.0331	0.0088	0.0350	0.0261	0.0360	0.0295	0.0406	0.0284
07701	0.0466	0.0060	0.0224	0.0385	0.0324	0.0406	0.0127	0.0307	0.0338	0.0388	0.0415	0.0404	0.0379
31420	0.0985	0.0268	0.0915	0.1108	0.0952	0.1125	0.0496	0.0920	0.0839	0.0941	0.0958	0.0978	0.1072
31432	0.1310	0.0219	0.0839	0.1347	0.1147	0.1328	0.0380	0.1329	0.0712	0.1356	0.1379	0.1291	0.1534
37110	0.0691	0.0188	0.0518	0.0604	0.0570	0.0654	0.0246	0.0664	0.0499	0.0585	0.0642	0.0688	0.0659
42731	0.0643	0.0226	0.0559	0.0726	0.0609	0.0640	0.0280	0.0614	0.0466	0.0644	0.0611	0.0698	0.0643
71523	0.0581	0.0138	0.0508	0.0643	0.0599	0.0604	0.0196	0.0533	0.0476	0.0617	0.0596	0.0544	0.0604
78331	0.0591	0.0103	0.0227	0.0505	0.0446	0.0575	0.0154	0.0495	0.0516	0.0539	0.0582	0.0564	0.0552
78374	0.0372	0.0074	0.0159	0.0318	0.0314	0.0280	0.0097	0.0375	0.0286	0.0307	0.0309	0.0351	0.0379
>90%	5	0	0	4	1	4	0	2	0	2	4	7	6

c: 500 Clusters:

>90%	2	0	0	10	1	1	0	3	0	2	7	7	6

Table 3

Non-hierarchical Clustering

Non-hierarchical clustering was performed using the Jarvis-Patrick clustering algorithm with the nearest neighbour list length varying from 14 to 20 and with common nearest neighbours 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).

#clusters	Tan 2095	Rus 2235	SM 2211	Bar 2097	Cos 2079	Ku2 2057	For 2616	Fos 2088	Sim 2456	Pea 2103	Yul 2273	Sti 2101	Den 2092
6233	0.0182	0.0139	0.0188	0.0186	0.0182	0.0181	0.0132	0.0182	0.013	0.0184	0.0164	0.0184	0.0186
6235	0.0190	0.0145	0.0194	0.0192	0.0190	0.0189	0.0136	0.0190	0.0135	0.0191	0.0171	0.0191	0.0192
6245	0.0082	0.0062	0.0084	0.0083	0.0082	0.0081	0.0058	0.0082	0.0057	0.0083	0.0073	0.0083	0.0083
7701	0.0089	0.0068	0.0092	0.0090	0.0089	0.0089	0.0065	0.0090	0.0063	0.0091	0.0080	0.0090	0.0091
31420	0.0272	0.0205	0.0279	0.0276	0.0272	0.0270	0.0198	0.0272	0.0194	0.0275	0.0246	0.0274	0.0277
31432	0.0215	0.0163	0.0222	0.0219	0.0215	0.0214	0.0157	0.0215	0.0153	0.0217	0.0193	0.0217	0.0219
37110	0.0198	0.0151	0.0204	0.0201	0.0198	0.0197	0.0144	0.0198	0.0142	0.0199	0.0180	0.0199	0.0201
42731	0.0273	0.0207	0.0281	0.0277	0.0273	0.0272	0.0198	0.0273	0.0194	0.0275	0.0246	0.0275	0.0276
71523	0.0176	0.0132	0.0181	0.0179	0.0176	0.0174	0.0126	0.0176	0.0124	0.0177	0.0158	0.0177	0.0179
78331	0.0154	0.0117	0.0158	0.0156	0.0154	0.0153	0.0112	0.0154	0.0110	0.0155	0.0138	0.0156	0.0155
78374	0.0099	0.0075	0.0102	0.0101	0.0099	0.0098	0.0070	0.0099	0.0070	0.0100	0.0090	0.0100	0.0101
>90%	11	0	11	11	11	11	0	11	0	11	0	11	11

Table 4

The performance varied with changes in the clustering parameters. This is illustrated in Table 5, which indicates the good performing coefficients for each set of parameters for all active classes. The final row shows the results where non-singleton clusters are included. The best-performing coefficient varies considerably with nearest neighbour list length. The increase has the effect of grouping the heterogeneous compounds more easily. These occur as singletons using shorter list lengths – as in the final row of Table 5. These compounds have improved performance under coefficients such as the Russell/Rao and Forbes – coefficients which have been shown to have bias towards larger or smaller compounds. Overall, however, there appears to be no best performer.

Non-singleton	Tan	Rus	SM	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
14-6	11	0	11	11	11	11	0	11	0	10	9	11	11
14-8	11	0	11	11	11	11	0	11	0	11	0	11	11
16-6	0	3	0	0	0	1	0	0	7	0	9	0	0
16-8	11	0	10	11	11	11	0	11	0	11	1	11	11
20-6	0	2	0	0	0	0	0	0	10	0	0	0	0
20-8	0	6	0	0	0	0	1	0	10	0	1	0	0
All													
14-8	0	4	0	0	0	0	6	0	9	0	1	0	0

Table 5

Compound Selection

In a further experiment, the same thirteen coefficients were used to select diverse subsets of the 20K MDDR compounds using a compound selection routine. Examination of the 20K dataset indicated that 591 different active classes were represented. This figure was chosen as the size of subset to be selected. The performance measure was the number of unique active classes represented by the subset. Ideally, this would then be 591, although this level of performance was clearly not expected. Figure 2 shows the results for the thirteen subsets.

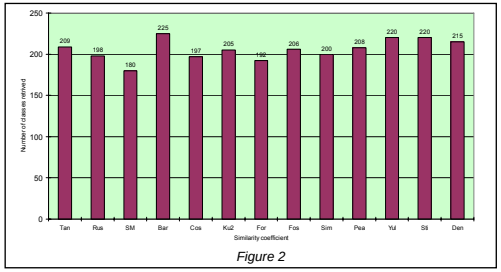

Figure 2

The Baroni-Urbani/Buser and the four correlation coefficients have been found to outperform the other coefficients. In particular, the Simple Match, Forbes, Russell/Rao and Cosine are all slightly less efficient than the other coefficients.

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, Stiles 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.

References

1. Holliday, J.D., Salim, N., Whittle, M. & Willett, P. Analysis and Display of the Size-Dependence of Chemical Similarity Coefficients. *Journal of Chemical Information and Computer Sciences*, 43, 2003, *819-828*.
2. Holliday, J.D., Salim, N. & Willett, P. On the magnitudes of coefficient values in the calculation of chemical similarity and dissimilarity. In: Lavine B. (Ed.), *Chemometrics and Chemoinformatics: ACS Symposium Series 894*, pp77-95, American Chemical Society, 2005.

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