Mol Divers
DOI 10.1007/s11030-014-9565-z

FULL-LENGTH PAPER

IMMAN: free software for information theory-based chemometric analysis

Ricardo W. Pino Urias · Stephen J. Barigye · Yovani Marrero-Ponce · César R. García-Jacas · José R. Valdes-Martiní · Facundo Perez-Gimenez

Received: 29 August 2014 / Accepted: 24 December 2014

Abstract The features and theoretical background of a new and free computational program for chemometric analysis denominated IMMAN (acronym for Information theory-based CheMoMetrics ANalysis) are presented. This is multi-platform software developed in the Java programming language, designed with a remarkably user-friendly graphical interface for the computation of a collection of information-theoretic functions adapted for rank-based unsupervised and supervised feature selection tasks. A total of 20 feature selection parameters are presented, with the unsupervised and supervised frameworks represented by 10 approaches in each case. Several information-theoretic parameters traditionally used as molecular descriptors (MDs) are adapted for use as unsupervised rank-based feature selection methods. On the other hand, a generalization scheme for the previously defined differential Shannon's entropy is discussed, as well as the introduction of Jeffreys information measure for supervised feature selection. Moreover, well-known information-theoretic feature selection parameters, such as information gain, gain ratio, and symmetrical uncertainty are incorporated to the IMMAN software (http://mobiosd-hub.com/imman-soft/), following an equal-interval discretization approach. IMMAN offers data pre-processing functionalities, such as missing values processing, dataset partitioning, and browsing. Moreover, single parameter or ensemble (multi-criteria) ranking options are provided. Consequently, this software is suitable for tasks like dimensionality reduction, feature ranking, as well as comparative diversity analysis of data matrices. Simple examples of applications performed with this program are presented. A comparative study between IMMAN and WEKA feature selection tools using the Arcene dataset was performed, demonstrating similar behavior. In addition, it is revealed that the use of IMMAN unsupervised feature selection methods improves the performance of both IMMAN and WEKA supervised algorithms.

Graphical abstract Graphic representation for Shannon's distribution of MD calculating software.

Electronic supplementary material The online version of this article (doi:10.1007/s11030-014-9565-z) contains supplementary material, which is available to authorized users.

R. W. P. Urias · Y. Marrero-Ponce · C. R. García-Jacas
Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR International), Cartagena de Indias, Bolívar, Colombia

R. W. P. Urias · J. R. Valdes-Martiní
Faculty of Mathematics Physics and Computation, Universidad Central "Marta Abreu" de Las Villas, Santa Clara 54830, Villa Clara, Cuba

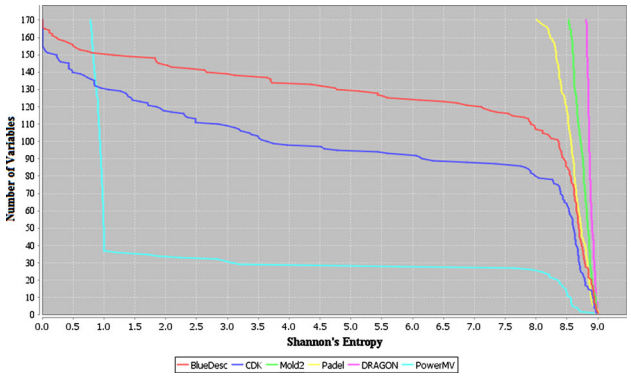
S. J. Barigye
Departamento de Química, Universidade Federal de Lavras, UFLA, Caixa Postal 3037, 37200-000 Lavras, MG, Brazil

Y. Marrero-Ponce (✉) · F. Perez-Gimenez
Facultad de Farmacia, Universitat de València, Burjasot, 46100 València, Spain
e-mail: ymarrero77@yahoo.es; ymponce@gmail.com
URL: http://www.uv.es/yoma/

Y. Marrero-Ponce
Grupo de Investigación en Estudios Químicos y Biológicos, Facultad de Ciencias Básicas, Universidad Tecnológica de Bolívar, Cartagena de Indias, Bolívar, Colombia

C. R. García-Jacas
Grupo de Investigación de Bioinformática, Centro de Estudio de Matemática Computacional (CEMC), Universidad de las Ciencias Informáticas, La Habana, Cuba

Published online: 26 January 2015



Keywords Computational program · Chemometric analysis · IMMAN · Information-theoretic function · Feature selection · Classification

Introduction

In recent years, there has been a significant upsurge in the number and diversity of molecular structure-characterizing features, also known as molecular descriptors (MDs), implemented in various educational and commercial computational programs [1–9]. This increase is, however, not necessarily all advantageous as it engenders high-dimensional space, which usually has a detrimental influence on the performance of regression and classification algorithms. Moreover, an exhaustive search of the entire MD space in search of subsets of features that best describe a specified molecular property comes along with high computational complexity, in addition to the fact that such exploration may lead to the selection of features that aggravate data overfitting [10]. The challenge of dealing with high-dimensional data is not limited to chemoinformatics. High-throughput data matrices obtained in genomics with microarray technology, metabolomics, proteomics, and texts analysis are typical examples of datasets characterized by the "*small sample-many features*" problem [10]. It is thus important to develop procedures that filter out noisy, redundant, or highly correlated variables without affecting the learning performance. It is known that usually, dimensionality reduction improves the quality of models (especially, their predictive power) and information extracted from models, in addition to permitting greater computational efficiency. Optimum classification models should ideally discriminate the molecules belonging to different classes, created on the basis of specified molecular properties or activities, the most common example being binary classifiers. Unfortunately, there exists no universally superior feature selection algorithm, since a method unsuitable for particular application may perform ideally in another. This illation is also known as the "*no free lunch theorem*" [11]. As a result, many computational methods for feature selection have been proposed in the literature and these are basically divided into filters, wrappers, and embedded methods. While wrapper and embedded methods incorporate learning algorithms in their settings, filter methods rely on the intrinsic tendencies of data as criteria for feature selection. Although it is known that filter methods may induce the selection of sets with redundant features, their key advantage is that the selected features are not adapted to a specific predictor algorithm and are thus suitable for dimensionality reduction tasks as well. Feature selection methods follow two primary objectives: (1) Obtain the finest low-dimensionality representation of data matrices, when no dependant (response) variables are available (or considered). The algorithms designed for this purpose are known as unsupervised methods. Examples of such methods include cluster analysis, principal component analysis, Shannon's entropy ranking, among others [1,12,13]. (2) Screen for features that best correlate with response variables for classification and regression. The procedures employed for this objective are collectively denominated as supervised methods. Typical examples include information gain, relief [14], Pearson correlation coefficients, and Fisher ratio, among others.

On the whole, feature selection methods are applications of diverse theoretical concepts and methods aimed at evaluating data patterns (or tendencies) as well as relations among features and/or instances. This article focuses exclusively on information-theoretic methods for feature selection from both a supervised and unsupervised perspective. Information-theoretic functions have increasingly deserved more attention as solid tools for various chemometric tasks. Consequently, a comprehensive review of the literature for all information theory-based unsupervised and supervised feature selection measures has been performed and all these algorithms condensed in a free computational program denominated IMMAN (acronym for Information theory-based CheMoMetrics ANalysis). In addition, information-theoretic parameters previously used to define the information content of a molecular graph are adapted for use as alternative criteria for rank-based feature selection approaches. Moreover, the concept of Symmetrical Kullback–Leibler Entropy (SKL), also known as *Jeffreys Information* [15,16], is introduced as an objective measure of the divergence between two probability distributions.

Program design

IMMAN offers a user-friendly graphical interface stratified in sections according to the different information-theoretic concepts and is developed using the JAVA programming language. The programs developed using this language may be executed in different architectures or operating systems, such as Windows, Linux, or OS X. The quantity of RAM necessary for the utilization of the IMMAN software depends on

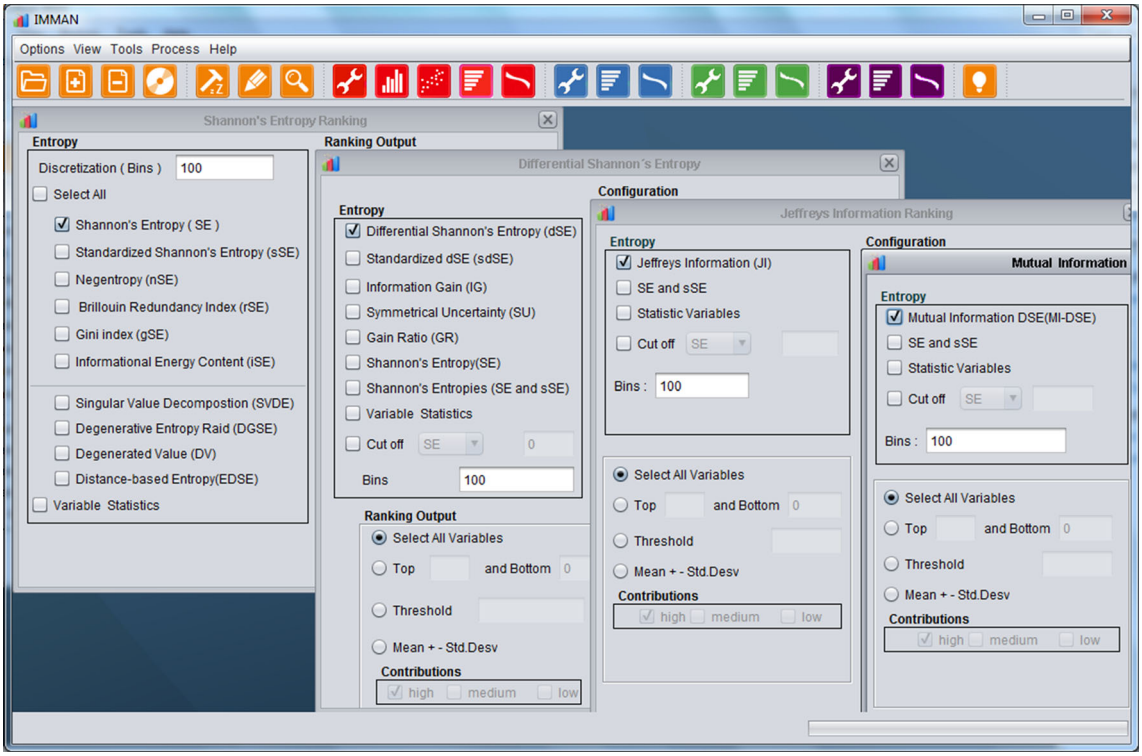


Fig. 1 IMMAN's graphical user interface (GUI)

the size of the data to be analyzed. Batch files (.bat) and shell scripts (.sh) are provided to augment the maximum heap size for the java engine. Depending on the RAM available on the system, a program could be executed via these files (or scripts) to award greater heap size to the java engine and avoid Out of Memory errors. The default setting for java lies in the range 16–64 MB, which is normally too small.

The IMMAN system comprises two major classes that manage the datasets over which computations are to be performed. The *DataSet* class is a representation of the files that are introduced to the system for processing. This class contains attributes, such as file name, an array for identification codes for the instances, the number of instances, and the number of variables contained, among others. Additionally, within this class there is a list of objects for the *Variable* class. These objects are attributes for the variables of the dataset to be analyzed and include variable name, position occupied by variable in the dataset, an array of instance values for the variable, and statistical parameters, such as the maximum, minimum, median, and the standard deviation which are calculated on creating the object. The *Variable* class is the central axis of the system; it is the one that contains all the operations related with the information-theoretic parameters. Overall, the system comprises a total of 64 classes contained in 17 packages.

The accepted input file formats for IMMAN are Tab and Comma Separated Value files (.txt, .csv). In the conception of this application, we considered it useful to provide for real-time computations of multiple dataset files, and thus comparisons of features or datasets from different sources are possible. Several data pre-processing functions, such as missing values replacement (i.e., with the minimum, maximum, mean, geometric mean, median, or with a user-defined numeric value) or feature deletion, data selection (in the sense that not all loaded datasets may be used for computations), data browsing or inspection, and feature-based dataset partitioning (in cases where the user desires to operate with a reduced number of features) are provided. In addition, for supervised feature selection tasks an option is provided for converting continuous response variables to categorical ones to serve as class variables. Figure 1 provides an image of the IMMAN's graphic user interface.

It is important to remark that IMMAN provides the possibility of performing single parameter ranking or ensemble (multi-criteria) ranking using the former as base ranking methods. Note that the ensemble ranking procedure could be performed using the scores (values for the analyzed features) or the positions they occupy on ranking, following specified amalgamation criteria (i.e., product, mean, geometric mean, or sum). In addition, other than the rank-based unsupervised and supervised computations, graphic visualizations for these outcomes could be employed. The unsupervised graphical analysis options include Shannon's distribution graph, histogram graph, importance and correla-

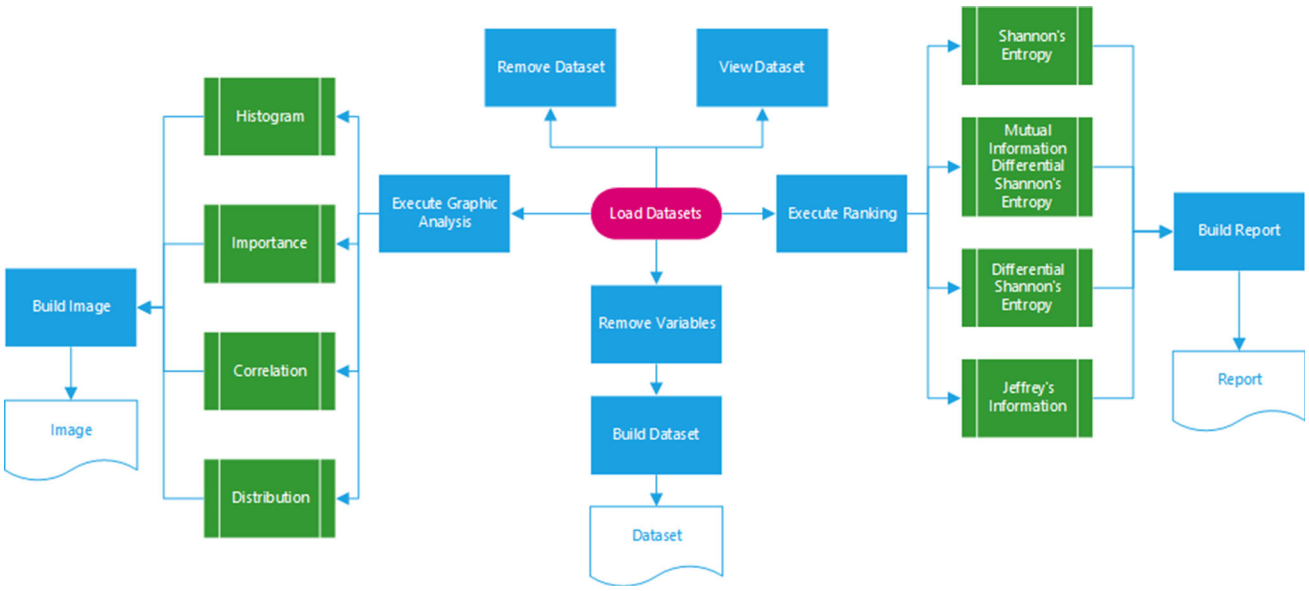


Fig. 2 Workflow for the operations carried out with the IMMAN program

tion graphs, respectively, while supervised graphic analysis allows for the importance and correlation graphs, exclusively.

Calculations performed may be exported to the report and, if desired, saved as Tab Separated Value files (.txt). Figure 2 is an illustrative work flow of operations performed with the IMMAN program. For more information see http://mobiosd-hub.com/imman-soft/.

Theory

Unsupervised variable ranking-based feature selection approaches

In unsupervised paradigms, the feature selection algorithms are unguided by objective functions but rather rely on mathematical quantification of intrinsic properties of datasets. Although in recent years, due to the enormous explosion of amounts of unlabeled datasets, there has been growing interest in unsupervised feature selection methods; these are simply a handful compared to the bulk of the supervised algorithms. Unsupervised feature selection algorithms ameliorate the performance of clustering algorithms, genomic microarray data analysis and similarity/dissimilarity studies of molecular compounds [17–19]. The unsupervised feature selection methods implemented in IMMAN are now briefly discussed.

Shannon's entropy and related entropic measures

Shannon's Entropy (SE) and Scaled Shannon's Entropy (sSE) are proposed by Godden and Bajorath [20,21] as parameters for the quantification of the information content, and thus the variability of MDs. These entropic measures may therefore be used as criteria to rank features in a dataset and if a threshold value is defined, a subset of features is retained for use in building correlation and/or classification models. A brief recapitulation of the theoretical aspects of this methodology is available as supporting information (SI1).

On the other hand, while SE has been previously used as a measure for structural and/composition diversity of molecules through the so-called information theory-based MDs (or information indices), there exists a series of other information-theoretic parameters, mathematically related to Shannon's entropy, that have been traditionally used to serve the same purpose, but not for feature selection tasks. These parameters include negentropy (nSE), Brillouin redundancy index (rSE), Gini index (gSE), and informational energy content (iSE) [1,22]. In this sense these parameters are likewise adapted for use in the evaluation of the information content of features. Table 1 shows the mathematical definitions of these parameters, as implemented in the IMMAN software.

These parameters follow a fixed, data independent discretization scheme (unsupervised equal-interval binning), where a predetermined number of bins is defined. Although the number of discrete intervals is user-defined, SE maximization is advised in that the number of bins should allow an equal distribution of instances (or an approximation) in the discrete intervals. It should be noted that although this discretization approach has been often criticized for being liable to bad cuts leading to uneven distribution of instances, studies have shown that equal-interval binning can yield excellent results, for example with the Naïve Bayes classifier [23].

Table 1 Information-theoretic parameters implemented in the IMMAN software

Parameter	Symbol	Formula
Negentropy	nSE	$nSE = n\log_2 n - \sum_{g=1}^{G} n_g \log_2 n_g$ n is the number of instances n_g is the number of instances in discrete interval g
Brillouin redundancy index	rSE	$rSE = 1 - \frac{SE}{\log_2 N} = 1 - sSE$ N is the number of discrete intervals (bins)
Gini index	gSE	$gSE = \sum_{g \neq k} p_g \cdot p_k$ p_g probability that a randomly selected instance belong to discrete interval g. p_k probability that a randomly selected instance belong to discrete interval k.
Informational energy content	iSE	$iSE = \sum p_g^2$

Singular value decomposition entropy (SVDEi)

This entropy measure evaluates the contribution of the i^{th} feature to the dataset entropy following a leave-one-out (LOO) setting [24,25]. Let S_j represent singular values of the matrix $\mathbf{A}_{[nxm]}$ of n instances and m features. Then S_j^2 denotes the eigenvalues of the $n \times n$ matrix $\mathbf{A}_*\mathbf{A}^t$. The dataset entropy is defined by

$$E(\mathbf{A}) = -\frac{1}{\log N} * \sum_{j=1}^{N} \frac{\boldsymbol{S}_j^2}{\boldsymbol{S}_T} \log \frac{\boldsymbol{S}_j^2}{\boldsymbol{S}_T}, \tag{1}$$

where $\boldsymbol{S}_T$ denotes the total sum of the S_j^2 values. Therefore, the contribution of the ith feature to the dataset entropy is defined as follows:

$$\mathrm{DSE_i} = E(\mathbf{A}) - E(\mathbf{A}'), \tag{2}$$

where $\mathbf{A}'$ denotes the matrix $\mathbf{A}$ without the analyzed feature. In this sense, features may be ranked according to their relative contribution to the dataset entropy.

Degenerative entropy raid (DGSE) and degenerated value (DV)

The DV is a diversity measure based on the number of instances characterized differently by features in a data matrix and is defined as

$$\mathrm{DV}(X) = \frac{\mathrm{Instances_{total}} - \mathrm{Instances_{different}}}{\mathrm{Instances_{total}}}. \tag{3}$$

The DV varies between 0 and 1, with the lower bound (DV = 0) corresponding to the ideal case where X assigns different values for all instances while the upper bound (DV = 1) stands for maximum degeneracy. Note that the DV is not strictly an entropic measure. Its inclusion is justified by its relationship with the DGSE as shown below. The DGSE evaluates the feature variability according to the degenerated value and is expressed as follows:

$$\mathrm{DGSE}(X) = \mathrm{DV}(X) * \frac{\mathrm{SE}(X)_{\mathrm{DV}(X)}}{\mathrm{SE}(X)_{\mathrm{inst}}}, \tag{4}$$

where DV (X) is the degenerated value for variable X, $\mathrm{SE}(X)_{\mathrm{DV}(X)}$, and $\mathrm{SE}(X)_{\mathrm{inst}}$ are Shannon's entropy for X using as the number of discrete intervals the DV and number of instances, respectively.

Euclidean distance-based entropy (EDSE)

The EDSE is computed on the distance among instances, as a measure of the clustering tendency of instances according to variable X, and is expressed as follows [26]:

$$\mathrm{EDSE(X)} = \sum_{\mathrm{i}} \sum_{\mathrm{j}} \left[D_{\mathrm{ij}} \log_2 D_{\mathrm{ij}} + (1 + D_{\mathrm{ij}}) \log_2 (1 - D_{\mathrm{ij}}) \right], \tag{5}$$

where D_{ij} is the normalized distance in the range [0.0–1.0] between the instances Xi and Xj. The EDSE is low for data with clustering tendency and high otherwise, a characteristic that makes it suitable for unsupervised feature selection procedure. Optimality is related with minimum EDSE for subsets of features.

Altogether, ten unsupervised rank-based feature selection methods are discussed. These are essentially divided into the discretization scheme-based algorithms (e.g., SE, sSE, nSE, rSE, gSE, and iSE), and those that do not follow any discretization procedure (e.g., SVDE, DGSE, DV, and EDSE). These algorithms provide an important arsenal of tools for unsupervised feature selection and dimensionality reduction.

Supervised feature selection algorithms

Supervised feature selection algorithms estimate the functional dependency between features and class labels. Based on the feature evaluation method, these algorithms are divided into two main groups: feature subset selection and feature ranking. While the former assesses the discrimination power of subsets of features, the latter evaluates individual features weighted by their degree of relevance. The IMMAN software supervised algorithms belong to the latter. Note that with this approach, a simple filtering criterion is followed (using a defined threshold value or the first k features) and no heuristic search strategy or learning scheme is employed.

Differential Shannon's entropy

To analyze the variability of compound populations, Godden and Bajorath [21] propose a parameter denominated "*Differential Shannon's entropy*." The initial definition was conceived by comparing two compound populations. In this report, however, this definition is generalized for n datasets, defined as

$$\mathrm{DSE} = \mathrm{SE}_{1,2,3\ldots n} - (\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3 + \ldots \mathrm{SE}_n)/n, \quad (6)$$

where "$\mathrm{SE}_{1,2,3\ldots n}$" is the SE calculated for the combination of n compound datasets under consideration. DSE is a measure of the complementarity of n compound collections with regard to the descriptor under analysis. The application of this measure in feature selection tasks is straight forward, in place of compound datasets, class-based partitions are considered. The default configuration is for binary class labels, adjustable to n classes. Note that IMMAN provides an option for transforming a continuous Y response into a categorical one, following a percentile-based rule. The usability of DSE in the identification of features useful in compound classification tasks has been demonstrated, see ref [27].

However, it should also be clarified that in strictly information-theoretic term, the terms DSE is used for entropy computations of continuous sources (or variables), rather than discrete variables as used in the initially proposed definition.

Mutual information differential Shannon's entropy (MI-DSE)

The MI-DSE is introduced as a modification of DSE to select class-specific features when there exists notable differences in compound class sizes [28]. The MI-DSE is mathematically expressed as follows:

$$\text{MI-DSE}(X) = \mathrm{SE}_{\mathrm{norm}}(X, Y) - \frac{\mathrm{SE}(X) - SE(Y)}{2}, \quad (7)$$

where $\mathrm{SE}_{\mathrm{norm}}$(X, Y) is the normalized entropy calculated on two compound classes X and Y combined, SE(X) and SE(Y) Shannon's entropy for X and Y, respectively.

Symmetric Kullback–Leibler entropy: Jeffreys information

Kullback–Leibler entropy or divergence is a heuristic measure of the "distance" between two probability distributions $f(x_i)$ and $g(x_i)$, and it is defined as [15,29,30]

$$D_{KL}(F \| G) = \sum_{i=1}^{m} f(x_i) log\left(\frac{f(x_i)}{g(x_i)}\right), \quad (8)$$

where $f(x_i)$ is the experimental (real) distribution and $g(x_i)$ is the theoretical (approximative) distribution. Viewed from a source coding perspective, Kullback–Leibler entropy defines the additional number of bits needed to codify independent draws of a discrete (o continuous) variable I with probability distribution $f(x_i)$, when a different distribution $g(x_i)$ is used [15,29,30]. Kullback–Leibler divergence, however, presents one shortcoming: it is a subjective parameter, i.e., its value depends on the choice of which variable is considered as experimental and the other theoretical. In other words, Kullback-Leibler divergence is asymmetrical and thus is not an ideal parameter for probability distributions comparisons in which a distinction between experimental and theoretical variables is inapplicable. To eliminate this limitation, a symmetric function of Kullback–Leibler divergence, denominated ***Jeffreys information (JI)***, is used and defined as [16]

$$\mathrm{JI}(F \| G) = D_{KL}(F \| G) + D_{KL}(G \| F). \quad (9)$$

Contrary to Kullback–Leibler divergence, JI is an "unbiased" measure of the distance/dissimilarity of variable distributions between compound classes and thus applicable in rank-based supervised feature selection tasks. High JI values are related with maximum functional dependence between the features and class labels, and low JI values otherwise.

Information gain (IG), gain ratio (GR), and symmetrical uncertainty (SU)

The IG of attribute X refers to the reduction in uncertainty about class attribute Y given that X is known and mathematically defined as follows [23]:

$$\mathrm{IG}(X|Y) = H(X) - H(X|Y), \quad (10)$$

where H(X) and H (X|Y) are entropy of attribute X and entropy of X given Y, respectively.

The GR is a normalization of the IG to compensate for the preference for the attribute with large number of values and is defined as [31]

$$\mathrm{GR}(X, Y) = \frac{\mathrm{IG}(X|Y)}{H(X, Y)}, \tag{11}$$

where IG (X|Y) and H(X, Y) are the information gain of X given Y and the intrinsic information (entropy of distribution of instances into branches), respectively.

The SU compensates for IG's bias toward attributes with more values and is mathematically expressed as [32]

$$\mathrm{SU}(X, Y) = \frac{\mathrm{IG}(X|Y)}{\mathrm{SE}(X) + \mathrm{SE}(Y)}. \tag{12}$$

Symmetrical Uncertainty normalizes its value to the range [0, 1], where 0 indicates that the attributes are completely independent while 1 indicates that each attribute predicts the values of the other.

It should be highlighted that while IG, GR, and SU have previously been reported in the literature and implemented in several feature selection software, the difference with the IMMAN's approach is that an unsupervised equal-interval discretization procedure is employed. Note that while the number of discrete intervals is pre-defined by the user, SE maximization is advised in that the number of bins should allow an equal distribution of instances (or an approximation) in the discrete intervals.

Sample case studies

The primary objective of these studies is to exemplify the practical utility of IMMAN in chemometric tasks. These studies are divided in three parts: Case studies I and II demonstrate the application of SE in comparative studies of families of molecular characterizing parameters and software for these, respectively, from an unsupervised feature selection perspective. On the other hand, case study III deals with the application of the unsupervised and supervised feature selection tools in classification tasks. Comparisons with WEKA software are made.

Case study I

In this section, we compare the performance of families of DRAGON's MDs [2] using SE measure, following the synthesis that high SE features are sensitive to progressive changes in chemical structures and thus generally suitable for correlation studies, while low SE features the contrary. For this study, the PrimScreen1 diversity dataset (available at http://www.otavachemicals.com/component/docman/doc_download/19-primscreen-1-db) was employed. Some MD families were grouped together into bigger families forming a total of 13 super-families. Using a discretization scheme of 1,000 bins, SE values were computed and the best 111 MDs in each family graphically represented for analysis. This cut-off value was not arbitrary chosen, but rather the family that presented the least number of variables determined this value. The probability-based normalization procedure is not advised when working with dataset files with differences in the number of variables as it gives a biased graphic impression of a dataset file with much more variables in respect to other datasets. Figure 3 shows a graphic representation of a family-wise SE distribution for DRAGON's MDs.

It is interesting to note that, generally 3D MD families show superior entropy distribution with 3D-molecule representations of structures based on electron diffraction (3D-MoRSE) MDs presenting the best performance while the worst entropy distribution is observed with fingerprint-based MDs (i.e., **2D Binary** and **Frequency fingerprints**, respectively) and the superfamily **0D-1D & Others** (molecular properties and charge descriptors), respectively. This is a logical result since 3D MDs take into account the nature and connectivity of the atoms, as well as the overall spatial configuration of the molecule, contrary to fingerprint-based MDs and 0D-1D MDs which are generally concerned with the identification of atom types, functional groups, or substituents of interest in a molecule and are thus insensitive to structural or conformational changes in molecular structures and therefore present high degeneracy. On the other hand, it would also be informative to assess the representativity of the overall best twenty MDs in terms of their SE values, of the commercial software DRAGON [2]. Table 2 shows the SE, sSE, nSE, rSE, gSE, iSE, DGSE, DV, and EDSE values, for the best twenty MDs ranked according to SE, for a discretization scheme of 1000 bins.

As it can be observed, the highest representativity is achieved with Randic Molecular Profiles (derived from the distance distribution moments of the geometry matrix) with five MDs (*i.e.*, DP12, DP14, DP15, SP14, and SP16), followed by four MDs for the Connectivity and Topological Index families, respectively, i.e., (X2, X3sol, XMOD, X0v) and (Xu, S1K, S2K, RHyDp). GETAWAY and Geometric indices are represented by two MDs each [(HTv, H3p) and (G1, G2), respectively] and finally one Molecular Property (i.e., AMR), Eigenvalue-based index (i.e., SEig) and Constitutional MD (i.e., Sp). These indices (or families) could be recommended as "ideal" features to be taken into account in molecular modeling and in the drug discovery process. It is not surprising that these MDs families have been successively applied in the modeling and prediction of a wide range of physicochemical, pharmacological, and toxicological properties of several molecular datasets [33–44]. The AMR, as an index in particular, is well known and frequently used in QSPR/QSAR models with comprehensible physicochemical interpretation.

Additionally, it is of interest to evaluate the degree of correlation of the unsupervised feature selection methods. To this end, for each feature selection method, the best 100 MDs

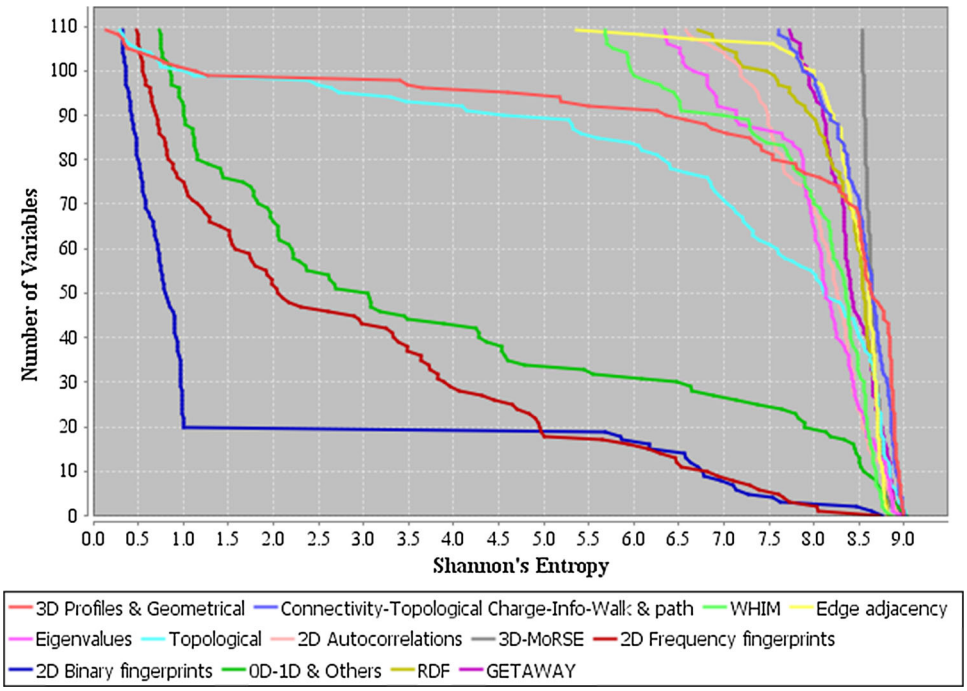


Fig. 3 Graphic representation for Shannon's entropy distribution of DRAGON's MDs

Table 2 The twenty best MDs of commercial software DRAGON, ranked according to Shannon's entropy

Feature	SE	sSE (10^{-1})	nSE (10^{3})	rSE (10^{-1})	gSE (10^{-3})	iSE (10^{-3})	DGSE (10^{-1})	DV (10^{-1})	EDSE (10^{5})	Ensemble (10^{2})[a]
AMR	9.022	9.053	9.022	9.466	1.180	2.170	9.966	9.970	6.847	15.993 (**20**)
X0v	9.012	9.043	9.012	9.573	1.165	2.224	9.521	9.550	6.879	16.332 (**14**)
Xu	9.006	9.036	9.006	9.635	1.244	2.228	9.471	9.550	6.805	16.328 (**15**)
S2K	9.004	9.035	9.004	9.646	1.327	2.206	8.921	9.040	6.689	16.161 (**19**)
G1	8.998	9.029	8.998	9.707	1.235	2.238	9.902	9.900	6.819	16.362 (**13**)
X2	8.998	9.029	8.998	9.708	1.210	2.226	9.336	9.360	6.808	16.274 (**18**)
DP12	8.989	9.020	8.989	9.798	1.336	2.234	9.545	9.600	6.605	16.284 (**17**)
XMOD	8.980	9.011	8.980	9.892	1.339	2.276	9.913	9.930	6.793	16.538 (**9**)
RHyDp	8.980	9.011	8.980	9.893	1.292	2.262	9.726	9.770	6.709	16.436 (**12**)
X3sol	8.979	9.010	8.979	9.898	1.242	2.264	9.432	9.460	6.773	16.447 (**11**)
DP14	8.979	9.009	8.979	9.906	1.293	2.280	9.652	9.680	6.792	16.560 (**7**)
Sp	8.978	9.009	8.978	9.909	1.165	2.278	8.002	8.140	6.780	16.543 (**8**)
H3p	8.972	9.003	8.972	9.974	1.341	2.250	6.954	7.190	6.659	16.304 (**16**)
SEig	8.969	9.000	8.969	10.002	1.332	2.290	9.927	9.970	6.612	16.579 (**6**)
SP16	8.965	8.996	8.965	10.042	1.321	2.352	9.614	9.660	6.900	17.005 (**1**)
G2	8.960	8.991	8.960	10.093	1.251	2.290	9.686	9.700	6.692	16.529 (**10**)
SP14	8.958	8.989	8.958	10.108	1.248	2.308	9.497	9.530	6.768	16.651 (**4**)
HTv	8.952	8.983	8.952	10.168	1.272	2.316	9.447	9.480	6.775	16.674 (**3**)
DP15	8.951	8.982	8.951	10.181	1.236	2.328	9.771	9.760	6.861	16.754 (**2**)
S1K	8.946	8.977	8.946	10.234	1.239	2.310	8.412	8.520	6.787	16.595 (**5**)

[a]Score-based Ensemble ranking using SE, sSE, nSE, and iSE as the base-ranking methods and the product as the amalgamation rule
Bold values indicate the position of variables ranked according to ensemble ranking

were selected and these used to create an incidence matrix (n × m), where n are the variables in the original dataset and m the feature selection parameters. Later, Pearson correlation analysis was performed. Table 3 shows the pair-wise correlation coefficients for the unsupervised feature selection methods.

As it can be observed, the parameters that follow an equal-interval discretization scheme (with the exception of gSE)

Table 3 Correlation coefficients between unsupervised feature selection parameters

	DGSE	DV	EDSE	SVDE	gSE	iSE	SE	nSE	rSE	sSE
DGSE	1.00	**0.96**	0.00	0.01	−0.05	0.19	0.16	0.16	0.16	0.16
DV	**0.96**	1.00	0.01	0.02	−0.05	0.21	0.18	0.18	0.18	0.18
EDSE	0.00	0.01	1.00	0.13	0.03	0.25	0.27	0.27	0.27	0.27
SVDE	0.01	0.02	0.13	1.00	−0.04	0.05	0.06	0.06	0.06	0.06
gSE	−0.05	−0.05	0.03	−0.04	1.00	−0.05	−0.05	−0.05	−0.05	−0.05
iSE	0.19	0.21	0.25	0.05	−0.05	1.00	**0.93**	**0.93**	**0.93**	**0.93**
SE	0.16	0.18	0.27	0.06	−0.05	**0.93**	1.00	**1.00**	**1.00**	**1.00**
nSE	0.16	0.18	0.27	0.06	−0.05	**0.93**	**1.00**	1.00	**1.00**	**1.00**
rSE	0.16	0.18	0.27	0.06	−0.05	**0.93**	**1.00**	**1.00**	1.00	**1.00**
sSE	0.16	0.18	0.27	0.06	−0.05	**0.93**	**1.00**	**1.00**	**1.00**	1.00

Bold values indicate the correlation coefficients values above cutoff of 0.9

are highly correlated, which a logical result is given that their mathematical definitions are generally related to SE. This result suggests that these parameters may not be used concurrently in simple feature ranking. Nonetheless, their utility is appreciated in an ensemble-based feature ranking scheme where the magnitudes (scores) or positions yielded by these parameters for a set of features influence the final ranking from a multi-criteria perspective. For example, if we consider the values of the highly correlated parameters SE, sSE, nSE, and igE of the 20 best MDs (ranked according to SE) in a score-based ensemble scheme, using the product as the amalgamation rule (rSE is left out because it has a negative relation to the relevance of the features), a rather different ranking pattern is achieved (see "Ensemble" column in Table 2). Similarly, as expected the DV and DGSE are highly correlated as well given that the latter is derived from the former. On the other hand, the parameters that do not follow the equal-interval discretization procedure are weakly correlated among themselves as well as to the bin-based unsupervised feature selection parameters. The orthogonality of the feature selection parameters is a desirable attribute as this enables the assessment of distinct tendencies (or patterns) of dataset matrices and thus retrieve dissimilar information.

Case study II

Secondly, we performed a study to compare the performance, in entropy terms, of the most prominent MD calculating software programs. This study is key as it takes into account that there exist several MD calculating programs, and such abundance presents a dilemma when it comes to choosing the "*ideal*" MD computing software for a particular study. Ideal in this sense refers to software with the most variable MDs. Like in the former case, using a discretization scheme of 1000 bins, SE values were calculated for each software (set of MDs previously computed on PrimScreen1 dataset). Figure 4 illustrates the graphic representation of Shannon's distribution of the best 170 variables (cut-off determined by BlueDesc) for each of the MD calculating computer programs.

As it can be observed, the most favorable distribution is provided by DRAGON software, with 100 % of the compared MDs presenting SE values greater than 8.7 bits, followed by MOLD2 and PADEL, respectively. This result suggests that DRAGON comprises a bigger pool of highly variable MDs, sensitive to molecular structural differences, than the rest of the MD calculating software. On the other hand, the worst Shannon's entropy distribution is demonstrated by POWER MV software. The apparently poor performance of POWER MV is attributed to the fact that the majority of the MDs contained in this software are mainly atom-type counts, and these possess weak discriminating power among similar molecular structures. It should be highlighted that the goal of these sample studies is give simple illustrations of possible case studies with IMMAN other than to establish strict ranking authority for the MD computing software. Stronger and more compelling inferences in this direction require studies with other sets of diverse databases.

Case study III: IMMAN versus WEKA

In this sub-section, we wish to compare the performance of the feature selection methods implemented in IMMAN and WEKA, on the basis of the significance of the subsets of features obtained with the different approaches. Two important differences exist between IMMAN and WEKA algorithms: (1) while IMMAN possesses both unsupervised and supervised methods, WEKA offers supervised algorithms, exclusively. (2) Another key difference lies with the discretization methods followed in each case. The IMMAN algorithms employ an equal-interval binning procedure [23], in which the instances are distributed in discrete intervals (*bins*) of equal size according to their numeric values. On the other hand, WEKA algorithms employ a supervised discretization scheme based on an entropy minimiza-

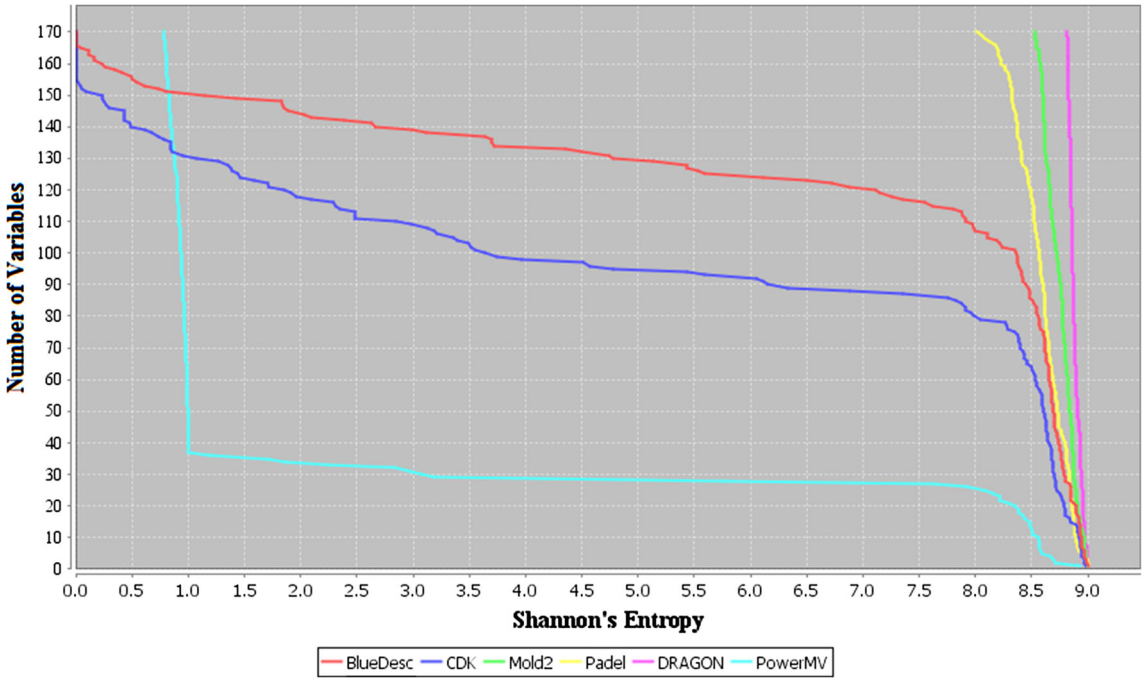


Fig. 4 Graphic representation for Shannon's entropy distribution of MD calculating software

tion heuristic following a recursive binary splitting procedure to obtain multiple intervals for continuous-valued attributes (multi-interval discretization). For details, see ref [45]. For this study the Arcene dataset is used. This dataset is freely available at the UCI Machine Learning Repository [46] and is one of the four datasets proposed in the NIPS 2003 Feature Selection Challenge [47]. The Arcene dataset exemplifies cases where the number of instances is small with respect to the features (high-dimensionality data); a common problem in cheminformatic and/or bioinformatics applications. Particularly, this dataset typifies a two-class classification problem aimed at distinguishing cancer patterns from normal ones in mass spectrometric data.

In the first study, IMMAN's unsupervised and supervised feature selection approaches as well as WEKA's algorithms are employed to obtain subsets of 15 variables, separately, and these are used to build classification models using Kth Nearest Neighbors (KNN1) and Support Vector Machine (SVM) classifiers, respectively, and the percentages of correct classification compared [23]. Tables 4 and 5 show the percentages of correct classification for IMMAN and WEKA, using KNN1 classifier.

As expected, generally supervised methods perform better than unsupervised methods using both KNN1 and SVM classifiers, as the former favor features that are linked to the class labels. Using KNN1, it is observed that generally the

Table 4 Comparison of the percentages of correct classification of KNN1-based classification models using IMMAN's and WEKA's feature selection approaches

Software	Method	Measure	Correct classification (%)
IMMAN	Unsupervised	EDSE	77
		rSE	74
		SE	74
		DGSE	70
		DV	68
		SVDE	64
		gSE	46
	Supervised	SU	81
		DSE	77
		IG	72
		MIDSE	72
		GR	67
		JI	67
WEKA	Supervised	W(SU)	83
		W(Significance)	82
		W(GR)	81
		W(IG)	81
		W(ChiSquare)	77
		W(OneR)	77
		W(Relieff)	70

Table 5 Comparison of the percentages of correct classification of SVM-based classification models using IMMAN's and WEKA's feature selection approaches

Software	Method	Measure	Correct classification (%)
IMMAN	Unsupervised	rSE*	71
		SE*	71
		DGSE	68
		DV	66
		EDSE	66
		gSE	59
		SVDE	52
	Supervised	IG*	82
		SU	79
		DSE	78
		MIDSE	66
		JI	56
		GR	54
WEKA	Supervised	W(GR)	81
		W(SU)	80
		W(ChiSquare)	77
		W(Significance)	77
		W(IG)	76
		W(OneR)	73
		W(Relieff)	70

WEKA's supervised feature selection algorithm depicts a slight edge over IMMAN methods (see Table 4), although there exist parameters that exhibit comparable performance with WEKA's algorithms, such as EDSE (77 %), SU (81 %), and DSE (77 %), with one of them being an unsupervised method (i.e., EDSE). As for SVM, IMMAN's IG offers the highest percentage of correct classification (82 %), followed by WEKA's GR (81 %) and SU (80 %), see Table 5. This result suggests that the IMMAN's algorithms are effective in the selection of subsets of features with good classification accuracy.

One of the key applications of unsupervised methods is in data dimensionality reduction. In the second experiment, we use the unsupervised rank-based methods as pre-filters. For each entropic measure, the mean is determined and used as threshold value (cut-off). The retained sets of features are then filtered using the IMMAN and WEKA supervised feature selection tools, separately, to obtain 15 variable subsets for each algorithm. The final subsets of variables are then validated using KNN1 and SVM classifiers. Tables 6 and 7 show the percentages of correct classification using combinations of supervised and unsupervised feature selection approaches, using KNN1 and SVM classifiers, respectively (note that only the pairings that yield the best percentages of

Table 6 Comparison of the percentages of correct classification for combinations of unsupervised and supervised methods using KNN1 Classifier

Software	Measure	Correct classification (%)
IMMAN	SU-SVDE	81
	DSE-gSE	78
	DSE-DGSE	77
	DSE-DV	77
	DSE-EDSE	77
	DSE-SE	77
	SU-EDSE	77
WEKA	W(Significance)-gSE	88
	W(IG)-gSE	86
	W(ChiSquare)-EDSE	85
	W(ChiSquare)-SE	85
	W(SU)-DGSE	85
	W(SU)-SE	85
	W(Relieff)-gSE	84
	W(Significance)-SVDE	84
	W(IG)-SVDE	83
	W(SU)-DV	83
	W(SU)-EDSE	83
	W(SU)-gSE	83
	W(Significance)-DV	82
	W(Significance)-EDSE	82
	W(GR)-DV	81

Table 7 Comparison of the percentages of correct classification for combinations of unsupervised and supervised methods using SVM Classifier

Software	Measure	Correct classification (%)
IMMAN	JI-DGSE	85
	JI-DV	85
	JI-gSE	85
	JI-SE	85
	IG-DGSE	82
	IG-DV	82
	IG-EDSE	82
	IG-SE	82
	IG-gSE	81
WEKA	W(GR)-gSE	83
	W(GR)-DGSE	82
	W(GR)-SE	82
	W(OneR)-DV	82
	W(OneR)-SE	82
	W(ChiSquare)-EDSE	81

Table 8 Percentages of correct classification obtained using the two-fold feature mixing scheme according to KNN1 and SVM classifiers

Cluster	Measure	KNN1		SVM	
		Singular	Two-fold[a]	Singular	Two-fold
1	W(Relieff)-DV	65	85^{3}	69	$83^{2,4}$
2	JI-DGSE	67	80^{4}	85	$85^{4,6}$
3	W(ChiSquare)-DGSE	80	$86^{4,13}$	79	84^{4}
4	W(SU)	83	87^{10}	80	86^{9}
5	W(IG)	81	86^{4}	76	84^{10}
6	JI	67	85^{4}	56	85^{2}
7	W(OneR)-DV	81	84^{1}	82	84^{15}
8	W(OneR)-SVDE	70	79^{5}	81	84^{2}
9	W(IG)-rSE	74	84^{4}	66	86^{4}
10	DSE-SVDE	69	87^{4}	70	84^{5}
11	SE	74	82^{4}	71	82^{4}
12	GR	67	$79^{3,4}$	54	82^{7}
13	SU-EDSE	77	86^{3}	75	84^{2}
14	MIDSE	72	83^{3}	66	80^{8}
15	DSE	77	84^{4}	78	84^{7}

[a]Subscript refers to cluster (measure) with which combination is made

correct classification are shown; for results of all combinations see supporting information SI2).

Generally, the use of unsupervised methods as pre-filters improves the performance of IMMAN and WEKA supervised feature selection tools, for both KNN1 and SVM classifiers. Nonetheless, it is worth noting that while the use of unsupervised parameters yields minimal improvements in the performance of IMMAN supervised algorithms with the KNN1 classifier, significant improvements are achieved with WEKA (compare Tables 4, 6). On the other hand, with SVM classification method, impressive improvements are observed with IMMAN supervised feature selection tools, for example the percentage of correct classification for JI rises from 56 % (see Table 5) to 85 % (see Table 7), yielding superior performance to WEKA algorithms. This trend (improvement) is evocative of the inexistence of a universally superior feature selection tool and advocates for the use of combinations unsupervised and supervised methods to obtain subsets with a more solid and easily interpretable knowledge structure amenable to greater classification accuracy.

In the third experiment, we evaluate the influence of combining variables filtered independently with IMMAN and WEKA feature selection tools (or pairings of unsupervised and supervised methods). To this end, an $n \times m$ incidence matrix is constructed, with n being the 1000 features in arcene dataset and m the feature selection tools, including combinations of unsupervised and supervised parameters. A k-means cluster analysis (using a k-value of 15) is performed for this data matrix and for each cluster the algorithm (variable) closest to centroid picked, obtaining a total of 15 dissimilar feature selection methods. Later, the sets of features originally filtered by the 15 algorithms from arcene data matrix are mixed in two- to seven-fold combinations and the resulting sets of variables used to build classification models using KNN1 and SVM methods. Table 8 shows the parameters (or combinations of these) selected for each cluster as well as comparisons of their percentages of correct classifications prior to and after performing two-fold variable mixing procedure. For a matrix showing the percentages of correct classification for all pair-wise combinations for the 15 feature selection tools, see Supporting Information (SI3). As it can be observed in Table 8, two-fold mixing of features obtained using dissimilar algorithms (from a clustering tendency perspective) yields significant improvements in classification accuracy of the models. This is a logical result since mixing features obtained with dissimilar algorithms provides an information structure with a wider span of the data structure patterns, which directly influences the performance of the classifier algorithms.

Figure 5 shows the percentages of correct classification obtained using up to seven-fold feature mixing schemes, using the KNN1 classifier. As it can be observed, in all the cases the percentages of correct classification improve achieving up to 92 % of correct classification for six- and seven-fold mixing schemes.

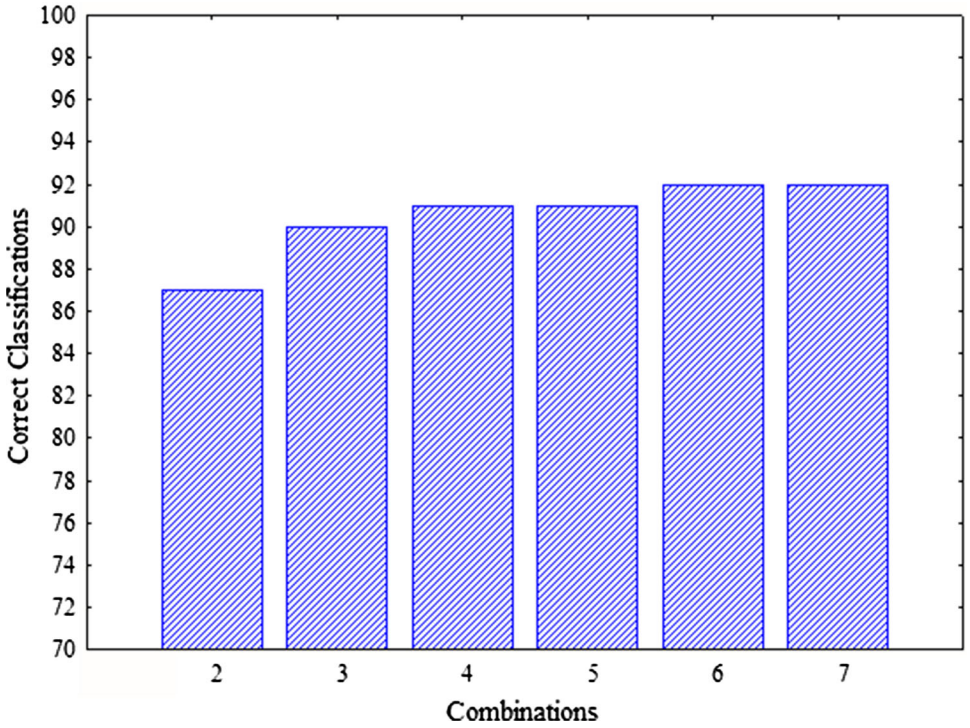


Fig. 5 Percentages of correct classification obtained using feature mixing schemes and KNN1 classifier

Conclusions

A free computational program for chemometric analysis designated IMMAN is developed to provide valuable information theory-based tools for unsupervised and supervised feature selection tasks. This software is developed in the Java programming language and can be executed in different operating systems. The usability of IMMAN has been demonstrated with sample case studies, demonstrating satisfactory behavior. In forthcoming releases, we intend to implement other information-theoretic formulations as measures of the correlation and dependence among variables to deal with the possible redundancy among features, a key handicap of rank-based methods. However, it is known that combinations of m top-ranked features considered individually do not necessarily yield the best subset of features, enunciated in the literature as "*the m best features are not the best m features*" [48,49]. Therefore, other goals include the incorporating feature subset selection algorithms to the IMMAN approach as well as learning algorithms in the feature selection procedure.

Supporting information available

A brief recapitulation of the theoretical aspects of SE method, the percentages of correct classification for all pair-wise combinations of supervised and unsupervised feature selection methods are freely available to all interested users as supplementary material via the Internet at http://www.link.springer.com/journal/11030. The IMMAN computational program, user manual, and codes may be freely downloaded via Internet at http://mobiosd-hub.com/imman-soft/.

Acknowledgments Barigye, S. J. acknowledges financial support from CNPq. Marrero-Ponce, Y. thanks the program 'International Visiting Professor' for a fellowship to work at *Universidad Tecnológica de Bolívar* (*Colombia*) in 2014. Finally, the authors are also indebted to the Molecular Diversity Editor in Chief Dr. Guillermo A. Morales for his comments and manuscript revision, as well as his kind attention.

References

1. Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics, vol 1. Wiley-VCH, Weinheim
2. Todeschini R, Consonni V, Pavan M (2002) DRAGON Software version 2.1. Milano Chemometric and QSAR Research Group. Milano
3. Guha R (1991) The CDK descriptor calculator, 0.94th edn. Indiana
4. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474. doi:10.1002/jcc.21707
5. Georg H (2008) BlueDesc-molecular descriptor calculator. University of Tübingen, Tübingen
6. Liu J, Feng J, Brooks A, Young S (2005) PowerMV. National Institute of Statistical Sciences, Research Triangle Park
7. ADRIANA. Code (2011) Molecular Networks. Erlangen, Germany
8. Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W (2008) Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Comput Sci 48:1337–1344. doi:10.1021/ci800038f
9. Kellogg GE (2001) Molconn-Z 4.0 edn. eduSoft, Virginia
10. Liu H, Motoda H (2008) Less is More. In: Liu H, Motoda H (eds) Computational methods of feature selection. Data mining and

knowledge discovery series. Taylor * Francis Group, Boca Raton, p 411
11. Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1:67–82. doi:10.1109/4235.585893
12. Venkatraman V, Dalby AR, Yang ZR (2004) Evaluation of mutual information and genetic programming for feature selection in QSAR. J Chem Inf Comput Sci 44:1686–1692. doi:10.1021/ci049933v
13. Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the Twentieth international conference on machine learning, Washington DC
14. Kira K, Rendell L (1992) The feature selection problem: traditional methods and a new algorithm. Association for the advancement of artificial intelligence. AAAI Press and MIT Press, Cambridge, pp 129–134
15. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86
16. Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proc Roy Soc A 186:453–461. doi:10.1098/rspa.1946.0056
17. Jennifer GD (2008) Unsupervised Feature Selection. In: Liu H, Motoda H (eds) Computational methods of feature selection. Data mining and knowledge discovery series. Taylor & Francis Group, Boca Raton, p 411
18. Varshavsky R, Gottlieb A, Linial M, Horn D (2006) Novel unsupervised feature filtering of biological data. Bioinformatics 22:e507–e513. doi:10.1093/bioinformatics/btl214
19. Maldonado AG, Doucet JP, Petitjean M, Fan B-T (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. Mol Divers 10:39–79. doi:10.1007/s11030-006-8697-1
20. Godden JW, Stahura FL (2000) Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. J Chem Inf Comput Sci 40:796–800. doi:10.1021/ci000321u
21. Godden JW, Bajorath J (2002) Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. J Chem Inf Comput Sci 42:87–93. doi:10.1021/ci0103065
22. Barigye SJ, Marrero-Ponce Y, Pérez-Giménez F, Bonchev D (2014) Trends in information theory-based chemical structure codification. Mol Divers 18:673–686. doi:10.1007/s11030-014-9517-7
23. Witten IH, Eibe F, Hall MA (2011) Data mining: practical machine learning tools and techniques. The Morgan Kaufmann series in data management systems, 3rd edn. Morgan Kaufmann, Burlington
24. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 97:10101–10106. doi:10.1073/pnas.97.18.10101
25. Devakumari D, Thangavel K (2010) Unsupervised adaptive floating search feature selection based on contribution entropy. In: 2010 international conference on communication and computational intelligence (INCOCCI), pp 623–627
26. Dash M, Choi K, Scheuermann P, Huan L (2002) Feature selection for clustering—a filter solution. In: Proceedings of the 2002 IEEE international conference on data mining (ICDM 2003), pp 115–122. doi:10.1109/icdm.2002.1183893
27. Stahura FL, Godden JW, Bajorath J (2002) Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations. J Chem Inf Comput Sci 42:550–558. doi:10.1021/ci010243q
28. Wassermann AM, Nisius B, Vogt M, Bajorath J (2010) Identification of descriptors capturing compound class-specific features by mutual information analysis. J Chem Inf Model 50:1935–1940. doi:10.1021/ci100319n
29. Cover TM, Thomas JA (1991) Elements of Information theory. Wiley, New York
30. Desurvire E (2009) Classical and quantum information theory. Cambridge University Press, New York
31. Quinlan JR (1983) Learning efficient classification procedures and their application to chess end games. In: Michalski R, Carbonell J, Mitchell T (eds) Machine learning. Symbolic computation. Springer, Berlin, pp 463–482. doi:10.1007/978-3-662-12405-5_15
32. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1988) Numerical recipes in C: the art of scientific computing. Cambridge University Press, New York
33. Consonni V, Todeschini R, Pavan M, Gramatica P (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. Part 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. J Chem Inf Comput Sci 42:693–705. doi:10.1021/ci0155053
34. Pérez González M, Terán C, Teijeira M, González-Moa MJ (2005) GETAWAY descriptors to predicting A2A adenosine receptors agonists. Eur J Med Chem 40:1080–1086. doi:10.1016/j.ejmech.2005.04.014
35. Saiz-Urra L, Pérez González M (2007) Quantitative structure-activity relationship studies of HIV-1 integrase inhibition.1. GETAWAY descriptors. Eur J Med Chem 42:64–70. doi:10.1016/j.ejmech.2006.08.005
36. Fedorowicz A, Singh H, Soderholm S, Demchuk E (2005) Structure–activity models for contact sensitization. Chem Res Toxicol 18:954–969. doi:10.1021/tx0497806
37. Saiz-Urra L, Pérez González M (2006) QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection. Bioorg Med Chem 14:7347–7358. doi:10.1016/j.bmc.2006.05.081
38. Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V (1996) Chemical information in 3Dspace. J Chem Inf Comput Sci 36:1030–1037. doi:10.1021/ci960343+
39. Gasteiger J, Schuur J, Selzer P, Steinhauer L, Steinhauer V (1997) Finding the 3D structure of a molecule in its IR spectrum. Fresen J Anal Chem 359:50–55. doi:10.1007/s002160050534
40. Schuur J, Selzer P, Gasteiger J (1996) The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. J Chem Inf Comput Sci 36:334–344. doi:10.1021/ci950164c
41. Baumann K (1999) Uniform-length molecular descriptors for quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR): classification studies and similarity searching. TRAC 18:36–46. doi:10.1016/S0165-9936(98)00075-2
42. Jelcic Z (2004) Solvent molecular descriptors on poly(D, L-lactide-co-glycolide) particle size in emulsification-diffusion process. Coll Surf A Physico-Chem Eng Asp 242:159–166. doi:10.1016/j.colsurfa.2004.03.027
43. Todeschini R, Bettiol C, Giurin G, Gramatica P, Miana P, Argese E (1996) Modeling and prediction by using WHIM descriptors in QSAR studies. Submitochondrial particles (SMP) as toxicity biosensors of chlorophenols. Chemosphere 33:71–79. doi:10.1016/0045-6535(96)00153-1
44. Randic M (1995) Molecular profiles. Novel geometry-dependent molecular descriptors. New J Chem 19:781–791
45. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th international joint conference on artificial intelligence, pp 1022–1027. http://dblp.uni-trier.de/db/conf/ijcai/ijcai93.html#FayyadI93

46. Newman DJ, Hettich S, Blake CL, Merz CJ (1998) UCI repository of machine learning databases. University of California, Department of Information and Computer Science, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html
47. Guyon I, Gunn SR, Ben-Hur A, Dror G (2004) Result analysis of the NIPS 2003 feature selection challenge. In: Advances in neural information processing systems, Vancouver, BC, pp 545–552. http://papers.nips.cc/paper/2728-result-analysis-of-the-nips-2003-feature-selection-challenge
48. Webb AR (2002) Statistical pattern recognition, 2nd edn. Wiley, Chichester
49. Cover TM (1974) The best two independent measurements are not the two best. IEEE Trans Syst Man Cybern 4:116–117. doi:10.1109/TSMC.1974.5408535