Rough Sets for Selection of Molecular Descriptors to Predict Biological Activity of Molecules

Pradipta Maji and Sushmita Paul

Abstract—**Quantitative structure activity relationship (QSAR) is one of the important disciplines of computer-aided drug design that deals with the predictive modeling of properties of a molecule. In general, each QSAR dataset is small in size with large number of features or descriptors. Among the large amount of descriptors presented in the QSAR dataset, only a small fraction of them is effective for performing the predictive modeling task. In this paper, a new feature selection algorithm is presented, based on rough set theory, to select a set of effective molecular descriptors from a given QSAR dataset. The proposed algorithm selects the set of molecular descriptors by maximizing both relevance and significance of the descriptors. An important finding is that the proposed feature selection algorithm is shown to be effective in selecting relevant and significant molecular descriptors from the QSAR dataset for predictive modeling. The performance of the proposed algorithm is studied using R^2 statistic of support vector regression method. The effectiveness of the proposed algorithm, along with a comparison with existing algorithms, is demonstrated on three QSAR datasets.**

Index Terms—**Drug design, feature selection, quantitative structure activity relationship (QSAR), rough set, support vector machine (SVM).**

I. Introduction

IN CONVENTIONAL drug design, the drug discovery proceeds largely by trial and error synthesizing thousands of molecules. Although this approach is the most effective method to discover drugs, it is very financially expensive and labor intensive. The conventional drug design method is improved by a nonconventional method, termed as computer-aided drug design (CADD) [1]. The CADD helps in predicting biological activity of a hypothetical molecule and guides scientists toward a specific direction to develop a drug by predicting a molecule with effective biological activity or molecular property against a target molecule. In effect, it minimizes both time and cost. Two well-known approaches are generally taken for prediction: structure-based method and quantitative structure activity relationship (QSAR) method [2]. In structure-based method, the procedure starts with the known 3-D structure of a target molecule, where the goal is to design a ligand or drug that can enhance or decrease the activity of the target molecule. Whereas the QSAR method predicts the activity of hypothetical compounds based on the assayed activity of previously synthesized one [3].

Manuscript received August 3, 2009; revised October 30, 2009 and January 18, 2010; accepted March 28, 2010. Date of publication May 18, 2010; date of current version October 15, 2010. This paper was recommended by Associate Editor R. Alhajj.

The authors are with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India (e-mail: pmaji@isical.ac.in; sushmita_t@isical.ac.in).

The QSAR is the process by which chemical structure is quantitatively correlated with a well-defined process such as biological activity or other molecular property. Biological activity can be expressed quantitatively as in the concentration of a substance required to give a certain biological response. Additionally, when physiochemical properties or structures are expressed by numbers, one can form a mathematical relationship or QSAR between the two. The mathematical expression can then be used to predict the biological response of other unknown chemical structures. The properties that describe the molecule quantitatively are known as molecular descriptors. Molecular descriptors can be obtained by calculated methods or experimental methods. In calculated method, a mathematical procedure is used that transforms chemical information into a number such as surface areas (polar, nonpolar), dipole moment, and volume. On the other hand, in experimental method, some standardized experiments are conducted to measure a molecular descriptor such as melting point, partition coefficients, and refractive index. The molecular descriptors describe different aspects of a molecule; compare different molecular structures, different conformations of same molecule, and database storage; and relate structure to activity [2], [4]–[6].

However, among the large amount of descriptors, only a small fraction is effective for performing the predictive modeling task. Also, a small subset of descriptors is desirable in developing QSAR data-based predicting tools for delivering precise, reliable, and interpretable results. With the descriptor selection results, the cost of biological experiment and decision can be greatly reduced by analyzing only the effective descriptors. Hence, identifying a reduced set of most relevant descriptors is the goal of descriptor selection. The small number of molecules and a large number of descriptors make this problem a more relevant and challenging problem in the QSAR method. This is an important problem in machine learning and referred to as feature selection [7].

Many approaches have been proposed to generate a error-free method for predicting biological activity or other chemical property of a molecule. Ozdemir *et al.* [8] used genetic algorithm to select a subset of molecular descriptors and the significance of these descriptors has been evaluated by a multilayer perceptron. Guha and Jurs [3], [9] used correlation, simulated annealing, and genetic algorithm to obtain the best subset of descriptors. Both linear and nonlinear predictive models have been used to establish the significance of selected descriptors. Similar type of work has been done by Leardi and Gonzalez [10], where genetic algorithm has been used for feature selection and partial least square method for prediction. Sventik *et al.* [11] used the concept of ensemble method for compound classification

and biological activity prediction. In [12], Jain *et al.* have used steric and polar descriptors to predict the biological activity. Tuppurainen *et al.* [13] and Turner *et al.* [14] have used an electronic eigenvalue molecular descriptor and a molecular vibration-based descriptor, respectively, to relate structure and activity of steroid data. Different 3-D molecular descriptors have been proposed in [15], [16] to forecast the biological activity. A different approach based on fuzzy regression has been used to predict the biological activity of persistent organic pollutants [17]. Kumar *et al.* [18] used a method based on fuzzy mappings for the QSAR modeling, while a neural and neuro-fuzzy model have been used in [19] for prediction of toxic action of phenols. On the other hand, Zhou *et al.* [20] have developed a robust boosting partial least square method for modeling the antagonisms of angiotensin II antagonists.

Rough set theory [21], [22] is a new paradigm to deal with uncertainty, vagueness, and incompleteness. It has been applied to fuzzy rule extraction, reasoning with uncertainty, fuzzy modeling, feature selection, and so forth [23]–[26]. It is proposed for indiscernibility in classification according to some similarity [21]. The theory of rough sets has been applied successfully to feature selection of discrete-valued data [27]. Given a dataset with discretized attribute values, it is possible to find a subset of the original attributes using rough set theory that is the most informative; all other attributes can be removed from the dataset with minimal information loss. From the dimensionality reduction perspective, informative features are those that are most useful in determining classifications from their values. Chouchoulas and Shen developed a rough-set-based feature selection algorithm, termed as rough set attribute reduction method [28], which is based on the idea of indiscernibility. The attributes are eliminated from the dataset in such a way that the reduced set, also termed as reduct, provides the same quality of classification as the original set. The dependency values of different combinations of equivalence relations are first calculated, and finally, the reduct with maximum dependency value is retrieved. However, this method is computationally costly as it has to generate several reducts. To alleviate this problem, quick reduct algorithm [28] first calculates the dependency or quality of approximation of single attribute with respect to class or decision attribute. After selecting the best attribute, other attributes are added to it to produce better quality. Addition of attributes is stopped when the final reduct has the same quality as that of maximum possible quality of the dataset.

A reduct with effective attributes can also be obtained from the discernibility matrix-based method [29]. The matrix is developed by considering those attributes that differentiate objects. A discernibility function can then be defined for discernibility matrix data. This function generates all the minimal reducts. However, this approach is computationally very costly. On the other hand, the variable precision rough-set-based attribute selection algorithm [30] is an important method with better generalization ability to produce effective reducts. The main idea here is to classify objects with minimal error. In this method, the relative classification error is calculated between the equivalence classes of condition and decision attributes. The dynamic reduct-based method [31] is an another rough-set-based attribute reduction algorithm, which is based on the idea that the reducts obtained from an information system are sensitive to changes in the system. This method generates a large number of reducts by randomly removing objects from the original data. The reducts, whose proportion of occurrence is more than a defined threshold, are considered as the dynamic reducts. The main drawback of this method is that a predefined threshold value is required. Also, the generation of all reducts is computationally very costly. Many algorithms have also been developed [32]–[34] to generate reducts. Different heuristic approaches based on rough set theory have also been developed for feature selection [35]. Combining rough sets and genetic algorithms, different algorithms have been proposed in [32], [34] to discover optimal or close to optimal reducts. However, the rough-set-based feature selection methods proposed in [28]–[34] select the relevant or predictive features of a dataset without considering the redundancy among them.

In this paper, a new feature selection method is proposed to select a set of molecular descriptors for predicting biological activity of molecules. It employs rough sets to provide a means by which discrete-valued data can be effectively reduced without the need for user-specified information. The proposed method selects a subset of molecular descriptors from the whole feature set by maximizing both relevance and significance of the selected descriptors. The relevance and significance of the descriptors are calculated based on rough set theory. Hence, the only information required in the proposed feature selection method is in the form of equivalence partitions for each attribute, which can be automatically derived from the given dataset. This avoids the need for domain experts to provide information on the data involved and ties in with the advantage of rough sets in that it requires no information other than the dataset itself. The performance of the proposed approach is compared with that of the existing approaches using the R^2 statistic of support vector regression method. An important finding is that the proposed approach is shown to be effective for selecting relevant and significant molecular descriptors from the QSAR datasets. The effectiveness of the proposed method, along with a comparison with other methods, is demonstrated on three QSAR datasets.

The rest of the paper is structured as follows: Section II introduces the necessary notions of rough sets. The proposed feature selection method is described in Section III for predicting biological activity of molecules. A few case studies and a comparison with other related methods are presented in Section IV. Finally, Section V concludes the paper.

II. Rough Sets

The theory of rough sets begins with the notion of an approximation space, which is a pair $\langle \mathbb{U}, \mathbb{A} \rangle$, where $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$ be a nonempty set, the universe of discourse, and $\mathbb{A}$ is a family of attributes, also called knowledge in the universe. V is the value domain of $\mathbb{A}$ and f is an information function $f : \mathbb{U} \times \mathbb{A} \rightarrow V$. An approximation space is also called an information system [21].

Any subset $\mathbb{P}$ of knowledge $\mathbb{A}$ defines an equivalence (also called indiscernibility) relation $IND(\mathbb{P})$ on $\mathbb{U}$

$$IND(\mathbb{P}) = \{(x_i, x_j) \in \mathbb{U} \times \mathbb{U} | \forall a \in \mathbb{P}, f(x_i, a) = f(x_j, a)\}.$$

If $(x_i, x_j) \in IND(\mathbb{P})$, then x_i and x_j are indiscernible by attributes from $\mathbb{P}$. The partition of $\mathbb{U}$ generated by $IND(\mathbb{P})$ is denoted as

$$\mathbb{U}/IND(\mathbb{P}) = \{[x_i]_{\mathbb{P}} : x_i \in \mathbb{U}\} \tag{1}$$

where $[x_i]_{\mathbb{P}}$ is the equivalence class containing x_i. The elements in $[x_i]_{\mathbb{P}}$ are indiscernible or equivalent with respect to knowledge $\mathbb{P}$. Equivalence classes, also termed as information granules, are used to characterize arbitrary subsets of $\mathbb{U}$. The equivalence classes of $IND(\mathbb{P})$ and the empty set $\emptyset$ are the elementary sets in the approximation space $\langle \mathbb{U}, \mathbb{A} \rangle$.

Given an arbitrary set $X \subseteq \mathbb{U}$, in general, it may not be possible to describe X precisely in $\langle \mathbb{U}, \mathbb{A} \rangle$. One may characterize X by a pair of lower and upper approximations defined as follows [21]:

$$\underline{\mathbb{P}}(X) = \bigcup \{[x_i]_{\mathbb{P}} | [x_i]_{\mathbb{P}} \subseteq X\} \quad \text{and}$$
$$\overline{\mathbb{P}}(X) = \bigcup \{[x_i]_{\mathbb{P}} | [x_i]_{\mathbb{P}} \cap X \neq \emptyset\}. \tag{2}$$

Hence, the lower approximation $\underline{\mathbb{P}}(X)$ is the union of all the elementary sets, which are subsets of X, and the upper approximation $\overline{\mathbb{P}}(X)$ is the union of all the elementary sets, which have a nonempty intersection with X. The tuple $\langle \underline{\mathbb{P}}(X), \overline{\mathbb{P}}(X) \rangle$ is the representation of an ordinary set X in the approximation space $\langle \mathbb{U}, \mathbb{A} \rangle$ or simply called the rough set of X. The lower (respectively, upper) approximation $\underline{\mathbb{P}}(X)$ (respectively, $\overline{\mathbb{P}}(X)$) is interpreted as the collection of those elements of $\mathbb{U}$ that definitely (respectively, possibly) belong to X. The lower approximation is also called positive region sometimes, denoted by $POS_{\mathbb{P}}(X)$. A set X is said to be definable (or exact) in $\langle \mathbb{U}, \mathbb{A} \rangle$ iff $\underline{\mathbb{P}}(X) = \overline{\mathbb{P}}(X)$. Otherwise, X is indefinable and termed as a rough set. $BN_{\mathbb{P}}(X) = \overline{\mathbb{P}}(X) \setminus \underline{\mathbb{P}}(X)$ is called a boundary set.

Definition 1: An information system $\langle \mathbb{U}, \mathbb{A} \rangle$ is called a decision table if the attribute set $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$, where $\mathbb{C}$ and $\mathbb{D}$ represent the condition and decision attribute sets, respectively. The dependency between $\mathbb{C}$ and $\mathbb{D}$ can be defined as

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|POS_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|} \tag{3}$$

where $POS_{\mathbb{C}}(\mathbb{D}) = \cup \underline{\mathbb{C}} X_i$, X_i is the ith equivalence class induced by $\mathbb{D}$ and $|\cdot|$ denotes the cardinality of a set.

Let $\mathbb{I} = \langle \mathbb{U}, \mathbb{A} \rangle$ be a decision table, where $\mathbb{U} = \{x_1, \ldots, x_7\}$ is a nonempty set of finite objects, the universe, and $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$ is a nonempty finite set of attributes. Here, $\mathbb{C} = \{\text{Age, LEMS}\}$ and $\mathbb{D} = \{\text{Walk}\}$ are the set of condition and decision attributes, respectively.

$x_i \in \mathbb{U}$	Age	LEMS	Walk
x_1	$16-30$	50	yes
x_2	$16-30$	0	no
x_3	$31-45$	$1-25$	no
x_4	$31-45$	$1-25$	yes
x_5	$46-60$	$26-49$	no
x_6	$16-30$	$26-49$	yes
x_7	$46-60$	$26-49$	no

$IND(\{\text{Age}\})$ creates the following partition of $\mathbb{U}$:

$$\mathbb{U}/IND(\{\text{Age}\}) = \{\{x_1, x_2, x_6\}, \{x_3, x_4\}, \{x_5, x_7\}\}$$

as the objects x_1, x_2, and x_6 are indiscernible with respect to the condition attribute set $\{\text{Age}\}$. Similarly, the partition of $\mathbb{U}$ generated by the condition attribute set $\{\text{LEMS}\}$ is given by

$$\mathbb{U}/IND(\{\text{LEMS}\}) = \{\{x_1\}, \{x_2\}, \{x_3, x_4\}, \{x_5, x_6, x_7\}\}$$

and the partition of $\mathbb{U}$ generated by the condition attribute set $\{\text{Age, LEMS}\}$ is as follows:

$$\mathbb{U}/\{\text{Age, LEMS}\} = \{\{x_1\}, \{x_2\}, \{x_3, x_4\}, \{x_5, x_7\}, \{x_6\}\}.$$

Similarly, the partition of $\mathbb{U}$ generated by the decision attribute set $\{\text{Walk}\}$ is given by

$$\mathbb{U}/\mathbb{D} = \mathbb{U}/\{\text{Walk}\} = \{\{x_1, x_4, x_6\}, \{x_2, x_3, x_5, x_7\}\}.$$

The positive region contains all objects of $\mathbb{U}$ that can be classified into classes of $\mathbb{U}/\mathbb{D}$ using the knowledge in attributes $\mathbb{C}$. Hence, for the above example, the positive region is as follows:

$$POS_{\mathbb{C}}(\mathbb{D}) = \bigcup \{\phi, \{x_1\}, \{x_2\}, \{x_5, x_7\}, \{x_6\}\}$$
$$= \{x_1, x_2, x_5, x_6, x_7\}.$$

The dependency between $\mathbb{C}$ and $\mathbb{D}$ is, therefore, given by

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{5}{7}.$$

An important issue in data analysis is discovering dependency between attributes. Intuitively, a set of attributes $\mathbb{D}$ depends totally on a set of attributes $\mathbb{C}$, denoted by $\mathbb{C} \Rightarrow \mathbb{D}$, if all attribute values from $\mathbb{D}$ are uniquely determined by values of attributes from $\mathbb{C}$. If there exists a functional dependency between values of $\mathbb{D}$ and $\mathbb{C}$, then $\mathbb{D}$ depends totally on $\mathbb{C}$. Dependency can be defined in the following way:

Definition 2: For $\mathbb{C}, \mathbb{D} \subseteq \mathbb{A}$, it is said that $\mathbb{D}$ depends on $\mathbb{C}$ in a degree κ $(0 \leq \kappa \leq 1)$, denoted by $\mathbb{C} \Rightarrow_{\kappa} \mathbb{D}$, if

$$\kappa = \gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|POS_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|}. \tag{4}$$

If $\kappa = 1$, $\mathbb{D}$ depends totally on $\mathbb{C}$, if $0 < \kappa < 1$, $\mathbb{D}$ depends partially (in a degree κ) on $\mathbb{C}$, and if $\kappa = 0$, then $\mathbb{D}$ does not depend on $\mathbb{C}$.

To what extent an attribute is contributing to calculate the dependency on decision attribute can be calculated by the significance of that attribute. The change in dependency when an

attribute is removed from the set of condition attributes is a measure of the significance of the attribute. The higher the change in dependency, the more significant the attribute is. If the significance is 0, then the attribute is dispensable.

Definition 3: Given $\mathbb{C}, \mathbb{D}$, and an attribute $a \in \mathbb{C}$, the significance of the attribute a is defined as follows:

$$\sigma_{\mathbb{C}}(\mathbb{D}, a) = \gamma_{\mathbb{C}}(\mathbb{D}) - \gamma_{\mathbb{C}-a}(\mathbb{D}). \tag{5}$$

Considering the above example, let $\mathbb{C} = \{a, b\}$, where $a = \{\text{Age}\}$, $b = \{\text{LEMS}\}$, and $\mathbb{D} = \{\text{Walk}\}$. The significance of the two attributes a and b is as follows:

$$\sigma_{\mathbb{C}}(\mathbb{D}, a) = \gamma_{\mathbb{C}}(\mathbb{D}) - \gamma_{\mathbb{C}-a}(\mathbb{D}) = \frac{5}{7} - \frac{2}{7} = \frac{3}{7}$$

$$\sigma_{\mathbb{C}}(\mathbb{D}, b) = \gamma_{\mathbb{C}}(\mathbb{D}) - \gamma_{\mathbb{C}-b}(\mathbb{D}) = \frac{5}{7} - \frac{2}{7} = \frac{3}{7}.$$

III. Proposed Feature Selection Algorithm

The main objective of the current research is to build a method that can effectively find out biological activity values of molecules provided with their molecular descriptors. In effect, it can help to decide which features of a molecule give rise to its overall activity and help to make modified compounds with enhanced properties.

In general, the QSAR dataset may contain a number of insignificant molecular descriptors. The presence of such irrelevant and insignificant molecular descriptors can produce inappropriate information. A standard descriptor set is the one that has high relevance with the activity values and high significance in the feature set. The molecular descriptors with high relevance are expected to predict the biological activity effectively. However, if insignificant descriptors are present in the subset, they may reduce the prediction capability. A feature set with high relevance and high significance enhances the predictive capability. Accordingly, a measure is required that can enhance the effectiveness of the descriptors. In this paper, the theory of rough sets is used to select the relevant and significant molecular descriptors from the QSAR dataset.

A. Maximum Relevance–Maximum Significance (MRMS)

Let $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$ be the set of n molecules and $\mathbb{M} = \{\mathbb{M}_1, \ldots, \mathbb{M}_j, \ldots, \mathbb{M}_m\}$ be the set of m molecular descriptors of a QSAR dataset. These molecules and descriptors form a table $\mathcal{T} = \{w_{ij} | i = 1, \ldots, n, j = 1, \ldots, m\}$, where $w_{ij} \in \Re$ is the measured value of the molecular descriptor $\mathbb{M}_j$ in the molecule x_i. Let $\mathbb{B} = \{\mathcal{B}_1, \ldots, \mathcal{B}_i, \ldots, \mathcal{B}_n\}$ represents the set of biological activity values of n molecules, where $\mathcal{B}_i \in \Re$ is the activity value of the molecule x_i. Hence, in terms of rough set theory, a QSAR dataset can be considered as a decision table $\mathbb{I} = \langle \mathbb{U}, \mathbb{M} \cup \mathbb{B} \rangle$, where $\mathbb{M}$ and $\mathbb{B}$ play the role of condition and decision attribute sets, respectively. However, the continuous values are discretized to compute the relevance and significance of descriptors using rough sets.

Let $\mathbb{S}$ be the set of selected descriptors with cardinality $d < m$. Define $\hat{f}(\mathbb{M}_i, \mathbb{B})$ as the relevance of the descriptor $\mathbb{M}_i$ with respect to the response variable or biological activity value $\mathbb{B}$ while $\tilde{f}(\mathbb{M}_i, \mathbb{M}_j)$ as the significance of the descriptor $\mathbb{M}_j$ with respect to the already selected descriptor $\mathbb{M}_i$. The total relevance of all selected descriptors is, therefore, given by

$$\mathcal{J}_{\text{relev}} = \sum_{\mathbb{M}_i \in \mathbb{S}} \hat{f}(\mathbb{M}_i, \mathbb{B}). \tag{6}$$

The task of descriptor or feature selection is to find a descriptor subset $\mathbb{S} \subseteq \mathbb{M}$ that maximizes the objective function $\mathcal{J}_{\text{relev}}$. In terms of rough set theory, the relevance $\hat{f}(\mathbb{M}_i, \mathbb{B})$ of a molecular descriptor $\mathbb{M}_i$ with respect to the biological activity $\mathbb{B}$ can be calculated using (4), i.e.

$$\mathcal{J}_{\text{relev}} = \sum_{\mathbb{M}_i \in \mathbb{S}} \gamma_{\mathbb{M}_i}(\mathbb{B}). \tag{7}$$

However, it is likely that descriptors selected according to the above criterion could have rich redundancy, that is, the dependency among these descriptors could be large. When two molecular descriptors highly depend on each other, the respective biological activity prediction power would not change much if one of them were removed. It follows that one descriptor is dispensable with respect to the other. The significance criterion defined in (5) is able to find out the dispensable descriptors. If the significance of a descriptor with respect to another descriptor is 0, then the descriptor is dispensable [21]. Therefore, the significance criterion can be added to select mutually exclusive descriptors. The total significance among the selected descriptors is

$$\mathcal{J}_{\text{signf}} = \sum_{\mathbb{M}_i \neq \mathbb{M}_j \in \mathbb{S}} \tilde{f}(\mathbb{M}_i, \mathbb{M}_j). \tag{8}$$

In the proposed feature selection method, the significance $\tilde{f}(\mathbb{M}_i, \mathbb{M}_j)$ of the descriptor $\mathbb{M}_j$ with respect to the already selected descriptor $\mathbb{M}_i$ is computed using (5), i.e.

$$\mathcal{J}_{\text{signf}} = \sum_{\mathbb{M}_i \neq \mathbb{M}_j \in \mathbb{S}} \sigma_{\mathbb{M}_i \cup \mathbb{M}_j}(\mathbb{B}, \mathbb{M}_j). \tag{9}$$

Therefore, the problem of selecting a set $\mathbb{S}$ of d relevant and significant descriptors from the whole set $\mathbb{M}$ of m descriptors is equivalent to maximize both $\mathcal{J}_{\text{relev}}$ and $\mathcal{J}_{\text{signf}}$, that is, to maximize the objective function $\mathcal{J}$, where

$$\mathcal{J} = \mathcal{J}_{\text{relev}} + \mathcal{J}_{\text{signf}} \tag{10}$$

$$\text{i.e., } \mathcal{J} = \sum_{\mathbb{M}_i \in \mathbb{S}} \gamma_{\mathbb{M}_i}(\mathbb{B}) + \sum_{\mathbb{M}_i \neq \mathbb{M}_j \in \mathbb{S}} \sigma_{\mathbb{M}_i \cup \mathbb{M}_j}(\mathbb{B}, \mathbb{M}_j). \tag{11}$$

Obviously, when d equals 1, the solution is the molecular descriptor that maximizes $\hat{f}(\mathbb{M}_i, \mathbb{B})$; $(1 \leq i \leq m)$. When $d > 1$, a simple incremental search scheme is to add one descriptor at one time. This type of selection is called the first-order incremental search. By definition of first-order search, it is assumed that the set of $(d-1)$ descriptors has already been obtained. The task is to select the optimal dth descriptor $\mathbb{M}_j$ from the remaining descriptors of the set $\mathbb{M}$ that contributes to the largest increase of the following condition:

$$\hat{f}(\mathbb{M}_j, \mathbb{B}) + \frac{1}{|\mathbb{S}|} \sum_{\mathbb{M}_i \in \mathbb{S}} \tilde{f}(\mathbb{M}_i, \mathbb{M}_j) \quad \text{where } |\mathbb{S}| = d - 1. \tag{12}$$

Hence, the following greedy algorithm is used to select relevant and significant descriptors from a QSAR dataset:

1) Initialize $\mathbb{M} \leftarrow \{\mathbb{M}_1, \ldots, \mathbb{M}_i, \ldots, \mathbb{M}_m\}, \mathbb{S} \leftarrow \emptyset$.
2) Calculate the relevance value $\hat{f}(\mathbb{M}_i, \mathbb{B})$ of each descriptor $\mathbb{M}_i \in \mathbb{M}$ with respect to the biological activity $\mathbb{B}$.
3) Select the descriptor $\mathbb{M}_i$ as the most relevant descriptor that has the highest relevance $\hat{f}(\mathbb{M}_i, \mathbb{B})$. In effect, $\mathbb{M}_i \in \mathbb{S}$ and $\mathbb{M} = \mathbb{M} \setminus \mathbb{M}_i$.
4) Repeat the following two steps until the desired number of descriptors are selected.
5) Calculate the significance of each of the remaining descriptors of $\mathbb{M}$ with respect to the already selected descriptors of $\mathbb{S}$.
6) From the remaining descriptors of $\mathbb{M}$, select descriptor $\mathbb{M}_j$ that maximizes

$$\hat{f}(\mathbb{M}_j, \mathbb{B}) + \frac{1}{|\mathbb{S}|} \sum_{\mathbb{M}_i \in \mathbb{S}} \bar{f}(\mathbb{M}_i, \mathbb{M}_j). \tag{13}$$

As a result of this, $\mathbb{M}_j \in \mathbb{S}$ and $\mathbb{M} = \mathbb{M} \setminus \mathbb{M}_j$.

In the proposed feature selection method, the relevance $\hat{f}(\mathbb{M}_i, \mathbb{B})$ of a molecular descriptor $\mathbb{M}_i$ with respect to the biological activity $\mathbb{B}$ is calculated using (4), while the significance $\bar{f}(\mathbb{M}_i, \mathbb{M}_j)$ of the descriptor $\mathbb{M}_j$ with respect to the already selected descriptor $\mathbb{M}_i$ is computed using (5).

B. Computational Complexity

The rough set theory-based proposed feature selection method has low computational complexity with respect to the number of descriptors in the original dataset. The computation of the relevance of m descriptors is carried out in step 2 of the proposed algorithm, which has $\mathcal{O}(m)$ time complexity. The selection of the most relevant descriptor from the set of m descriptors, that is step 3, has also a complexity $\mathcal{O}(m)$. There is only one loop in the proposed feature selection process, which is executed $(d-1)$ times, where d represents the number of selected features. Each iteration of the loop takes only a constant amount of time. The complexity to calculate the significance of a descriptor with respect to the already selected descriptors is $\mathcal{O}(\acute{m})$, where $\acute{m}$ is the cardinality of the already selected descriptor set. In effect, the selection of a set of d relevant and significant descriptors from the whole set of m descriptors using the proposed first-order incremental search method has an overall computational complexity of $(\mathcal{O}(m) + \mathcal{O}(d\acute{m})) = \mathcal{O}(m)$ as $d, \acute{m} \ll m$.

C. Generation of Equivalence Classes

In QSAR dataset, the molecular descriptor values as well as the biological activity values of different molecules are continuous. Hence, to measure both the relevance and significance of molecular descriptors using rough set theory, the continuous descriptor values of a molecule are usually divided into several discrete partitions to generate equivalence classes. The discretization method reported in [36] is employed to discretize the continuous descriptor values. The values of a descriptor or an attribute are discretized using mean μ and standard deviation σ computed over n values of that attribute: any value larger than $(\mu + \frac{\sigma}{2})$ is transformed into state 1; any value between $(\mu - \frac{\sigma}{2})$ and $(\mu + \frac{\sigma}{2})$ is transformed into state 0; and any value smaller than $(\mu - \frac{\sigma}{2})$ is transformed into state -1 [36]. The equivalence classes are then generated to compute both the relevance and significance of molecular descriptors.

IV. Experimental Results

The performance of the proposed rough-set-based MRMS method is extensively studied and compared with that of some existing algorithms. All the algorithms are implemented in C language and run in LINUX environment having machine configuration Pentium IV, 2.8 GHz, 1 MB cache, and 512 MB RAM. To analyze the performance of different algorithms, the experimentation is done on three QSAR datasets. The major metric for evaluating the performance of different algorithms is the R^2 statistic of support vector regression method.

A. QSAR Datasets

In this paper, the following three QSAR datasets are used that are available at http://www.cheminformatics.org.

1) Steroid Dataset: This dataset contains 31 steroid molecules presented in MOL format, which is used in cheminformatics applications for storing atomic coordinates, chemical bond information, and metadata of the 3D structure of a single chemical compound in plain text tabular format. The $\log(1/k)$ values of these molecules are also given. All these molecules are categorized into three activity classes. Among them, 11 are reported as high-activity molecules, 9 moderate, and the rest 11 as the lowest activity molecules.

2) Small Dopamine Dataset: It contains 26 dopamine molecules given in MOL format. The biological activity of these molecules is also available.

3) Large Dopamine Dataset: This dataset consists of 116 dopamine molecules that are given along with their molecular descriptors in binary form. The continuous-valued biological activity of each molecule is also given.

Both steroid and small dopamine datasets are available in MOL format. The molecular descriptors of these datasets are obtained using MODEL software [6], which calculates approximately 4000 molecular descriptors for each molecule. The calculated descriptors cover different aspects of the molecular structure including topological, electronic, constitutional, geometrical, and physical descriptors.

B. Support Vector Regression Method

The support vector machine (SVM) [37] is a relatively new and promising classification and regression method. It is a margin classifier that draws an optimal hyperplane in the feature vector space; this defines a boundary that maximizes the margin between data samples in different classes, therefore, leading to good generalization properties. A key factor in the SVM is to use kernels to construct nonlinear decision boundary. In this paper, radial basis function kernels

TABLE I
PERFORMANCE ON NUMBER OF EQUIVALENCE CLASSES

Data Set	Experiment	$c = 2$	$c = 3$	$c = 5$
Steroid	10-fold CV	0.12	0.89	0.81
	LOOCV	0.33	0.88	0.84
Small Dopamine	10-fold CV	0.18	0.37	0.24
	LOOCV	0.39	0.45	0.27
Large Dopamine	10-fold CV	0.52	0.52	0.52
	LOOCV	0.53	0.53	0.53

are used. The source code of the SVM is downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

The performance of the SVM is analyzed using R^2 statistic or coefficient of determination value. The R^2 statistic tells about the goodness of fit of a model and how well a regression approximates its attributes. The value of R^2 statistic ranges from 0 to 1. The near the value reaches to 1, the better is the approximation. The R^2 statistic can be calculated as follows:

$$R^2 = 1 - \frac{SS_{\text{err}}}{SS_{\text{tot}}} \tag{14}$$

where $SS_{\text{tot}} = \Sigma_i (y_i - \bar{y})^2$ represents the total sum of squares, which is proportional to the sample variance, and $SS_{\text{err}} = \Sigma_i (y_i - f_i)^2$ is the sum of squared errors, also called the residual sum of squares. Here, $\bar{y}$ represents the mean of the observed data, while y_i and f_i are the ith observed and modeled or predicted values, respectively.

C. Optimum Number of Equivalence Classes

In QSAR dataset, both molecular descriptors and biological activity values are continuous. Hence, to measure the relevance and significance of descriptors using rough set theory, the continuous values have to be divided into several discrete partitions to generate equivalence classes. In the proposed method, the continuous values are discretized into three ($c = 3$) states as per the procedure reported in Section III-C.

In order to establish the effectiveness of the three ($c = 3$) states discretization procedure, the extensive experiments are carried out on different QSAR datasets. The performance of the proposed feature selection method for $c = 3$ is compared with that for $c = 2$ and 5. For $c = 2$, any value larger than mean is transformed into one state, while others to another state. On the other hand, for $c = 5$, the intermediate state of $c = 3$ is partitioned into three states, while other two states remain unaltered, therefore, leading to total five states. Table I reports the comparative performance of the proposed method for $c = 2$, 3, and 5 with respect to the R^2 statistic of the SVM. To compute the R^2 statistic, both leave-one-out cross validation (LOOCV) and 10-fold cross validation (CV) are performed on each QSAR dataset. All the results reported in Table I establish the fact that the performance of the proposed rough-set-based feature selection method is significantly better in case of $c = 3$ than that of $c = 2$ and 5.

D. Performance Analysis

The experimental results on three QSAR datasets are presented in Figs. 1–9. Subsequent discussions analyze the results

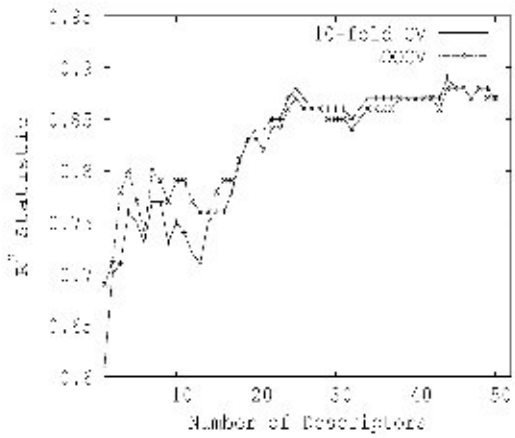

Fig. 1. Results on steroid molecules obtained by 10-fold CV and LOOCV.

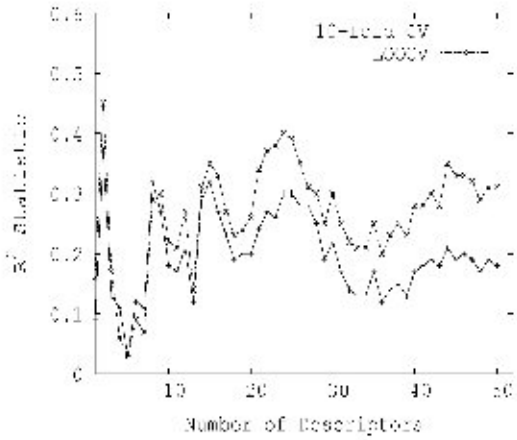

Fig. 2. Results on small dopamine obtained by 10-fold CV and LOOCV.

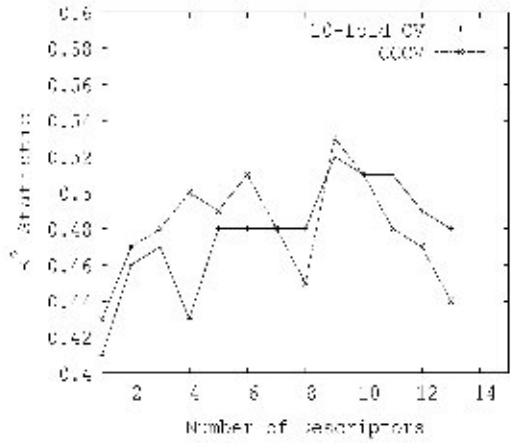

Fig. 3. Results on large dopamine obtained by 10-fold CV and LOOCV.

with respect to the R^2 statistic of the SVM. To compute the R^2 statistic of the SVM, both LOOCV and 10-fold CV are performed on each QSAR dataset. The number of molecular descriptors selected ranges from 1 to 50.

Fig. 1 presents the performance of the proposed MRMS method on steroid molecules obtained by both 10-fold CV and LOOCV, while Figs. 2 and 3 depict that for small and large

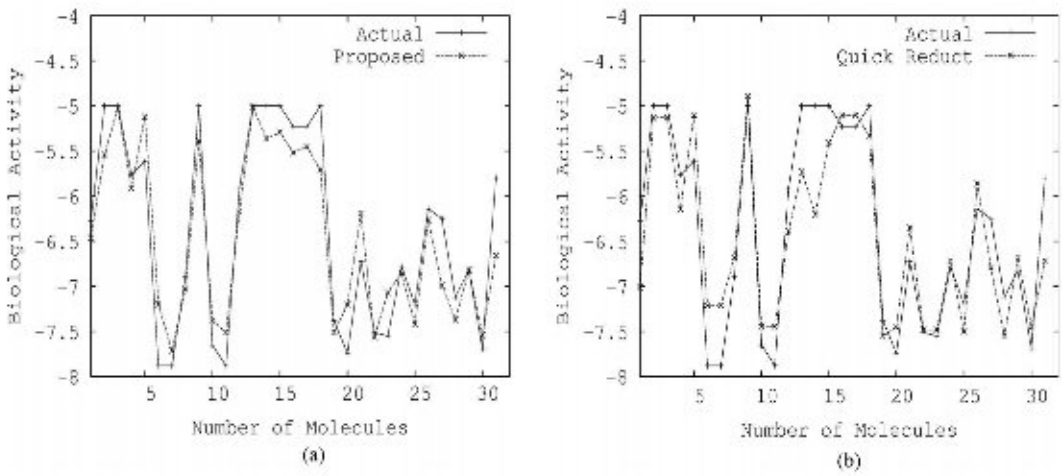

Fig. 4. Results for steroid molecules obtained by leave-one-out CV. (a) Proposed method. (b) Quick reduct algorithm.

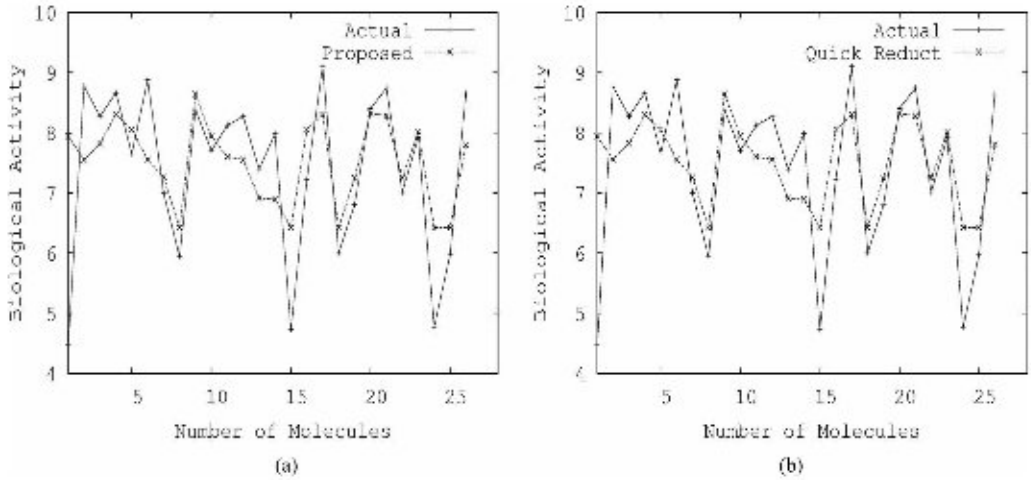

Fig. 5. Results for small dopamine molecules obtained by leave-one-out CV. (a) Proposed method. (b) Quick reduct algorithm.

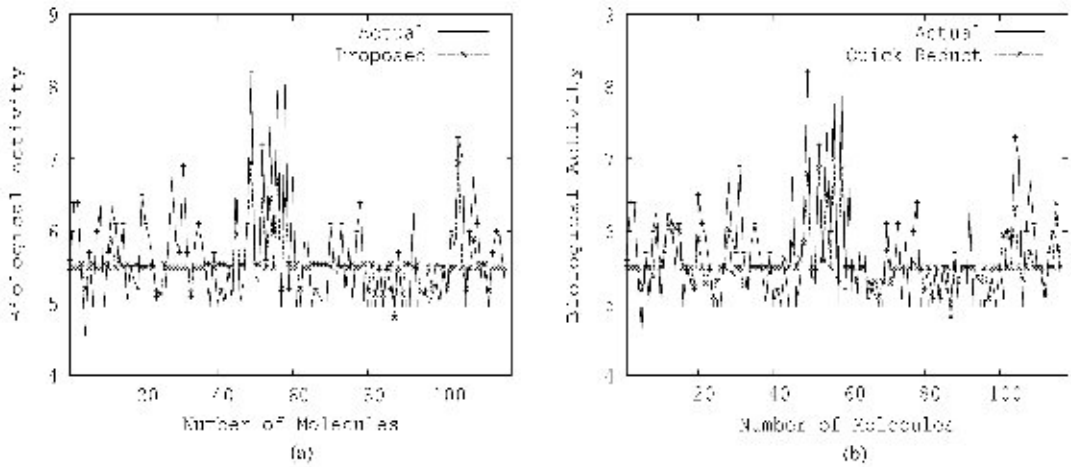

Fig. 6. Results for large dopamine molecules obtained by leave-one-out CV. (a) Proposed method. (b) Quick reduct algorithm.

dopamine molecules, respectively. In Fig. 1, it is seen that as the number of selected descriptors of steroid molecules ranges from 1 to 15, the R^2 statistic of the SVM fluctuates in case of both 10-fold CV and LOOCV. It indicates that the proposed MRMS method gets stuck into local minima of the search space for this range. However, the R^2 statistic continuously increases with the increase in number of selected descriptors for more than 15. Finally, the proposed method attains its maximum R^2 statistic of 0.88 and 0.89 using only 44 descriptors for the LOOCV and ten-fold CV, respectively. That is, the MRMS method is able to find

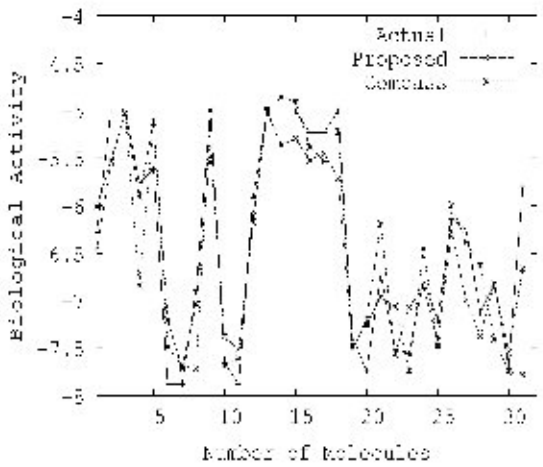

Fig. 7. Results of MRMS and Compass on steroid molecules using LOOCV.

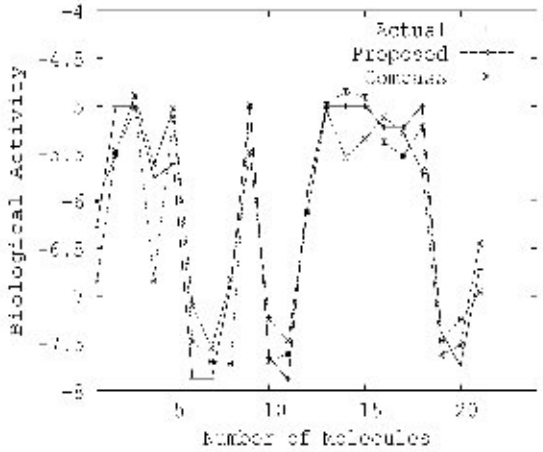

Fig. 8. Results of MRMS and Compass on 21 training steroid molecules.

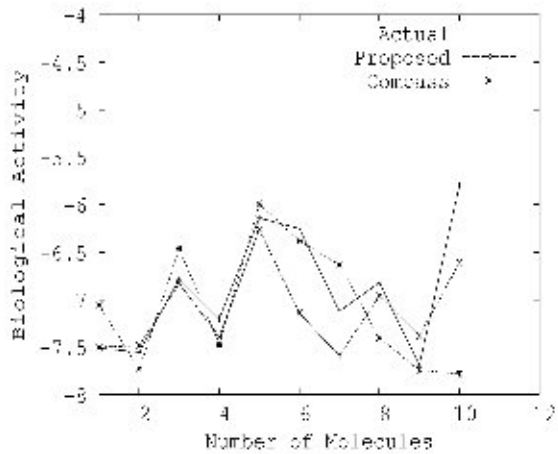

Fig. 9. Results of MRMS and Compass on 10 test steroid molecules.

out an optimum or near-optimum solution using 44 descriptors for both 10-fold CV and LOOCV. On the other hand, in Fig. 2, it can be seen that in case of small dopamine molecules, two most relevant and significant descriptors are sufficient to achieve the maximum R^2 statistic values of 0.45 and 0.37 of the proposed method for the LOOCV and 10-fold CV, respectively. Finally,

TABLE II
EXECUTION TIME OF DIFFERENT ALGORITHMS

Data Set	Quick Reduct	Discern. Matrix	MRMS
Steroid	383253	55061	3498
Small Dopamine	350015	54044	4299
Large Dopamine	487745	35027	1755

Fig. 3 depicts the results for large dopamine molecules. From the results presented in Fig. 3, it is seen that the proposed method attains maximum R^2 statistic of 0.53 with nine descriptors using the LOOCV, while for 10-fold CV, the best R^2 statistic is 0.52 with the same number of descriptors. In other words, the MRMS method is able to find out optimum or near-optimum solutions using two and nine molecular descriptors for small and large dopamine molecules, respectively.

Figs. 4–6 present the comparative performance analysis of the proposed MRMS method and one of the most popular rough-set-based algorithms, called quick reduct algorithm [28]. All the results are reported for three QSAR datasets based on the LOOCV. The actual and obtained biological activity values of different molecules for three QSAR datasets are reported for comparison. The R^2 statistic values of quick reduct algorithm are 0.82, 0.45, and 0.56 for steroid, small dopamine, and large dopamine molecules, respectively. For 10-fold CV, the R^2 statistic values of quick reduct algorithm are 0.83, 0.37, and 0.52 on steroid, small dopamine, and large dopamine, respectively. From the results reported in Figs. 4–6, it is seen that the performance of the proposed MRMS method is better than the quick reduct algorithm in case of steroid dataset and comparable with the quick reduct algorithm for both small and large dopamine molecules. In this regard, it should be noted that another rough-set-based algorithm, called discernibility-matrix-based method [29], attains the R^2 statistic values of 0.79, 0.43, and 0.39 for steroid, small dopamine, and large dopamine molecules, respectively, using 10-fold CV, while the corresponding values for the LOOCV are 0.79, 0.61, and 0.41, respectively. However, as the computational complexity of both quick reduct method [28] and discernibility-matrix-based method [29] is very high, they require significantly higher execution time compared to that of the MRMS algorithm.

Table II compares the execution time (in millisecond) of the proposed MRMS algorithm and that of the quick reduct algorithm [28] and discernibility-matrix-based method [29] for three QSAR datasets. From the results reported in Table II, it is seen that the execution time required for the proposed MRMS algorithm is significantly lower than that of other two algorithms, irrespective of the datasets used. As the computational complexity of both quick reduct algorithm and discernibility-matrix-based method is exponential in nature [28], [29], they require significantly higher execution time compared to that of the proposed algorithm. The significantly lesser execution time of the proposed algorithm is achieved due to its low computational complexity.

E. Comparative Performance Analysis

The proposed MRMS method performs significantly better than different existing QSAR methods. To establish the

TABLE III
RESULT ON TRAINING SET OF STEROID DATA

	Methods	R^2 statistic
Existing Models	CoMFA	0.69
	Compass	0.89
Proposed	MRMS	0.97

TABLE IV
RESULT ON TEST SET OF STEROID DATA

	Methods	R^2 statistic
Existing Models	Compass	0.76
	MS-WHIM	0.28
	PARM	0.33
	TQSAR	0.16
	SOMFA	0.20
	EVA	0.36
	CoMFA	0.25
	COMSA	0.09
	MEDV	0.45
	QS-SM	0.36
	EEVA	0.36
Proposed	MRMS	0.67

superiority of the proposed method, extensive experimentation is carried out on different QSAR datasets. Fig. 7 presents the predicted biological activity values of the proposed method and Compass [12], a well-known existing QSAR model, along with the actual activity values. Results are reported based on the LOOCV. The R^2 statistic values corresponding to the proposed method and Compass are 0.89 and 0.79, respectively. Next, the steroid dataset is divided into two sets: training set of 21 molecules and test set of 10 molecules. The LOOCV results of 21 molecules obtained by the proposed method as well as two well-known existing approaches, namely Compass [12] and CoMFA [14], are reported in Table III. Figs. 8 and 9 depict the actual and predicted values of the proposed method and Compass [12] for 21 training and 10 test steroid molecules, respectively. A detailed comparison of the proposed method with other existing 3D QSAR methods, namely Compass [12], MS-WHIM [15], PARM [38], TQSAR [16], SOMFA [39], EVA [14], CoMFA [14], COMSA [40], MEDV [41], QS-SM [42], and EEVA [13], is presented in Table IV on test set of steroid data, that is, molecules 22–31.

From the R^2 statistic reported in Tables III and IV, along with the results reported in Figs. 7–9, it can be seen that the proposed MRMS method outperforms different existing QSAR approaches in case of steroid dataset. Also, the proposed method predicts biological activity of 21 training and 10 test molecules significantly better than the Compass [12]. Moreover, the model building phase of Compass takes about 1 minute per molecule for steroid dataset [12], which is significantly higher than that of the proposed method.

Among 2901 molecular descriptors of steroid dataset, 44 relevant and significant descriptors obtained using the proposed MRMS method can predict biological activity values of steroid molecules accurately. All these 44 descriptors can be grouped into one of the following four descriptor types, namely topological, geometrical, electronic, and charge. By analyzing the R^2 statistic values of steroid dataset, one can deduce that the topological, geometrical, electronic, and charge descriptors do favorably affect the biological activities of these molecules, while thermodynamic and constitutional descriptors can make adverse affect on biological activities.

Finally, the 10-fold CV result of the MRMS method for large dopamine data is compared with the existing approach Boosting of Sventik *et al.* [11]. While the proposed method achieves the R^2 value of 0.52 with nine attributes, the best result obtained by the Boosting method is 0.48, that is, the proposed method performs significantly better than the existing method.

V. CONCLUSION AND FUTURE DIRECTION

This paper introduces a new feature selection algorithm based on rough set theory in order to identify the relevant and significant molecular descriptors from high-dimensional QSAR datasets. It presents the results of selecting effective molecular descriptors for predicting biological activity of molecules.

The MRMS framework is proposed here as the molecular descriptor selection method. The performance of the proposed method is evaluated by the R^2 statistic of support vector regression method. For all datasets, significantly better results are found for the proposed method, compared to different existing QSAR models. The results obtained on real datasets demonstrate that the proposed method can bring a remarkable improvement on descriptor selection problem, and therefore, it can be a promising alternative to existing QSAR models for prediction of biological activity of molecules. All the results reported in this paper demonstrate the feasibility and effectiveness of the proposed method. The new method is capable of identifying effective molecular descriptors that may contribute to revealing underlying molecular structures, providing a useful tool for exploratory analysis of QSAR data.

As the current study is done with just a single conformation of datasets, a further study can be done using the proposed method with multiple conformations. An extensive analysis is also required to understand how these molecules or ligands bind to their respective target biomolecules.

REFERENCES

[1] J. Bajorath, T. E. Klein, T. P. Lybrand, and J. Novotny, "Computer-aided drug discovery: From target proteins to drug candidates," *in Pac. Symp. Biocomput.*, 1999, vol. 4, pp. 413–414.

[2] A. R. Leach, *Molecular Modelling: Principles and Applications*, vol. 2. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[3] R. Guha and P. C. Jurs, "Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors," *J. Chem. Inf. Comput. Sci.*, vol. 44, pp. 2179–2189, 2004.

[4] I. V. Tetkoa, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, and V. V. Prokopenko, "Virtual computational chemistry laboratory design and description," *J. Comput.-Aided Mol. Des.*, vol. 19, pp. 453–463, 2005.

[5] A. R. Katritzky, V. Lobanov, and M. karelson, "Comprehensive descriptors for structural and statistical analysis version 1.1," Tech. Rep., Univ. Florida, 1994.

[6] Z. R. Li, L. Y. Han, Y. Xue, C. W. Yap, H. Li, L. Jiang, and Y. Z. Chen, "MODEL—molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds," *Biotechnol. Bioeng.*, vol. 97, pp. 389–396, 2007.

[7] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1982.
[8] M. Ozdemir, M. J. Embrechts, F. Arciniegas, C. M. Breneman, L. Lockwood, and K. P. Bennett, "Feature selection for in-silico drug design using genetic algorithms and neural networks," in *Proc. IEEE Mt Workshop Soft Comput. Ind. Appl.*, 2001, pp. 25–27.
[9] R. Guha and P. C. Jurs, "Development of QSAR models to predict and interpret the biological activity of artemisinin analogues," *J. Chem. Inf. Comput. Sci.*, vol. 44, pp. 1440–1449, 2004.
[10] R. Leardi and A. L. Gonzalez, "Genetic algorithms applied to feature selection in PLS regression: How and when to use them," *Chemometrics Intell. Lab. Syst.*, vol. 41, pp. 195–207, 1998.
[11] V. Sventik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan, and Q. Song, "Boosting: An ensemble learning tool for compound classification and QSAR modeling," *J. Chem. Inf. Model.*, vol. 45, no. 3, pp. 786–799, 2005.
[12] A. N. Jain, K. Koile, and D. Chapman, "Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark," *J. Med. Chem.*, vol. 37, pp. 2315–2327, 1994.
[13] K. Tuppurainen, M. Viisas, R. Laatikainen, and M. Peräkylä, "Evaluation of a novel electronic eigenvalue (EEVA) molecular descriptor for QSAR/QSPR studies: Validation using a benchmark steroid data set," *J. Chem. Inf. Comput. Sci.*, vol. 42, pp. 607–613, 2002.
[14] D. B. Turner, P. Willett, A. M. Ferguson, and T. W. Heritage, "Evaluation of a novel molecular vibration-based descriptor (EVA) for QSAR studies: 2. Model validation using a benchmark steroid dataset," *J. Comput.-Aided Mol. Des.*, vol. 13, pp. 271–296, 1999.
[15] G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, and A. Zaliani, "MS-WHIM: New 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids," *J. Comput.-Aided Mol. Des.*, vol. 11, pp. 79–92, 1997.
[16] D. Robert, L. Amat, and R. Carbo-Dorca, "Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures: Prediction of the corticosteroid-binding globulin binding affinity for a steroid family," *J. Chem. Inf. Comput. Sci.*, vol. 39, pp. 333–344, 1999.
[17] V. Uddameri and M. Kuchanur, "Fuzzy QSARs for predicting log K_{oc} of persistent organic pollutants," *Chemosphere*, vol. 54, pp. 771–776, 2004.
[18] M. Kumar, K. Thurow, N. Stoll, and R. Stoll, "Robust fuzzy mappings for QSAR studies," *Eur. J. Med. Chem.*, vol. 42, pp. 675–685, 2007.
[19] C. D. N. Neagu, A. O. Aptula, and G. Gini, "Neural and neuro-fuzzy models of toxic action of phenols," in *Proc. First Int. IEEE Symp. Intell. Syst.*, 2002, vol. 1, pp. 283–288.
[20] Y. P. Zhou, C. B. Cai, S. Huan, J. H. Jiang, H. L. Wu, G. L. Shen, and R. Q. Yu, "QSAR study of angiotensin II antagonists using robust boosting partial least squares regression," *Analytica Chimica Acta*, vol. 593, pp. 68–74, 2007.
[21] Z. Pawlak, *Rough Sets, Theoretical Aspects of Resoning About Data*. Dordrecht, The Netherlands: Kluwer, 1991.
[22] A. Skowron, R. Swiniarski, and P. Synak, "Approximation spaces and information granulation," in *Transactions on Rough Sets III*, vol. 3400, J. F. Peters, A. Skowron, and A. D. Van, Eds. Heidelberg, Germany: Springer-Verlag, 2005, pp. 175–189.
[23] W. Xiaolong, Q. Chen, and D. S. Yeung, "Mining pinyin-to-character conversion rules from large-scale corpus: A rough set approach," *IEEE Trans. Syst., Man Cybern., Part B, Cybern.*, vol. 34, no. 2, pp. 834–844, Apr. 2004.
[24] F. Fernandez-Riverola, F. Diaz, and J. M. Corchado, "Reducing the memory size of a fuzzy case-based reasoning system applying rough set techniques," *IEEE Trans. Syst., Man, Cybern., Part C: Appl. Rev.*, vol. 37, no. 1, pp. 138–146, Jan. 2007.
[25] P. Maji and S. K. Pal, "Rough set based generalized fuzzy C-means algorithm and quantitative indices," *IEEE Trans. Syst., Man Cybern., Part B, Cybern.*, vol. 37, no. 6, pp. 1529–1540, Dec. 2007.
[26] Q. E. Wu, W. Tuo, H. Y. Xuan, and L. J. Sheng, "Topology theory on rough sets," *IEEE Trans. Syst., Man Cybern., Part B, Cybern.*, vol. 38, no. 1, pp. 68–77, Feb. 2008.
[27] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approach," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1457–1471, Dec. 2004.
[28] A. Chouchoulas and Q. Shen, "Rough set-aided keyword reduction for text categorisation," *Appl. Artif. Intell.*, vol. 15, pp. 843–873, 2001.
[29] A. Skowron and C. Rauszer, "The discernibility matrices and functions in information systems," in *Intelligent Decision Support*, R. Slowinski, Ed. Dordrecht, The Netherlands: Kluwer, 1992, pp. 331–362.
[30] W. Ziarko, "Variable precision rough set model," *J. Comput. Syst. Sci.*, vol. 46, pp. 39–59, 1993.
[31] J. Bazan, A. Skowron, and P. Synak, "Dynamic reducts as a tool for extracting laws from decision tables," in *Proc. 8th Symp. Methodologies Intell. Syst.*, 1994, pp. 346–355.
[32] A. Bjorvand and J. Komorowski, "Practical applications of genetic algorithms for efficient reduct computation," in *Proc. 15th IMACS World Congr. Sci. Comput., Model. Appl. Math.*, 1997, vol. 4, pp. 601–606.
[33] D. Ślęzak, "Approximate reducts in decision tables," in *Proc. 6th Int. Conf. Inf. Process. Manage. Uncertainty Knowl.-Based Syst.*, 1996, pp. 1159–1164.
[34] J. Wróblewski, "Finding minimal reducts using genetic algorithms," in *Proc. 2nd Annu. Joint Conf. Inf. Sci.*, 1995, pp. 186–189.
[35] N. Zhong, J. Dong, and S. Ohsuga, "Using rough sets with heuristics for feature selection," *J. Intell. Inf. Syst.*, vol. 16, pp. 199–214, 2001.
[36] P. Maji, "f-information measures for efficient selection of discriminative genes from microarray data," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1063–1069, Apr. 2009.
[37] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[38] H. Chen, J. Zhou, and G. Xie, "PARM: A genetic algorithm to predict bioactivity," *J. Chem. Inf. Comput. Sci.*, vol. 38, pp. 243–250, 1998.
[39] D. Robinson, P. Winn, P. Lyne, and W. Richards, "Self-organizing molecular field analysis: A tool for structure-activity studies," *J. Med. Chem.*, vol. 42, pp. 573–583, 1999.
[40] J. Polanski and B. Walczak, "The comparative molecular surface analysis (COMSA): A novel tool for molecular design," *Comput. Chem.*, vol. 24, pp. 615–625, 2000.
[41] S. S. Liu, C. S. Yin, Z. L. Li, and S. X. Cai, "QSAR study of steroid benchmark and dipeptides based on MEDV-13," *J. Chem. Inf. Comput. Sci.*, vol. 41, pp. 321–329, 2001.
[42] L. Amat, E. Besalu, and R. Carbo-Dorca, "Identification of active molecular sites using quantum-self-similarity matrices," *J. Chem. Inf. Comput. Sci.*, vol. 41, pp. 978–991, 2001.

Pradipta Maji received the B.Sc. (Honors) degree in physics, the M.Sc. degree in electronics science, and the Ph.D. degree in computer science from Jadavpur University, Kolkata, India, in 1998, 2000, and 2005, respectively.

He is currently an Assistant Professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. He is the author or coauthor of around 60 papers published in international journals and conferences. He is also a reviewer of many international journals. His research interests include pattern recognition, computational biology and bioinformatics, medical image processing, cellular automata, soft computing, and so forth.

Dr. Maji was the recipient of the 2006 Best Paper Award of the International Conference on Visual Information Engineering from The Institution of Engineering and Technology, U.K., the 2008 Microsoft Young Faculty Award from the Microsoft Research Laboratory India Pvt., and the 2009 Young Scientist Award from the National Academy of Sciences, India, and has been selected as the 2009 Associate of the Indian Academy of Sciences, India.

Sushmita Paul received the B.Sc. degree in biotechnology from Rajasthan University, Jaipur, India, in 2005, and the M.Sc. degree in bioinformatics from Banasthali Vidyapith, Jaipur, India, in 2007.

She is currently a Research Scholar in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. Her research interests include computational biology and bioinformatics, pattern recognition, soft computing, and so forth.

Ms. Paul was the recipient of the 2009 Best Paper Award of the International Conference on Information Technology from the Orissa Information Technology Society, India.