JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

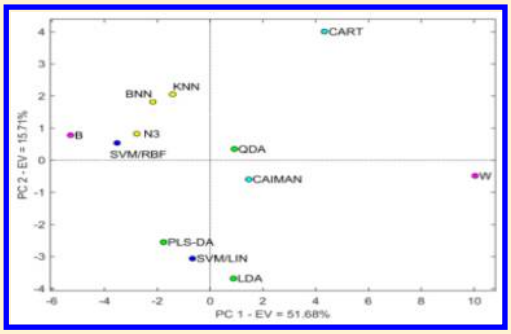
N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers

Roberto Todeschini,* Davide Ballabio, Matteo Cassotti, and Viviana Consonni

Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.zza della Scienza, 1, 20126 Milan, Italy

Supporting Information

ABSTRACT: Two novel classification methods, called N3 (N-nearest neighbors) and BNN (binned nearest neighbors), are proposed. Both methods are inspired by the principles of the K-nearest neighbors (KNN) method, being both based on object pairwise similarities. Their performance was evaluated in comparison with nine well-known classification methods. In order to obtain reliable statistics, several comparisons were performed using 32 different literature data sets, which differ for number of objects, variables and classes. Results highlighted that N3 on average behaves as the most efficient classification method with similar performance to support vector machine based on radial basis function kernel (SVM/RBF). The method BNN showed on average higher performance than the classical K-nearest neighbors method.



1. INTRODUCTION

A large number of data sets consist of input variables (predictors) and one or more categorical variables (classes). The identification of functional relationships between predictors and categorical response is the aim of the so-called supervised pattern recognition methods (or simply, classification methods). These methods are applied in a number of scientific fields such as analytical chemistry, food chemistry, toxicology, QSAR/QSPR, image analysis, process and environmental monitoring, social, medical and economic sciences.

Given a set of training data that belong to G classes, classification methods address the problem of assigning a new object to one of the G classes on the basis of a classification rule, which has been inferred from the training data whose memberships of the G classes are known.[1]

Existing classification algorithms are based on a variety of approaches that confer on them different characteristics and properties.[2−11] Classification methods can be characterized as (a) local or global, if only few or all training objects are considered for class assignment; (b) class-modeling, when a delimited space for each class is defined and objects falling outside are not classified; (c) parametric or nonparametric, if they derive an analytical function; (d) distance-based, when distances between objects are used to classify; (e) linear or nonlinear, on the basis of how class boundaries are estimated; (f) probabilistic if based on estimations of probability distributions.

Despite this varied panorama, the question of what is the best classification method is unanswered. In fact, a more appropriate interrogation would ask what is the best classification method for a particular problem.[9] Among classifiers, the K-nearest neighbor (KNN)[2] is one of most simple and common algorithms: an object is classified according to the classes of the k nearest objects, i.e. it is classified according to the majority class of its K-nearest neighbors in the data space. Therefore, from the computational point of view, one only needs to calculate and analyze the distances between all the possible pairs of objects. KNN has other advantages: it does not assume a form of the underlying probability density function and can handle multiclass problems. KNN has been suggested as a standard comparative method for more sophisticated classification techniques.[2] On the other hand, it is very sensitive to the choice of distance measure and data scaling procedure.[12]

A number of variants of the KNN method were proposed in literature. Some variants are devoted to reduce the computing time[13] by representing each class by its average and this corresponds to the well-known method called "nearest mean classifier", or representing each class with a group of n averages calculated by means of clustering procedures. However, the local characteristics (object distribution and shape) of the data are lost and results are worse than those obtained by the original version of KNN. Other approaches are aimed at assigning the class to the target object by using not the majority criterion but weighting the k neighbors by their similarity/diversity measures or by other functions.[3,14] Other KNN variants are based on algorithm for finding a set of k variable neighbors on the training set or by searching the optimal set of weights for the variables; in both cases, a large number of parameters should be optimized.[15] Finally, other variants attempt to transform KNN into a modeling classifier by selecting a suitable threshold on the similarity/diversity

Received: May 28, 2015
Published: October 19, 2015

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374

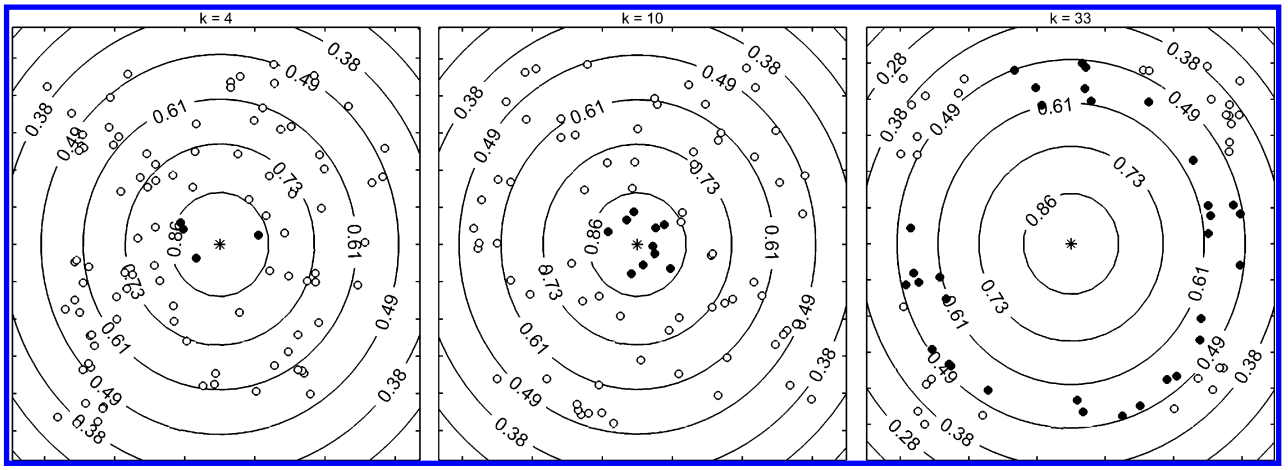


Figure 1. BNN method applied to three different targets (marked by * in the center of the plots). The first nonempty similarity bin 1.0−0.86 contains 4 objects (black circles) in the left plot and 10 objects in the central plot; for the third target, 33 objects are selected in the fourth similarity bin 0.61−0.49.

measure in order to detect outliers or objects outside the applicability domain.[16]

In this paper two new classification methods are proposed, namely the N-nearest neighbors (N3) and the binned nearest neighbors (BNN). These methods are based on a similarity approach like the K-nearest neighbors method, but based on different neighbors identification approaches. Their predictive performance in classification was compared with those of nine well-known linear and nonlinear classification algorithms by means of 32 literature data sets.

2. THEORY

Two new classification methods N3 and BNN are described in the following sections.

N3 Classifier. The N3 (N-nearest neighbors) method is based on the principle of the KNN algorithm which considers only local information to perform the classification of each object.

Unlike KNN, which searches for the best number (k) of neighbors to be considered, N3 method takes into account all the $n-1$ objects to classify the ith object. The $n-1$ neighbors are sorted from the most similar to the least similar to the target and the corresponding similarity rank vector is obtained; then, the neighbor contributions to class assignment exponentially decrease as the similarities diminish, since they are weighted by the rank r, whose role is modulated by an α exponent to be optimized.

For each ith object, the gth class weight is calculated on the basis of the following function:

$$w_{ig} = \frac{1}{\hat{n}_g} \cdot \sum_{\substack{j=1 \\ j \neq i}}^{n-1} \frac{s_{ij}}{r_{ij}^{\alpha}} \cdot \delta_j \qquad \delta_j = \begin{cases} 1 & \text{if } c_j = g \wedge \dfrac{s_{ij}}{r_{ij}^{\alpha}} > \varepsilon \\ 0 & \text{otherwise} \end{cases}$$
$$\hat{n}_g = \sum_j \delta_j \tag{1}$$

where s_{ij} is the similarity between the ith and jth object, r_{ij} is the similarity rank of the jth object with respect to the ith object, calculated as tied-rank; α is a real-valued parameter to be optimized in the range $[0.1, 2.5]$; δ_j is the Dirac delta that equals 1 if the jth object belongs to the gth class and its contribution (the numerator) to the class weight is greater than ε (e.g., 10^{-7}); c_j is the class of the jth object and $\hat{n}_g$ is the number of neighbors that contribute to the class weight.

Similarity s_{ij} between the ith and jth objects is here calculated from the distance measure d_{ij} as

$$s_{ij} = \frac{1}{1 + d_{ij}} \tag{2}$$

to ensure that similarity values range in-between 0 and 1. If the distance measure is already bounded in the range $[0,1]$, then the similarity can be calculated as

$$s_{ij} = 1 - d_{ij} \tag{3}$$

The weights are finally normalized according to the following:

$$w'_{ig} = \frac{w_{ig}}{\sum_g w_{ig}} \tag{4}$$

and the ith object is assigned to the class having the maximum w'_{ig} value. These normalized weights can be interpreted as fuzzy measures of the class membership of each object.

The optimal α value is searched for as the value giving the lowest classification error by a validation protocol.

The ranks used in the calculation of the class weights are a sequence of natural numbers, which makes the N3 function having some analogy with other series that constitute the basic principles of the theory of numbers. Among these series, there is the harmonic series, defined as

$$H_n = \sum_{n=1}^{\infty} \frac{1}{n} \qquad n \in \mathbb{N} \tag{5}$$

where n are the natural numbers, which is divergent, while the Riemann zeta function, defined as

$$\zeta(\alpha) = \sum_{n=1}^{\infty} \frac{1}{n^{\alpha}} \qquad n \in \mathbb{N} \tag{6}$$

converges for $\alpha > 1$.

Both series are special cases of the Dirichlet series, which is defined as

$$f(\alpha) = \sum_{n=1}^{\infty} \frac{g_n}{n^{\alpha}} \qquad n \in \mathbb{N} \tag{7}$$

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374

Table 1. Similarity Values of the 17 Bins for Some Selected α Values

bin no.	similarity vector	exponent α						
(m)	(S_m)	0.1	0.3	0.5	0.7	0.9	1.2	1.5
bin 1	1.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
bin 2	0.9	0.9895	0.9689	0.9487	0.9289	0.9095	0.8812	0.8538
bin 3	0.8	0.9779	0.9352	0.8944	0.8554	0.8181	0.7651	0.7155
bin 4	0.7	0.9650	0.8985	0.8367	0.7791	0.7254	0.6518	0.5857
bin 5	0.6	0.9502	0.8579	0.7746	0.6994	0.6314	0.5417	0.4648
bin 6	0.5	0.9330	0.8123	0.7071	0.6156	0.5359	0.4353	0.3536
bin 7	0.4	0.9124	0.7597	0.6325	0.5266	0.4384	0.3330	0.2530
bin 8	0.3	0.8866	0.6968	0.5477	0.4305	0.3384	0.2358	0.1643
bin 9	0.2	0.8513	0.6170	0.4472	0.3241	0.2349	0.1450	0.0894
bin 10	0.1	0.7943	0.5012	0.3162	0.1995	0.1259	0.0631	0.0316
bin 11	10^{-2}	0.6310	0.2512	0.1000	0.0398	0.0158	0.0040	0.0010
bin 12	10^{-3}	0.5012	0.1259	0.0316	0.0079	0.0020	0.0003	0.0000
bin 13	10^{-4}	0.3981	0.0631	0.0100	0.0016	0.0003	0.0000	0.0000
bin 14	10^{-5}	0.3162	0.0316	0.0032	0.0003	0.0000	0.0000	0.0000
bin 15	10^{-6}	0.2512	0.0158	0.0010	0.0001	0.0000	0.0000	0.0000
bin 16	10^{-7}	0.1995	0.0079	0.0003	0.0000	0.0000	0.0000	0.0000
bin 17	10^{-8}	0.1585	0.0040	0.0001	0.0000	0.0000	0.0000	0.0000

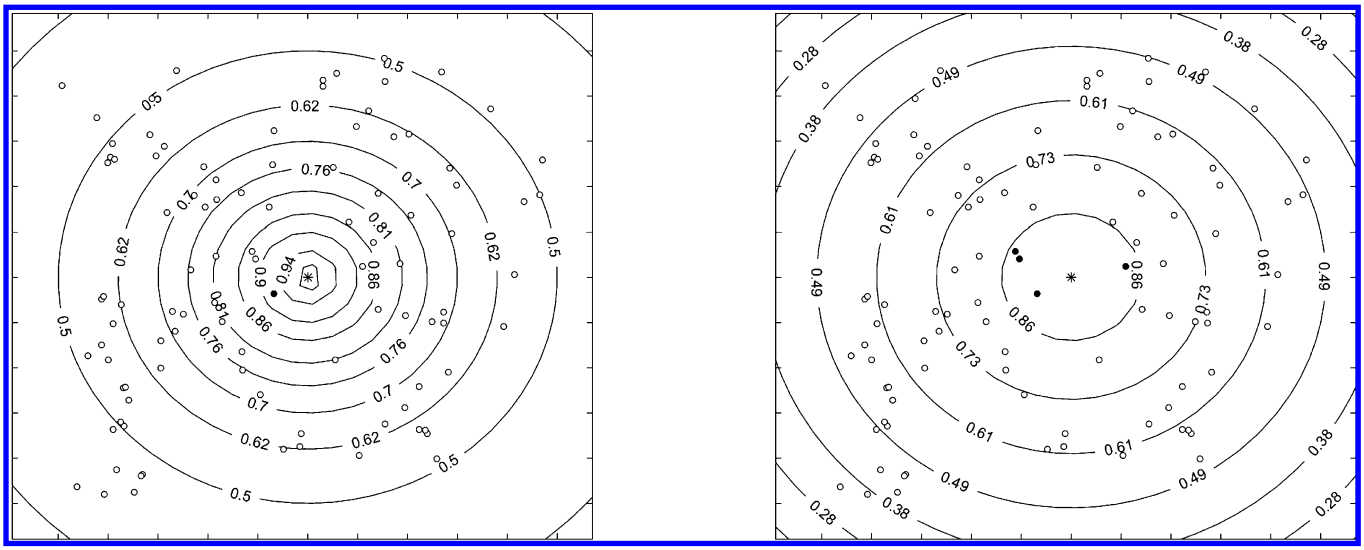


Figure 2. Binning sequence for $\alpha = 0.5$ (left side) and $\alpha = 1.4$ (right side). In the first case, the first nonempty similarity bin is (0.94−0.90) with a width of 0.04 containing one object; in the second case, the first nonempty similarity bin is (1.00−0.86) with a width of 0.14 containing 4 objects.

where g_n is a function. This series converges for $\alpha > 0$ if the function g_n tends to zero, i.e. $g_{n\to\infty} \to 0$; this is the case of the similarity function, $s_{ij\to\infty} = 0$. Hence, the strong analogy with the N3 function can be easily noted.

BNN Classifier. Together with N3, another classification method is here proposed and compared with the traditional classifiers.

The binned nearest neighbors (BNN) method predicts the target response by means of a variable number k of neighbors according to the criterion of majority vote. The main idea is to consider for prediction all those neighbors that have the largest and comparable similarity to the target. To select the most similar neighbors, similarity intervals (i.e., bins) are predefined and the neighbors are distributed into these intervals according to their similarity to the target. All the neighbors falling into the bin with the largest similarity are considered for prediction (see Figure 1).

The BNN method implies first a binning procedure of the similarity measure as defined in the following.

Similarity bins are determined by optimization of a parameter α, which defines the bin width:

$$\text{bin}_m^\alpha = [S_m^\alpha, S_{m+1}^\alpha] \qquad m = 1, 2, 3, \ldots \tag{8}$$

where S are different similarity values defined in decreasing order with step of 0.1 in the range $[0.1, 1]$ and with step of 10^{-1} (e.g., 0.01, 0.001, etc.) in the range $[0, 0.1]$. The last seven bins were added to avoid a too large last bin, when dealing with small values of α (e.g., $\alpha = 0.1$, Bin 10 = 0.7943; see Table 1), and thus to avoid the risk that all the $n - 1$ objects fall together in the last bin.

The optimal exponent α used to define the bin thresholds is searched for in the range $[0.1, 1.5]$ with a step of 0.05 and is selected as the value giving the lowest classification error by a validation protocol. Table 1 shows the similarity values for the different bins and selected α values.

It can be derived that when α is in the range between 0 and 1, a concave distribution of bins is obtained and bins have increasing width (i.e., the first bins corresponding to the largest

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374

similarity values are narrower), while when α is greater than 1, a convex distribution is obtained and bins have decreasing width (i.e., the first bins are wider). In Figure 2, two examples are shown for $\alpha = 0.5$ and $\alpha = 1.4$.

Once bins have been defined, the BNN algorithm classifies objects on the basis of the following scheme:

1. similarity to the target is calculated for each object;
2. objects are distributed into the similarity bins according to their similarity to the target;
3. only objects in the first nonempty bin are selected as the nearest neighbors to be used for prediction;
4. prediction is taken as majority vote; when more classes share the same number of neighbors, the target object is assigned to the class having the maximum sum of the similarity contributions.

It can be observed that both N3 and BNN can be also used as modeling classifiers if a threshold on the minimum similarity is settled up, i.e. the similarity of the closest object to the target must be larger than a predefined threshold in order to get a reliable prediction. However, this approach was not investigated in this paper.

Classification Methods for Comparison. N3 and BNN were compared with the following nine linear and nonlinear classification methods.

KNN[2] is the classical method based on local similarities; it was implemented on the average Euclidean distance to measure the dissimilarities between objects and the optimal k was selected in the range [1, 10] by means of a validation procedure. Among the different weighted variants of KNN, the exponentially weighted KNN[3] (wKNN) was considered; it selects the optimal number k of neighbors to be considered for classification and neighbor contributions to class assignment exponentially decrease as the distances to the target increase; then, the membership function for the gth class is calculated as follows:

$$w_{ig} = \frac{\sum_{j=1}^{k} \delta_j \cdot e^{-d_{ij}}}{\sum_{j=1}^{k} e^{-d_{ij}}} \qquad \delta_j = \begin{cases} 1 & \text{if } c_j = g \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where d_{ij} is the Euclidean distance of the jth neighbor from the ith target and c_j is the true class of the jth object.

Partial least squares−discriminant analysis (PLSDA[7]) was applied by searching the optimal number of latent variables by means of a validation procedure. When dealing with PLSDA, class thresholds were calculated by means of Bayes Theorem.[17] It derives that objects could be not assigned when dealing with classification problems based on more than two classes.

Support vector machine (SVM[8]) with both linear kernel (LIN) and nonlinear kernel (radial basis function, RBF) were applied. The optimal c and γ parameters were searched for again by means of a validation procedure.

CAIMAN[5] is a nonlinear classification method; it requires the optimization of a α parameter, which was carried out by means of a cross-validation protocol.

Finally, linear and quadratic discriminant analysis (LDA and QDA[9]) and classification trees (CART[6]) were applied for comparison with the proposed methods.

Model Evaluation and Validation. Classification algorithms were compared on the basis of their predictive classification performance, which was evaluated by means of nonerror rate (NER%), that is, the average of class sensitivities:

$$\text{NER\%} = 100 \cdot \frac{\sum_{g=1}^{G} \frac{n'_g}{n_g}}{G} \tag{10}$$

where G is the number of classes, n'_g is the number of objects correctly classified in the gth class and n_g the total number of objects actually belonging to the gth class.

In order to maximize the comparability among methods, the leave-one out (LOO) validation procedure was adopted. To further evaluate the method performance, a test set was selected for data sets with more than 150 objects. In these cases, objects were randomly divided into training and test sets, containing 80% and 20% of the total number of objects, respectively. The selection was performed maintaining the class proportions, that is, the number of test objects of each class was proportional to the number of training objects of that class. The training set was used to optimize and build the classification models, while objects of the test set were used just to evaluate the predictive ability of the trained models.

Comparison of Classification Methods. Two graphical tools were used to analyze and compare the classification results of all the tested classification methods achieved on all the data sets: principal component analysis (PCA) and minimum spanning tree (MST). Two theoretical classification methods were added to highlight the quality of the methods under comparison: the methods BEST (B) and WORST (W), which take the best and worst nonerror rate values provided by the 11 classification methods on each data set, respectively.

PCA was performed on the data matrix including the NER% values obtained by the 11 classification methods plus the BEST and WORST methods (13 rows) on the 32 data sets (columns); MST was derived from the distance matrix collecting the pairwise Euclidean distances between methods calculated on NER% values of the 32 data sets.

3. DATA SETS

In order to compare the different classification methods, 32 multivariate data sets were considered. The data sets have a different number of classes and objects, as well as different partitions of the objects into the classes. Data characteristics and literature references are collected in Table 2. These are classical benchmarks data sets taken from literature and from very different scientific fields. Among them, BIODEG is a chemoinformatic data set including 12 selected molecular descriptors. Indeed, for those chemoinformatic data sets that have a huge number of descriptors, dimensionality reduction (e.g., by PCA) or variable selection algorithms can be previously applied, in order to get data sets constituted by few descriptors.

Since QDA and CAIMAN methods require the inversion of the class covariance matrices, the data dimensionality was reduced using the first principal components with eigenvalues greater than one, when dealing with data sets for which the number of variables was larger than the number of class objects or when the inverse of the class covariance matrices could not be calculated due to the correlation. In Table 2, the number of retained components is reported within brackets in the column no. var.

All data sets were range scaled between 0 and 1 prior to classification modeling.

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374

Table 2. Characteristics of the Considered Benchmark Data Sets[a]

id	data set	no. obj	no. var	no. class	rel dev %	class partition
1	IRIS[18]	150	4	3	0.0	50 50 50
2	WINES[19]	178	13	3	32.4	59 71 48
3	PERPOT[20]	100	2	2	0.0	50 50
4	ITAOILS[21]	572	8	9	87.9	25 56 206 36 65 33 50 50 51
5	SULFA[22]	50	7	2	61.1	14 36
6	DIABETES[23]	768	8	2	46.4	268 500
7	BLOOD[24]	748	4	2	68.8	178 570
8	VERTEBRAL[25]	310	6	2	52.4	100 219
9	SEDIMENTS[26]	1413	9	2	84.0	1218 195
10	BIODEG[27]	837	12 (5)	2	48.6	553 284
11	DIGITS[5]	500	7	10	27.6	47 42 49 57 54 42 58 45 56 50
12	APPLE[28]	508	15 (3)	2	64.5	133 375
13	TOBACCO[29]	26	6	2	0.0	13 13
14	SCHOOL[30] (p 567)	85	2	3	16.1	31 28 26
15	BANK[30] (p 564)	46	4	2	16.0	21 25
16	HIRSUTISM[31]	133	7	2	75.7	107 26
17	THIOPHENE[32] (p 493)	24	3	3	0.0	8 8 8
18	SUNFLOWERS[33]	70	21 (5)	2	40.9	44 26
19	VINAGRES[34]	66	20 (6)	3	75.8	33 25 8
20	CHEESE[35]	134	21 (7)	4	72.1	68 19 27 20
21	ORUJOS[36]	120	9	2	69.6	28 92
22	MEMBRANE[32] (p 520)	36	2	3	0.0	12 12 12
23	METHACYCLINE[37]	22	4	2	16.7	12 10
24	SIMUL4[38]	32	2	2	0.0	16 16
25	VEGOIL[39]	83	7	4	73.0	37 25 11 10
26	CRUDEOIL[40]	56	5 (2)	3	81.6	7 11 38
27	SAND[41]	81	2	2	27.7	34 47
28	HEMOPHILIA[42]	75	2	2	33.3	30 45
29	COFFEE[43]	43	13 (3)	2	80.6	36 7
30	OLITOS[44]	120	25 (7)	4	78.0	50 25 34 11
31	FISH[45]	27	10 (4)	2	7.1	13 14
32	HEARTHDISEASE[1]	462	7	2	47.0	160 302

[a]In the different columns, the data set name and the scientific reference, the total number of objects, variables, and classes are reported; in column rel dev %, the relative deviation class size obtained as percentage of the relative difference between the number of objects belonging to the greatest class and those belonging to the smallest class; in the last column, the class partitions are reported.

Table 3. Comparison of the Class Sensitivities of N3, BNN, KNN, and wKNN Methods for Some Data Sets

data	class	n_g	sensitivities N3	BNN	KNN	wKNN
IRIS	C1	50	100.0	100.0	100.0	100.0
	C2	50	96.0	96.0	98.0	98.0
	C3	50	92.0	94.0	92.0	92.0
WINES	C1	59	100.0	100.0	100.0	100.0
	C2	71	88.7	95.7	93.0	93.0
	C3	48	100.0	100.0	100.0	100.0
BLOOD	C1	178	69.7	32.0	33.7	35.4
	C2	570	66.1	92.5	90.9	91.8
VERTEBRAL	C1	210	79.5	84.3	82.4	82.4
	C2	100	82.0	79.0	78.0	78.0
SEDIMENTS	C1	1218	92.0	97.7	97.2	97.2
	C2	195	85.6	80.0	82.6	82.6
CHEESE	C1	68	98.5	98.5	100.0	100.0
	C2	19	63.2	63.2	68.4	68.4
	C3	27	77.8	81.5	74.1	74.1
	C4	20	65.0	70.0	70.0	75.0
OLITOS	C1	50	80.0	90.0	86.0	86.0
	C2	25	100.0	92.0	92.0	92.0
	C3	34	85.3	85.3	85.3	85.3
	C4	11	90.9	27.3	18.2	18.2

Table 4. Comparison of the Number of Neighbors of BNN and KNN for Some Data Sets Obtained by Leave-One-Out Validation[a]

data set	BNN median (k)	BNN NER%	KNN k	KNN NER%
WINES	5	98.6	10	97.7
VERTEBRAL	3	81.6	1	80.2
SEDIMENT	1	88.9	1	89.9
DIGITS	20	72.3	8	73.6
APPLE	15	92.3	6	91.9
BANK	3	91.2	4	86.9
THIOPHENE	2	83.3	5	83.3
VINAGRES	2	91.7	7	95.8
METHACYCLINE	2	86.7	1	82.5
CRUDEOIL	4	84.8	3	87.9
HEMOPHILIA	12	85.6	4	82.8
HEARTDESEASE	35	65.2	3	63.2

[a]See Table 5. In boldface are the cases for which the NER% of BNN is 1% higher than KNN, and in italics are the opposite case.

4. SOFTWARE

All models were calibrated in MATLAB.[46] The Classification Toolbox[47] for MATLAB was used to calibrate models by means of LDA, QDA, PLSDA, CART, KNN. N3 and BNN models were calculated by means of the N3/BNN toolbox for MATLAB, written by the authors; this toolbox is available for download at http://michem.disat.unimib.it/chm/download/n3info.htm. The LibSVM library[48] was used for SVM models with both linear and RBF kernel. PCA and MST were carried out in MATLAB by means of routines written by the authors.

5. RESULTS AND DISCUSSION

Model optimization was performed by means of a leave-one-out validation procedure for those classifiers which depend on the selection of a parameter. For N3 classifier, the following 11 values of the α parameter were used: {0.10, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50}. For BNN classifier, similarity values were calculated as for N3, i.e. as a nonlinear inverse function of the average Euclidean distance. The computational time of the two new methods is not significant for the studied data sets, ranging from 1 to 2 s for the smallest data sets to 56 (BNN) and 65 s (N3) for the largest (sediment data set), while this is not the case for SVM/RBF which requires significant larger computational time. In any case, it is known that KNN-based approaches are computationally demanding for huge data sets, but all the actual software languages allow parallelization, thus considerably smoothing this problem.

Table S1 (Supporting Information) collects the optimized parameters for the different classification methods. Afterward, classification algorithms were applied on the data sets and the

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374

Table 5. Nonerror Rates in Leave-One-Out Cross-Validation for the 32 Considered Data Sets[a]

Id	Data set	N3	BNN	KNN	wKNN	LDA	QDA	PLSDA	CART	CAIMAN	SVM /LIN	SVM /RBF
1	IRIS	96.0	96.7	96.7	96.7	**98.0**	97.3	90.2	94.0	**98.0**	97.3	97.3
2	WINES	96.2	98.6	97.7	98.0	99.1	**99.5**	**99.5**	86.2	98.7	99.1	**99.5**
3	PERPOT	99.0	99.0	99.0	99.0	85.0	92.0	86.0	97.0	97.0	87.0	**100**
4	ITAOILS	**96.2**	95.2	94.7	94.7	94.7	95.9	95.9	87.2	82.8	94.7	95.9
5	SULFA	77.4	73.8	73.8	73.8	45.2	69.4	74.0	81.5	58.7	50.0	**88.7**
6	DIABETES	73.6	71.1	70.5	70.5	72.7	69.6	**75.1**	68.8	73.5	72.3	72.8
7	BLOOD	67.9	62.2	62.3	63.6	53.7	54.5	**68.7**	62.1	59.3	50.0	64.1
8	VERTEBRAL	80.8	81.6	80.2	80.2	80.7	84.0	82.1	76.9	56.0	83.3	**84.3**
9	SEDIMENTS	88.9	88.9	**89.9**	**89.9**	66.9	69.4	79.4	84.3	61.1	50.5	69.9
10	BIODEG	84.5	85.3	85.4	**85.5**	77.0	78.6	79.9	79.6	65.6	81.5	83.8
11	DIGITS	74.2	72.3	73.6	73.7	74.0	68.6	41.0	65.2	**77.3**	74.9	74.5
12	APPLE	94.0	92.3	91.9	91.9	91.9	87.6	**95.4**	92.1	83.9	94.4	92.3
13	TOBACCO	92.3	92.3	92.3	92.3	84.6	80.8	88.5	**96.2**	92.3	92.3	92.3
14	SCHOOL	95.3	**96.6**	96.2	96.2	90.8	95.2	89.4	86.8	95.0	94.0	96.4
15	BANK	86.9	**91.2**	86.9	86.9	86.5	88.5	84.9	86.5	88.5	88.5	88.9
16	HIRSUTISM	88.3	90.1	90.0	90.0	55.4	81.4	84.1	70.5	52.9	72.7	**93.8**
17	THIOPHENE	83.3	83.3	83.3	83.3	79.2	79.2	**90.5**	58.3	83.3	83.3	83.3
18	SUNFLOWERS	92.3	90.4	91.2	91.2	87.8	90.8	92.7	82.1	88.9	90.8	**96.9**
19	VINAGRES	**100**	95.8	95.8	95.8	**100**	87.5	**100**	67.3	**100**	**100**	**100**
20	CHEESE	76.1	78.3	78.1	79.4	78.8	82.9	84.7	63.9	77.5	76.2	**85.6**
21	ORUJOS	98.2	**98.4**	98.2	98.2	92.6	94.1	93.9	88.4	62.5	95.7	98.2
22	MEMBRANE	94.4	94.4	94.4	94.4	88.9	94.4	**96.7**	91.7	94.4	91.7	94.4
23	METHACYCLINE	82.5	**86.7**	82.5	82.5	45.8	81.7	55.8	65.8	80.0	54.2	82.5
24	SIMUL4	**100**	**100**	**100**	**100**	28.1	**100**	46.9	90.6	93.8	34.4	**100**
25	VEGOIL	99.0	**100**	99.0	99.0	98.0	82.2	99.0	99.3	89.9	99.3	**100**
26	CRUDEOIL	89.2	84.8	87.9	87.9	85.2	73.6	**89.7**	64.9	78.4	85.3	84.8
27	SAND	93.9	**94.9**	93.9	93.9	93.9	93.9	93.9	81.9	93.9	**94.9**	**94.9**
28	HEMOPHILIA	85.6	85.6	82.8	82.8	85.6	83.9	85.6	78.9	**86.7**	**86.7**	85.6
29	COFFEE	**100**	**100**	**100**	**100**	**100**	92.9	**100**	**100**	**100**	**100**	**100**
30	OLITOS	89.1	73.6	70.4	70.4	83.1	80.0	**94.0**	58.0	77.2	87.6	87.6
31	FISH	92.6	92.9	92.9	92.9	96.4	85.2	**100**	88.7	89.0	**100**	**100**
32	HEARTHDISEASE	**69.9**	65.2	63.2	63.2	68.8	66.2	69.7	66.1	67.3	68.0	68.0

[a]The gray background indicates that calculations were performed on principal components for QDA and CAIMAN classifiers and that more than 5% of objects were unclassified for PLSDA.

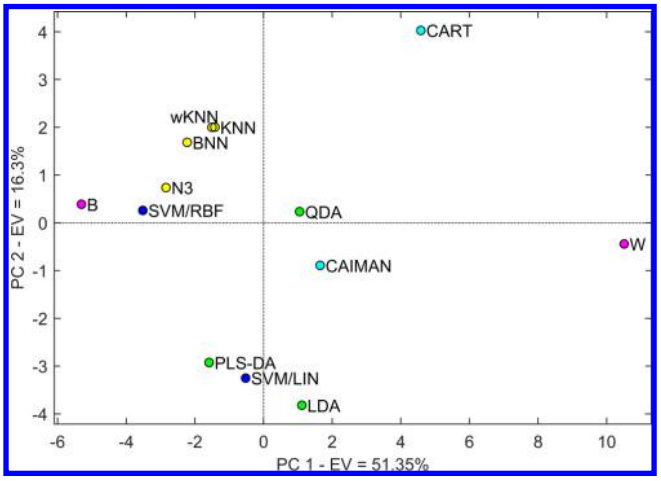


Figure 3. Score plot of the first two principal components obtained from NER% values collected in Table 5, including also BEST (B) and WORST (W) theoretical methods.

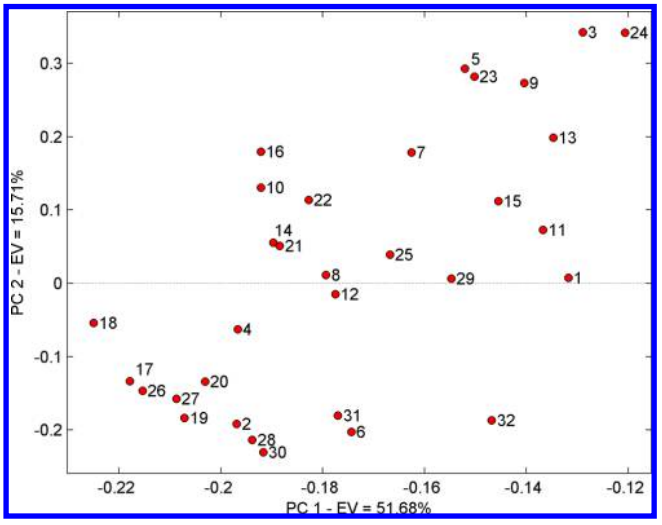


Figure 4. Loading plot of the first two principal components obtained from NER% values collected in Table 5.

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374

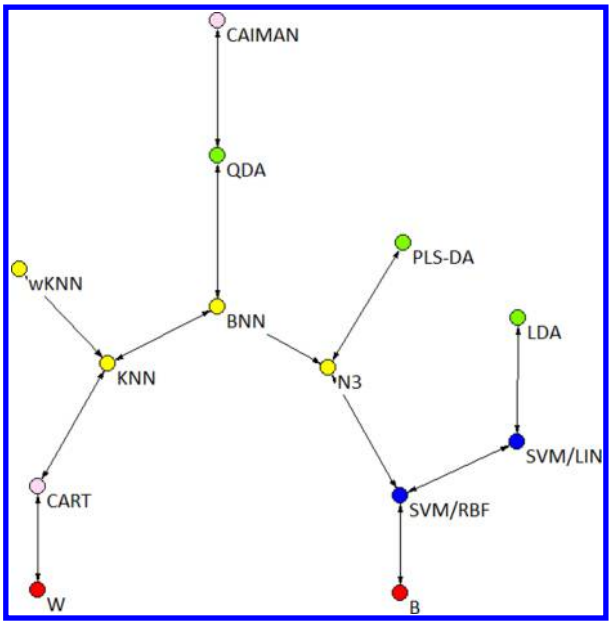


Figure 5. Minimum spanning tree obtained by Euclidean distances calculated on NER% values collected in Table 5, including also BEST (B) and WORST (W) theoretical methods.

leave-one-out procedure was performed to evaluate their performance in cross-validation.

First of all, in order to better understand the behavior of the four methods based on local similarity (N3, BNN, KNN, and wKNN), sensitivities of each class are reported for some data sets (Table 3).

As it can be easily observed, the main difference between N3 from one side and BNN, KNN, and wKNN from the other side is the ability of N3 to classify in a better way small classes. This behavior assumes astonishing proportions in class C4 of OLITOS (91 vs 18) and class C1 of BLOOD (70 vs 34), but it is evident for most of the cases. This characteristic of N3 is particularly interesting because it is unusual for a classifier. It is due to the denominator of the N3 function, which transforms the sum at the numerator in unit of objects which contribute to the summation.

The main difference between BNN and KNN (and wKNN) is related to the number of neighbors which are considered for the class assignment. While in KNN the value of neighbors is optimized by validation and then maintained constant, in BNN the number of neighbors is variable for each object to be classified. A comparison between the number of optimal k neighbors of KNN and the median of the neighbors used by BNN for classification is shown in Table 4 for some data sets, together with the corresponding nonerror rates.

Nonerror rates in LOO cross-validation obtained for each data set are collected in Table 5. The best result obtained for each data set is highlighted in bold. For QDA and CAIMAN methods, a gray background indicates the cases where models were obtained by a dimension reduction based on principal components. In several cases PLSDA could not classify all objects: for example, in four cases (THIOPHENE, CHEESE, CRUDEOIL, OLITOS), PLSDA gave very good results associated with 17%, 21%, 13%, and 18% of unclassified objects, respectively. A gray background in the PLSDA column indicates data sets where more than 5% of objects were unclassified. Note that, among methods based on local similarities, both N3 and BNN are the best methods which require the optimization of only one parameter; in several cases they gave better results than classical KNN, while wKNN generally behaves like KNN giving slightly better results only in few cases (e.g., BLOOD and CHEESE).

Sensitivities and specificities of all the classes of each data set are provided in Table S2 of the Supporting Information.

In order to compare the different classification results, PCA was run on the data matrix constituted by the transposed NER % data set collected in Table 5, including also BEST (B) and WORST (W) theoretical methods. In Figures 3 and 4, the score and loading plots of the first two components are shown, respectively. The first component (PC1) is related to the overall performance of the methods, while the second component (PC2) conveys information on the sensitivity of each method on the particular data set: methods close to zero on this component are less sensitive to different data sets.

On the basis of the results collected in Table 5 and Figure 3, it can be said that the methods with the best overall performance were SVM/RBF and N3, followed by BNN, wKNN, KNN, and PLSDA. The most robust methods were QDA, SVM/RBF, N3, CAIMAN; whereas CART, LDA, SVM/LIN, and PLSDA seem very sensitive to the data set, i.e. they can achieve (very) good or poor results on the basis of the data set to which they are applied.

From the data collected in Table 5, the Euclidean distances between the classifiers were calculated and the MST reported in Figure 5 was obtained. MST synthesizes the information obtained by PCA and confirms the considerations drawn from the PCA regarding the overall performance of the methods. Along the main path from the worst (W) to the best (B) theoretical methods, there are CART, KNN, BNN, N3, and SVM/RBF that are ranked according to their overall performance.

Table 6. Main Statistics of the 11 Classification Methods Derived from Their Performance on 32 Data Sets[a]

stats	N3	BNN	KNN	wKNN	LDA	QDA	PLSDA	CART	CAIMAN	SVM/LIN	SVM/RBF
best	6	8	3	4	3	2	6[b]	2	5	5	13
best (alone)	3	4	0	1	0	0	2[b]	1	1	0	6
mean	2.8	3.6	4.1	4.0	11.2	7.7	6.9	11.4	10.1	9.3	2.2
std dev	3.0	4.5	4.7	4.7	16.1	6.3	12.0	9.7	11.3	15.0	3.8
max	11.3	20.4	23.6	23.6	71.9	20.5	53.1	36.0	40.9	65.6	20.0

[a]Best: no. of times a classifier gives the best result. Best (alone): no. of times a classifier is the unique classifier giving the best result. Mean: mean of the differences between the best result and the result of the classifier for the 32 data sets. Std dev: standard deviation of the differences between the best result and the result of the classifier. Max: maximum difference between the best result and the result of the classifier. [b]For PLSDA the best results obtained when too many objects were excluded (gray cells in Table 5) are not taken into account.

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374

Table 7. NER% Values Obtained on the Test Sets of 11 Data Sets

id	data set	N3	BNN	KNN	LDA	QDA	PLSDA	CART	CAIMAN	SVM/LIN	SVM/RBF
1	IRIS	100.0	100.0	100.0	100.0	100.0	90.5	100.0	97.4	100.0	100.0
2	WINES	90.5	95.2	92.9	97.6	100.0	97.0	90.1	98.9	92.9	95.2
4	ITAOILS	88.7	91.8	90.9	79.7	90.7	61.1	81.1	87.1	94.6	94.6
6	DIABETES	70.9	64.0	66.6	72.3	70.9	75.1	71.8	76.4	71.3	70.4
7	BLOOD	65.0	56.7	65.1	50.0	55.5	66.4	55.5	76.5	50.0	61.8
8	VERTEBRAT	84.3	78.2	83.0	84.3	80.7	85.7	73.1	71.2	92.9	84.4
9	SEDIMENTS	86.2	82.8	86.7	74.6	70.1	80.2	85.4	86.2	50.0	78.7
10	BIODEG	85.3	81.8	83.1	78.8	78.0	82.5	80.1	72.6	81.8	83.2
11	DIGITS	70.1	67.8	68.9	72.0	67.4	50.0	63.9	70.1	72.0	71.9
12	APPLE	93.0	96.8	93.1	97.5	97.5	97.3	88.7	90.1	100.0	97.5
32	HEARTHDISEASE	64.4	62.8	62.7	69.0	72.8	76.5	53.4	71.1	67.4	67.4

Table 8. Statistics for the Test Set Validation Results[a]

statistics	N3	BNN	KNN	LDA	QDA	PLSDA	CART	CAIMAN	SVM/LIN	SVM/RBF
mean	5.7	7.5	6.2	7.7	7.0	9.0	10.7	5.8	8.0	5.1
std dev	4.5	6.4	4.7	7.8	6.8	10.2	7.9	6.8	12.3	4.7
max	12.1	19.8	13.8	26.5	21.0	33.5	23.1	21.7	36.7	14.7

[a]Mean: mean of the differences between the best result and the result of the classifier for the 11 data sets. Std dev: standard deviation of the differences between the best result and the result of the classifier. Max: maximum difference between the best result and the result of the classifier.

In Table 6, some statistical parameters are derived from the estimated NER% reported in Table 5. In particular, for each method, the number of times each method achieved the best result (best) and the number of times each method was the unique to give the best result (best (alone)) are reported. Moreover, for each classifier and each data set, the difference between the obtained NER% and the best NER% achieved was calculated; from these differences, mean (mean), standard deviation (std dev), and maximum difference (max) were calculated and reported.

SVM/RBF (13), BNN (8), N3 (6), and PLSDA (6) classifiers gave the maximum number of best results, both shared and alone. It can also be observed that classifiers based on local similarity (N3, BNN, KNN, and wKNN) gave the "best alone" results 8 times over 32 (25% of the cases).

The lowest differences, on average, from the best results were achieved by SVM/RBF and N3; these two classifiers were also characterized by the lowest standard deviations (N3 = 3.0 and SVM/RBF = 3.8), indicating their regular behavior in approximating the best results. The maximum difference between the best NER and the actual NER of the methods is minimum for N3 (11.3), and quite low for SVM/RBF, BNN, QDA, KNN, and wKNN (20.0, 20.4, 20.5, 23.6, 23.6, respectively).

Finally, for 11 data sets with more than 150 objects, data were partitioned into a training set and a test set comprised of 20% objects randomly selected, respecting the class proportions; models were optimized and calibrated with the training objects and then applied on the test objects. Being the 20% training/test splitting performed only once, the comparison of methods cannot be generalized and data sets having fewer objects are not considered.

The NER% values obtained by each classifier on the test sets are reported for the 11 data sets in Table 7. The results of wKNN were not reported since they are strong correlated to those of KNN. The previous relative behavior in comparing the results of the classifiers is well confirmed, in spite of the fact that all the methods based on local similarities (KNN, BNN, and N3) and support vectors (SVM/RBF and SVM/LIN) are in the worst condition; indeed, these methods do not provide any analytical model but their models are constituted by the objects themselves and their most reliable condition to estimate the prediction ability is achieved when all objects are present in the training set. As for the analysis of the results for the 32 data sets (Table 6), the same statistics were also calculated on the test set validation results (Table 8). As it can be easily observed, the average differences with best results, their standard deviations and the maximum differences give similar information on the model performance. However, classification performance of BNN slightly decreased with respect to the LOO validation procedure. This fact could derive by the use of BNN without any similarity threshold to avoid unreliable class predictions when neighbors are actually dissimilar to the target.

6. CONCLUSIONS

This study deals with the proposal of two novel classification methods inspired by the principles of the K-nearest neighbors, which were exhaustively compared to other well-known classifiers over 32 real and simulated multivariate data sets.

In spite of its simplicity, N3 seems to perform very well, especially when dealing with data sets with less represented classes; indeed, even in the cases when it does not achieve the highest NER%, its results still maintain acceptable performance on almost all the considered data sets.

Analogous considerations can be done for BNN, although its performance is, in general, slightly lower than N3 and SVM/RBF but better than KNN. It seems to be useful for particular data sets where also nonstrictly local information may be relevant for classification, that is several neighbors are useful to correctly classify a target.

SVM/RBF demonstrated to achieve overall the best results; however in one case it gave a considerably lower NER% than N3 (19% lower). Moreover, SVM/RBF requires the optimization of two parameters by a design of experiments procedure, whereas only one parameter needs to be optimized for N3 and BNN.

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374

SVM/RBF and N3 not only provided the overall best results but also resulted to be not very sensitive to the data set, i.e. their performance is stable regardless of the data set.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00326.

Table S1: optimized values of the parameters of the classification methods for each data set. Table S2: sensitivity and specificity values for the classes of all the 32 data sets for N3, BNN, and KNN (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: roberto.todeschini@unimib.it.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY (USA), 2009.

(2) Kowalski, B. R.; Bender, C. F. The K-Nearest Neighbor Classification Rule (pattern Recognition) Applied to Nuclear Magnetic Resonance Spectral Interpretation. *Anal. Chem.* **1972**, *44*, 1405−1411.

(3) Nigsch, F.; Bender, A.; van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. O. Melting Point Prediction Employing *k*-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J. Chem. Inf. Model.* **2006**, *46*, 2412−2422.

(4) McLachlan, G. *Discriminant Analysis and Statistical Pattern Recognition*; Wiley: New York, NY (USA), 1992.

(5) Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A.; Pavan, M. CAIMAN (Classification And Influence Matrix Analysis): A New Approach to the Classification Based on Leverage-Scaled Functions. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 3−17.

(6) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth Inc.: Monterey, California (US), 1984.

(7) Ståhle, L.; Wold, S. Partial Least Squares Analysis with Cross-Validation for the Two-Class Problem: A Monte Carlo Study. *J. Chemom.* **1987**, *1*, 185−196.

(8) Brereton, R. G.; Lloyd, G. R. Support Vector Machines for Classification and Regression. *Analyst* **2010**, *135*, 230−267.

(9) Hand, D. J. *Construction and Assessment of Classification Rules*; Wiley: Chicester (UK), 1997.

(10) Frank, I.; Friedman, J. H. Classification: Oldtimers and Newcomers. *J. Chemom.* **1989**, *3*, 463−175.

(11) Derde, M. P.; Massart, D. L. UNEQ: A Disjoint Modelling Technique for Pattern Recognition Based on Normal Distribution. *Anal. Chim. Acta* **1986**, *184*, 33−51.

(12) Todeschini, R. K-Nearest Neighbour Method: The Influence of Data Transformations and Metrics. *Chemom. Intell. Lab. Syst.* **1989**, *6*, 213−220.

(13) Cartmell, J. *Smart Average K-Nearest Neighbors Algorithm*; Colorado State University, 2009.

(14) Parvin, H.; Alizadeh, H.; Minati, B. A Modification on K-Nearest Neighbor Classifier. *Glob. J. Comput. Sci. Technol.* **2010**, *10*, 37−41.

(15) Voulgaris, Z.; Magoulas, G. D. Extensions of the K Nearest Neighbour Methods for Classification Problems. In *AIA '08 Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications*; ACTA Press: Anaheim, CA (USA), 2008; pp 23−28.

(16) Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Assessing the Validity of QSARs for Ready Biodegradability of Chemicals: An Applicability Domain Perspective. *Curr. Comput.-Aided Drug Des.* **2014**, *10*, 137−147.

(17) Pérez, N. F.; Ferré, J.; Boqué, R. Calculation of the Reliability of Classification in Discriminant Partial Least-Squares Binary Classification. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 122−128.

(18) Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **1936**, *7*, 179−188.

(19) Forina, M.; Armanino, C.; Castino, M.; Ubigli, M. Multivariate Data Analysis as a Discriminating Method of the Origin of Wines. *Vitis* **1986**, *25*, 189−201.

(20) Forina, M. *Artificial Data Set*; University of Genoa, 2005.

(21) Forina, M.; Armanino, C.; Lanteri, S.; Tiscornia, E. Classification of Olive Oils from Their Fatty Acid Composition. *Food Research Data Analysis Proceedings IUFoST Symposium*, Oslo, Norway, Sep 20−23, 1982; Martens, H., Russwurm, H., Jr., Eds.; 1983.

(22) Miyashita, Y.; Takahashi, Y.; Takayama, C.; Ohkubo, T.; Funatsu, K.; Sasaki, S.-I. Computer-Assisted Structure/taste Studies on Sulfamates by Pattern Recognition Methods. *Anal. Chim. Acta* **1986**, *184*, 143−149.

(23) Smith, J. W.; Everhart, J. E.; Dickson, W. C.; Knowler, W. C.; Johannes, R. S. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computational Applied Medical Care*, 1988; pp 261−265.

(24) Baggerly, K. A.; Morris, J. S.; Edmonson, S. R.; Coombes, K. R. Signal in Noise: Evaluating Reported Reproducibility of Serum Proteomic Tests for Ovarian Cancer. *J. Natl. Cancer Inst.* **2005**, *97*, 307−309.

(25) Berthonnaud, E.; Dimnet, J.; Roussouly, P.; Labelle, H. Analysis of the Sagittal Balance of the Spine and Pelvis Using Shape and Orientation Parameters. *J. Spinal Disord. Technol.* **2005**, *18*, 40−47.

(26) Alvarez-Guerra, M.; Ballabio, D.; Amigo, J. M.; Viguri, J. R.; Bro, R. A Chemometric Approach to the Environmental Problem of Predicting Toxicity in Contaminated Sediments. *J. Chemom.* **2010**, *24*, 379−386.

(27) Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative Structure-Activity Relationship Models for Ready Biodegradability of Chemicals. *J. Chem. Inf. Model.* **2013**, *53*, 867−878.

(28) Ballabio, D.; Consonni, V.; Costa, F. Relationships between Apple Texture and Rheological Parameters by Means of Multivariate Analysis. *Chemom. Intell. Lab. Syst.* **2012**, *111*, 28−33.

(29) Forina, M. *Tobacco Data Set*; University of Genoa.

(30) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*, 3rd ed.; Prentice-Hall/Pearson, 1992.

(31) Armanino, C.; Lanteri, S.; Forina, M.; Balsamo, A.; Migliardi, M.; Cenderelli, G. Hirsutism: A Multivariate Approach of Feature Selection and Classification. *Chemom. Intell. Lab. Syst.* **1989**, *5*, 335−341.

(32) Mager, P. P. *Design Statistics in Pharmacochemistry*; Research Studies Press: Letchworth, 1991.

(33) Saviozzi, A.; Lotti, G.; Piacenti, D. La Composizione Amminoacidica Delle Farine Di Girasole. *Riv. Della Soc. Ital. Sci. Dell'Alimentazione* **1986**, *15*, 437−444.

(34) Benito, M. J.; Ortiz, M. C.; Sánchez, M. S.; Sarabia, L. A.; Iñiguez, M. Typification of Vinegars from Jerez and Rioja Using Classical Chemometric Techniques and Neural Network Methods. *Analyst* **1999**, *124*, 547−552.

(35) Resmini, P.; Pellegrino, L.; Bertuccioli, M. Moderni Criteri per La Valutazione Chimico-Analitica Della Tipicità Di Un Formaggio: L'esempio Del Parmigiano-Reggiano. *Riv. Della Soc. Ital. Sci. Dell'Alimentazione* **1986**, *15*, 315−326.

(36) Ortiz, M. C.; Saez, J. A.; Palacios, J. L. Typification of Alcoholic Distillates by Multivariate Techniques Using Data from Chromatographic Analyses. *Analyst* **1993**, *118*, 801−805.

(37) Worth, A. P.; Cronin, M. T. D. Embedded Cluster Modelling−A Novel Method for Analysing Embedded Data Sets. *Quant. Struct.-Act. Relat.* **1999**, *18*, 229−235.

(38) Todeschini, R. *Artificial Data Set*; University of Milano - Bicocca.

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374

(39) Brodnjak-Vončina, D.; Kodba, Z. C.; Novič, M. Multivariate Data Analysis in Classification of Vegetable Oils Characterized by the Content of Fatty Acids. *Chemom. Intell. Lab. Syst.* **2005**, *75*, 31−43.
(40) Gerrild, P. M.; Lantz, R. J. *Chemical Analysis of 75 Crude Oil Samples from Pliocene Sand Units, Elk Hills Oil Field, California*; Open-File Report; USGS Numbered Series 69−105; U.S. Geological Survey, 1969.
(41) Hamilton, L. J. Cross-Shelf Colour Zonation in Northern Great Barrier Reef Lagoon Surficial Sediments. *Aust. J. Earth Sci.* **2001**, *48*, 193−200.
(42) Habbema, J. D. F.; Hermans, J.; Van den Broek, K. A Step-Wise Discriminant Analysis Program Using Density Estimation. *CompStat 1974, Proc. Computational Statistics*, 1974; pp 101−110.
(43) Streuli, H. Mathematische Modelle Für Die Chemische Zusammensetzung von Libensmitteln Und Ihre Bedeutung Für Deren Beurteilung. *Lebensm.-Wiss. Technol.* **1987**, *20*, 203−211.
(44) Armanino, C.; Leardi, R.; Lanteri, S.; Modi, G. Chemometric Analysis of Tuscan Olive Oils. *Chemom. Intell. Lab. Syst.* **1989**, *5*, 343−354.
(45) Forina, M.; Armanino, C.; Lanteri, S. Acidi Grassi Degli Animali Acquatici: Uno Studio Chemiometrico. *Riv. Della Soc. Ital. Sci. Dell'Alimentazione* **1982**, *11*, 15−22.
(46) *MATLAB*; MathWorks Inc.: Natick, MA, USA.
(47) Ballabio, D.; Consonni, V. Classification Tools in Chemistry. Part 1: Linear Models. PLS-DA. *Anal. Methods* **2013**, *5*, 3790−3798.
(48) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans Intell Syst. Technol.* **2011**, *2*, 1−27.

DOI: 10.1021/acs.jcim.5b00326
J. Chem. Inf. Model. 2015, 55, 2365−2374