A comparative study using different topological representations in Pattern Recognition based Drug Activity Characterization

Francesc J. Ferri, Wladimiro Diaz-Villanueva and Maria J. Castro

Abstract—The use of certain machine learning and pattern recognition tools for automated pharmacological drug design has been recently introduced. Different families of learning algorithms have been applied to the task of associating observed chemical properties and pharmacological activities to certain kinds of representations of the candidate compounds. In this work, several families of molecular descriptors are considered in order to establish the appropriateness of these families for a particularly challenging drug design task consisting of characterizing the analgesic properties of a relatively large number of compounds. As a second goal, the composite use of descriptors from different families and a first attempt to select the best attributes from these families is considered. As a conclusion, relatively good discrimination results can be obtained by combining the best descriptors of the different families considered.

I. INTRODUCTION

The design of new medical drugs with desired chemical properties is a challenging and very important problem in the pharmaceutical industry. The traditional approach for formulating new compounds requires the designer to test a very large number of molecular compounds, to select them in a blind way, and to look for the desired pharmacological property. Therefore, it is very useful to have tools to discriminate the pharmacological activity of a given molecular compound so that the laboratory experiments can be directed to those molecular groups in which there is a high probability of finding new compounds with the desired properties.

All methods developed for this purpose are based on the fact that the activity of a molecule derives from its structure and therefore it is possible to find a relationship between this structure and the properties that the molecule exhibits [19]. Thus, the way the molecular structure is represented has special relevance.

In Chemical Graph Theory, molecular structures are represented as doubly labeled graphs which can be conveniently characterized by a number of specific topological indices [12]. A great number of topological indices have been proposed, but only a small subset is widely used in common QSAR studies. In this work, three sets or families of topological indices are considered. A first reduced set of 62 topological-structural indices [15], the well-known Kier-Hall set of 116 indices has been selected from the three families underlying the properties of the compound. In this work, a number of specific topological indices developed by Galvez et al. [9].

These or similar representations have already been applied to different discrimination problems in drug design (analgesic, antidiabetic, antibacterial, etc.). In the particular case of antibacterial activity, very good classification results have been reported using multilayer perceptrons (MLP) [3], [14] and Support Vector Machines [8] using topological descriptors.

Also important in the above mentioned drug design application is the cost/benefit problem and the corresponding discrimination thresholds that have to be used to maximize the outcomes of the learned classifiers from the point of view of the pharmacological problem. The use of Receiver Operating Characteristic (ROC) curves [4] has been shown to be a valuable tool in this particular context to evaluate the classifier in a wide range of practical situations.

II. AUTOMATIC DRUG DESIGN

A. Drug Characterization

The so-called quantitative structure-activity relationship (QSAR) models are currently used in the computer aided design of new medical drugs with desired chemical properties. As an alternative to the methods based on the “exact” description of the electronic properties of a molecule calculated by mechanical-quantum methods, the molecular topology describes the molecule as a set of indices which are in fact graph invariants. These topological indices are numerical descriptors that encode information about the number of atoms and their structural environment. This representation is derived from the hydrogen-suppressed molecular formula seen as a graph [1], [19].

The molecular topology considers a molecule as a planar graph where atoms are represented by vertices and chemical bonds are represented by edges. The chosen set of molecular descriptors should adequately capture the phenomena underlying the properties of the compound. In this work, a set of 116 indices has been selected from the three families considered that we will refer to as topological, Kier-Hall and electro-topological.

The topological indices consist of 62 simple integers. Fourteen of these are related to the molecular attributes of the compound; for example, the total number of atoms of a certain element (carbon, nitrogen, oxygen, sulphur, fluorine, chlorine, . . . ), the total number of bonds of a certain type (simple, double or triple), the number of atoms with a specific vertex degree, distance between the bonds, etc. The remaining forty-eight indices include different topological information, such as the number of double bonds at distance .
1 or 2, and the minimum distance between pairs of atoms, which are counted as the number of bonds between atoms.

The Kier-Hall indices describe the branching (an thus the shape and size) of a molecule by taking into account the contribution of each one of the possible subgraphs of a certain number of nodes that can be obtained from the hydrogen-suppressed graph of the whole molecule. These indices have been widely used in QSAR analysis for series of biologically active molecules in several fields (enzymatic inhibition, analgesic activity, etc.). There are a total of 22 real-valued descriptors of this kind.

Finally, the 32 real-valued electro-topological or charge indices try to describe the charge transfer between pairs of atoms and in some extent the total charge transfer in the molecule.

These molecular representations have shown their ability for discriminating and predicting different kinds of pharmacological properties. Nevertheless, it is known that certain indices are more important than others for detecting particular cases. Obviously, the QSAR studies rely on the key fact that the activity of a molecule directly derives from its structure or, more precisely, from certain aspects of it. The better the chosen set of indices captures these particular aspects, the better the (blind) machine learning methods will characterize the activity of the molecule. As the molecular descriptors or indices have to be general in order to be applied in a wide range of drug design contexts, the ability of the particular learning methods used to capture non linear relations and high order dependencies among them becomes a key fact in the whole process.

B. Cost/benefit in automated drug design

It is important to note that we are interested not only in achieving a high accuracy in classification but also a convenient compromise between true positive and false alarm rates. The high economical costs due to the pharmacological tests on each candidate molecule in drugs research makes an important issue to keep the number of false positives as low as possible, even if this implies to reject some true positives.

Given a particular classifier whose output consists of a continuous value in a specified interval (as in the cases considered in this work), the Receiver Operating Characteristic (ROC) curve is defined as the plot of the true positive rate (TP) against false positive rate (FP) considering the threshold used in the classifier as a parameter. The so-called ROC space is given by all possible results of such a classifier in the form (FP,TP). The performance of any classifier (with the corresponding threshold included) can be represented by a point in the ROC space. ROC curves move from the “all-inactive” point (0,0) which corresponds to the highest value of the threshold to the “all-active” point (1,1) given by the lowest value for the threshold. The straight line between these two trivial points in the ROC space corresponds to the family of random classifiers with different a priori probabilities for each class. The more a ROC curve separates from this line, the better the corresponding classification scheme is. As ROC curves move away from this line, they approach the best possible particular result that corresponds to the point (0,1) in the ROC space which means no false alarms and highest possible accuracy in the active class.

The ROC curve is a perfect tool to find the best trade-off between true positives and false positives and to compare classifiers in a range of different situations. A number techniques to obtain different measures from ROC curves have been developed [7], [6] In particular, we will use the Area Under the Curve (AUC) measure to assess the different classifiers. The AUC is a common method used in ROC analysis to give a global measure of classifier performance. When normalized, this is an scalar value in the range [0, 1].

III. Machine Learning Techniques for Antibacterial Activity Discrimination

The particular discrimination problem was to determine whether a molecule has analgesic properties or not. To this end, a particularly difficult database has been compiled by putting together 111 compounds with known analgesic properties and 862 compounds with no analgesic properties. We will refer here to active and inactive compounds, respectively.

As in previous works, different classification approaches have been considered. In particular, Linear Discrimination Analysis [15], Multilayer Perceptrons [3], Support Vector Machines [2], [8] and k-Nearest Neighbors [5] as a reference.

A. Linear discriminant analysis

Linear discriminant analysis (LDA) has been widely used for this and similar problems in the specific literature [15]. The method consists of finding the optimal separation hyperplane. It is well known that the LDA solution can be very far from the optimal one in the case of highly nonlinear relations among data. This kind of limitation is shared by all linear classification methods.

B. Support vector machines and Nearest Neighbors

Support vector machines (SVM) are considered in this work as a reference given that this was the best option when studying the discrimination of drugs with or without antibacterial activity [8].

From a technical point of view, SVM are linear classifiers that operate in an appropriately transformed space that allows a wide range of nonlinear discrimination possibilities [18].

The radial basis function (RBF) or proximity kernel parameterized by an influence parameter, \( \gamma \), has been considered for the experiments. This parameter along with the soft margin or regularization parameter \( C \) have been chosen by looking at values similar to the ones in previous related works [8].

When using proximity kernels, SVM are closely related to distance-based classifiers whose main representative is the k-Nearest Neighbor classifier. This classifier has also been considered in this work as a reference. The parameter \( k \) has been internally estimated in all cases by using a Leaving-one-out procedure considering training data only.
C. Multilayer Perceptron

Multilayer Perceptrons (MLP) have been extensively used in a wide range of applications that require a nonlinear approach. This classifier consists of a variable number of layers composed by neurons each of which is in fact a linear classifier. Neurons are nonlinearly connected by using a nonlinear sigmoid-like activation function. MLPs can be adaptively trained by suboptimal gradient descent methods.

The training of the MLPs was carried out here by using the neural software package “SNNS: Stuttgart Neural Network Simulator” [20]. The network topology (layers and connections among them), training algorithm and parameter settings were tuned in the vicinity of the ones used in similar studies [3], [14] where significantly easier problems of the same kind were considered.

In particular, the results presented have been obtained by using the Backpropagation algorithm with a learning rate equal to 0.001 and a momentum term of 0.001. The first (input) layer consists of a number of neurons equal to the dimensionality of the particular representation space (number of different attributes used or selected). Only one hidden layer with a specified number of neurons has been considered. In the third (output) layer only one neuron is needed that gives a value in the range $[-1, 1]$ that indicates whether the feature vector at the input is inactive or active, respectively.

When training each neural network, a fixed and relatively large number of presentations has been fixed. A validation set (taken out from the training set) has been used to select the final trained network and prevent overtraining.

The hyperbolic tangent function was used as activation function for all neurons in order to keep outputs in the interval $[-1, 1]$ as in the original LDA experiments [15]. The number of neurons in the hidden layer has been set to 0 (linear), 2, 6, 8 and 12. In the forthcoming results only some of these are shown due to space limitations.

IV. Data Preparation and Experiments

For the experiments presented in this work, a dataset of 973 samples with potential pharmacological activity has been considered. Out of these, 111 molecules are known to have analgesic properties while the other 862 compounds do not have these properties at all. These unbalanced proportions are not necessarily related to the a priori probability of activity in real pharmacological design trials. It is worth noting that one of the main drawbacks to obtain good and stable results with some of the methods considered in this work, comes from the small representativeness of the active class in the data set available.

In order to obtain results as significant as possible, 10-fold stratified cross-validation (CV) has been used to compute all accuracies and performance measures shown.

All 116 molecule descriptors were used to obtain feature vectors in which values were linearly normalized to the interval $[-1, 1]$ in an independent way. Each feature vector was then labeled either with 1 (indicating that the molecule has analgesic properties) or -1 (the molecule is inactive).

V. Feature Selection Experiments

In order to assess different kinds of molecular descriptors and combinations among them, several feature selection (FS) experiments have been carried out using the whole dataset and all 116 molecular descriptors. In this particular work, only suboptimal sequential algorithms have been considered. FS is mainly applied in this work as a filter [13] and not as a wrapper (apart from the LDA case) due to the computational burden involved. A filter FS algorithm uses a given and easy to compute criterion (the Mahalanobis distance assuming Gaussian classes in our case) to select features which only approximates the desired behavior of the final classifier. A wrapper FS algorithm uses the estimated error of the same final classifier as criterion to select features.

Sequential algorithms in which individual features are added or discarded in an incremental and greedy way, have the advantage of being computationally feasible compared to other approaches. In particular, and according to several recent and broad studies [10], [11] the Sequential Forward Floating Search (SFFS) algorithm has been considered [17]. This algorithm and the whole sequential family have an extra advantage in the case of monotonic criteria. Moreover, SFFS has been shown to be able to give very convenient tradeoffs also when used with nonmonotonic criteria or as a wrapper [16].

A number of experiments have been carried out but only two particular ones are reported here. First, SFFS with the Mahalanobis distance has been used as a filter to obtain a sequence of subsets of increasing cardinality to be used when learning the different classifiers. Second, SFFS is used as a wrapper along with the LDA classifier. The expected accuracy taking into account the a priori probability of each class has been considered to assess the different subsets of features. The sequence of subsets obtained in this second experiment has been used also with the other classifiers to compare this two ways of selecting features.

Figure 1 shows the respective criterion values of the two sequence of subsets with regard to the subset cardinality as the search progresses. In the first case using Mahalanobis distance, SFFS led to a more exhaustive exploration and finished after 211 iterations. In the latter, it took only 89 iterations when it was wrapped along with the LDA classifier.

From the particular point of view of the application, it is interesting to see how automatic FS algorithms pick their features from each one of the different families. For each cardinality in each of the two sequences of subsets obtained, the relative amount of features from each one of the three families of descriptors has been computed. This proportions are shown in Figure 2.

It is interesting to see that in all cases, the relative proportions of different families “converge” to their proportion in the whole set of features. This behavior is smoother and its convergence quicker in the case of the subsets obtained when using SFFS as a filter.
VI. EMPIRICAL ASSESSMENT

The above classification methods have been applied to the training sets taken from the available data set and the corresponding (continuous) outputs have been obtained for the test data following the usual CV procedure. For each partition into train and test, a ROC curve is obtained. The threshold averaged ROC curves of all 10 folds have been computed as explained in [6] and intervals for a 95% confidence level have been computed along the two axes.

Figure 3 shows the averaged ROC curves corresponding to representative examples of some of the classification schemes considered. Even though these curves correspond to experiments in this section using feature subsets of different cardinalities (27, 24 and 30 for these particular curves), the curves are representative of the general tendency obtained with LDA and MLP classifiers which were the best suited ones for the task considered in this paper. The difference among these classifiers is of special importance in the zone corresponding to low values of false positive rate which is critical for the application of the results in drug design procedures. The results with k-NN were highly variable and are shown without error bars as a reference.

The experimental procedure has been applied first using some fixed subsets of features corresponding to different families of descriptors. In particular, topological descriptors (topo), Kier-Hall indices (kh), and electro-topological descriptors (ele) along with the combination of the first two (tkh), have been considered. For each one of the classifiers, Table I contains the performance measured as AUC corresponding to these fixed subsets of features along with the same measures using the whole set of features (all).

The same performance measure has been obtained for the different classifiers using each one of the feature subsets in the sequences obtained in the previous feature selection experiments with SFFS. Only the results corresponding to LDA, k-NN and MLP are shown in Figure 4 to keep the graph (relatively) clear. The SVM gave the worst results in this experiment by a significant amount and are not shown.

From the particular results obtained in this experiment and shown in Figure 4 it is worth mentioning the ones obtained with MLP using 27 and 13 features filtered by SFFS using Mahalanobis distance. The first of these is the one whose ROC curve is shown in Figure 3. This value of AUC obtained using MLP using 27 and 13 features filtered by SFFS is basically the same obtained using MLP on the whole set of 116 descriptors which is shown in Table I. This constitutes another confirmation about the abilities of MLP networks to adapt to highly complex scenarios when tuned properly.

The best results using k-NN and LDA classifiers correspond to subsets of features of cardinalities 30 and 24, respectively. In both cases and in general, filtered features

![Fig. 1. Criterion values corresponding to the subset search carried out by SFFS in the two experiments mentioned in the text.](image)

![Fig. 2. Relative proportions of features from each one of the families (topological, Kier-Hall and electro-topological) in each of the feature subsets obtained by SFFS both as a filter (F) and as a wrapper (W). Dotted lines represent the proportion of each family in the whole set of features.](image)

![Fig. 3. Averaged ROC curves for MLP, LDA and k-NN with their respectively best subsets of features of cardinalities 27, 24 and 30 as filtered by SFFS.](image)

### Table I

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>dim</th>
<th>LDA</th>
<th>k-NN</th>
<th>SVM</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>top</td>
<td>62</td>
<td>0.82±0.09</td>
<td>0.81±0.07</td>
<td>0.84±0.11</td>
<td>0.88±0.04</td>
</tr>
<tr>
<td>kh</td>
<td>22</td>
<td>0.79±0.06</td>
<td>0.79±0.07</td>
<td>0.77±0.06</td>
<td>0.88±0.05</td>
</tr>
<tr>
<td>ele</td>
<td>82</td>
<td>0.73±0.08</td>
<td>0.79±0.09</td>
<td>0.76±0.07</td>
<td>0.84±0.08</td>
</tr>
<tr>
<td>tkh</td>
<td>84</td>
<td>0.82±0.04</td>
<td>0.78±0.11</td>
<td>0.84±0.07</td>
<td>0.88±0.06</td>
</tr>
<tr>
<td>all</td>
<td>116</td>
<td>0.84±0.05</td>
<td>0.81±0.06</td>
<td>0.80±0.09</td>
<td>0.96±0.08</td>
</tr>
</tbody>
</table>
using Mahalanobis distance gives the better results.

We have used filter FS in this paper except in the case of LDA. Even in this case and contrary to expectation, filtering with a different criterion gives significant and consistently better results along the different cardinalities considered.

VII. CONCLUDING REMARKS AND FURTHER WORK

In this work a classical FS and ROC analysis have been performed on a particular drug activity discrimination problem. The results presented are preliminary and show that very competitive results can be obtained with a very significantly reduced number of descriptors. This is important both from the pattern recognition and computational point of view. But from the point of view of the application, the discovery of different interactions among descriptors from different families can be of capital importance for the research line related to finding new and better descriptors.

Multilayer Perceptron has been shown to be better than other approaches in a wide range of situations. It has been a relatively surprise the poor results of the SVM approach. Apart from doing a more exhaustive experimentation (which is on its way), we have a strong evidence that the intrinsic difficulty of the task and its severe imbalance is at the root of these poor results.

Also to some extent unexpected was the fact that SFFS as a filter using the (simplest) Mahalanobis distance gives the best results in the experiments presented. The reason for this is two fold. On one hand the criterion used when wrapping SFFS (weighted accuracy) is not exactly the same that has been used for final assessment (AUC). Second, the small size of the test sets used when doing 10-fold CV, makes the possible values of the FS criterion very reduced and some of the flexibility of the SFFS is lost.

Further work is now being done in order to properly deal with the natural imbalance of this problem both in the classifiers considered and in the algorithms and criteria to use. Also, the potential impact of the presented FS results on drug activity characterization from a chemical point of view are currently being investigated.

ACKNOWLEDGMENTS

We are grateful to F. Pérez and M.T. Salabert, from the Universitat de València, for supplying the datasets and for their supervision.

REFERENCES