Chemometrics and Intelligent Laboratory Systems 151 (2016) 71–77

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

A novel local manifold-ranking based K-NN for modeling the regression between bioactivity and molecular descriptors

Liang Shen [a,1], Dongsheng Cao [b,1], Qingsong Xu [c,*], Xin Huang [d], Nan Xiao [c], Yizeng Liang [d]

[a] School of Science, Qingdao University of Technology, Qingdao 266520, PR China
[b] School of Medicine, Central South University, Changsha 410083, PR China
[c] School of Mathematical and Statistics, Central South University, Changsha 410083, PR China
[d] Research Center of Modernization of Traditional Chinese Medicines, Central South University, Changsha 410083, PR China

ARTICLE INFO

Article history:
Received 10 February 2015
Received in revised form 9 December 2015
Accepted 10 December 2015
Available online 21 December 2015

Keywords:
Chemical similarity
Manifold-ranking
k-nearest neighbors (K-NN)
Quantitative structure–activity relationship (QSAR)

ABSTRACT

In the present study, we propose a novel local regression algorithm based on manifold-ranking and *k*-nearest neighbors (MRKNN for short). Under the framework of kernel methods, the group relationship shared among multiple molecules is firstly captured by the graph where nodes represent molecules and edges represent pairwise relations. Then, manifold ranking algorithm is developed for query-oriented extractive summarization, where the influence of query is propagated to other molecules through the structure of the constructed graph. When evaluated on four SAR datasets, MRKNN algorithm can provide a feasible way to exploit the intrinsic structure of similarity relationships. Results have validated the efficacy of the proposed algorithm.

1. Introduction

Quantitative structure–activity relationship (QSAR) technology is capable of modeling and prediction of the relationship between response-variable and molecular predictors. In QSAR modeling, the predictors consist of physico-chemical properties or theoretical molecular descriptors of chemicals, while the response-variable could be a biological activity of the chemicals. Up to now, many modeling methods have been applied to describe the studies of QSAR, including multivariate linear regression (MLR) [1–3], partial least-squares regression (PLS) [4–6], principal component analysis (PCA) [7], neural networks (NN) [8–9], linear discrimination analysis (LDA) [10–13], classification and regression tree (CART) [14], and random forests (RF) [15–16], etc. A common feature underlying models built by these techniques is that the entire training data set is considered for building those models. Such models were usually termed as "global" models. However, if the relationship between biological activity and chemical structures is overly complex, the global methods usually perform bad [17].

As a result, methods which can focus on structure–activity trends that are not necessarily global can be used to build QSAR models. Intuitively, one can develop a local model by considering subsets of a large data set that have similar compounds and then build models on the subsets. Actually, one of the most commonly used local methods in QSAR studying is *k*-nearest neighbors (K-NN) [18]. It is one of the most fundamental and simple classification or regression methods for a QSAR study when there is little or no prior knowledge about the distribution of the data.

Nowadays, similarity queries (retrieval) is a versatile primitive for molecule databases [19]. It plays an important role in applications, such as molecule or gene ranking. In general, a molecule query aims at retrieving similar molecules, which means detecting those molecules that are functionally related to it. The output of a molecule query contains molecules ranked in descending order of their similarity. High-ranked objects are likely to have similar properties to the query, and thus be of interest for property prediction.

The problem that arises here is that the outcome of a query usually returns a vast amount of results. The order of the returned results plays a very important role to the user's needs. For instance, every user would wish the first result also to be the desirable one. Inspired from algorithms used in web search engines, such as the well-known Google's PageRank [20], many methods have been developed that rank results according to their importance [21–22]. The need for ranking exists in biological databases, as well [23–24]. Therefore, given a query molecule, how to get an accurate ranking order and then use these ranking results to modeling has drawn attention to most of us.

In the present study, we construct a novel local algorithm, which obtains a prediction for a query molecule using its local neighborhood given by manifold-ranking approach. This method is termed as MRKNN

* Corresponding author.
E-mail address: qsxu@csu.edu.cn (Q. Xu).
[1] These authors contributed equally to this paper.

http://dx.doi.org/10.1016/j.chemolab.2015.12.005
0169-7439/

method. A significant advantage of MRKNN is that it can avoid the risks caused by the pair-wise distance and provide a feasible way to exploit the intrinsic structure of similarity relationships in SAR datasets. The remainder of this paper is organized as follows. Section 2 presents three commonly used similarity metrics in K-NN as well as manifold-ranking algorithm. In Section 3, we introduce several QSAR datasets and assessment of predictive accuracy. Section 4 shows the results of the comparison among these local algorithms on four datasets. Section 5 gives the conclusions.

2. Methodology

2.1. Similarity metrics

The notion of chemical (molecular) similarity searching is one of the most important concepts in chemoinformatics. It plays an important role in modern approaches to predicting the properties of chemical compounds, designing chemicals with a predefined set of properties and, especially, in conducting drug design studies by screening large databases containing structures of available (or potentially available) chemicals. The base of such relationships is on an assumption that compounds of similar structure will exhibit similar bioactivities or physicochemical properties [19]. That is to say, a data set with very similar chemical structures should give accurate prediction of analogous molecular property when used to establish a QSAR model. A simple strategy hence involves computing the similarity between the known reference structure and each of the molecules in a database, ranking the database molecules in decreasing order of the computed similarities and then carrying out real screening on just the top-ranked database molecules.

Chemical similarity (molecular similarity) is often described as an inverse of a measure of distance in descriptor space. A similarity coefficient can be converted to a distance by taking its "complement" Distance = 1 − Similarity. Three distance-based approaches have been found to be the most useful in QSAR research, namely, the Euclidean, Manhattan and Canberra distance measures [25–26]. The Euclidean distance is the square root of the squared differences between corresponding elements of the rows (or columns) in the distance matrix. The Manhattan distance is the sum of the absolute differences between corresponding elements of the rows (or columns) in the distance matrix. Details of three distanced-based approaches are shown in Table 1. Thus far, as noted in the introduction, we focus here mainly on quantitative characteristics according to the type of molecular representation.

2.2. k-nearest neighbors algorithm

The k-nearest neighbors algorithm (K-NN) is an intuitive method commonly used for classification and regression problems. It predicts objects "values" or class memberships based on the k closest training examples in the feature space. Given a dataset, K-NN works by selecting the k closest samples from a set of well-known classified data (training data) and choosing the class with the most representatives in the set. Generally, the neighborhoods can be selected according to the similarity metrics introduced in Table 1. Formally, the upper limit of k is the total number of compounds in the training dataset ($k < n$). Specifically, the k-nearest neighbors fit for dependent variable $\hat{y}$ is defined as follows:

$$\hat{y}(x) = \sum_{i=1}^{NN_k(x_i)} w_i y_i \tag{1}$$

where $NN_k(x_i)$ is the neighborhood of x_i, w_i represents the weight of x_i.

In general, choosing a proper neighborhood may contribute to reduce the complexity of computation and get a more accurate model. The optimal neighborhood k is often determined by cross-validation. However, although K-NN is valid and convenient for QSAR modeling, it also has some shortcomings. Firstly, it detects the similarity mainly by pairwise distance rather than by structure. Thus, it can not efficiently capture the intrinsic structure hidden in the data set. Secondly, when the structure of data shows complex or overlap, it inevitably results poor prediction.

Assume we want to select the top five most similar data to the query (the red triangle), then Fig. 1(A) is obtained if using Euclidean distance. However, three points of them are not the most similar data to the query in real sense, since they belong to the class of "rectangular". Actually, what we want is to select the five most similar data which belong to the same class as the query (or the same structure to the query), this could be explained in Fig. 1(B). Thus, exploring a proper similarity method can make the model more accurately.

2.3. Manifold-ranking method

Manifold-ranking, originally developed by Zhou et al., is a semi-supervised algorithm [27–28]. It has been used in the field of document retrieval, image processing and face recognition, etc. Specifically, the goal of learning to rank is to derive a ranking function f which can determine relative preference between two data or variables. The detailed description of manifold-ranking is described as follows.

Given a set of points $\boldsymbol{\chi} = \{x_1, \cdots, x_q, x_{q+1}, \cdots, x_n\} \subset \mathbb{R}^m$, the first q points are the queries and the rest are the points which we want to rank according to their relevance to the queries. Let $d: \boldsymbol{\chi} \times \boldsymbol{\chi} \to \mathbb{R}$ denote a metric on $\boldsymbol{\chi}$, such as Euclidean distance, which assigns each pair of points x_i and x_j a distance $d(x_i, x_j)$. Let $f: \boldsymbol{\chi} \to \mathbb{R}$ denote a ranking function which assigns to each point x_i a ranking value f_i. We can view $\mathbf{f}$ as a vector $\mathbf{f} = [f_1, \cdots, f_n]^T$. We also define a vector $\mathbf{y} = [y_1, \cdots, y_n]^T$, where $y_i = 1$ if x_i is a query, and $y_i = 0$ otherwise.

Step 1. Sort the pairwise distance among points in ascending order, then repeat connecting the two points with an edge according the order to get a connected graph.

Step 2. Establish the affinity matrix $\mathbf{W}$ defined by kernel function [29],

$$\mathbf{W} = \left\{ w_{ij} = \langle \phi(x_i), \phi(x_j) \rangle \right\}, i = 1, \cdots, n; j = 1, \cdots, n$$

if there is an edge linking x_i and x_j. Note that $\mathbf{W}_{ii} = 0$ because there are no loops in the graph. Compared with the original algorithm proposed by Zhou, we also use another Laplacian kernel method to construct the affinity matrix among data sets.

Step 3. Symmetrically normalize $\mathbf{W}$ by $\mathbf{K} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ in which $\mathbf{D}$ is the diagonal matrix with (i,i)-element equal to the sum of the i–th row of $\mathbf{W}$.

Step 4. Iterate

$$f(t+1) = \alpha \mathrm{f}(t) + (1-\alpha) y \tag{2}$$

until convergence, where α is a parameter in $[0,1)$.

Step 5. Let f_i^* denote the limit of the sequence $\{f_i(t)\}$, then rank each point x_i according to its ranking scores f_i^* (largest ranked first).

Table 1
Chemical similarity coefficients for molecular descriptors.

Name(s) of similarity metric	For quantitative characteristics
Euclidean distance	$\sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$
Manhattn distance	$\sum_{i=1}^{d}\lvert x_i - y_i\rvert$
Canberra distance	$\sum_{i=1}^{d}\frac{\lvert x_i - y_i\rvert}{\lvert x_i + y_i\rvert}$

Notes: Assume that two molecules, A and B, assume further that $[x_1, x_2, \cdots, x_d]'$ (or $[y_1, y_2, \cdots, y_d]'$) are set to one in the descriptors for A (or B).

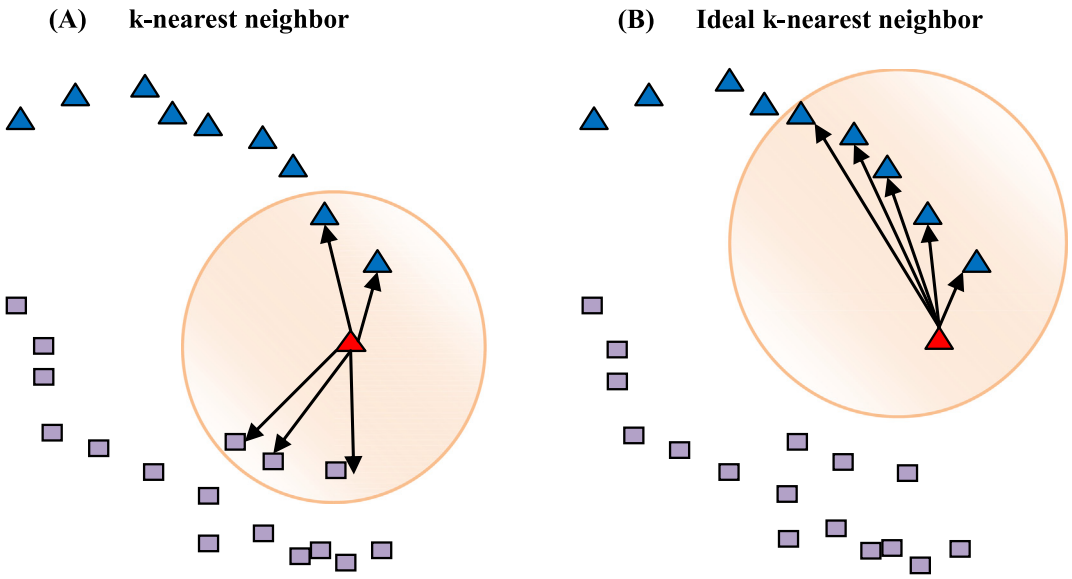

Fig 1. Two patterns: triangle and square. The query sample (red, solid triangle) should be classified by local methods. Assume $k = 5$, Left panel (A): K-NN, Right panel (B): Ideal K-NN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To illustrate the manifold-ranking algorithm, let us consider a spiral dataset generated by two intertwining moons in Fig. 2. Fig. 2(A) represents the original data which belong to two classes (structures). The data of upper moon belong to one class, while the data of lower moon belong to another class. In Fig. 2(B), the blue star represents the query. Then given the query, what we want is to obtain the ranking of similarities for the remaining data. By using manifold ranking algorithm, the ranking results are shown in Fig. 2(B), where the area of the circle represents the size of the similarity to the query. Therefore, we can see that manifold algorithm ranks the neighbors mainly by the structure rather than solely by pair-wise distance.

2.4. Building MRKNN using manifold ranking

In this section, we will discuss our semi-supervised manifold-ranking guided k-nearest neighbors (MRKNN) algorithm. As the structure of data usually shows complex, using traditional approaches can not detect the molecular similarities inherently. With the manifold-ranking algorithm, the rank scores can be propagated among all the data and then the similarity can be found. That is to say, this algorithm ranks the data with respect to the intrinsic cluster structure. Under the framework of kernel methods, we then constructed a novel algorithm which combines manifold ranking with k-nearest neighbors

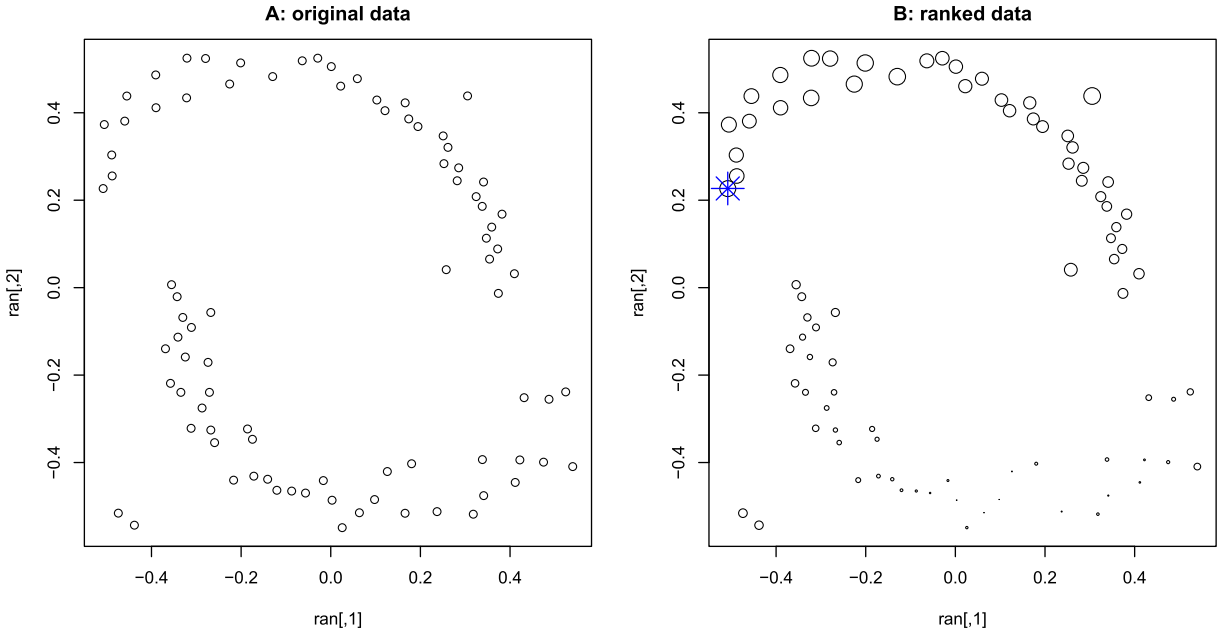

Fig 2. Ranking on the two moons pattern. The blue star point represents the query. (**A**) the original data set with the query; (**B**) the results using manifold-ranking. The area of the circle represents the size of the similarity to the query for each node. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Pseudocode for the MRKNN algorithm.

Input	Data $Z=\{(x_1,y_1),\cdots,(x_n,y_n)\}$

1:	Initialization: $f_1(0)=1; f_i(0)=0$
	for $t=0,1,2,\cdots$do
	for $i=2$ to m do
	$f_i(t+1)\leftarrow K_{1i}+\alpha\sum_{j=2}^{m}K_{ji}f_j(t)$
	end for
	Until convergence:$f_i^*\leftarrow\{f_i(t)\}$

2:	Use cross-validation to select the optimal k-nearest neighbor for KNN after ranking.
3:	Prediction: $\hat{y}_i=\sum_{i=1}^{k}w_i y_{NN(x_i')}$
Output	The neighborhood k, $\hat{y}_i$, RMSECV, Q-squared, $i=1,2,\cdots,n$

(MRKNN for short). Pseudocode of the MRKNN algorithm is provided in Table 2.

As the local MRKNN model has to consider the neighborhood of the queries, then the reliable QSAR predictions are limited generally to the chemicals that are structurally similar to the queries. The chemicals that satisfy the scope of the model are considered as within the Applicability Domain (AD). There are many methods for defining Applicability Domain, such as Range-Based and Geometric Methods, Distanced-Based Methods, Probability Density Distribution-Based Method, etc. [30] In the present paper, the nearest neighbor approach which belongs to the Distance-Based methods is used to define the AD. Within the user defined threshold of nearest neighbors, the query chemical with higher similarity is indicated to have a proper number of training neighbors and therefore, is considered to be reliably predicted. For the effectiveness of computation, we set the threshold of nearest neighbor as 30.

3. Data sets, software and performance evaluation

3.1. QSAR data sets and software

Four datasets were used for this study. According to ref. [31], whole-molecule 3D descriptors such as molecular volume and charged partial surface area (CPSA) descriptors have been calculated for ACE, GPB and THERM data sets. These descriptors are calculated using Gasteiger-Marsili charges implemented in Cerius2 (the Polygraph set) and the CORINA structures [31–33] generated from SMILES strings. Because the charges and structures are determined with a straightforward and unambiguous approach, it is referred to as 2.5D descriptors. The fourth surface tension (ST) data set was extracted from JASPAR (JASPAR 1972).

(1) *ACE*. A set of 114 angiotensin converting enzyme (ACE) inhibitors has been taken from the work of Depriest et al. [34], which describes their use for CoMFA modeling. Activities are spread over a wide range, with pIC_{50} values ranging from 2.1 to 9.9 (μmol/L).
(2) *GPB*. A set of 66 inhibitors of glycogen phosphorylase b (GPB) have pK_i values ranging from 1.3 to 6.8 (μmol/L) [35].
(3) *THERM*. A set of 76 thermolysin inhibitors (THERM) have pK_i values ranging from 0.5 to 10.2 (μmol/L) [36].
(4) *ST*. Surface tension is a property of the surface of a liquid that allows it to resist an external force. The surface tension at 25 °C for 1416 chemicals was obtained from the data compilation of JASPAR [37]. The estimated experimental surface tension value is only used if the closest experimental data point is within 10 °C of 25 °C. The modeled property was the surface tension in dyn/cm.

As many of the raw descriptors contain little information or are correlated with other descriptors, thus, the following step is to reduce the number of descriptors via objective feature selection. This procedure is used to reduce the initial descriptor pool to a more manageable size by using statistical methods which ignore the dependent variable. Objective feature selection involves the use of a correlation testing (where if a given pair of descriptors have a Pearson correlation greater than a user specified cutoff, a random member of the pair is deleted) and an identical testing (where descriptors contain a user specified percentage of identical values are deleted). In this study, the correlation cutoff and identical cutoff were set to 0.75 and 0.75, respectively.

After that, the stage of subjective feature selection was employed to search for optimal descriptor subsets. The genetic algorithm (GA) of optimization routines coupled with the stepwise multiple linear regression was used to find models [38–42]. By this optimization method, a list of top performing descriptor subsets are generated. After that, the optimal subset of variables was an eight-descriptor model, a six-descriptor model, a three-descriptor model and a five-descriptor model for ACE, GPB, THERM and ST, respectively. The descriptors selected by GA-lm for four sets are listed in Table 3.

3.1.1. Software

The algorithm used in the present study, together with other programs, was written in R environment (version 3.1.1), and run on a personal computer (Intel Pentium processors4/2.6GHZ 4.00 GB RAM). The MRKNN was performed with the "kernlab" package. The R scripts used in this study are available upon request.

3.2. Performance evaluation

Validation is a crucial aspect of any quantitative structure–activity relationship (QSAR) modeling. Several validation techniques have been proposed in order to estimate the model prediction capability in chemometrics. Basically, RMSE is a frequently used measure of the standard deviation differences between predicted values and observed values. The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions. Generally, there are two forms of RMSE, including RMSECV and RMSEP. Herein, RMSECV is used to measure the root-mean-square-error of cross validation, while RMSEP is used to measure the root mean square error of independent validation set.

$$RMSECV=\sqrt{\frac{1}{N}\sum_{i=1}^{N_{training}}(y_i-\hat{y}_i)^2} \tag{3}$$

$$RMSEP=\sqrt{\frac{1}{N_{test}}\sum_{i=1}^{N_{test}}(y_i-\hat{y}_i)^2} \tag{4}$$

where y_i, $\hat{y}_i$ and $\overline{y}$ are the observed value, predicted value and the mean value of observed value, respectively. $N_{training}$, N_{test} and N represent the

Table 3
Description of QSAR data sets on 2.5D descriptors used as inputs for models.

	ACE	GPB	THERM	ST
No. of molecules	114	66	76	1416
No. of variables	55	55	55	188
Descriptor selected	S_dO, CHI.V.2, JX, CHI.0, Jurs.FPSA.2, PMI.mag, Fh2o, Shadow.Zlength	S_ssNH, Shadow.XYfra, Kappa-2	S_ssO, IC, Rotlbonds, AlogP98, Jurs.PPSA.1	a_ICM, balabanJ, GCUT_SLOGP_2, GCUT_SMR_1, KierFlex, weinerPath, weinerPol

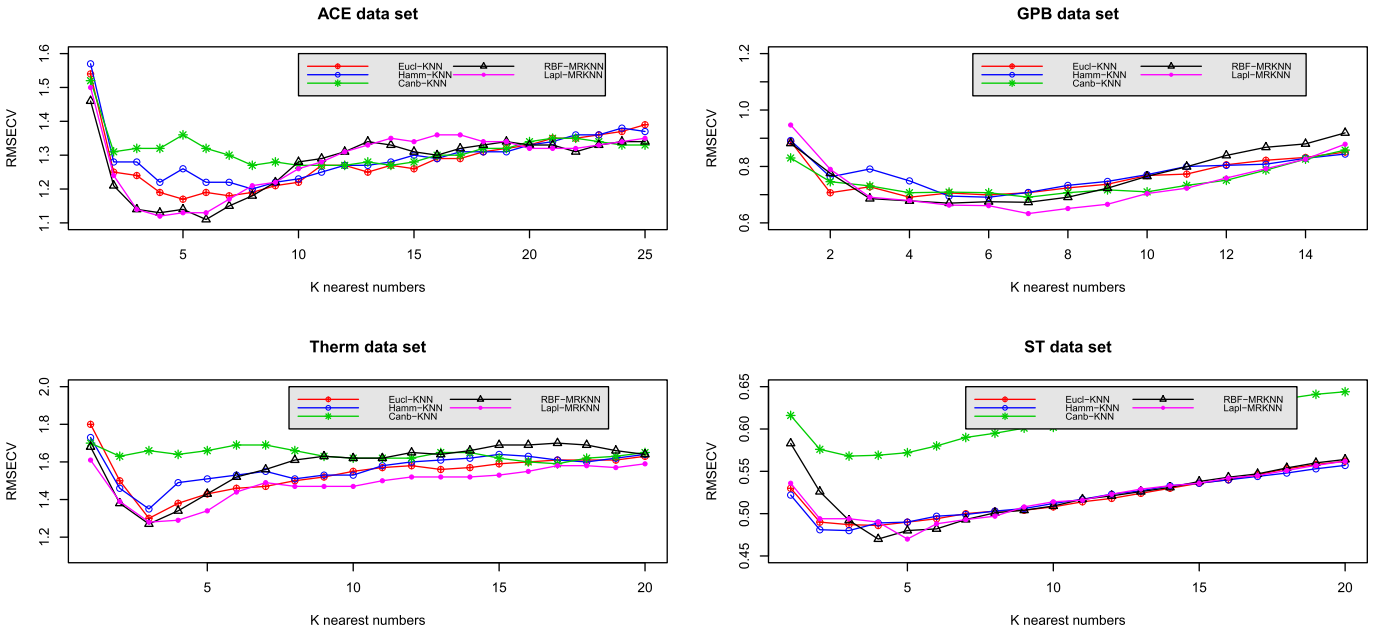

Fig. 3. RMSECV vs. k nearest numbers of the five different models for ACE, GPB, THERM and ST.

number of training samples, the number of testing samples and the whole size of data set, respectively.

For regression problems in chemometrics, another commonly used measure employed to evaluate the model behavior is Q-squared (q^2) [43–46],

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}. \quad (5)$$

Often, a high value of this statistical characteristic ($q^2 > 0.5$) is considered as a proof of the high predictive ability of the model. To assess the effect of outliers on the value of q^2, compounds with residuals more than 3 standard deviations from the average residuals were identified and excluded from the calculation.

4. Results and discussion

To assess the performance of our proposed MRKNN approach in SAR studies, two kernel functions (Gauss Radial Basis and Laplacian) were used to give the model, denoted as RBF-MRKNN and Lapl-MRKNN. As a comparison, Euclidean distance with K-NN (Euc-KNN), Manhattan distance with K-NN (Man-KNN), and Canberra distance with K-NN (Can-KNN) were used as baseline methods to give the prediction. For each data set, we considered three-fold internal cross-validation for the whole data set as well as the case independent validation set. Each predictor of four datasets is scaled to have zero mean and unit variance.

To increase the accuracy and relevance of the similarity measure, we used weighted mean in formula $\hat{y}_i = \sum_{i=1}^{k} w_i y_{NN(x_i')}$ of Table 2. In this formula, w_i represents the ranking score of x_i in MRKNN-related method; while in KNN-related method, w_i represents the inverse of the distance between the query and the sample x_i.

4.1. Internal validation results for three QSAR datasets

Cross-validation (CV), the most commonly used method for internal validation, is a statistical technique in which different proportions of chemicals are iteratively held-out from the training set used for model development and "predicted" as new by the developed model in order to verify internal "predictivity". For three-fold cross-validation, the original whole dataset is randomly partitioned into three roughly equal-sized parts. Of the three parts, the two parts are used as training data to fit the model, and the remaining single part is retained as the validation data for testing the model. The cross-validation process is repeated three times so that every part can be predicted as a validation set. The composition of the sets is summarized in Table 3. Herein, the iteration times t in Table 2 is set to 400 in order to reach convergence.

For Euc-KNN, Man-KNN and Can-KNN, the optimal parameter of nearest neighbor k was determined by cross validation and grid-based search. For RBF-MRKNN and Lapl-MRKNN, the parameter of nearest neighbor k and the width of the Radial basis and Laplacian kernel function δ were also determined by cross validation and grid-based search.

The trends of RMSECV with the increasing of nearest neighborhoods for each data set are shown in Fig. 3. In general, RMSECV reaches a first minimum, then rises again with increasing model complexity (nearest

Table 4
Predictive results of four datasets by 3-fold cross-validation.

	Euc-KNN	Man-KNN	Can-KNN	RBF-MRKNN	Lapl-MRKNN
ACE					
param (k/δ)	5	8	12	6/0.01	4/0.007
q^2_{CV}	0.73	0.72	0.69	0.77	0.76
RMSECV	1.17	1.20	1.27	1.11	1.12
GPB					
param (k/δ)	4	6	7	5/0.0025	7/0.003
q^2_{CV}	0.57	0.57	0.57	0.60	0.64
RMSECV	0.69	0.70	0.69	0.67	0.63
THERM					
param (k/δ)	3	3	17	3/0.01	3/0.01
q^2_{CV}	0.49	0.45	0.33	0.51	0.50
RMSECV	1.30	1.35	1.59	1.27	1.28
ST					
param (k/δ)	4	3	3	4/1	5/1
q^2_{CV}	0.76	0.76	0.67	0.78	0.78
RMSECV	0.48	0.48	0.56	0.47	0.47

For RBF-MRKNN and Lapl-MRKNN, the number of nearest neighbors is followed by the width of RBF or Laplacian kernel function.

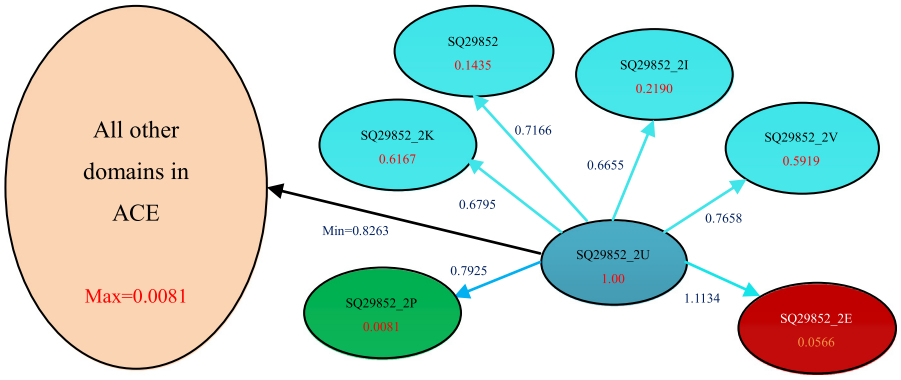

Fig. 4. Visualization of part of the similarity network. Shown is a small part of the molecular similarity network, where the navy blue node SQ29852_2U is the query, and the domains which are directed by light blue edges from the query are its homologs. Note that the red scores inside the nodes are the manifold ranking activation values, while the edges from the query to the nodes are labeled with Euclidean distance in KNN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

numbers). The number of neighbors with the lowest RMSECV was selected as the optimal complexity. Statistics are reported for the most predictive combination of parameters and model complexity in Table 4.

For the ACE set, RBF-MRKNN minimizes RMSECV at six nearest neighborhoods ($q^2_{CV} = 0.77$) while Lapl-MRKNN minimizes RMSECV at four nearest neighborhoods ($q^2_{CV} = 0.76$). Both MRKNN methods perform better than K-NN related methods. For the GPB set, Lapl-MRKNN yields the best results at seven nearest neighborhoods, with RBF-MRKNN only slightly less predictive. The other three pair-wise distance-based KNN methods perform no better than RBF-MRKNN and Lapl-MRKNN. For the THERM, Euc-KNN performs the best in the traditional distance-based KNN methods, but still less predictive than other two MRKNN related methods. For the ST set, both RBF-MRKNN and Lapl-MRKNN perform better than the baseline methods. On the whole, the predictive performance by MRKNN related methods is better or comparative compared with the K-NN related methods. These results have validated the efficacy of the proposed algorithm.

In the following, an example was taken to illustrate why this might be the case. Two similarity methods, the Euclidean distance and manifold-ranking with RBF kernel, were used to give the similarity ranking for phosphate in ACE inhibitors, respectively. The results of ranking by two methods are shown in Fig. 4. Shown is a small part of the molecular similarity network, where the (navy blue) node SQ29852_2U is the query, and the domains which are directed by (light blue) edges from the query are its homologs. The large (pink) node at the left represents all other domains. Here, the top five most similar molecules to the query were selected by these two methods. According to the query, RBF-MRKNN selected SQ29852_2K, SQ29852_2I, SQ29852_2V, SQ29852 and SQ29852_2P to give the model, while Euc-KNN selected SQ29852_2I, SQ29852_2K, SQ29852, SQ29852_2P and SQ29852_2E to give the model. Note that SQ29852_2V is assigned a higher score by Euclidean than SQ29852_2I by manifold-ranking, even though the Euclidean values assigned indicate the opposite between them. Thus, together with the performance in Table 4, we can conclude that MRKNN can efficiently and directly measure the similarities of molecules by a series of local information hidden in the molecules.

Table 5
Predictive results of three datasets by independent validation set.

Methods	ACE		GPB		THERM		ST	
	RMSEP	q^2	RMSEP	q^2	RMSEP	q^2	RMSEP	q^2
Euc-KNN	1.08	0.71	0.53	0.65	1.4	0.44	0.49	0.75
Man-KNN	1.17	0.65	0.55	0.64	1.46	0.43	0.48	0.77
Can-KNN	1.06	0.72	0.63	0.64	1.67	0.31	0.58	0.65
RBF-MRKNN	1.05	0.74	0.52	0.66	1.28	0.51	0.47	0.77
Lapl-MRKNN	1.07	0.72	0.51	0.67	1.29	0.50	0.47	0.77

4.2. External validation results for three QSAR datasets

Although the above discussion indicates that MRKNN algorithm can improve the prediction capability of regression, we would also wish the model to be validated by some other new molecules. Thus, we partitioned the dataset randomly into the training set of 80% and the independent test set of 20%, where the training set is used for selecting the optimal parameter values of the model, and the independent (external) validation set is merely used for evaluating the performance of the model. Table 5 lists the prediction results of five methods on independent validation set. One can clearly see that RBF-MRKNN and Lapl-MRKNN obtain satisfactory predictive results, and the results from the independent validation set are very close to those by three-fold cross-validation. Together with the results of internal cross-validation in Table 4, the results by external validation set indicate that MRKNN algorithm could be considered novel and competitive.

5. Conclusions

In this work, we aim at constructing a novel local algorithm, which combines the manifold-ranking with *k*-nearest neighbors (MRKNN). Then, under the framework of kernel methods, MRKNN can efficiently model the relationship between molecular structures and bioactivities of compounds. The critical innovation that led to the success of MRKNN is its ability to exploit inherent similarity structure for the given queries. Four QSAR datasets collectively demonstrated the predictive ability. Our proposed MRKNN algorithm can be regarded as a novel and promising modeling technique for QSAR problems.

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China (Grants no. 11271374), Shandong Provincial Natural Science Foundation of China (Grants no. ZR2015AQ009), National Bureau of Statistcs of the People's Republic of China (Grant no. 2015LY79), the Mathematics and Interdisciplinary Sciences Project and the Innovation Program of Central South University (Grants no. 90700-505019112).

References

[1] J.E. Pedhazur (Ed.), Multiple Regression in Behavioral Research: Explanation and Prediction, second ed.Holt, Rinehart and Winston, New York, 1982.
[2] H. Goldstein, Multilevel mixed linear model analysis using iterative generalized least squares, Biometrika 73 (1) (1986) 43–56.
[3] T.L. Lai, H. Robbins, C.Z. Wei, Strong consistency of least squares estimates in multiple regression, Proc. Natl. Acad. Sci. 75 (7) (1978) 3034–3036.
[4] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2) (2001) 109–130.

[5] S. Rannar, F. Lindgren, P. Geladi, S. Wold, A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: theory and algorithm, J. Chemom. 8 (2) (1994) 111–125.
[6] H. Abdi, Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 97.
[7] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52.
[8] H.K.D.H. Bhadeshia, Neural networks in materials science, ISIJ Int. 39 (1999) 966–979.
[9] M. Egmont-Petersen, D. de Ridder, H. Handels, Image processing with neural networks—a review, Pattern Recogn. 35 (2002) 2279–2301.
[10] A. Speck-Planche, V.V. Kleandrova, F. Luan, M.N.D.S. Cordeiro, A ligand-based approach for the in silico discovery of multi-target inhibitors for proteins associated with HIV infection, Mol. BioSyst. 8 (2012) 2188–2196.
[11] F. Luan, M.N.D.S. Cordeiro, N. Alonso, X. Garcia-Mera, O. Caamano, F.J. Romero-Duran, M. Yanez, H. Gonzalez-Diaz, TOPS-MODE model of multiplexing neuroprotective effects of drugs and experimental-theoretic study of new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases, Bioorg. Med. Chem. 21 (2013) 1870–1879.
[12] A.S. Planche, V.V. Kleandrova, M.N.D.S. Cordeiro, Chemoinformatics for rational discovery of safe antibacterial drugs: simultaneous predictions of biological activity against streptococci and toxicological profiles in laboratory animals, Eur. J. Pharm. Sci. 48 (2013) 812–818.
[13] N. Alonso, O. Caamano, F.J. Romero-Duran, F. Luan, M.N.D.S. Cordeiro, M. Yanez, H. Gonzalez-Diaz, X. Garcia-Mera, Model for high-throughput screening of multitarget drugs in chemical neurosciences: synthesis, assay, and theoretic study of rasagiline carbamates, ACS Chem. Neurosci. 4 (10) (2013) 1393–1403.
[14] D.S. Cao, Q.S. Xu, Y.Z. Liang, X.A. Chen, H.D. Li, Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity, Chemometr. Intell. Lab. Syst. 103 (2010) 129–136.
[15] M.A. Aizerman, E.A. Braverman, L. Rozonoer, Theoretical foundations of the potential function method in pattern recognition learning, Autom. Remote. Control. 25 (1964) 821–837.
[16] L. Breiman, Random For. Mach. Learn. 45 (1) (2001) 5–32.
[17] R. Guha, D. Dutta, P.C. Jurs, T. Chen, Local lazy regression: making use of the neighborhood to improve QSAR predictions, J. Chem. Inf. Model. 46 (2006) 1836–1847.
[18] L.E. Peterson, K-nearest neighbor, Scholarpedia 4 (2) (2009) 1883.
[19] M.A. Johnson, G.M. Maggiora, Concepts and Applications of Molecular Similarity, John Willey & Sons, New York, 1990.
[20] S. Brin, L. Page, Comput. Netw. ISDN Syst. 30 (1998) 107.
[21] W. Wang, S.J. Li, J.W. Li, W.J. Li, F.R. Wei, Exploring hypergraph-based semi-supervised ranking for query-oriented summarization, Inf. Sci. 237 (2013) 271–286.
[22] J. Yang, B. Xu, B.B. Lin, X.F. He, Multi-query parallel field ranking for image retrieval, Neurocomputing 135 (2014) 192–202.
[23] J. Weston, C. Leslie, E.e.D. Zhou, A. Elisseeff, W.S. Noble, Semi-supervised protein classification using cluster kernels, Bioinformatics 21 (15) (2005) 3241–3247.
[24] J. Weston, A. Elisseeff, D. Zhou, C. Leslie, W.S. Noble, Protein ranking: from local to global structure in the protein similarity network, Proc. Natl. Acad. Sci. (PNAS) 101 (17) (2004) 6559–6563.
[25] N. Nikolova, J. Jaworska, Approaches to measure chemical similarity—a review, QSAR Comb. Sci. 22 (2003) 106–1026.
[26] M. Kokare, B.N. Chatterji, P.K. Biswas, Comparison of similarity metrics for texture image retrieval, Proceedings of the IEEE Region 10 Annual International Conference, 3 2003, pp. 571–575.
[27] D. Zhou, J. Weston, A. Gretton, O. Bousquet, B. Schölkopf, Ranking on data manifolds, Adv. Neural Inf. Process. Syst. 16 (2004) 169–176.
[28] D. Zhou, B. Schölkopf, Learning from labeled and unlabeled data using random walks, Pattern Recognition: Proceedings of the 26th DAGM Symposium 2004, pp. 237–244.
[29] T.M. Huang, V. Kecman, I. Kopriva, Kernel Based Algorithms for Mining Huge Data Sets, in Supervised, Semi-Supervised, and Unsupervised Learning, Springer-Verlag, Berlin, 2006.
[30] S. Faizan, M. Kamel, B. Davide, M. Andrea, C. Viviana, T. Roberto, Comparison of different approaches to define the applicability domain of QSAR models, Molecules 17 (2012) 4791–4810.
[31] D.T. Stanton, P.C. Jurs, Development and use of charged partial surface-area structural descriptors in computer-assisted quantitative structure property relationship studies, Anal. Chem. 62 (21) (1990) 2323–2329.
[32] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges, Tetrahedron 36 (1980) 3219–3288.
[33] J. Gasteiger, C. Rudolph, J. Sadowski, Automatic generation of 3D-atomic coordinates for organic molecules, Tetrahedron Comput. Methodol. 3 (1990) 537–547.
[34] S.A. Depriest, D. Mayer, C.B. Naylor, G.R. Marshall, 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determined active site geometries, J. Am. Chem. Soc. 115 (1993) 5372–5384.
[35] H. Gohlke, G. Klebe, DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein, J. Med. Chem. 45 (2002) 4153–4170.
[36] G. Klebe, U. Abraham, T. Mietzner, Molecular similarity indexes in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological-activity, J. Med. Chem. 37 (1994) 4130–4146.
[37] J.J. Jasper, The surface tension of pure liquid compounds, J. Phys. Chem. Ref. Datas 1 (1972) 841–1009.
[38] J.J. Sutherland, L.A. O'Brien, D.F. Weaver, A comparison of methods for modeling quantitative structure—activity relationships, J. Med. Chem. 47 (22) (2004) 5541–5554.
[39] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, V. Consonni, Quantitative structure–activity relationship models for ready biodegradability of chemicals, J. Chem. Inf. Model. 53 (4) (2013) 867–878.
[40] R. Guha, 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. 44 (6) (2004) 2179–2189.
[41] S.K. Jain, A.K. Yadav, P. Nayak, 3D QSAR analysis on oxadiazole derivatives as anticancer agents, Int. J. Pharm. Sci. Drug Res. 3 (3) (2011) 230–235.
[42] D. Broadhurst, R. Goodacre, A. Jones, . Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression,with applications to pyrolysis mass spectrometry, Anal. Chim. Acta 348 (1997) 71–86.
[43] Q.S. Xu, Y.Z. Liang, H.L. Shen, Generalized PLS regression, J. Chemom. 15 (2001) 135–148.
[44] X. Huang, Q.S. Xu, Y.Z. Liang, PLS regression based on sure independence screening for multivariate calibration, Anal. Methods 4 (9) (2012) 2815–2821.
[45] A. Golbraikh, A. Tropsha, Beware of q^2! J. Mol. Graph. Model. 20 (4) (2002) 269–276.
[46] D. Coomans, C. Smyth, I. Lee, T. Hancock, J. Yang, Unsupervised data mining: introduction, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier, Oxford, UK 2009, pp. 559–576.