

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/221971909>

# Dunn's index for cluster tendency assessment of pharmacological data sets

Article in *Canadian Journal of Physiology and Pharmacology* · March 2012

DOI: 10.1139/y2012-002 · Source: PubMed

CITATIONS

15

READS

3,300

5 authors, including:



**Oscar Miguel Rivera-Borroto**

37 PUBLICATIONS 211 CITATIONS

[SEE PROFILE](#)



**Mónica Rabassa-Gutiérrez**

Universidad Central "Marta Abreu" de las Villas

1 PUBLICATION 15 CITATIONS

[SEE PROFILE](#)



**Yovani Marrero-Ponce**

Universidad San Francisco de Quito (USFQ)

272 PUBLICATIONS 5,020 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



CTMC special issue [View project](#)



Combined QSAR, TOMOCOMD-CARD, and chemometric tools to discover novel drug candidates against *Trichomonas Vaginalis* [View project](#)

# Dunn's index for cluster tendency assessment of pharmacological data sets

Oscar Miguel Rivera-Borroto, Mónica Rabassa-Gutiérrez,  
Ricardo del Corazón Grau-Ábalo, Yovani Marrero-Ponce, and  
José Manuel García-de la Vega

**Abstract:** Cluster tendency assessment is an important stage in cluster analysis. In this sense, a group of promising techniques named visual assessment of tendency (VAT) has emerged in the literature. The presence of clusters can be detected easily through the direct observation of a dark blocks structure along the main diagonal of the intensity image. Alternatively, if the Dunn's index for a single linkage partition is greater than 1, then it is a good indication of the blocklike structure. In this report, the Dunn's index is applied as a novel measure of tendency on 8 pharmacological data sets, represented by machine-learning-selected molecular descriptors. In all cases, observed values are less than 1, thus indicating a weak tendency for data to form compact clusters. Other results suggest that there is an increasing relationship between the Dunn's index as a measure of cluster separability and the classification accuracy of various cluster algorithms tested on the same data sets.

**Key words:** cluster analysis, cluster tendency, VAT techniques, Dunn's index, pharmacological data sets, clusters overlap, classification accuracy.

**Résumé :** L'évaluation de la tendance à l'agrégation constitue une étape importante de l'analyse de grappes. En ce sens, un groupe de techniques prometteuses appelées *visual assessment of tendency* (« VAT ») est apparu dans la littérature. La présence de grappes peut être facilement détectée par l'observation directe d'une structure en blocs foncée parallèlement à la diagonale principale d'intensité d'image. Alternativement, si l'indice de Dunn d'une partition de liaison unique est supérieur à 1, ceci constitue une bonne indication de la présence d'une structure en bloc. Dans ce rapport, l'indice de Dunn a été appliqué comme nouvelle mesure de tendance sur 8 ensembles de données pharmacologiques, représentés par des descripteurs moléculaires sélectionnés par apprentissage automatique. Dans tous les cas, les valeurs observées étaient inférieures à 1, indiquant ainsi une faible tendance des données à former des grappes compactes. D'autres résultats suggèrent qu'il existe une relation croissante entre l'indice de Dunn comme mesure de la séparabilité des grappes et la précision de la classification de différents algorithmes de grappes testés dans les mêmes ensembles de données.

**Mots-clés :** analyse de grappe, tendance à l'agrégation, techniques VAT, indice de Dunn, ensemble de données pharmacologiques, empiètement des grappes, précision de la classification.

[Traduit par la Rédaction]

## Introduction

In recent years, pharmaceutical companies have been increasing and diversifying their corporate databases through either compound acquisition from compound vendors or through proprietary synthesis of combinatorial libraries. In either case, large numbers of compounds need to be analysed for either their internal diversity or the diversity in which they "add" to the current corporate compounds. Such deci-

sions are commonly made through the use of clustering applications (MacCuish et al. 2001).

Cluster analysis groups data objects into clusters so that objects belonging to the same cluster are similar, whereas those belonging to different ones are dissimilar. The clustering process can be divided into the following stages: data collection, initial screening, representation, clustering tendency, clustering strategy, validation, and interpretation (Jain and Dubes 1988).

Received 28 September 2011. Accepted 19 December 2011. Published at www.nrcresearchpress.com/cjpp on 23 March 2012.

**O.M. Rivera-Borroto.** Laboratorio de Bioinformática, Centro de Estudios de Informática, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830 Villa Clara, Cuba; Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba.

**M. Rabassa-Gutiérrez and R.C. Grau-Ábalo.** Laboratorio de Bioinformática, Centro de Estudios de Informática, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830 Villa Clara, Cuba.

**Y. Marrero-Ponce.** Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba.

**J.M. García-de la Vega.** Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain.

**Corresponding author:** Oscar Miguel Rivera-Borroto (e-mail: oscarrb@uclv.edu.cu).

The problem of determining whether clusters are present as a step prior to actual clustering is called the *assessment of clustering tendency*. For this purpose, tendency (or *clusterability*) techniques provide an a priori quantitative measure to check the presence of nonrandom groupings of molecules (Willett 1985). This tendency for data points to group together on the basis of chemical properties may provide new insights into the underlying chemical relationships. Also, forcing unstructured data into clusters would not only waste time and effort, but could lead to erroneous conclusions about data organization (Lawson and Jurs 1990).

Various formal techniques (most of them statistically based), as well as other less formal ones, have been proposed in the literature (Everitt 1978; Fernández Pierna and Massart 2000; Jain and Dubes 1988; Lawson and Jurs 1990; Massey 2002; Veenman et al. 2002). However, subsequent studies have shown that they have limitations (Forina et al. 2001; Hodes 1992; Rządca and Ferri 2003). Alternatively, visual techniques for various data analysis problems have been studied in the last 25 years (Cleveland 1993; Tukey 1977). In this direction, a group of promising techniques named visual assessment of tendency (VAT) has emerged in the literature because of their effectiveness and easy interpretability (Bezdek and Hathaway 2002).

Clustering tendency is a stage that is often ignored, especially in the presence of large data sets. Also, this tool is rarely available in statistical software commonly used for cluster analysis. It provides an additional impetus for the development and introduction, from other branches of science, of novel tendency measures. This work's objective is to introduce the VAT approach in pharmacological informatics tasks. Specifically, we introduce the Dunn's index, directly linked to VAT techniques, as a novel measure of data-set clusterability. Its use is exemplified on 8 medicinal chemistry repositories of international interest, represented by machine-learning-selected real molecular descriptors.

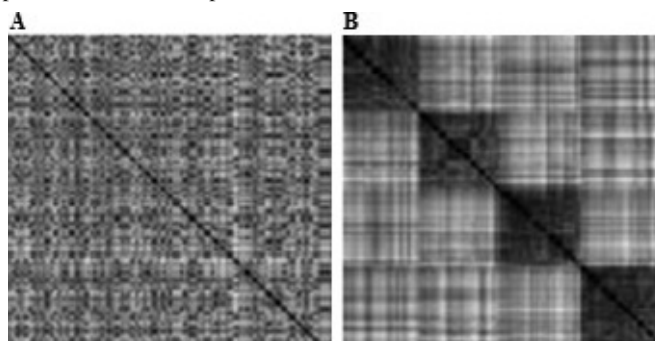
## Experimental

### Theoretical background

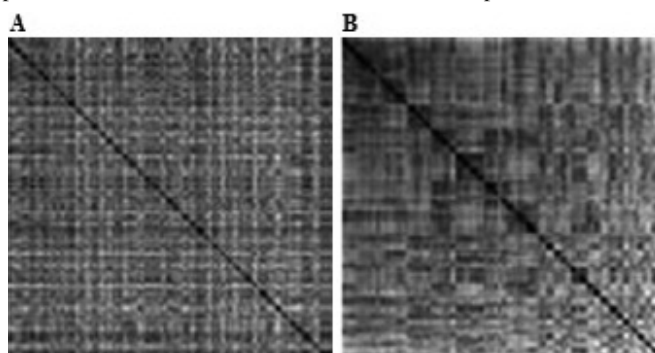
Basically, the VAT approach is based on pairwise dissimilarity relationships between  $n$  objects or relational data. VAT presents the pairwise dissimilarity information as a square digital image with  $n^2$  pixels of grey tonalities, the *intensity image*, so that the black colour corresponds to the minimum dissimilarity and the white colour corresponds to the maximum dissimilarity. Later, the objects are suitably reordered so that the image is better able to highlight potential cluster structure through the direct observation of a well-defined structure of dark blocks along the main diagonal of the intensity image (see Figs. 1 and 2) (Bezdek and Hathaway 2002).

However, even for a data set of moderate size, the computation time,  $O(n^2)$ , for reordering the image matrix becomes expensive (Huband et al. 2004). Furthermore, the dimensions of the image matrix may exceed the resolution of the display monitor, requiring compression or swapping to view the entire image (Hathaway et al. 2006; Huband et al. 2005). To solve this problem, those authors proposed the visual inspection of *tendency curves* (Hu and Hathaway 2008). Alternatively, they found that single linkage (SL) partitions are *always* aligned partitions of the VAT-reordered objects. How-

**Fig. 1.** Dissimilarity image for a hypothetical relational data set where a cluster structure is latent. (A) Data set presented with the original random ordering. (B) Data set presented after the visual assessment of tendency (VAT) reordering technique, indicating the presence of 4 well-separated clusters.



**Fig. 2.** Dissimilarity image for a hypothetical relational data set where a cluster structure is lacking. (A) Data set presented with the original random ordering. (B) Data set presented after the visual assessment of tendency (VAT) reordering technique, indicating the typical case of data sets that do not have well-separated clusters.



ever, the  $c$  SL clusters do not always appear as  $c$  separate dark blocks in the VAT image, even in cases where the “preferred” partition is found by SL (Havens et al. 2009). A related study demonstrates that the Dunn's index (DI), traditionally used as an internal index of cluster validity (Stein et al. 2003), for any SL partition is a measure of the “blockiness” of the VAT image, so that the greater the value of DI, the greater the contrast of the blocks in the VAT image. A value greater than 1 is an indication of the natural tendency for data to form clusters (Havens et al. 2008).

Let  $C = \{C_1, C_2, C_3, \dots, C_k\}$  be a partition of a set of molecular objects  $\mathbf{M}$ ; let  $\delta : C \times C \rightarrow \mathbb{R}^+$  be a cluster-to-cluster distance measure; and let  $\Delta : C \rightarrow \mathbb{R}^+$  be a cluster diameter measure. Therefore, DI is calculated as

$$[1] \quad DI(C) = \frac{\min_{i \neq j} \{\delta(C_i, C_j)\}}{\max_{1 \leq l \leq k} \{\Delta(C_l)\}}$$

where  $\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{d(x, y)\}$ ,  $\Delta(C_l) = \max_{x, y \in C_l} \{d(x, y)\}$ , and  $d : C \times C \rightarrow \mathbb{R}^+$  is a function that measures the distance between objects of  $\mathbf{M}$  (Stein et al. 2003).

### Application data sets

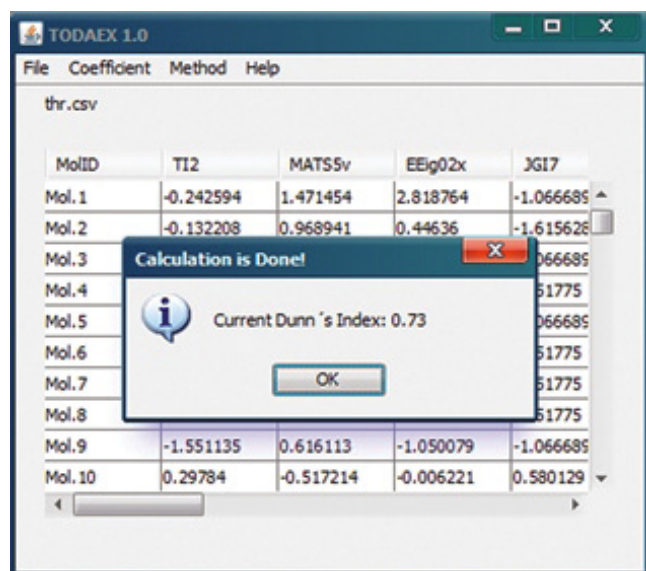
In this preliminary study, 8 different medicinal chemistry data sets were taken from the original work of Sutherland et al. (2004). These chemical repositories have also been used

**Table 1.** Description of medicinal chemistry data sets used in this study.

Parameter		No. of compounds	Pharmacokinetic variable <sup>b</sup>	Range of values
Data set <sup>a</sup>	Pharmacological target			
ACE	Angiotensin converting enzyme inhibitors	114	pIC <sub>50</sub>	2.1–9.9
AchE	Acetyl-cholinesterase inhibitors	111	pIC <sub>50</sub>	4.3–9.5
BZR	Ligands for the benzodiazepine receptor	163	pIC <sub>50</sub>	5.5–8.9
COX-2	Cyclooxygenase-2 inhibitors	322	pIC <sub>50</sub>	4.0–9.0
DHFR	Dihydrofolate reductase inhibitors	397	pIC <sub>50</sub>	3.3–9.8
GPB	Glycogen phosphorylase b inhibitors	66	pK <sub>i</sub>	1.3–6.8
THER	Thermolysin inhibitors	76	pK <sub>i</sub>	0.5–10.2
THR	Thrombin inhibitors	88	pK <sub>i</sub>	4.4–8.5

<sup>a</sup>Data sets are presented in the order of the original source (Sutherland et al. 2004). These are freely available at <http://www.cheminformatics.org/datasets/index.shtml>.

<sup>b</sup>pIC<sub>50</sub> =  $-\log IC_{50}$ , where IC<sub>50</sub> is half the maximum inhibitory concentration, and it is used as a measure of the drug potency, pK<sub>i</sub> =  $-\log K_i$ , where K<sub>i</sub> is the inhibition constant of the drug, also used as a measure of the drug potency.

**Fig. 3.** Software interface for the Dunn's index calculation.

by other researchers in quantitative structure–activity relationship (QSAR) studies (Bruce et al. 2007). A brief description of these data sets is shown in Table 1.

### Molecular representation and computational strategies

Closely allied with the notion of molecular similarity is that of a *chemical space*. Chemical spaces provide a means for conceptualizing and visualizing molecular similarity. A chemical space consists of a set of molecules and a set of associated relations (e.g., similarities, dissimilarities, or distances) among the molecules that give the space a “structure” (Johnson 1989).

A very large number of descriptors that can be used in similarity calculations have been developed. They are typically designed to provide a molecular description that is transferable, in an information-preserving representation, to an abstract descriptor space (Bender and Glen 2004). One solution for solving this problem is to select a particular set of descriptors for which it was shown that it performs well on a certain problem. A further strategy is to first calculate a large number of descriptor values and later on to remove

those descriptors from the set that show a correlation coefficient above a certain value. A different approach is to let the computer choose the optimal combination of descriptors for a given problem (Böcker et al. 2004).

In our study, molecules were represented as real vectors of the type  $\mathbf{m} = (I_1, I_2, \dots, I_n)$ , where  $I_k$ s are the molecular indices. A collection of such vectors constituted the data set matrix  $\mathbf{M}$ , which after an appropriate conditioning was susceptible to the cluster tendency assessment. A description for obtaining the final numerical data sets is given in the following paragraphs.

Data sets were handled with the JChem for Excel utility (ChemAxon 2010). Each set was reoptimized with CORINA 3D structure generator software (Sadowski et al. 1994; Anonymous 2011b). The most relevant software parameters fixed in this process were: wh, write added hydrogen atoms; rs, remove small fragments; and neu, neutralize formal charges. Output files were loaded in the software for the DRAGON molecular descriptors calculation (Talet SRL 2007). In this stage, all available molecular descriptor families were computed (a total of 3224 descriptors) and then binary features were removed. The resulting files were loaded into the Weka data mining software (Hall et al. 2009; Anonymous 2011a) and then were subjected to a treatment including prefiltration, rescaling, and feature selection processes. In this stage, nominal attributes were removed with the RemoveType filter, attributes that did not vary at all or that varied too much were removed with the RemoveUseless filter while keeping default parameters, and resulting numerical attributes and the class attribute were standardized to have zero mean and unit variance with the Standardize filter. Feature selection was performed by using the AttributeSelection filter; here, CfsSubsetEval was set with the rest of the default parameters where it selects subsets of features that are highly correlated with the class while having low intercorrelation among them (Hall 1998). Our criterion was that a supervised correlation-based subset evaluator is important for cluster analysis of chemoinformatic data sets because it helps to warrant that similar chemical structures group in the same cluster, whereas dissimilar molecules group in different clusters. Only linearly dependent descriptors with the class satisfy the neighborhood principle (Nikolova and Jaworska 2003). Additionally, we choose the CfsSubsetEval evaluator because of its



simplicity, and because it has been used by other researchers on the same data sets, obtaining relatively good results in the accuracy scores of the compared classifiers (Johansson et al. 2009; Sönströd et al. 2009).

Finally, to calculate the DI the resulting files were loaded in our TODAEX 1.0 software for chemoinformatic data analysis (Rabassa-Gutiérrez 2010). This process includes the internal parameters SL as the clustering strategy and the closest pair as the fusion strategy (Murtagh 1983). Additionally, among available options we set the Euclidean distance as the proximity measure and  $c = 2$  clusters for the user-defined cut-off technique. This partition was first proposed in a QSAR study on a comparison machine of learning classifiers based on the balanced distribution of the numerical classes (Bruce et al. 2007). Also, it agrees with Mojena's criterion for the optimum number of clusters for the best performing algorithms on the same data sets (Mojena 1977). Once the Euclidean distance matrix and the cluster assignation of molecules were available, the DI was finally calculated.

### TODAEX software description

TODAEX is the acronym of Tools for data set exploration. It is a graphical application (see Fig. 3) that internally performs a variety of techniques from statistics, machine learning, and database management to aid chemoinformatics researchers discovering structures, anomalous behaviors, tendencies, and general underlying knowledge in their molecular descriptors data sets. The current version performs cluster tendency assessment by the innovative use of the DI. Further releases are planned to include other cluster tendency, outlier detection, clustering algorithms, optimum number of cluster detection, and virtual-screening techniques.

To calculate the DI, the user should follow the following steps: (i) load the data matrix in the Open Data of the File menu, (ii) select the dissimilarity measure in the Coefficient menu to be used by the SL cluster algorithm (the current version includes the Euclidian, squared Euclidean, Manhattan, Angular (1-Pearson correlation), Soergel, and Camberra dissimilarities), (iii) select Dunn's index in the Method menu; the software automatically will display an option window with a variety of cut-off techniques for SL tree pruning (the current version includes Mojena's (Mojena 1977), Podani's (Podani 1998), and the user-defined partitions. The software, comprised of the application and a user manual, is available by request from the corresponding author.

## Results and discussion

### Selection of molecular descriptors

Linearly relevant descriptors corresponding to each pharmacological activity (medicinal chemistry data set) selected by the Weka CfsSubsetEval evaluator are shown in Table A1 in the Appendix. The percentages of dimensionality reduction for binary descriptors removal (Weka selection) stage corresponding to each data set were ACE, 55.18% (98.27%); AchE, 56.27% (97.94%); BZR, 54.28% (98.98%); COX-2, 52.70% (98.62%); DHRF, 53.07 (98.94%); GBP, 55.89% (99.30%); THERM, 56.51% (98.93%); and THR, 56.36% (98.93%); the geometric mean of these values being 55.01% (98.74%). These results indicate that ~55% of the molecular

descriptors of the DRAGON software are binary and, thus, were discarded in this study. The subsequent steps of cleansing and selection in Weka resulted in 98.74% of nonrelevant features being removed. This significant value suggests a high degree of specificity in molecular features – biological activity relationships.

This result is particularly important because as the dimensionality of the data increases, the data become increasingly sparse in the occupied space. This can lead to big problems for both supervised and unsupervised learning. In the literature, this phenomenon is referred to as the curse of dimensionality (Janecek et al. 2008). Moreover, a large number of features or descriptors may contain irrelevant or weak relevant features that negatively affect the accuracy of prediction algorithms (John et al. 1994). The extreme case of this phenomenon is depicted in Watanabe's ugly duckling theorem (Watanabe 1969). Basically, if one considers the universe of object features and has no preconceived bias about which features are better, no matter which two objects one compares, all will be equally similar (or dissimilar) (Watanabe 1969).

### Clusterability of data sets

The DI values obtained for the 8 pharmacological data sets are ACE, 0.65; AchE, 0.68; BZR, 0.85; COX-2, 0.37; DHRF, 0.35; GPB, 0.51; THER, 0.61; and THR, 0.73; respectively. Taking into account this criterion ( $DI > 1$ ), one can infer that natural clusters are lacking in these repositories and therefore any cluster algorithm would fail to provide a nonrandom and meaningful grouping of molecules. Other authors have also noticed various counterproductive examples like these in the literature (Willett 1985). However, in his original paper Dunn (1974) showed that  $DI > 1$  is a necessary and sufficient condition to the data set being partitioned into  $k$  compact and well-separated clusters (CWS), while as DI decreases below 1, data set partitions becomes increasingly fuzzy in the sense that the values of the membership functions depart significantly from 0 or 1. The validity of this criterion is then supported by the visual evidence that comes from the VAT technique. However, there is still no rigorous proof to use this threshold as a critical value to infer on "significant clusters", since the corresponding statistical distribution is lacking. As a consequence, previous results should be reinterpreted as pharmacological data sets under study can be organized preferably into fuzzy clusters. In any case, being consistent with the VAT theory, the corresponding intensity image for each data set would have a poor resolution of a dark blocklike structure along the main diagonal of the intensity matrix. Interestingly, in most cases this value is above 0.6, which suggests that, to obtain a better resolution of clusters, a more complex evaluator for the feature selection stage is probably needed, while keeping the linearity with the class. Also, one could argue that the DI is a *very strong* criterion as a tendency measure (see eq. [1] because it demands that the separation between the closest clusters is at least in the order of magnitude of the diameter of the biggest one, and this condition may not reflect some high-throughput-screening results in which active candidates are structurally very similar to their inactive counterparts or decoys, so their corresponding clusters are very close to each other.

**Table 2.** Performance of combinatorial clustering methods and the relationship with cluster tendency.

Data sets	Algorithm				
	CL <sup>a</sup>	MSSN	MVN	MADN	MISS
ACE	88.60 <sup>b</sup>	88.60	87.72	88.60	87.72
AchE	55.86	51.35	55.86	55.86	76.58
BZR	49.69	77.91	73.01	73.62	73.62
COX-2	59.94	60.87	67.39	66.46	72.36
DHFR	52.14	78.59	69.27	69.02	75.57
GBP	77.27	71.21	56.06	56.06	65.15
THERM	51.32	78.95	80.26	78.95	76.32
THR	69.32	75.00	64.77	65.91	67.05
Pearson <sup>c</sup>	0.013 <sup>d</sup>	0.150 <sup>d</sup>	0.144 <sup>d</sup>	0.188 <sup>d</sup>	0.085 <sup>d</sup>

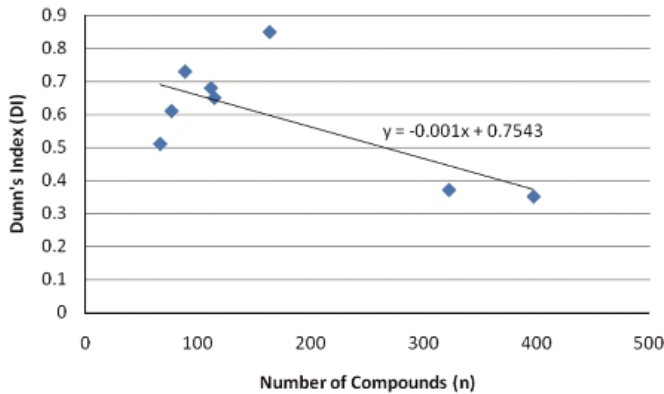
<sup>a</sup>Algorithms compared on the same data sets as the present study: CL, complete linkage; MSSN, minimum sum of squares of new cluster; MVN, minimum variance of new cluster; MADN, minimum average distance within new cluster; MISS, minimum increase of sum of squares or Ward's algorithm (Rivera Borroto et al. 2011).

<sup>b</sup>Figures reflecting the overall percentage of correct classification (Qt) as a measure of classification accuracy.

<sup>c</sup>Pearson's correlation coefficient between each algorithm's performance and the cluster tendency of data sets given by Dunn's index.

<sup>d</sup>No statistically significant relationship at significance level  $\alpha = 0.05$ .

**Chart 1.** Tendency line for the relationship between Dunn's index and the number of pharmacological compounds.



**Cluster separability and classification accuracy**

Previous studies have suggested that data overlap has a strong influence on the accuracy of prediction algorithms. In a recent report, García et al. (2007) studied the behavior of various machine-learning algorithms when classifying imbalanced, overlapped artificial data sets. Their results indicate that the class imbalance is not a problem by itself, but the loss of performance (degradation of the accuracy) of the models is also related to other critical factors such as the class overlap and the small disjoints. Some years earlier, Mojena (1977) introduced a novel measure of statistical overlap among populations to study the behavior of combinatorial cluster algorithms (including Ward's algorithm) at detecting the optimum number of clusters on simulated data sets. His experiments were conducted on data sets with varying numbers of populations and degrees of cluster overlap. The obtained results indicated that the mean performance of the grouping algorithms tends to decrease as either the degree of overlap or the number of populations increases.

Taking these precedents into account and that eq. [1] can be interpreted as a proximity measure among latent clusters in the data sets, with clusters representing samples of the underlying populations generating the very data sets, it is interesting to study the relationship between the DI as a direct measure of cluster separability and the accuracy of classification algorithms. In this direction, along with former results of clusterability, we also used some results from a previous study on clustering methodologies comparison performed in our laboratory using the same data sets (see Table 2) (Rivera-Borroto et al. 2011).

Pearson's correlation coefficient values reflected in Table 2 indicate that there is an increasing linear relationship between the classification accuracy of each cluster's algorithm and the cluster's separability given by DI, which is consistent with previous regularities found by other authors. However, none of those relationships are statistically significant, which suggests that DI is not able to describe all the distributional complexity of data sets, that is, the dispersion heterogeneity, asymmetry, and kurtosis of the analyzed data. On the other hand, there is a statistically significant relationship between DI and the number of members or compound *n* of the data sets (Pearson's  $r = -0.685$ ,  $\alpha = 0.05$ ).

Chart 1 illustrates the decreasing linear relationship between these two variables, suggesting a decrease in the minimum distance intercluster and an increase in the maximum diameter intracluster caused by the rise in the number of cluster members provoking the cluster's expansion, and so a higher degree of clusters overlap.

**Conclusions**

In this paper it is proposed that the DI traditionally used in cluster validity studies also serves as a novel measure for assessing cluster tendency in pharmacological data sets. This index provides a measure of contrast between the blocks on the VAT image diagonal and the background regions, the

greater the value of DI the greater the contrast of the blocks in the VAT image. For any SL partition, a DI value greater than *one* indicates a natural tendency for data to form clusters. This index was implemented in Java, and the software has a user friendly interface. Results obtained for 8 pharmacological repositories of international interest, represented by machine-learning-selected descriptors, suggest that data sets show a weak tendency to form natural groups. However, from the original definition of DI it is clear that  $DI > 1$  should be better understood as a criterion that separates compact data structures from fuzzy ones, so data sets under study should be preferably organized into fuzzy clusters. Experimental data considering the direct link between this novel criterion and the clusters separability suggest that DI (or its modification) has potential applications to explain and predict the performance of prediction algorithms on overlapped data sets. Further research on the theoretical relationship among the DI and other tendency measures, their relationship to the concept of statistical overlap of populations, and a more refined comparison using simulations as well as a larger number of standard chemical data sets is necessary to set the true applicability of this novel approach in pharmacological informatics.

## Acknowledgments

First author (O.M.R.-B.) wishes to express his gratitude to Christof H. Schwab from Molecular Networks GmbH (Germany) for his kind support with the CORINA software. He is also grateful to Noel Ferro from the University of Hannover (Germany), Nelaine Mora-Diez from Thompson Rivers University (Canada), and Lourdes Casas-Cardoso from Cadiz University (Spain) for providing him with important bibliographical materials. This research was partially supported by the Universidad Central de Las Villas (UCLV) and the Vlaamse Interuniversitaire Raad - Institutional University Cooperation (VLIR-IUS) Collaboration Programme. The Universidad Autónoma de Madrid - Universidad Central de Las Villas Scholarship Programme under the auspices of Caja Madrid, Spain, also supported part of this research.

## References

- Anonymous. 2011a. Weka is a collection of machine learning algorithms for data mining tasks. The software Weka version 3-6-4 is available from the Machine Learning Group at the University of Waikato, Hamilton, New Zealand, at <http://www.cs.waikato.ac.nz/ml/weka/> [accessed 27 July 2011].
- Anonymous. 2011b. The CORINA 3D structure generator is available from Molecular Networks GmbH, Erlangen, Germany, at <http://www.molecular-networks.com> [accessed 27 July 2011].
- Bender, A., and Glen, R.C. 2004. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2**(22): 3204–3218. doi:10.1039/b409813g. PMID:15534697.
- Bezdek, J.C., and Hathaway, R.J. 2002. VAT: A tool for visual assessment of (cluster) tendency. Paper presented at the 2002 International Joint Conference on Neural Networks (IJCNN'02), Piscataway, N.J.
- Böcker, A., Schneider, G., and Teckentrup, A. 2004. Status of HTS data mining approaches. *QSAR Comb. Sci.* **23**(4): 207–213. doi:10.1002/qsar.200330860.
- Bruce, C.L., Melville, J.L., Pickett, S.D., and Hirst, J.D. 2007. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **47**(1): 219–227. doi:10.1021/ci600332j. PMID:17238267.
- ChemAxon. 2010. JChem for Excel. Version 5.3.8 (166) [computer program]. JChem for Excel is a Microsoft Excel integrated tool enabling scientists to manage and analyze chemical structures and their data. The software is available from ChemAxon Kft, Budapest, Hungary. Available from <http://www.chemaxon.com> [accessed 27 July 2011].
- Cleveland, W.S. 1993. Visualizing data. Hobart Press, Summit, N.J.
- Dunn, J.C. 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybernetics*, **3** (3): 32–57. doi:10.1080/01969727308546046.
- Everitt, B.S. 1978. Graphical techniques for multivariate data. North Holland Publishing Company, New York.
- Fernández Pierna, J.A., and Massart, D.L. 2000. Improved algorithm for clustering tendency. *Anal. Chim. Acta*, **408**(1–2): 13–20. doi:10.1016/S0003-2670(99)00879-X.
- Forina, M., Lanteri, S., and Esteban Díez, I. 2001. New index for clustering tendency. *Anal. Chim. Acta*, **446**(1–2): 59–70. doi:10.1016/S0003-2670(01)01033-9.
- García, V., Sánchez, J.S., and Mollineda, R.A. 2007. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. *In* Lecture notes in computer science. Edited by L. Rueda, D. Mery, and J. Kittler. Springer-Verlag, Berlin, Heidelberg. pp. 397–406.
- Hall, M.A. 1998. Correlation-based feature subset selection for machine learning. The University of Waikato, Hamilton, New Zealand.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. 2009. The WEKA data mining software: an update. *SIGKDD Explorations Newsl.* **11**(1): 10–18. doi:10.1145/1656274.1656278.
- Hathaway, R.J., Bezdek, J.C., and Huband, J.M. 2006. Scalable visual assessment of cluster tendency for large data sets. *Pattern Recognit.* **39**(7): 1315–1324. doi:10.1016/j.patcog.2006.02.011.
- Havens, T.C., Bezdek, J.C., Keller, J.M., and Popescu, M. 2008. Dunn's cluster validity index as a contrast measure of VAT images. Paper presented at the 19th International Conference on Pattern Recognition (ICPR'08), Tampa, Florida, 8–11 December 2008.
- Havens, T., Bezdek, J., Keller, J., Popescu, M., and Huband, J. 2009. Is VAT really single linkage in disguise? *Ann. Math. Artif. Intell.* **55**(3–4): 237–251. doi:10.1007/s10472-009-9157-2.
- Hodes, L. 1992. Limits of classification. 2. Comment on Lawson and Jurs. *J. Chem. Inf. Model.* **32**(2): 157–166. doi:10.1021/ci00006a007.
- Hu, Y., and Hathaway, R.J. 2008. Tendency curves for visual clustering assessment. Paper presented at the WSEAS International Conference on Applied Computing, Mathematics and Computers in Science and Engineering, Stevens Point, Wisconsin.
- Huband, J.M., Bezdek, J.C., and Hathaway, R.J. 2004. Revised visual assessment of (cluster) tendency (reVAT). *In* Proceedings of the North American Fuzzy Information Processing Society (NAFIPS). Edited by S. Dick, L. Kurgan, P. Musilek, W. Pedrycz, and M. Reformat. IEEE, Banff, Alberta, Canada. pp. 101–104.
- Huband, J.M., Bezdek, J.C., and Hathaway, R.J. 2005. bigVAT: Visual assessment of cluster tendency for large data sets. *Pattern Recognit.* **38**(11): 1875–1886. doi:10.1016/j.patcog.2005.03.018.
- Jain, A.K., and Dubes, R.C. 1988. Algorithms for clustering data. Prentice-Hall, Englewood Cliffs, N.J.
- Janecek, A., Gansterer, W.N., Demel, M., and Ecker, G. 2008. On the relationship between feature selection and classification accuracy. *J. Mach. Learn. Res. Proceedings Track.* **4**: 90–105.
- Johansson, U., Löfström, T., and Norinder, U. 2009. Evaluating ensembles on QSAR classification. Paper presented at the 3rd Skövde Workshop on Information Fusion Topics 2009



- (SWIFT'09), Skövde, Sweden, Oct. 12–13, 2009. Available from <http://bada.hb.se/handle/2320/5901>.
- John, G.H., Kohavi, R., and Pfleger, K. 1994. Irrelevant features and the subset selection problem. Paper presented at the Eleventh International Conference on Machine Learning (ICML'94), 10–13 July 1994. Rutgers University, New Brunswick, N.J.
- Johnson, M.A. 1989. A review and examination of mathematical spaces underlying molecular similarity analysis. *J. Math. Chem.* **3**(2): 117–145. doi:10.1007/BF01166045.
- Lawson, R.G., and Jurs, P.C. 1990. New index for clustering tendency and its application to chemical problems. *J. Chem. Inf. Comput. Sci.* **30**(1): 36–41. doi:10.1021/ci00065a010.
- MacCuish, J., Nicolaou, C., and MacCuish, N.E. 2001. Ties in proximity and clustering compounds. *J. Chem. Inf. Comput. Sci.* **41**(1): 134–146. doi:10.1021/ci000069q. PMID:11206366.
- Massey, L. 2002. Determination of clustering tendency with ART neural networks. Paper presented at the 4th International Conference on Recent Advances in Soft Computing, Nottingham, UK, 12–13 December 2002.
- Mojena, R. 1977. Hierarchical grouping methods and stopping rules: an evaluation. *Comput. J.* **20**(4): 359–363. doi:10.1093/comjnl/20.4.359.
- Murtagh, F. 1983. A survey of recent advances in hierarchical clustering algorithms. *Comput. J.* **26**(4): 354–359.
- Nikolova, N., and Jaworska, J. 2003. Approaches to measure chemical similarity – a review. *QSAR Comb. Sci.* **22**(9–10): 1006–1026. doi:10.1002/qsar.200330831.
- Podani, J. 1998. Explanatory variable in classification and the detection of the optimum number of cluster. In *Data science, classification, and related methods*. Edited by C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, and H.H. Bock. Springer, Tokyo. pp. 125–132.
- Rabassa-Gutiérrez, M. 2010. TODAEX. Version 1.0 [computer program]. Centro Nacional de Derecho de Autor (CENDA), Santa Clara, Cuba.
- Rivera-Borroto, O.M., Marrero-Ponce, Y., García de la Vega, J.M., and Grau-Ábalo, R.C. 2011. Comparison of combinatorial clustering methods on pharmacological data sets represented by machine learning-selected real molecular descriptors. *J. Chem. Inf. Model.* **51**(12): 3036–3049. doi:10.1021/ci2000083.
- Rzadca, K., and Ferri, F. 2003. Incrementally assessing cluster tendencies with a maximum variance cluster algorithm. In *Lecture notes in computer science*. Edited by F.J. Perales. Springer, Berlin, Heidelberg. pp. 868–875.
- Sadowski, J., Gasteiger, J., and Klebe, G. 1994. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **34**(4): 1000–1008. doi:10.1021/ci00020a039.
- Sönströd, C., Johansson, U., and Norinder, U. 2009. Generating comprehensible QSAR models. Paper presented at the 3rd Skövde Workshop on Information Fusion Topics 2009 (SWIFT'09), Skövde, Sweden, 12–13 October 2009. Available from <http://bada.hb.se/handle/2320/5911>.
- Stein, B., Meyer zu Eissen, S., and Wißbrock, F. 2003. On cluster validity and the information need of users. Paper presented at the 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA'03), Benalmádena, Spain, 8–10 September 2003.
- Sutherland, J.J., O'Brien, L.A., and Weaver, D.F. 2004. A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem.* **47**(22): 5541–5554. doi:10.1021/jm0497141. PMID:15481990.
- Talete SRL. 2007. DRAGON for Windows. Version 5.5 [computer program]. The software for DRAGON molecular descriptors calculations is available from Talete SRL, Milan, Italy, at <http://www.talete.mi.it> (accessed 27 July 2011).
- Tukey, J.W. 1977. *Exploratory data analysis*. Addison-Wesley, Co., Reading, Mass.
- Veenman, C.J., Reinders, M.J.T., and Backer, E. 2002. A maximum variance cluster algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(9): 1273–1280. doi:10.1109/TPAMI.2002.1033218.
- Watanabe, S. 1969. *Knowing and guessing: a quantitative study of inference and information*. John Wiley and Sons Inc., New York.
- Willett, P. 1985. Clustering tendency in chemical classifications. *J. Chem. Inf. Comput. Sci.* **25**(2): 78–80. doi:10.1021/ci00046a004.

## Appendix



**Table A1.** List of molecular descriptors selected by the Weka feature selection filter CfsSubsetEval.

Data	Descriptor	Family <sup>a</sup>	Dim. <sup>b</sup>
ACE	MAXDP	Topological descriptors	2D
	PW4	Topological descriptors	2D
	Lop	Topological descriptors	2D
	BIC5	information indices	2D
	ATS4m	2D Autocorrelations	2D
	MATS8m	2D Autocorrelations	2D
	MATS3p	2D Autocorrelations	2D
	EEig03d	Edge adjacency indices	2D
	EEig11d	Edge adjacency indices	2D
	EEig12d	Edge adjacency indices	2D
	DISPp	Geometrical descriptors	3D
	RDF035u	RDF descriptors	3D
	RDF035m	RDF descriptors	3D
	RDF035e	RDF descriptors	3D
	RDF035p	RDF descriptors	3D
	Mor23m	3D-MoRSE descriptors	3D
	Mor26v	3D-MoRSE descriptors	3D
	Mor26p	3D-MoRSE descriptors	3D
	E3u	WHIM descriptors	3D
	E1p	WHIM descriptors	3D
	C-006	Atom-centred fragments	1D
	C-026	Atom-centred fragments	1D
	ALOGP2	Molecular properties	Others
	F03[O-O]	2D Frequency fingerprints	2D
	F06[O-O]	2D Frequency fingerprints	2D
AChE	D/Dr07	Topological descriptors	2D
	IC4	Information indices	2D
	SIC5	Information indices	2D
	BIC5	Information indices	2D
	MATS4m	2D Autocorrelations	2D
	MATS4p	2D Autocorrelations	2D
	GATS4m	2D Autocorrelations	2D
	GATS6e	2D Autocorrelations	2D
	GATS5p	2D Autocorrelations	2D
	JGI10	Topological charge indices	2D
	RDF045u	RDF descriptors	3D
	RDF090u	RDF descriptors	3D
	RDF155u	RDF descriptors	3D
	RDF090m	RDF descriptors	3D
	RDF090e	RDF descriptors	3D
	RDF155e	RDF descriptors	3D
	Mor22m	3D-MoRSE descriptors	3D
	Mor11e	3D-MoRSE descriptors	3D
	Mor32e	3D-MoRSE descriptors	3D
	E3u	WHIM descriptors	3D
	G2m	WHIM descriptors	3D
	G3v	WHIM descriptors	3D
	G1e	WHIM descriptors	3D
	E3p	WHIM descriptors	3D
	R6e+	GETAWAY descriptors	3D
	nR = Cs	Functional group counts	1D
	nArCONR2	Functional group counts	1D
	H-053	Atom-centred fragments	1D
	O-058	Atom-centred fragments	1D

**Table A1 (continued).**

Data	Descriptor	Family <sup>a</sup>	Dim. <sup>b</sup>
BZR	TI2	topological descriptors	2D
	Vindex	Information indices	2D
	ATS7e	2D Autocorrelations	2D
	J3D	Geometrical descriptors	3D
	HOMA	Geometrical descriptors	3D
	RDF020u	RDF descriptors	3D
	RDF030m	RDF descriptors	3D
	RDF055m	RDF descriptors	3D
	RDF020p	RDF descriptors	3D
	RDF030p	RDF descriptors	3D
	Mor09u	3D-MoRSE descriptors	3D
	Mor04v	3D-MoRSE descriptors	3D
	G3p	WHIM descriptors	3D
	H7m	GETAWAY descriptors	3D
	R6u	GETAWAY descriptors	3D
	R3m	GETAWAY descriptors	3D
	C-005	Atom-centred fragments	1D
	H-047	Atom-centred fragments	1D
	N-072	Atom-centred fragments	1D
	Hy	Molecular properties	Others
COX-2	F01[O-S]	2D frequency fingerprints	2D
	F07[N-F]	2D frequency fingerprints	2D
	RDF055m	RDF descriptors	3D
	Lop	Topological descriptors	2D
	D/Dr09	Topological descriptors	2D
	X1A	Connectivity indices	2D
	G(N.O)	Geometrical descriptors	3D
	RDF060m	RDF descriptors	3D
	RDF060v	RDF descriptors	3D
	Mor12u	3D-MoRSE descriptors	3D
	Mor30m	3D-MoRSE descriptors	3D
	Mor08v	3D-MoRSE descriptors	3D
	Mor30v	3D-MoRSE descriptors	3D
	Mor12e	3D-MoRSE descriptors	3D
	E3u	WHIM descriptors	3D
	P1v	WHIM descriptors	3D
	E1e	WHIM descriptors	3D
	R6u+	GETAWAY descriptors	3D
	R3m+	GETAWAY descriptors	3D
	H-049	Atom-centred fragments	1D
DHFR	O-058	Atom-centred fragments	1D
	F03[N-O]	2D frequency fingerprints	2D
	F05[N-N]	2D frequency fingerprints	2D
	F07[N-F]	2D frequency fingerprints	2D
	nR05	Constitutional descriptors	0D
	D/Dr10	Topological descriptors	2D
	GATS7m	2D autocorrelations	2D
	GATS6p	2D autocorrelations	2D
	BELm2	Burden eigenvalues	2D
	BELe1	Burden eigenvalues	2D
	RCI	Geometrical descriptors	3D
	Mor10u	3D-MoRSE descriptors	3D
	Mor03m	3D-MoRSE descriptors	3D
	Mor04m	3D-MoRSE descriptors	3D
	Mor09e	3D-MoRSE descriptors	3D
	R5u	GETAWAY descriptors	3D
	C-033	Atom-centred fragments	1D
	O-057	Atom-centred fragments	1D
	F04[C-N]	2D Frequency fingerprints	2D
	F04[N-O]	2D Frequency fingerprints	2D

**Table A1** (concluded).

Data	Descriptor	Family <sup>a</sup>	Dim. <sup>b</sup>
GBP	X5A	Connectivity indices	2D
	BIC1	Information indices	2D
	MATS8v	2D autocorrelations	2D
	MATS7e	2D autocorrelations	2D
	Mor13m	3D-MoRSE descriptors	3D
	R5m+	GETAWAY descriptors	3D
	C-006	Atom-centred fragments	1D
	H-046	Atom-centred fragments	1D
	F02[N-O]	2D Frequency fingerprints	2D
	F07[O-O]	2D Frequency fingerprints	2D
THERM	X5v	Connectivity indices	2D
	IC1	Information indices	2D
	GATS5m	2D Autocorrelations	2D
	GATS7p	2D Autocorrelations	2D
	RDF065m	RDF descriptors	3D
	Mor17m	3D-MoRSE descriptors	3D
	Mor31m	3D-MoRSE descriptors	3D
	Mor16e	3D-MoRSE descriptors	3D
	Du	WHIM descriptors	3D
	R5p	GETAWAY descriptors	3D
THR	nCt	Functional group counts	1D
	nROH	Functional group counts	1D
	F01[O-S]	2D Frequency fingerprints	2D
	F03[C-N]	2D Frequency fingerprints	2D
	F09[C-N]	2D Frequency fingerprints	2D
	TI2	Topological descriptors	2D
	MATS5v	2D Autocorrelations	2D
	EEig02x	Edge adjacency indices	2D
	JGI7	Topological charge indices	2D
	P2u	WHIM descriptors	3D
	E2p	WHIM descriptors	3D
	E3s	WHIM descriptors	3D
	H8m	GETAWAY descriptors	3D
	HATS6v	GETAWAY descriptors	3D
	nCq	Functional group counts	1D
	nHDon	Functional group counts	1D
	H-053	Atom-centred fragments	1D
	Hy	Molecular properties	Others
	F08[C-S]	2D Frequency fingerprints	2D
	F08[N-O]	2D Frequency fingerprints	2D

<sup>a</sup>Classification according to the descriptor family (or block as in the DRAGON software),

<sup>b</sup>Classification according to the dimensionality or complexity of molecular representation.