MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **87** (2022) 273-298
ISSN: 0340-6253
doi: 10.46793/match.87-2.273T

Evaluation of Classification Performances of Minimum Spanning Trees by 13 Different Metrics

Roberto Todeschini*, Cecile Valsecchi

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

(Received August 5, 2021)

Abstract

Minimum Spanning Tree (MST) is a well-known clustering algorithm that provides a graphical tree representation of the objects in a data set by exploiting local information to link each pair of similar objects. The *a-posteriori* analysis of this tree in terms of nodes and edges provides the basis to derive simple classifiers, namely semi-supervised classification approaches based on the minimum spanning tree approach. In this work, we propose different metrics to evaluate the MST ability to group objects of the same *a-priori* known classes. The classification capability of the proposed approach, using 13 different distance measures, was compared with that of classical supervised classification approaches such as N-Nearest Neighbour (N3), Binned Nearest Neighbour (BNN), Partial Least Squares-Discriminant Analysis (PLS-DA), K-Nearest Neighbour (KNN), exponentially weighted K-Nearest Neighbour (wKNN) and Support Vector Machine with radial functions (SVM-RBF) on 31 data sets. The proposed approach resulted to be competitive and comparable with the considered classical supervised classification methods. Finally, we analysed the role of the 13 different measures in terms of performance and percentage of not-assigned objects.

1 Introduction

In classification problems, a typical data set consists 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

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

methods (or simply, classification methods). These methods are applied in different scientific fields such as analytical chemistry, food chemistry, toxicology, QSAR/QSPR, image analysis, process and environmental monitoring, social, medical and economical 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 to the *G* classes are known [1].

Existing classification algorithms are based on a variety of approaches that confer them different characteristics and properties [1–3]. We can distinguish (i) probabilistic methods, such as Kernel Density Estimators, (ii) methods intrinsically based on Principal Components Analysis, such as SIMCA, (iii) methods based on discriminant analysis, such as, for example, linear, quadratic, regularized discriminant analysis, (iv) methods based on local analysis of the variable space, such as, for example, KNN and N3, and so on, (v) fuzzy methods, where to each object is assigned a probability to belong to each class and (vi) classification tree methods, such as CART, where the classification proceeds hierarchically by successively binary splitting of the objects and producing easily interpretable graphs. All these methods are supervised classification methods, where the knowledge of the class partition of the objects influences the model development.

Minimum Spanning Tree (MST) is a clustering algorithm that provides a graphical tree representation of the objects in a data set by exploiting local information to link each pair of objects, namely analysing the object distance matrix.

The idea to use the Minimum Spanning Tree (MST) to perform classification was pursued by different authors, for instance, by applying to the MST graph some rules to classify labelled vertices.

In the paper of Zhou et. al. [4] the MST is combined with the KNN method. The assignment is initially performed by analysing the classes of the first neighbours; in case of tied results, the next paths are also taken into account for all the objects and so on, until the object is classified by majority voting. The MST is built by the Euclidean metric and allows to predict one unknown sample at a time. Juszczak et al. [5] applied the MST approach to asymmetric one-class problems, thus, a weighted spanning tree is built only using the training objects of the class of interest. Also in this case, the method exploits the Euclidean metric as in several contributions provided by Vitale, Cesa-Bianchi and co-workers [6]. Chakrabarty and Roy [7] proposed an optimized KNN classifier based on the MST algorithm to automatically classify email documents with an initially unknown number of clusters filtering the emails in a two-class problem by using the Jaccard-Tanimoto metric.

La Grassa et al. [8] also worked on one-class problems combining MST and KNN in different ways and performing an extended comparison with other classifiers on 6 data sets. Also in this case, the MST is built by the Euclidean metric. Yang-Min Zhang et al. [9] proposed a different algorithm to obtain a tree from a graph minimizing the sum of weights of the cuts, where the cut is the link between two vertices having different labels.

In all the cited works, only a few numbers of data sets were used for comparisons with other approaches and the role of the different metrics was never evaluated.

In this paper, a classification approach is proposed, namely a semi-supervised classification, based on the *a-posteriori* analysis of a Minimum Spanning Tree (MST) obtained by using different rules for links and nodes of the graph and different metrics.

An example of unsupervised classification methods was already proposed in literature, such as the counter propagation artificial neural networks (CP-ANN), based on the self-organizing maps (SOM), also called Kohonen maps [10]. In this method, the objects are clustered into a map of N *p*-dimensional neurons, where *p* is the number of input variables of each object and N the size of the map. During the clustering procedure, the class information of each object is projected into an output map (Grossberg layers), without any feedback on the map.

To reach our goal by using MST for classification purposes, two approaches are proposed, which, after the building of the MST graph, analyse the graph topology to obtain measures of the MST classification performance from a link- and node-based points of view.

2 Theory

The Minimum Spanning Tree algorithm provides a bi-dimensional plot of the n objects described by their pairwise distance matrix $\mathbf{D}$ ($n \times n$).

The graph is constituted by a set of straight-line segments joining the n objects such that:

1. every point is connected to every other point by a set of lines (a *path*);
2. each point is visited by at least one line, its *vertex degree* is at least equal to 1;
3. no closed circuits appear in the structure, and thus, it is a *tree*.

Among the several algorithms proposed to build a MST, the most popular is an iterative algorithm where, at any step, the segments belong to two sets A and B: set A contains the set of segments (i.e., object pairs, the pairwise links) assigned to the MST (initially it is empty), and set B, those not yet assigned to MST. Then, the algorithm assigns to A the shortest distance

in B that does not form a closed loop with any of the segments already present in A. The iterations stop when the set A contains $n - 1$ elements (links).

Then, the MST can be thought of as a distance selection approach, where only a subset of $n - 1$ distances are used over the $n \times (n - 1)/2$ distances, that is, a percentage equal to $2/n$ is used to build the tree.

As each segment joins a pair of objects that are the most similar between them, MST is used, *de facto*, as clustering method. For the same data set, several MSTs can be obtained for metrics calculated by different distance measures.

2.1 Link-based strategy

Once the MST was calculated, the knowledge of the classes to which the objects belong to allows to (1) assess the discriminant power of the MST and (2) use MST as a classification method. To reach this purpose, a measure of the MST discriminant power is needed.

The first step is based on a very simple rule: each link of two objects belonging to different classes is considered an error, except $G - 1$ links which is the minimum number of links to maintain a tree structure having G classes. Mathematically, it is defined as the following.

Let **A** be the $n \times n$ symmetric binary matrix (i.e., the adjacency matrix) containing the links of each object with the other ones. The overall tree classification rate (T) can be defined as:

$$T = 1 - \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} \cdot \delta_{ij} - (G-1)}{L} \qquad \delta_{ij} = \begin{cases} 1 & if\ c_i \neq c_j \\ 0 & if\ c_i = c_j \end{cases} \tag{1}$$

where c are the classes of the objects and L the total number of links, which is equal to $n - 1$ in the MST; a_{ij} is the entry of **A** and it is equal to 1 if objects i and j are connected, 0 otherwise. The Dirac delta function δ_{ij} is equal to 1 if the classes of the j-th and i-th objects differ, and 0 otherwise. In other words, T provides the accuracy of the MST partition where the second term of Eq. (1) is the fraction of the links between objects belonging to different classes, which represent the errors in the MST partitions. As already explained above, the term G –1 is the minimum number of transition links between objects belonging to different classes for data partitioned into G classes. An overall tree classification rate equal to 1 indicate perfect partitions of the classes by MST.

To get a classification rate measure for each class (or class sensitivities) we considered the same very simple rule: each link of the obtained tree, where the two objects belong to different classes, is considered as error. Considering n objects partitioned into G classes, a link-based

confusion matrix **C** ($G \times G$) can be constructed where all the diagonal elements are filled by the number of links between objects belonging to the same class. All the links between objects belonging to two different classes are stored in the corresponding off-diagonal cell with a score of 0.5, in a symmetric way. Finally, a quantity of 0.5 is subtracted to all the off-diagonal elements greater than zero of the link-based confusion matrix. This symmetric correction plays the same role of the $G - 1$ correction term in the formula (1).

More formally, the diagonal values C_{gg} (i.e., correct assignations for the g-th class) are defined as:

$$C_{gg} = \sum_{i \in g} \sum_{j=1}^{n} a_{ij} \cdot \delta_{jg} \qquad \delta_{jg} = \begin{cases} 1 & if\ c_j = g \\ 0 & if\ c_j \neq g \end{cases} \tag{2}$$

where c_j is the class of the j-th object. The Dirac delta function δ_{ig} is equal to 0 if the class of the j-th object differs from g, and 1 otherwise. The off-diagonal values of C are defined as:

$$C_{gg'} = C_{g'g} = \sum_{i \in g} \sum_{j \in g'} a_{ij} \cdot 0.5 \qquad g' \neq g \tag{3}$$

From the link-based confusion matrix **C**, traditional classification metrics can be computed. The MST ability to identify the g-th class can thus be quantified by the sensitivity (Sn_g) value as:

$$Sn_g = \frac{RS_g - ER_g}{RS_g} \tag{4}$$

where:

$$ER_g = ER_{g'} = \sum_{g' \neq g} \left(C_{gg'} - 0.5 \right) \cdot \delta_{gg'} \qquad \delta_{gg'} = \begin{cases} 1 & if\ C_{gg'} > 0 \\ 0 & if\ C_{gg'} = 0 \end{cases} \tag{5}$$

$$RS_g \equiv C_{gg} + \sum_{\substack{g'=1 \\ g' \neq g}}^{G} C_{gg'} \tag{6}$$

and the total sum of the entries equal to the number of links can be expressed as:

$$\sum_{g=1}^{G} RS_g = n - 1 \tag{7}$$

Finally, a link-based Non-Error Rate (L) is computed as the average of the class sensitivities:

$$L = \frac{\sum_{g=1}^{G} Sn_g}{G} \tag{8}$$

For example, for a data set constituted by 30 objects in class C1 and 20 objects in class C2 and a perfect MST (i.e., a tree able to separate the two classes), the following confusion matrix is obtained:

$\begin{vmatrix} 29 & 0.5 \\ 0.5 & 19 \end{vmatrix}$, which is successively corrected subtracting 0.5 from the off-diagonal values to: $\begin{vmatrix} 29 & 0 \\ 0 & 19 \end{vmatrix}$. In this case, the 48 total links constituted by the objects belonging to the same class are recognized and both L and T measures are equal to 1.

2.2 Node-based strategy

Changing perspective from links to nodes allows to calculate a membership matrix **M** (*n* x *G*) and to assign a new object to a class by a majority voting criterion. Starting from the adjacency matrix **A** *(n x n),* the elements of the membership matrix are calculated as:

$$m_{ig} = \frac{\sum_{j=1}^{n} a_{ij} \cdot \delta_j}{\sum_{j=1}^{n} a_{ij}} \qquad \delta_j = \begin{cases} 1 & if\ c_j = g \\ 0 & if\ c_j \neq g \end{cases} \qquad i = 1,n \quad g = 1,G \tag{10}$$

where c_j is the class the *j*-th object linked to *i* belongs to.

If the $\mathbf{m}_i$ membership vector has a unique highest value, the object is assigned to the corresponding class, otherwise is considered not-classified.

However, to take into account the real space defined by distances, a new membership matrix **M** (*n* x *G*), which replaces the previous one, can be obtained. In other words, the spanning tree defines the topology, and the distances are used to classify the objects using the given topology. This choice allows to take full advantage of the peculiarities and fuzzy differences of the metrics used.

Starting from the adjacency matrix **A**, the elements of the membership matrix are calculated as:

$$m_{ig} = \frac{\sum_{j=1}^{n} a_{ij} \cdot exp\left(-d_{ij}\right) \cdot \delta_j}{\sum_{j=1}^{n} a_{ij} \cdot exp\left(-d_{ij}\right)} \qquad \delta_j = \begin{cases} 1 & if\ c_j = g \\ 0 & if\ c_j \neq g \end{cases} \qquad i = 1,n \quad g = 1,G \tag{11}$$

where d_{ij} is the actual distance between objects *i* and *j*.

Then, each object is assigned to the class when the two highest membership values differ more than a membership threshold *t*, here selected as 0.05, otherwise the object is considered not-classified, or more formally:

$$i \in g_1 \quad if\ m_{ig_1} - m_{ig_2} > t \tag{12}$$

where g_1 and g_2 are the classes for which the *i*-th object has the first and second maximum membership values.

A threshold equal to zero is also taken into account to obtain a crisp classification rule.

This approach with a distance-based membership matrix can be considered as a variant of the exponentially weighted KNN (wKNN) [11], where the variable k values are given by the tree topology, that is by the vertex degree of each object. This approach is different from the original wKNN method (where, as usual, the optimal k value is searched for by validation) being the number of k neighbours assigned *a priori* by the tree topology, and thus not dependent of the knowledge of the classes. The wKNN method has been demonstrated to perform as the classical KNN method and, in some cases, to slightly outperform it [3].

Therefore, from the membership function, an object-based confusion matrix **C** can be defined with diagonal and off-diagonal values are defined as:

$$C_{gg} = \sum_{i \in g} max\left(m_{ig}\right) \cdot \delta_{ig} \qquad \delta_{ig} = \begin{cases} 1 & if\ max\left(m_{ig}\right) = g \\ 0 & if\ max\left(m_{ig}\right) \neq g \end{cases} \tag{13}$$

$$C_{gg'} = \sum_{i \in g} max\left(m_{ig'}\right) \cdot \delta_{ig'} \qquad \delta_{ig'} = \begin{cases} 1 & if\ max\left(m_{ig'}\right) = g' \\ 0 & if\ max\left(m_{ig'}\right) \neq g' \end{cases} \tag{14}$$

Similarly, to L, also in this case sensitivities and a node-based non-error rate with threshold t (NER_t) can be defined in the classical way as:

$$Sn_g = \frac{c_{gg}}{n_g} \tag{15}$$

$$NER_t = \frac{\sum_{g=1}^{G} Sn_g}{G} \tag{16}$$

where n_g is the number of object belonging to class g. In the following analysis we considered two values for t: 0 (NER(0)) and 0.5 (NER(0.05).

3 Prediction

The proposed approaches allow also predicting the class of new objects by (i) adding the object to the training set, one at a time, (ii) calculating the distance matrix, (iii) calculating the MST of $n + 1$ objects and (iv) evaluating its class by a link or node-based strategy.

An example of prediction for a new object is given in Figure 1. In the link-based strategy the number of links to the objects belonging to different classes is counted and the unknown object is assigned to the class having the maximum number of links (e.g., *a* and *c* in Figure 1). In this case, an object can also result as unclassified when the number of links with two different classes is equal (cases *b* and *d* in Figure 1). The prediction algorithm can be considered as a KNN classification method with a variable number of k, where k is obtained by the topological

structure of the spanning tree, i.e., each vertex is characterized by the number of its links, or, in other words, by its vertex degree.

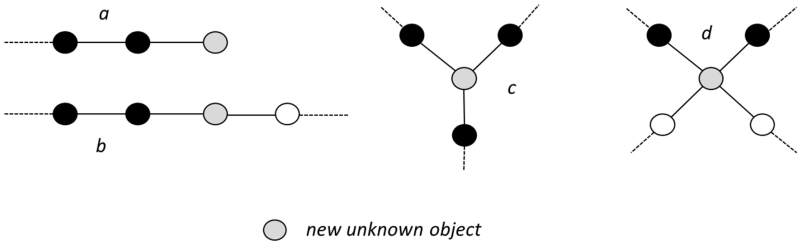


Figure 1. Examples of prediction of a new object (in grey colour) in a MST with two classes (white and black objects).

The case of type *b* in the MST graph is usually very common and, consequently, the number of not-classified objects is often quite high. To avoid this drawback, in the node-based strategy, we took into account the real space defined by the distances and introducing a threshold *t*. Also in this case, the additional presence of not-classified objects is provided. For the cases *a* and *c*, the membership is equal to 1, while, for cases *b* and *d*, the calculated maximum membership value is considered, according to formula (11).

Moreover, selecting as a distance threshold the maximum distance among the MST links of the training set, that is coming back from the topological space of MST to the real space of the distances, the MST can be transformed into a semi-supervised modelling method. In this case a new object will be excluded from the applicability domain if the minimum distance of its links is greater than the distance threshold.

4 Data sets

The 31 data sets used for evaluating the performances of MST as classification method are collected in Table 1. The 13 distance measures studied in this work are collected in Table 2. All the distances are calculated after a range scaling of the data.

Table 1. Characteristics of the considered benchmark data sets. In the different columns, the data set name with the original reference, the total number of objects (N.obj.), variables (N.var.) and classes (N.class) 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 biggest class and those belonging to the smallest class; in the last column the class partitions are reported.

Id	*Data set*	*N. obj.*	*N. var.*	*N. class*	*rel. dev.%*	*class partition*
1	IRIS [12]	150	4	3	0.0	50 50 50
2	WINES [13]	178	13	3	32.4	59 71 48
3	PERPOT [14]	100	2	2	0.0	50 50
4	ITAOILS [15]	572	8	9	87.9	25 56 206 36 65 33 50 50 51
5	SULFA [16]	50	7	2	61.1	14 36
6	VINEGARS [17]	66	20	3	75.8	33 25 8
7	CHEESE [18]	134	21	4	72.1	68 19 27 20
8	OLITOS [19]	120	25	4	78.0	50 25 34 11
9	COFFEE [20]	43	13	2	80.6	36 7
10	DIGITS [21]	500	7	10	27.6	47 42 49 57 54 42 58 45 56 50
11	VEGOIL [22]	83	7	4	73.0	37 25 11 10
12	CRUDEOIL [23]	56	5	3	81.6	7 11 38
13	APPLE [24]	508	15	2	64.5	133 375
14	TOBACCO [25]	26	6	2	0.0	13 13
15	METHACYCLINE [26]	22	4	2	16.7	12 10
16	DIABETES [27]	768	8	2	46.4	268 500
17	THIOPHENE [28]	24	3	3	0.0	8 8 8
18	SAND [29]	81	2	2	27.7	34 47
19	HEARTDISEASE [30]	462	7	2	47.0	160 302
20	BIODEG [31]	837	12	2	48.6	553 284
21	SCHOOL [32]	85	2	3	16.1	31 28 26
22	ORUJOS [33]	120	9	2	69.6	28 92
23	HIRSUTISM [34]	133	7	2	75.7	107 26
24	SUNFLOWERS [35]	70	21 (5)	2	40.9	44 26
25	BLOOD [36]	748	4	2	68.8	178 570
26	VERTEBRAL [37]	310	6	2	52.3	210 100
27	BANK [32]	46	4	2	16.0	21 25
28	MEMBRANE	36	2	3	0.0	12 12 12
29	HEMOPHILIA [38]	75	2	2	33.3	30 45
30	FISH [39]	27	10	2	7.1	13 14
31	SEDIMENT [40]	1413	9	2	84.0	1218 195

Table 2. The definitions of the 13 distances used in this work The symbols x and y represent two objects and p is the total number of variables.

Distance	*Symbol*	*Definition*	*Range*	*Average*
Euclidean	EUC	$D_{xy}^{EUC} = \sqrt{\sum_{j=1}^{p}\left(x_j - y_j\right)^2}$	$0 \le D_{xy}^{EUC} < \infty$	$\bar{D}_{xy}^{EUC} = \frac{D_{xy}^{EUC}}{\sqrt{p}}$
Canberra	CAN	$D_{xy}^{CAN} = \sum_{j=1}^{p}\frac{\lvert x_j - y_j\rvert}{\lvert x_j\rvert + \lvert y_j\rvert}$	$0 \le D_{xy}^{CAN} \le p$	$\bar{D}_{xy}^{CAN} = \frac{D_{xy}^{CAN}}{p}$

Lance-Williams	LW	$D_{xy}^{LW} = \frac{\sum_{j=1}^{p} \lvert x_j - y_j \rvert}{\sum_{j=1}^{p} \left(\lvert x_j \rvert + \lvert y_j \rvert \right)}$	$0 \le D_{xy}^{LW} \le 1$	$\bar{D}_{xy}^{LW} = D_{xy}^{LW}$
Manhattan	MAN	$D_{xy}^{MAN} = \sum_{j=1}^{p} \lvert x_j - y_j \rvert$	$0 \le D_{xy}^{MAN} < \infty$	$\bar{D}_{xy}^{MAN} = \frac{D_{xy}^{MAN}}{p}$
Lagrange	LAG	$D_{xy}^{LAG} = \max_j \lvert x_j - y_j \rvert$	$0 \le D_{xy}^{LAG} < \infty$	$\bar{D}_{xy}^{LAG} = D_{xy}^{LAG}$
Clark	CLA	$D_{xy}^{CLA} = \sqrt{\sum_{j=1}^{p} \left(\frac{x_j - y_j}{\lvert x_j \rvert + \lvert y_j \rvert} \right)^2}$	$0 \le D_{xy}^{CLA} \le p$	$\bar{D}_{xy}^{CLA} = \frac{D_{xy}^{CLA}}{\sqrt{p}}$
Soergel	SOE	$D_{xy}^{SOE} = \frac{\sum_{j=1}^{p} \lvert x_j - y_j \rvert}{\sum_{j=1}^{p} \max\left(x_j, y_j \right)}$	$0 \le D_{xy}^{SOE} \le 1$	$\bar{D}_{xy}^{SOE} = D_{xy}^{SOE}$
Bhattacharyya	BHA	$D_{xy}^{BHA} = \sqrt{\sum_{j=1}^{p} \left(\sqrt{x_j} - \sqrt{y_j} \right)^2}$	$x, y \ge 0$ $0 \le D_{xy}^{BHA} < \infty$	$\bar{D}_{xy}^{BHA} = \frac{D_{xy}^{BHA}}{\sqrt{p}}$
Wave-Edge	WE	$D_{xy}^{WE} = \sum_{j=1}^{p} \left(1 - \frac{\min\left(x_j, y_j \right)}{\max\left(x_j, y_j \right)} \right)$	$0 \le D_{xy}^{WE} \le p$	$\bar{D}_{xy}^{WE} = \frac{D_{xy}^{WE}}{p}$
Jaccard-Tanimoto	JT	$D_{xy}^{JT} = \sqrt{1 - \frac{\sum_{j=1}^{p} x_j \cdot y_j}{\sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} y_j^2 - \sum_{j=1}^{p} x_j \cdot y_j}}$	$0 \le D_{xy}^{JT} \le 1$	$\bar{D}_{xy}^{JT} = D_{xy}^{JT}$
Cosine	COS	$D_{xy}^{CD} = 1 - \frac{\sum_{j=1}^{p} x_j \cdot y_j}{\sqrt{\sum_{j=1}^{p} x_j^2 \cdot \sum_{j=1}^{p} y_j^2}}$	$0 \le D_{xy}^{CD} \le 1$	$\bar{D}_{xy}^{CD} = D_{xy}^{CD}$
Dehmer	DEH	$D_{xy}^{DEM} = p - \sum_{j=1}^{p} e^{-\left(\frac{x_j - y_j}{s_j} \right)^2}$	$0 \le D_{xy}^{DEM} \le p$	$\bar{D}_{xy}^{DEM} = \frac{D_{xy}^{DEM}}{p}$
Intersection	INT	$D_{xy}^{INT} = 1 - \frac{\sum_{j=1}^{p} min\left\{ x_i, y_i \right\}}{max\left\{ \sum_{j=1}^{p} x_j, \sum_{j=1}^{p} y_j \right\}}$	$0 \le D_{xy}^{INT} \le 1$	$\bar{D}_{xy}^{INT} = D_{xy}^{INT}$

5 Software

The software code for the calculation of MST was developed by the Authors in MATLAB [41], as well as the software code for the evaluation of the link and node measures. For the MST

graphs, the free available software Pajek [42] was used. Moreover, a toolbox, designed in MATLAB, is also available which allows to compare the 13 distances and different data scaling methods on custom data providing a tree graph and predictions for a new object (https://michem.unimib.it/download/matlab-toolboxes/mst-viewer-for-matlab/).

6 Results and discussion

The performances in terms of the two link-measures (T and L), and of the two node measures (NER(0) and NER(0.05)) for the 31 data sets and the considered 13 distance measures are collected in Appendix A, Appendix B, Appendix C and Appendix D, respectively. In Appendix E, the percentages of not-classified objects in case of NER (0.05) are also collected.

In the paragraph *Comparison of the models* the models obtained are discussed and compared with other classification methods, together with some comments about the role of the different metrics, while in the paragraph *Comparison of the metrics* an extended multivariate comparison of the metrics is performed.

6.1 Comparison of the models

The comparison of the classification performances is performed with the Non-Error Rate (NER) of six other methods studied in [3] for the same data sets.

The considered benchmark methods are N-Nearest Neighbour (N3) [3], which resulted as the best method in the comparison with other 10 classification methods; Partial Least Squares Discriminant Analysis (PLS-DA) [1], which is the unique method among the methods studied in [3] also giving not-classified objects, K-Nearest Neighbour (KNN) [3], Binned Nearest Neighbour (BNN) [3], Support Vector Machine with radial function (SVM-RBF) [1]. Finally, the exponentially weighted K-Nearest Neighbour (wKNN) [3] was also considered, being the most similar classification method to the proposed approach; indeed, as already noted above, wKNN differs from the NER(t) approach by using a fixed optimal k value of neighbours, found by a validation procedure [11].

Looking at the metrics giving the best result for each data set reported in Appendix D for NER(0.05), the performances of the proposed measure appear immediately as quite satisfactory. Figure 2 shows the performance in terms of NER(0.05) as filled black circles, while inner white circles represent the proportion of not classified objects.

For eight data sets, different metrics achieved equal performance as in the case of the Vinagres and Vergoil data sets where 10 metrics provided a NER(0.05) of 100%. For the remaining 23

data sets, one metric showed an absolute best value in terms of NER(0.05), which was not always balanced by a low percentage of unassigned items. For example, with the Bhattacharyya metric we achieved the best NER(0.05) for the Thiophene data set (87.8%) with a high percentage of unclassified objects (25%), while using the Euclidean or Jaccard-Tanimoto metrics with only 8.3% of unclassified objects we achieved a slightly lower NER(0.05) of 85.7%.

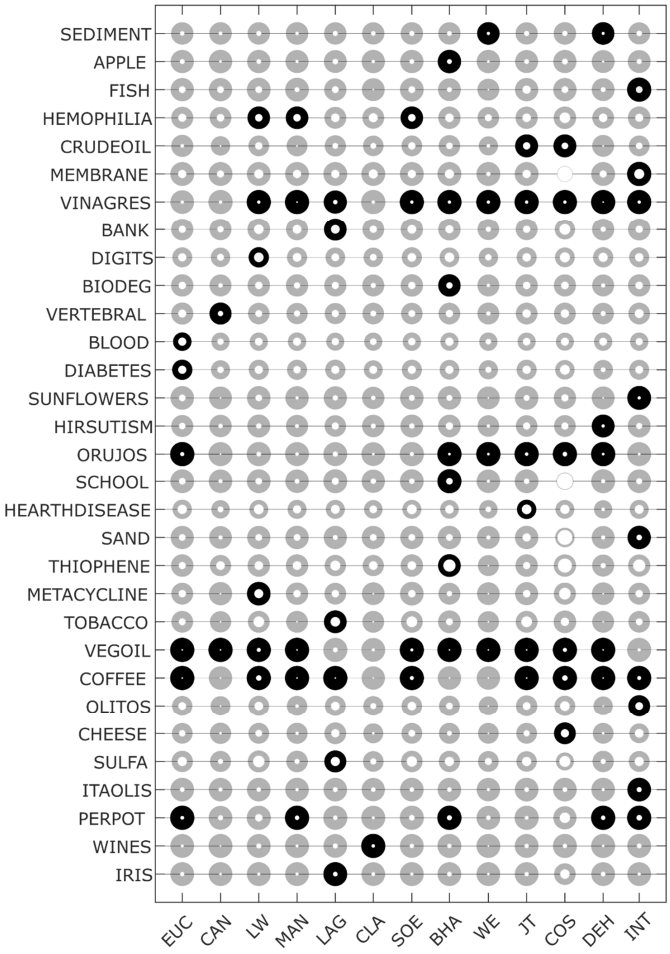


Figure 2. Graphical representation as filled and empty circles of NER(0.05) (Appendix D) and not classified objects (Appendix E), respectively. Maximum values of NER(0.05) are highlighted.

A comparison of the weighted results was also performed with the NER(0) crisp results (reported in Appendix B) obtained with a membership threshold equal to zero, thus avoiding not-classified objects. In the 86% of the cases, the presence of not-classified objects allowed increasing (or equal) performances; only in 56 cases over 403 (13 x 31), the weighted performances with a membership threshold of 0.05 resulted worse than those without a membership threshold equal to 0, with only 13 values with differences lower than -1.

In Figure 3 the score plot of a PCA performed on the best results achieved by the presented methods (NER(0.05), NER(0), L and T) and by traditional approaches (PLS-DA, SVM-RBF, N3, BNN, KNN and wKNN) is showed. Performances for traditional approaches are expressed as NER calculated with a leave one out strategy and the optimized parameters are given in ref [3]. Two theoretical metrics were also added to the rows, called B (best) and W (worst), defined by the best and worst results obtained for each data set. These added theoretical metrics allow stretching the first component for a better evaluation of the behaviour of the approaches. A full comparison of the proposed measures with other classification methods is reported in Appendix F.

T and L are calculated in a link-based strategy, while the other performances are computed on objects and thus fully comparable. Considering the first component, which can be related to the overall quality of the approaches, the node-based strategy with a threshold of 0.05 (NER(0.05)) achieved satisfying performances comparable with SVM-RBF results and better than N3, BNN and KNN. The second component explains the consistency of the results given by the approaches on different data set. The approach showing more dataset-dependent result variation is PLS-DA.

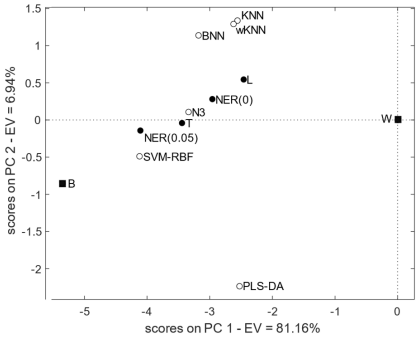


Figure 3. The score plot of the first two PCs for the best results achieved by the presented methods (filled circles, NER(0.05), NER(0), L and T) and by traditional approaches (empty circles, PLS-DA, SVM-RBF, N3, BNN, KNN and wKNN). B and W (filled squares) are the best and worst metrics.

Summarizing the results, it can be said that the unsupervised approaches to classification problems are rightfully considered competitive and comparable with the other classical classification tools. Moreover, a not marginal advantage is the possibility to obtain a graphical efficient representation of the results by means of the minimum spanning tree graph.

Two examples of MST graphs for the data set Tobacco are reported in Figure 4, for Bhattacharyya and Soergel metrics. In this case the former metric provided better results than the latter metric, with a higher NER(0.05) of 87.5% and a lower percentage of not assigned objects (7.7%).

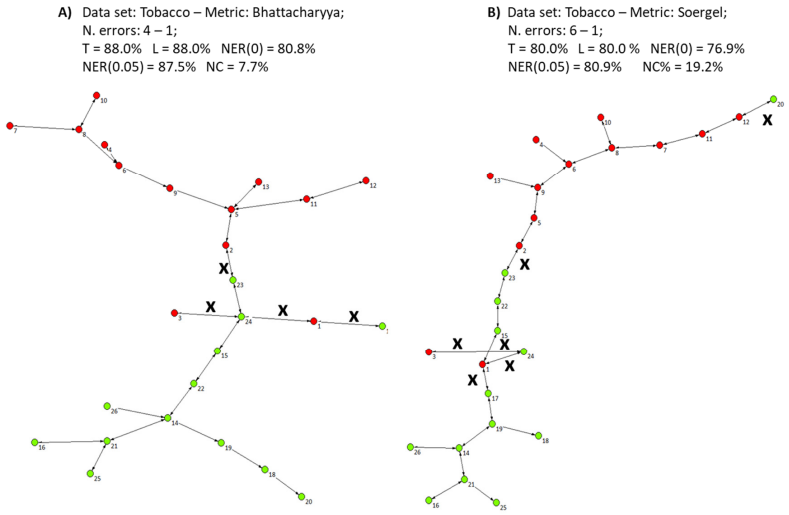


Figure 4. MST of the data set Tobacco (2 classes), with the Bhattacharyya metric (A), and Soergel metric (B). The symbol **x** denotes the links considered as errors.

6.2 Comparison of the metrics

Being the proposed approaches unsupervised, we produced several spanning trees with different metrics and evaluated *a-posteriori* the best topological space for classification.

The 13 metrics were arbitrarily represented by different symbols as the following:

1) EUC, MAN, LAG, BHA, JT in the first group (up triangles)
2) CAN, LW, CLA, COS in the second group (down triangles)
3) SOE, WE, INT in the third group (empty circles)
4) DEH in the fourth group (filled circle).

The first group is constituted by Minkowski-like metrics (Euclidean, Manhattan and Lagrange), together with Euclidean-like metrics such as Jaccard-Tanimoto and Bhattacharyya. The second group is constituted by metrics based on ratio of sums or sum of ratios, such as Canberra, Lance-Williams, Clark and Cosine metrics. The third group is constituted by metrics where min/max functions are present in their definition, while the Dehmer metric is isolated, being the unique based on the exponential function and fully invariant to scaling.

A PCA was performed on a matrix where the rows are the 13 metrics, the columns are the 31 data sets and each entry is the calculated NER(0.05) measure (Figure 5A). Looking at the first component (explaining 86.2% of the total variance), the best metrics are the Lance-Williams and Soergel metrics.

It can be noted that here, for the first time, the Dehmer metric is compared with other classical metrics [43]. Indeed, proposed as a measure of difference between two molecules described by a topological descriptor of a molecular graph [44], DEH has been here extended into a multivariate distance measure.

The second component (explaining 4.1% of the total variance) is related to the different behaviour of the metrics for the diverse data sets. In particular, WE, CLA and CAN are more sensible to the data sets.

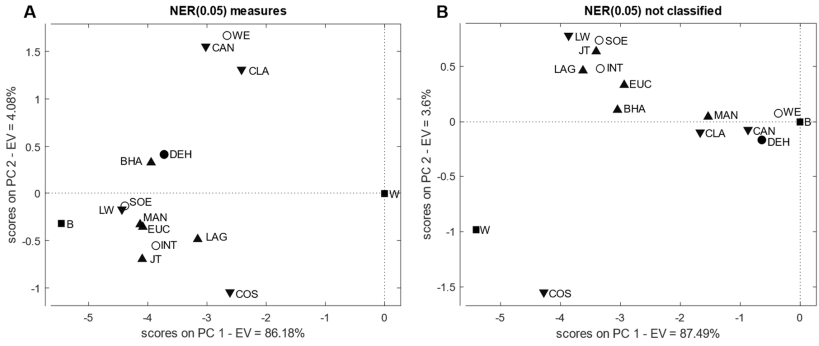


Figure 5. The score plot of the first two PCs for the metrics based on the NER(0.05) measures (A) and not classified objects (B). B and W are the theoretical best and worst metrics.

The analysis of the behaviour of the metrics with respect to the percentage of not-classified objects (Figure 5B) was also performed with the purpose to highlight the metrics giving, in the average, the minimum number of not-classified objects. Two cases B (best) and W (worst) are added to the 13 metrics representing the two extreme behaviours: B is always zero implying

that no not-classified object is given, while W is 24-dimensional vector with the maximum percentage of not-classified objects for each data set. The first component of PCA explains the 87.5% of the total variance and the three metrics nearest to the best are DEH, WE and CAN, indicating that these metrics, on average, give a small percentage of not-classified objects. CLA and MAN are not so far from the first ones. Metrics that, on average, give high percentages of not-classified objects are COS, LW, LAG, SOE, JT, INT, followed by BHA and EUC metrics. The COS metric appears isolated on the bottom of the figure being its standard deviation twice the other greatest standard deviations (last row of Appendix E).

The same approach was applied to a matrix where the row are the 13 metrics, the column are the 31 data sets and each entry is the calculated L measure (Appendix G). Considerations about this plot are analogous to those given for Fig. 5A.

7 Conclusions

In general, both link-based and node-based measures show good performances for classification purposes and can be considered as a possible reliable complement to other classical classification tools.

Moreover, the use of 13 different metrics enlarges the possibilities to obtain reliable models, also increasing the knowledge about the optimal topological space to perform classification. In particular, the Dehmer metric has been here extended into a multivariate distance measure and compared with other classical metrics. Manhattan, Lance-Williams, Soergel and Dehmer metrics seem to perform quite well in several cases. Among the best metrics, Dehmer and Manhattan give, on average, small reasonable percentages of not-classified objects.

A not marginal advantage of this approach, shared with the SOM method used for classification, is that the results can be graphically analysed by the obtained spanning tree. Moreover, being a semi-supervised classification method, an informative representation of a class is not necessary, since in this case a class can be also represented by a singleton. Indeed, in the link-based approach, if the object appears as a terminal node, being the number of classes minus one subtracted to the total link errors, it is correctly viewed as a singleton. This kind of classification can be also possible also for the node-based approach, but at least two objects of the same class must be present in a terminal chain.

References

[1] K. Varmuza, P. Filzmoder, *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, Boca Raton, 2009.

[2] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Springer, New York, 2008.

[3] R. Todeschini, D. Ballabio, M. Cassotti, V. Consonni, N3 and BNN: Two new similarity based classification methods in comparison with other classifiers, *J. Chem. Inf. Model.* **55** (2015) 2365–2375.

[4] C. Zhou, L. Wan, Y. Liang, A hybrid algorithm of minimum spanning tree and nearest neighbor for classifying human cancers, ICACTE 2010 - 2010 3rd Int. Conf. Adv. *Comput. Theory Eng. Proc.* **5** (2010) V5-585-V5-589.

[5] P. Juszczak, D.M.J. Tax, E. Pękalska, R.P.W. Duin, Minimum spanning tree based one-class classifier, *Neurocomputing* **72** (2009) 1859–1869.

[6] F. Vitale, N. Cesa-Bianchi, C. Gentile, G. Zappella, See the tree through the lines: The SHAZOO algorithm, *Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011*, 2011, pp. 1–9.

[7] A. Chakrabarty, S. Roy, An optimized k-NN classifier based on minimum spanning tree for email filtering, *2014 2nd Int. Conf. Bus. Inf. Manag. ICBIM 2014*, 2014, pp. 47–52.

[8] R. La Grassa, I. Gallo, A. Calefati, D. Ognibene, Binary classification using pairs of minimum spanning trees or n-ary trees, in: M. Vento, G. Percannella (Eds.), *Computer Analysis of Images and Patterns*, Springer, New York, 2019, pp. 365–376.

[9] Y. M. Zhang, K. Huang, C. L. Liu, Fast and robust graph-based transductive learning via minimum tree cut, *Proc. - IEEE Int. Conf. Data Mining, ICDM*, (2011) 952–961.

[10] J. Zupan, M. Novič, I. Ruisánchez, Kohonen and counterpropagation artificial neural networks in analytical chemistry, *Chemom. Intell. Lab. Syst.* **38** (1997) 1–23.

[11] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, J. B. O. Mitchell, Melting point prediction employing *k*-nearest neighbor algorithms and genetic parameter optimization, *J. Chem. Inf. Model.* **46** (2006) 2412–2422.

[12] R. A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.* **7** (1936) 179–188.

[13] M. Forina, C. Armanino, M. Castino, M. Ubigli, Multivariate data analysis as discriminating method of the origin of wines, *Vitis* **25** (1986) 189–201.

[14] M. Forina, *Artificial Data Set*, Univ. Genoa, Genoa, 2005.

[15] M. Forina, C. Armanino, S. Lanteri, E. Tiscornia, Classification of olive oils from their fatty acid composition, in: H. Martens, H. Russwurm, Jr. (Eds.), *Food Research and Data Analysis*, Appl. Sci. Pub., London, 1983, pp. 189–214.

[16] Y. Miyashita, Y. Takahashi, C. Takayama, T. Ohkubo, K. Funatsu, S. I. Sasaki, Computer-assisted structure/taste studies on sulfamates by pattern recognition methods, *Anal. Chim. Acta.* **184** (1986) 143–149.

[17] M. J. Benito, M. C. Ortiz, M. S. Sànchez, L. A. Sarabia, M. Iñiguez, Typification of vinegars from Jerez and Rioja using classical chemometric techniques and neural network methods, *Analyst* **124** (1999) 547–552.

[18] P. Resmini, L. Pellegrino, M. Bertuccioli, Moderni criteri per la valutazione chimico-analitica della tipicità di un formaggio: l'esempio del Parmigiano-Reggiano, *Riv. Soc. It. Sci. Alim.* **15** (1986) 315–326.

[19] C. Armanino, R. Leardi, S. Lanteri, G. Modi, Chemometric analysis of tuscan olive oils, *Chemom. Intell. Lab. Syst.* **5** (1989) 343–354.

[20] H. Streuli, Mathematische Modelle für die chemische Zusammensetzung von Libensmitteln und ihre Bedeutung für deren Beurteilung, *Lebensmittel-Technologie* **20** (1987) 203–211.

[21] R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan, CAIMAN (Classification And Influence Matrix Analysis): A new approach to the classification based on leverage-scaled functions, *Chemom. Intell. Lab. Syst.* **87** (2007) 3–17.

[22] D. Brodnjak-Voncina, Z. C. Kodb, M. Novič, Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids, *Chemom. Intell. Lab. Syst.* **75** (2005) 31–43.

[23] P. M. Gerrild, R. J. Lantz, *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.

[24] D. Ballabio, V. Consonni, F. Costa, Relationships between apple texture and rheological parameters by means of multivariate analysis, *Chemom. Intell. Lab. Syst.* **111** (2012) 28–33.

[25] M. Forina, Tobacco Data Set, (n.d.).

[26] A. P. Worth, M. T. D. Cronin, Embedded cluster modelling - A novel method for analysing embedded data sets, *Quant. Struct. Rel.* **18** (1999) 229–235.

[27] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, *Proc. Symp. Comput. Appl. Med. Care.* **9** (1988) 261–265.

[28] P. P. Mager, *Design Statistics in Pharmacochemistry*, Res. Stud. Press, Letchworth, 1991.

[29] L. J. Hamilton, Cross-shelf colour zonation in northern Great Barrier Reef lagoon surficial sediments, *Aust. J. Earth Sci.* **48** (2001) 193–200.

[30] U.C. Centre, Heart disease Data Set, (n.d.). https://myweb.uiowa.edu/pbreheny/uk/764/data/heart.txt.

[31] 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** (2013) 867–878.

[32] R. A. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall/Pearson, 1992.

[33] M. C. Ortiz, J. A. Saez, J. L. Palacios, Typification of alcoholic distillates by multivariate techniques using data from chromatographic analyses, *Analyst* **118** (1993) 801–805.

[34] C. Armanino, S. Lanteri, M. Forina, A. Balsamo, M. Migliardi, G. Cenderelli, Hirsutism: a multivariate approach of feature selection and classification, *Chemom. Intell. Lab. Syst.* **5** (1989) 335–341.

[35] A. Saviozzi, G. Lotti, D. Piacentini, La Composizione Amminoacidica Delle Farine Di Girasole, *Riv. Soc. It. Sci. Alim.* **15** (1986) 437–444.

[36] K. A. Baggerly, S. Morris, S. R. Edmonson, K. R. Coombes, Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer, *J. Nat. Canc. Inst.* **97** (2005) 307–309.

[37] E. Berthonnaud, J. Dimnet, P. Roussouly, H. Labelle, Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters, *J. Spinal Disord. Tech.* **18** (2005) 40–47.

[38] J. D. F. Habbema, J. Hermans, K. Van den Broek, A step-wise discriminant analysis program using density estimation, *Proc. Comput. Stat.* (1974) 101–110.

[39] M. Forina, C. Armanino, S. Lanteri, Acidi grassi degli animali acquatici: uno studio chemiometrico, *Riv. Soc. It. Sci. Alim.* **11** (1982) 15–22.

[40] M. Alvarez-Guerra, D. Ballabio, J. M. Amigo, J. R. Viguri, R. Bro, A Chemometric approach to the environmental problem of predicting toxicity in contaminated sediments, *J. Chemom.* **24** (2010) 379–386.

[41] MATLAB R2018b, (2018). https://mathworks.com/products/matlab.html.

[42] V. Batagelj, A. Mrvar, Pajek, (n.d.).

[43] M. M. Deza, E. Deza, *Encyclopedia of Distances*, Springer, Dordrecht, 2009.

[44] M. Dehmer, F. Emmert-Streib, Y. Shi, Interrelations of graph distance measures based on topological indices, *PLoS One* **9** (2014) #94985.

Appendix A

The overall classification rate (T) values for each metric and each data set. In bold, the best results of the T measure.

DATA	EUC	CAN	LW	MAN	LAG	CLA	SOE	BHA	WE	JT	COS	DEH	INT
IRIS	95.3	95.3	95.3	94.6	**96.6**	94.0	95.3	94.6	95.3	95.3	76.5	94.6	94.0
WINES	94.9	96.6	94.4	96.6	92.7	**98.3**	94.4	96.6	95.5	93.8	96.6	96.6	94.9
PERPOT	**99.0**	93.9	94.9	**99.0**	97.0	94.9	94.9	98.0	93.9	96.0	84.8	**99.0**	96.0
ITAOLIS	95.8	94.0	**96.1**	96.0	94.4	93.0	**96.1**	95.1	94.0	96.0	95.8	96.0	96.0
SULFA	75.5	75.5	**77.6**	75.5	73.5	**77.6**	**77.6**	75.5	73.5	73.5	67.3	75.5	**77.6**
VINAGRES	96.9	98.5	**100.0**	98.5	93.8	98.5	**100.0**	98.5	98.5	**100.0**	98.5	98.5	**100.0**
CHEESE	82.0	**83.5**	82.7	82.0	82.0	76.7	82.7	80.5	81.2	82.7	82.7	82.7	76.7
OLITOS	79.8	74.8	79.8	80.7	73.9	69.7	79.8	79.8	75.6	**82.4**	79.8	76.5	79.8
COFFEE	**100.0**	97.6	**100.0**	**100.0**	**100.0**	97.6	**100.0**	97.6	97.6	**100.0**	**100.0**	**100.0**	**100.0**
DIGITS	63.9	62.1	**65.5**	65.1	63.3	63.1	**65.5**	62.5	61.1	64.7	62.3	63.7	63.3
VEGOIL	**100.0**	97.6	**100.0**	**100.0**	98.8	92.7	**100.0**	**100.0**	98.8	**100.0**	**100.0**	**100.0**	98.8
CRUDEOIL	**92.7**	81.8	85.5	87.3	89.1	83.6	85.5	90.9	80.0	89.1	**92.7**	89.1	87.3
APPLE	91.5	**93.7**	92.3	92.3	90.5	92.5	92.3	93.1	93.5	90.7	90.7	91.5	90.3
TOBACCO	84.0	84.0	80.0	84.0	**88.0**	84.0	80.0	**88.0**	84.0	80.0	84.0	**88.0**	80.0
METACYCLINE	76.2	**85.7**	81.0	81.0	76.2	81.0	81.0	76.2	**85.7**	81.0	76.2	81.0	71.4
DIABETES	70.1	64.1	69.8	70.1	69.2	62.8	69.8	67.5	65.6	68.8	65.2	**70.3**	67.9
THIOPHENE	**87.0**	73.9	**87.0**	**87.0**	73.9	73.9	**87.0**	78.3	73.9	**87.0**	78.3	**87.0**	82.6
SAND	90.0	88.8	88.8	91.3	91.3	91.3	88.8	91.3	88.8	88.8	61.3	90.0	**92.5**
HEARTHDISEASE	63.8	60.7	63.6	**64.6**	63.3	62.3	63.6	60.3	61.0	62.9	62.0	62.9	63.6
BIODEG	81.6	81.5	**82.8**	82.3	79.2	81.3	**82.8**	82.1	80.4	81.7	80.5	81.0	81.2
SCHOOL	96.4	94.0	95.2	95.2	95.2	94.0	95.2	**97.6**	94.0	95.2	50.0	**97.6**	96.4
ORUJOS	**97.5**	95.0	94.1	95.0	95.0	93.3	94.1	95.8	95.0	**97.5**	**97.5**	95.8	95.0
HIRSUTISM	83.3	82.6	86.4	85.6	81.1	84.8	86.4	84.8	81.8	84.1	85.6	86.4	**87.1**
SUNFLOWERS	89.9	85.5	89.9	89.9	89.9	85.5	89.9	91.3	85.5	**92.8**	84.1	89.9	**92.8**
BLOOD	67.3	66.3	67.6	67.6	66.3	65.5	67.5	67.2	66.5	66.0	65.5	**68.0**	66.9
VERTEBRAL	79.6	**82.2**	77.3	77.3	76.1	79.9	77.3	79.3	80.9	78.0	78.3	79.9	76.4
BANK	82.2	**86.7**	84.4	82.2	82.2	80.0	84.4	84.4	**86.7**	80.0	68.9	82.2	80.0
MEMBRANE	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	42.9	**94.3**	**94.3**
HEMOPHILIA	78.4	79.7	**83.8**	**83.8**	78.4	78.4	**83.8**	79.7	78.4	82.4	81.1	78.4	**83.8**
FISH	88.5	**92.3**	88.5	**92.3**	80.8	**92.3**	88.5	84.6	88.5	88.5	88.5	84.6	**92.3**
SEDIMENT	89.9	89.9	**90.4**	**90.4**	89.9	89.6	**90.4**	89.5	90.2	89.8	89.9	89.7	89.4

Appendix B

The link-based non-error rate (L measure) values for each metric and each data set. In bold, the absolute best results.

DATA	EUC	CAN	LW	MAN	LAG	CLA	SOE	BHA	WE	JT	COS	DEH	INT
IRIS	95.3	95.3	95.3	94.6	**96.6**	94.0	95.3	94.6	95.3	95.3	76.5	94.6	94.0
WINES	95.0	96.9	94.7	96.7	92.7	**98.4**	94.7	96.7	95.8	94.0	96.7	96.7	95.2
PERPOT	**99.0**	93.9	94.9	**99.0**	97.0	94.9	94.9	98.0	93.9	96.0	84.8	**99.0**	96.0
ITAOLIS	95.2	92.5	**95.7**	95.4	93.5	91.2	**95.7**	94.2	92.8	95.4	95.1	95.2	95.2
SULFA	71.7	70.6	72.5	71.7	70.7	**73.6**	72.5	71.7	68.8	67.5	59.1	72.6	**73.6**
VINAGRES	95.8	98.8	**100**	98.8	91.3	98.8	**100**	98.8	98.8	**100**	98.8	98.8	**100**
CHEESE	76.1	**78.8**	76.3	75.0	77.1	68.8	76.3	74.3	75.0	76.0	**78.8**	75.3	66.8
OLITOS	77.4	71.9	75.6	78.0	71.4	68.2	75.6	78.4	70.2	**80.5**	76.8	70.9	79.6
COFFEE	**100**	95.1	**100**	**100**	**100**	95.1	**100**	95.1	95.1	**100**	**100**	**100**	**100**
DIGITS	69.2	66.5	**70.3**	69.5	69.3	67.6	**70.3**	68.2	65.9	70.1	67.8	69.0	68.7
VEGOIL	**100**	97.6	**100**	**100**	99.2	93.5	**100**	**100**	99.2	**100**	**100**	**100**	99.2
CRUDEOIL	**88.9**	68.7	75.9	78.8	83.0	72.0	75.9	86.1	64.7	82.7	**88.9**	82.4	78.8
APPLE	88.7	**91.7**	89.8	89.6	87.3	90.3	89.8	90.9	91.5	87.7	88.0	88.5	87.3
TOBACCO	83.9	84.0	80.0	84.0	**88.0**	84.0	80.0	**88.0**	84.0	80.0	84.0	**88.0**	80.0
METACYCLINE	76.1	**85.7**	81.0	81.0	76.1	80.8	81.0	76.1	**85.7**	81.0	76.1	81.0	71.4
DIABETES	**66.5**	60.1	66.2	66.0	65.7	58.9	66.2	63.7	61.6	65.7	62.0	65.6	64.6
THIOPHENE	**87.1**	79.3	86.8	**87.1**	77.8	79.3	86.8	82.1	79.1	86.8	78.3	**87.1**	86.4
SAND	89.5	88.3	88.2	90.9	90.9	90.8	88.2	90.9	88.3	88.2	60.4	89.5	**92.2**
HEARTHDISEASE	59.4	57.2	59.0	**59.6**	59.4	58.4	59.0	56.1	57.6	59.1	58.0	57.0	59.1
BIODEG	80.0	79.6	**81.3**	80.9	78.0	79.6	**81.3**	80.4	78.7	80.0	78.9	79.4	79.6
SCHOOL	96.4	94.0	95.1	95.1	95.2	93.9	95.1	**97.5**	94.0	95.1	50.9	**97.5**	96.3
ORUJOS	**96.5**	92.9	91.6	92.9	92.7	90.5	91.6	94.2	92.9	**96.5**	**96.5**	94.2	93.1
HIRSUTISM	78.4	76.3	80.8	80.8	75.2	79.2	80.8	79.2	75.0	77.4	78.5	81.6	**81.7**
SUNFLOWERS	88.6	84.7	88.8	88.6	88.8	84.4	88.8	90.7	84.2	**92.0**	83.0	88.6	**92.0**
BLOOD	**57.3**	55.3	56.8	57.1	55.3	54.9	56.7	55.7	55.8	55.4	54.1	57.2	55.6
VERTEBRAL	77.1	**80.1**	74.6	74.5	73.0	77.1	74.6	76.5	78.5	75.4	75.8	77.4	73.9
BANK	82.2	**86.6**	84.4	82.2	82.2	79.9	84.4	84.4	**86.6**	80.0	68.6	82.2	79.9
MEMBRANE	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	**94.3**	45.2	**94.3**	**94.3**
HEMOPHILIA	77.5	78.7	**82.9**	**82.9**	77.2	77.2	**82.9**	78.7	77.5	81.6	80.3	77.2	**82.9**
FISH	88.5	**92.3**	88.4	**92.3**	80.8	**92.3**	88.4	84.6	88.5	88.4	88.4	84.6	**92.3**
SEDIMENT	78.4	78.6	**79.5**	**79.5**	78.8	78.0	**79.5**	77.8	78.9	78.2	79.1	77.9	77.1

Appendix C

The node-based non-error rate (NER(0) measure) values for each metric and each data set without a membership threshold. In bold, the absolute best results.

DATA	EUC	CAN	LW	MAN	LAG	CLA	SOE	BHA	WE	JT	COS	DEH	INT
IRIS	95.3	95.3	96.0	95.3	**97.3**	94.0	96.0	94.0	95.3	94.7	78.7	94.7	94.7
WINES	95.1	97.2	95.7	96.7	93.7	**98.6**	95.7	97.0	97.2	95.7	96.9	98.1	95.2
PERPOT	**100**	96.0	98.0	99.0	99.0	98.0	98.0	**100**	96.0	99.0	82.0	98.0	99.0
ITAOLIS	**94.7**	92.6	93.7	94.5	93.2	91.9	93.7	93.5	92.4	**94.7**	94.2	93.9	**94.7**
SULFA	73.8	76.6	72.4	78.8	**79.6**	67.5	72.4	67.5	71.6	68.8	58.9	75.2	71.6
VINAGRES	91.7	98.7	**100**	**100**	91.7	98.7	**100**	**100**	98.7	95.8	**100**	**100**	**100**
CHEESE	72.5	77.2	74.2	73.9	74.2	67.1	74.2	73.3	73.3	70.1	**79.0**	72.5	65.6
OLITOS	65.2	67.1	66.1	71.9	67.5	67.4	66.1	70.9	64.1	74.5	69.9	63.4	**75.8**
COFFEE	**100**	92.9	**100**	**100**	**100**	92.9	**100**	92.9	92.9	**100**	**100**	**100**	**100**
DIGITS	65.0	61.9	65.6	65.3	64.3	61.9	65.6	62.2	61.8	65.4	62.9	**66.6**	64.4
VEGOIL	**100**	**100**	**100**	**100**	99.0	94.8	**100**	**100**	**100**	**100**	**100**	**100**	99.0
CRUDEOIL	83.1	63.6	73.6	74.4	76.2	70.1	73.6	80.1	61.4	82.2	**87.9**	81.4	74.4
APPLE	89.6	**93.0**	89.8	90.2	88.3	91.8	89.8	91.6	92.6	88.8	89.2	89.0	87.9
TOBACCO	76.9	76.9	76.9	80.8	76.9	80.8	76.9	80.8	76.9	76.9	**84.6**	80.8	76.9
METACYCLINE	82.5	**91.7**	**91.7**	82.5	82.5	87.5	**91.7**	82.5	87.5	90.8	82.5	86.7	82.5
DIABETES	68.2	61.2	66.8	66.7	68.2	59.7	66.8	64.0	61.9	**68.3**	63.4	65.9	66.7
THIOPHENE	**79.2**	70.8	**79.2**	**79.2**	70.8	70.8	**79.2**	75.0	75.0	**79.2**	70.8	**79.2**	**79.2**
SAND	87.3	86.9	88.4	88.4	89.9	86.9	88.4	87.3	86.9	87.3	66.0	88.4	**90.9**
HEARTHDISEASE	59.3	58.3	57.7	59.8	56.6	58.2	57.7	59.6	58.1	**60.1**	58.5	56.8	59.7
BIODEG	82.3	81.1	82.4	82.5	80.7	80.6	82.4	**82.8**	81.1	**82.8**	80.6	81.0	82.1
SCHOOL	92.9	94.0	94.0	94.0	92.9	91.5	94.0	**95.1**	94.0	94.0	50.0	**95.1**	92.7
ORUJOS	**100**	98.2	97.7	97.7	94.1	94.1	97.7	98.2	**100**	**100**	**100**	**100**	97.1
HIRSUTISM	82.4	83.8	89.1	84.8	76.2	85.8	89.1	87.7	78.1	79.0	79.9	**89.6**	89.1
SUNFLOWERS	90.0	88.5	91.2	93.1	90.8	83.6	91.2	90.0	87.8	92.0	84.4	89.2	**95.8**
BLOOD	55.9	55.9	55.8	56.6	54.7	55.3	55.8	54.5	**57.0**	54.4	53.4	56.7	56.4
VERTEBRAL	77.7	**81.8**	75.7	75.7	76.9	77.4	75.7	77.9	80.6	77.2	76.6	77.9	76.7
BANK	78.9	80.5	84.9	80.9	82.9	78.5	84.9	76.5	80.5	76.5	64.2	80.9	**86.9**
MEMBRANE	94.4	**97.2**	94.4	94.4	94.4	94.4	94.4	94.4	**97.2**	94.4	47.2	94.4	91.7
HEMOPHILIA	79.4	80.0	**85.0**	83.9	77.8	78.3	**85.0**	80.6	80.0	78.3	**85.0**	76.7	81.7
FISH	89.0	92.9	92.9	92.9	81.6	89.0	92.9	89.0	81.6	92.9	85.4	85.4	**100**
SEDIMENT	85.4	85.9	85.9	85.9	85.1	83.4	85.9	84.9	85.8	85.0	**86.6**	**86.6**	83.8

Appendix D

The weighted node based non-error rate (NER(0.05) measure) for each metric and each data set by using a membership threshold of 0.05. In bold, the absolute best results.

DATA	EUC	CAN	LW	MAN	LAG	CLA	SOE	BHA	WE	JT	COS	DEH	INT
IRIS	95.9	96.0	96.6	95.9	**97.2**	95.2	96.6	95.9	96.0	95.9	79.4	95.8	95.3
WINES	95.9	97.2	96.4	97.6	95.0	**99.5**	96.4	97.3	97.2	96.5	98.0	98.1	96.9
PERPOT	**100**	95.8	98.9	**100**	99.0	98.0	98.9	**100**	95.9	99.0	91.6	**100**	**100**
ITAOLIS	94.8	93.6	94.5	94.7	94.2	92.0	94.5	93.9	92.8	95.1	94.6	94.6	**95.5**
SULFA	77.1	75.1	81.6	78.9	**82.4**	71.9	78.5	76.9	73.2	71.0	53.4	77.5	73.9
VINAGRES	95.2	**100**	**100**	**100**	91.7	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
CHEESE	72.7	76.9	73.4	75.4	80.3	67.7	73.4	74.7	73.3	73.5	**81.3**	73.4	66.0
OLITOS	70.1	67.1	72.1	70.8	73.3	68.3	72.1	73.5	64.1	77.6	73.2	63.4	**81.0**
COFFEE	**100**	92.9	**100**	**100**	**100**	92.9	**100**	92.9	92.9	**100**	**100**	**100**	**100**
DIGITS	68.0	64.7	**70.2**	69.1	66.2	65.0	69.2	65.9	63.9	68.7	64.8	70.0	68.0
VEGOIL	**100**	**100**	**100**	**100**	99.0	96.8	**100**	**100**	**100**	**100**	**100**	**100**	99.0
CRUDEOIL	83.1	63.6	73.1	74.4	76.7	71.9	73.1	78.6	62.1	**90.0**	**90.0**	81.4	72.9
APPLE	91.5	93.4	92.1	90.4	89.1	92.4	92.1	**93.5**	93.0	90.0	91.4	89.2	89.2
TOBACCO	80.1	76.9	80.9	80.8	**88.5**	80.8	80.9	87.5	76.9	80.9	87.1	80.8	79.2
METACYCLINE	80.8	91.7	**95.0**	85.0	80.0	87.5	**95.0**	81.4	87.5	89.9	81.9	86.7	81.3
DIABETES	**69.7**	61.6	68.8	68.4	69.6	60.7	68.8	66.3	62.0	68.7	64.8	65.9	67.6
THIOPHENE	85.7	76.6	85.7	85.7	77.8	76.6	85.7	**87.8**	75.0	85.7	**87.8**	82.7	83.8
SAND	91.3	87.8	91.4	92.8	91.8	91.0	91.4	92.9	87.8	92.6	64.6	89.2	**93.1**
HEARTHDISEASE	63.2	57.8	60.5	61.3	61.2	58.3	60.5	59.3	58.0	**63.5**	59.9	56.7	61.0
BIODEG	84.9	82.1	85.5	84.8	83.2	82.3	85.6	**85.9**	81.8	85.5	83.2	82.0	84.5
SCHOOL	96.1	93.8	96.1	93.8	97.4	92.3	96.2	**98.6**	93.8	96.2	50.8	97.3	94.7
ORUJOS	**100**	98.2	99.5	97.6	97.4	95.8	99.5	**100**	**100**	**100**	**100**	**100**	97.1
HIRSUTISM	82.3	85.2	89.2	83.9	78.8	85.6	89.2	87.2	78.9	78.9	80.6	**89.3**	88.8
SUNFLOWERS	94.5	88.5	94.6	92.9	91.6	84.8	94.6	92.8	87.8	94.6	85.6	89.2	**96.9**
BLOOD	**58.1**	56.2	56.7	57.7	55.1	55.9	57.2	56.0	56.9	55.9	55.0	57.1	56.7
VERTEBRAL	82.9	**83.0**	78.5	76.9	76.2	79.2	78.5	79.6	81.4	80.2	81.0	78.6	80.1
BANK	83.3	81.6	87.2	82.5	**87.6**	80.9	85.0	85.4	79.6	81.0	69.9	84.2	85.7
MEMBRANE	97.0	97.2	97.0	97.0	97.0	97.0	97.0	97.0	97.2	97.0	42.9	97.0	**100**
HEMOPHILIA	82.6	80.1	**87.1**	**87.1**	81.5	81.4	**87.1**	84.8	79.2	83.9	86.0	80.4	86.7
FISH	92.0	92.3	96.2	96.2	84.7	89.0	96.2	88.5	81.6	92.3	92.3	88.5	**100**
SEDIMENT	85.4	86.1	85.2	86.4	85.2	84.2	85.5	84.6	**86.7**	84.2	86.2	**86.7**	82.7

Appendix E

The percentage of not-classified objects compared with the PLS/DA. Bold character highlights the percentages of not-classified objects corresponding to the best NER results.

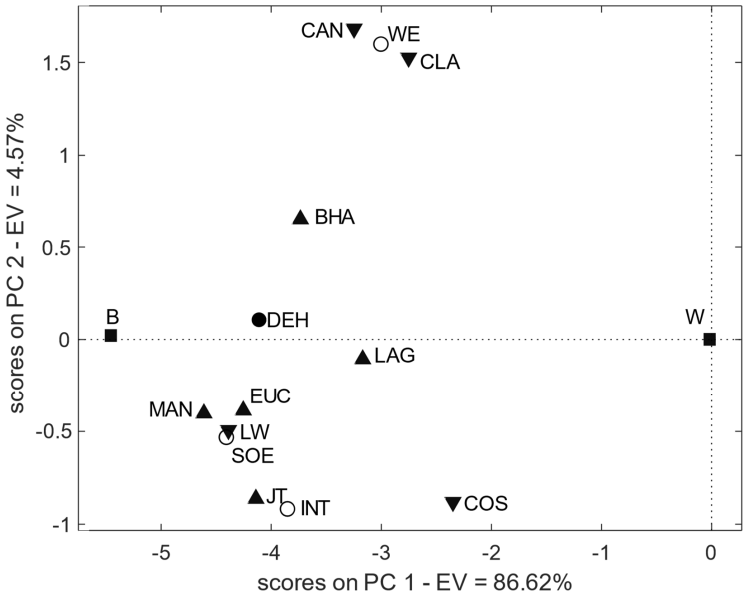
DATA	EUC	CAN	LW	MAN	LAG	CLA	SOE	BHA	WE	JT	COS	DEH	INT	PLS/DA
IRIS	2.7	**0.7**	1.3	1.3	3.3	2.0	1.3	3.3	**0.7**	2.7	14.0	3.3	1.3	23.2
WINES	1.1	**0.0**	3.9	1.1	3.9	1.7	2.8	2.8	**0.0**	2.8	2.2	0.6	3.4	0
PERPOT	2.0	4.0	9.0	3.0	3.0	**1.0**	5.0	4.0	3.0	2.0	17.0	3.0	5.0	0
ITAOLIS	3.3	2.4	3.0	1.4	3.7	3.1	3.0	3.1	**1.2**	3.3	3.3	**1.2**	4.0	0
SULFA	14.0	**4.0**	22.0	**4.0**	14.0	6.0	18.0	14.0	8.0	14.0	18.0	**4.0**	14.0	0
VINAGRES	1.5	1.5	1.5	**0.0**	3.0	1.5	1.5	1.5	1.5	1.5	1.5	**0.0**	1.5	4.5
CHEESE	6.7	0.7	7.5	1.5	11.2	0.7	7.5	3.7	**0.0**	6.0	10.4	0.7	12.7	20.9
OLITOS	10.0	**0.0**	12.5	1.7	12.5	5.8	12.5	7.5	**0.0**	10.0	15.0	**0.0**	10.0	18.3
COFFEE	**0.0**	**0.0**	4.7	**0.0**	**0.0**	**0.0**	2.3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	2.3	0
DIGITS	11.2	7.8	15.4	12.6	10.0	7.4	12.4	15.0	**6.6**	11.8	11.6	14.0	11.8	88
VEGOIL	**0.0**	**0.0**	2.4	**0.0**	3.6	1.2	1.2	**0.0**	**0.0**	**0.0**	1.2	**0.0**	1.2	3.6
CRUDEOIL	1.8	**0.0**	8.9	**0.0**	5.4	5.4	8.9	7.1	1.8	8.9	5.4	**0.0**	7.1	12.5
APPLE	3.7	0.8	3.5	1.2	4.9	1.4	3.5	3.7	0.8	3.5	5.3	**0.4**	4.1	0
TOBACCO	7.7	**0.0**	19.2	**0.0**	15.4	**0.0**	19.2	7.7	**0.0**	19.2	11.5	**0.0**	7.7	0
METACYCLINE	9.1	**0.0**	13.6	9.1	9.1	**0.0**	9.1	4.5	**0.0**	9.1	4.5	**0.0**	9.1	0
DIABETES	11.2	6.8	11.1	9.0	10.9	10.4	10.8	10.9	6.0	11.2	15.6	**3.6**	10.2	0
THIOPHENE	8.3	12.5	12.5	12.5	16.7	12.5	12.5	25.0	**0.0**	8.3	25.0	4.2	25.0	16.7
SAND	7.4	**2.5**	7.4	7.4	3.7	7.4	7.4	7.4	**2.5**	11.1	37.0	**2.5**	4.9	0
HEARTHDISEASE	14.5	5.6	16.0	11.0	13.0	9.5	15.8	12.1	5.0	15.8	14.3	**1.7**	13.2	0
BIODEG	9.1	3.1	9.1	6.3	8.4	4.2	8.8	6.6	**1.6**	9.1	9.1	2.7	9.4	0
SCHOOL	7.1	2.4	7.1	3.5	8.2	4.7	4.7	7.1	**1.2**	4.7	43.5	2.4	5.9	30.6
ORUJOS	2.5	**0.0**	3.3	1.7	4.2	0.8	2.5	0.8	0.8	1.7	1.7	0.8	0.8	0
HIRSUTISM	6.8	3.0	4.5	2.3	8.3	3.8	4.5	9.0	**1.5**	8.3	8.3	**1.5**	6.0	0
SUNFLOWERS	7.1	**0.0**	7.1	1.4	4.3	2.9	5.7	5.7	**0.0**	5.7	7.1	**0.0**	1.4	0
BLOOD	11.0	6.4	9.8	10.6	8.7	7.1	9.0	11.1	**4.9**	9.4	10.6	10.0	8.3	0
VERTEBRAL	10.6	4.5	7.7	6.1	8.4	5.5	7.7	9.0	**2.9**	11.6	11.6	3.9	13.2	0
BANK	8.7	6.5	15.2	13.0	13.0	8.7	13.0	10.9	**4.3**	8.7	19.6	**4.3**	8.7	0
MEMBRANE	11.1	8.3	13.9	11.1	11.1	13.9	11.1	13.9	**2.8**	11.1	38.9	8.3	16.7	41.7
HEMOPHILIA	10.7	9.3	9.3	9.3	12.0	13.3	9.3	12.0	**8.0**	12.0	18.7	12.0	10.7	0
FISH	7.4	7.4	7.4	3.7	7.4	**0.0**	7.4	3.7	**0.0**	11.1	7.4	3.7	7.4	0
SEDIMENT	3.0	2.0	3.9	1.6	3.8	3.1	4.0	3.5	1.4	4.0	4.2	**1.3**	4.7	0
AVERAGE	6.8	3.3	8.8	4.8	7.9	4.7	7.8	7.3	2.1	7.7	14.3	2.9	7.8	8.4
STD. DEVIATION	4.1	3.4	5.3	4.5	4.3	4.1	4.9	5.4	2.5	4.8	11.1	3.6	5.4	18.3

Appendix F

Comparison of the results obtained by T, L, NER(0) and NER(0.05) measures, and by the NER of wKNN, KNN, N3, PLS/DA, SVM and BNN classification methods, assumed as references. In bold are shown the best results and in italics the best results.

DATA	T	L	NER(0)	NER(0.05)	wKNN	KNN	N3	PLS-DA	SVM	BNN
IRIS	96.6	96.6	**97.3**	97.2	96.7	96.7	96.0	90.2	**97.3**	96.7
WINES	98.3	98.4	98.6	**99.5**	98.0	97.7	96.2	**99.5**	**99.5**	98.6
PERPOT	99.0	99.0	**100**	**100**	99.0	99.0	99.0	86.0	**100**	99.0
ITAOLIS	96.1	95.7	94.7	95.5	94.7	94.7	**96.2**	95.9	95.9	95.2
SULFA	77.6	73.6	79.6	82.4	73.8	73.8	77.4	74.0	**88.7**	73.8
CHEESE	83.5	78.8	79.0	81.0	79.4	78.1	76.1	84.7	**85.6**	78.3
OLITOS	82.4	80.5	75.8	81.0	70.4	70.4	89.1	**94.0**	87.6	73.6
COFFEE	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
VEGOIL	**100**	**100**	**100**	**100**	99.0	99.0	99.0	99.0	**100**	**100**
TOBACCO	88.0	88.0	84.6	88.5	**92.3**	**92.3**	**92.3**	88.5	**92.3**	**92.3**
METACYCLINE	85.7	85.7	91.7	**95.0**	82.5	82.5	82.5	55.8	82.5	86.7
THIOPHENE	87.0	87.1	79.2	87.8	83.3	83.3	83.3	**90.5**	83.3	83.3
SAND	92.5	92.2	90.9	93.1	93.9	93.9	93.9	93.9	**94.9**	**94.9**
HEARTHDISEASE	64.6	59.6	60.1	63.5	63.2	63.2	**69.9**	69.7	68.0	65.2
SCHOOL	97.6	97.5	95.1	**98.6**	96.2	96.2	95.3	89.4	96.4	96.6
ORUJOS	97.5	96.5	**100**	**100**	98.2	98.2	98.2	93.9	98.2	98.4
HIRSUTISM	87.1	81.7	89.6	89.3	90.0	90.0	88.3	84.1	**93.8**	90.1
SUNFLOWERS	92.8	92.0	95.8	**96.9**	91.2	91.2	92.3	92.7	**96.9**	90.4
DIABETES	70.3	66.5	68.3	69.7	70.5	70.5	73.6	**75.1**	72.8	71.1
BLOOD	68.0	57.3	57.0	58.1	63.6	62.3	67.9	**68.7**	64.1	62.2
VERTEBRAL	82.2	80.1	81.8	83.0	80.2	80.2	80.8	82.1	**84.3**	81.6
BIODEG	82.8	81.3	82.8	**85.9**	85.5	85.4	84.5	79.9	83.8	85.3
DIGITS	65.5	70.3	66.6	70.2	73.7	73.6	74.2	41.0	**74.5**	72.3
BANK	86.7	86.6	86.9	87.6	86.9	86.9	86.9	84.9	88.9	**91.2**
VINAGRES	**100**	**100**	**100**	**100**	95.8	95.8	**100**	**100**	**100**	95.8
MEMBRANE	94.3	94.3	97.2	**100**	94.4	94.4	94.4	96.7	94.4	94.4
CRUDEOIL	**92.7**	88.9	87.9	90.0	87.9	87.9	89.2	89.7	84.8	84.8
HEMOPHILIA	83.8	82.9	85.0	**87.1**	82.8	82.8	85.6	85.6	85.6	85.6
FISH	92.3	92.3	**100**	**100**	92.9	92.9	92.6	**100**	**100**	92.9
APPLE	93.7	91.7	93.0	93.5	91.9	91.9	94.0	**95.4**	92.3	92.3
SEDIMENT	**90.4**	79.5	86.6	86.7	89.9	89.9	88.9	79.4	69.9	88.9

Appendix G



The score plot of the first two PCs for the metrics based on the L measures. B and W are the theoretical best and worst metrics.