Pattern Anal Applic (2007) 10:189–201
DOI 10.1007/s10044-007-0061-2

THEORETICAL ADVANCES

An analysis of how training data complexity affects the nearest neighbor classifiers

J. S. Sánchez · R. A. Mollineda · J. M. Sotoca

Received: 14 November 2005 / Accepted: 22 December 2006 / Published online: 16 February 2007

Abstract The k-nearest neighbors (k-NN) classifier is one of the most popular supervised classification methods. It is very simple, intuitive and accurate in a great variety of real-world domains. Nonetheless, despite its simplicity and effectiveness, practical use of this rule has been historically limited due to its high storage requirements and the computational costs involved. On the other hand, the performance of this classifier appears strongly sensitive to training data complexity. In this context, by means of several problem difficulty measures, we try to characterize the behavior of the k-NN rule when working under certain situations. More specifically, the present analysis focuses on the use of some data complexity measures to describe class overlapping, feature space dimensionality and class density, and discover their relation with the practical accuracy of this classifier.

1 Introduction

One of the most widely studied non-parametric classification approaches corresponds to the k-NN decision rule [4]. In brief, given a set of n previously labeled examples (training set, TS), say $\mathcal{X} = \{(x_1, \omega_1), (x_2, \omega_2), \ldots, (x_n, \omega_n)\}$, the k-NN classifier consists of assigning a new input sample $\mathbf{x}$ to the class most frequently represented among the k closest instances in the TS, according to a certain dissimilarity measure (generally, the Euclidean distance metric). A particular case is when $k = 1$, in which an input sample is decided to belong to the class indicated by its closest neighbor.

Several properties make the k-NN classifier quite attractive, including the fact that the asymptotic risk (i.e., when $n \rightarrow \infty$) converges to the optimal Bayes risk as $k \rightarrow \infty$ and $k/n \rightarrow 0$ [3]. If $k = 1$, the upper bound of the classification error rate is approximately twice the optimal Bayes error under the assumption of an infinite number of training examples [4]. The optimal behavior of this rule in asymptotic classification performance along with a conceptual and implementational simplicity make it a powerful classification technique capable of dealing with arbitrarily complex problems, provided that there is a large enough number of training instances available.

However, in many practical situations, such a theoretical maximum can hardly be achieved due to certain inherent weaknesses that significantly reduce the effective applicability of k-NN classifiers. For example, the performance of these rules, as with most non-parametric classification approaches, is extremely sensitive to data complexity. In particular, class overlapping, class density, high data dimensionality and incorrectness or imperfections in the TS can negatively affect the behavior of these classifiers [5, 6]. Also, class imbalance (i.e., high differences in class distributions) has been reported to be an obstacle in applying distance-based algorithms to real-world problems [9, 13, 21].

Analogously, on the other hand, the accuracy of k-NN classifiers significantly drops in domains where many data attributes are irrelevant [17]. Such attributes inappropriately affect the values returned by most dissimilarity metrics. On the other hand, these classifiers cannot be straightforwardly employed in domains with

J. S. Sánchez (✉) · R. A. Mollineda · J. M. Sotoca
Departamento de Llenguatges i Sistemes Informátics,
Universitat Jaume I, Av. Vicent Sos Baynat s/n,
12071 Castelló, Spain
e-mail: sanchez@uji.es

missing attributes [15] because most distance metrics can only be used if each example can be interpreted as a point in the feature space [14]. Another problem using the k-NN rule refers to the seeming necessity of a lot of memory and computational resources (especially, in applications with a huge number of training examples), since it is necessary to search the entire TS to identify the nearest neighbors to the test sample [5].

Friedman [7] establishes that the combination of the bias and variance components of the estimation error can be more significant for classification than the probabilities themselves. This analysis is supported by an evaluation of the k-NN classifier, showing that certain types of very high bias produced by the curse of dimensionality can be compensated by a low variance to produce good classification results. In particular, the experiments study class density and dimensionality in a controlled domain with a very simple decision boundary and no overlapping.

Nevertheless, it is well known that class overlapping negatively affects the performance of the k-NN classifiers, and this has been widely proved in many empirical studies (e.g., see [18]). Analogously, the effect of feature space dimensionality on the k-NN performance has also been extensively investigated in many works. For example, Beyer et al. [2] showed that under certain broad conditions, as dimensionality increases, the distance of the nearest neighbor approaches the distance of the farthest neighbor, that is, the contrast in distances to different data points becomes nonexistent.

This paper focuses on the analysis of three practical scenarios where the application of the k-NN classifiers can become less effective: class overlapping, high data dimensionality and class density. This is not just another work to show how all these situations produce an important degradation in classifier performance. Briefly, we are interested in systematically characterizing the effects of these situations on the k-NN classification results. The aim is to provide analytical measures of overlap, dimensionality and density, and relate them to the expected accuracy of the k-NN classifier. This could help to explain the practical behavior of this decision rule under the above-mentioned conditions.

The rest of the paper is organized as follows. In Sect. 2, we briefly describe a set of data complexity measures that will be used in the subsequent empirical analysis. Section 3 presents the experiments carried out over different synthetic databases and discusses the corresponding results. Also, three real data sets are used in Sect. 3 to validate the conclusions drawn in the case of the artificial domains. Finally, the main conclusions and possible directions for future research are outlined in Sect. 4.

2 Measures of data complexity

In general, the behavior of classifiers is strongly dependent on specific data characteristics. Usual theoretical analyses consist of searching accuracy bounds, most of them supported by impractical conditions. Meanwhile, empirical studies are commonly based on weak comparisons of classifier accuracies over a reduced collection of unexplored data sets. Such studies usually ignore the particular statistical and geometrical descriptions of class distributions to explain classification results. Recently, several research papers [1, 8, 10–12, 19, 20] have introduced the employment of a number of measures to characterize the data complexity (or problem difficulty), thus trying to relate such descriptions to practical classifier performance.

In the present work, some data complexity measures are used to analyze the behavior of the k-NN rule on a number of situations. From an analytical point of view, these measures provide a qualitative description of the training data characteristics and correspondingly, they could help to explain the performance of the k-NN classifier under certain practical conditions.

In this section, we review a number of measures from literature. Most of the them have been originally defined only for two-class discrimination, although in many cases it is possible to produce a generalization for the C-class problem [16]. Moreover, two new measures, the *volume of local neighborhood* and the *class density of overlap region* (see Sect. 2.4 and Sect. 2.5, respectively), will be introduced in the present work.

2.1 Fisher's discriminant ratio (F1)

The plain version of this well-known measure computes how separated are two classes according to a given feature. In other words, it analyzes the effectiveness of a single feature in separating the corresponding problem classes:

$$\mathrm{F1} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}, \tag{1}$$

where μ_1, μ_2, σ_1^2, and σ_1^2 are the means of the two classes and their variances, respectively.

This measure, as defined in Eq. 1, is relative to one feature dimension and two classes. Thus for a multidimensional two-class problem, it is usual to take the maximum value over all the feature dimensions. On the other hand, a possible generalization for C classes, which also considers all feature dimensions, can be stated as follows [16]:

$$\mathrm{F1} = \frac{\sum_{i=1}^{C} n_i \cdot \delta(\mu, \mu_i)}{\sum_{i=1}^{C} \sum_{j=1}^{n_i} \delta(x_j^i, \mu_i)}, \tag{2}$$

where n_i denotes the number of samples in class i, δ is some metric, μ is the overall mean, μ_i corresponds to the mean of class i, and $x^i{}_j$ represents the sample j belonging to class i.

2.2 Volume of overlap region (F2)

This measure [10] computes, for each feature f_h, the length of the overlap range normalized by the length of the total range in which all values of both classes are distributed. Then for a d-dimensional problem, the volume of the overlap region for two classes is obtained as the product of normalized lengths of overlapping ranges for all features.

$$\mathrm{F2} = \prod_{h=1}^{d} \frac{\min\max_h - \max\min_h}{\max\max_h - \min\min_h}, \tag{3}$$

where

$$\begin{aligned} \min\max_h &= \min\{\max(f_h, c_1), \max(f_h, c_2)\} \\ \max\min_h &= \max\{\min(f_h, c_1), \min(f_h, c_2)\} \\ \max\max_h &= \max\{\max(f_h, c_1), \max(f_h, c_2)\} \\ \min\min_h &= \min\{\min(f_h, c_1), \min(f_h, c_2)\} \end{aligned}$$

being max(f_h,c_i) and min(f_h,c_i) the maximum and minimum values of feature f_h in class c_i (i = 1, 2), respectively.

Note that a simple generalization of F2 for the general C-class problem can be defined as a sum of the plain measure (the one given in Eq. 3) for all possible pairs of classes [16].

2.3 Non-parametric separability of classes (N2)

This measure [10] evaluates to what extent two classes are separable by examining the existence and shape of the class boundary. More specifically, N2 is the ratio of the average distance to the intra-class nearest neighbor and the average distance to the inter-class nearest neighbor. It compares the intra-class dispersion with the inter-class separability. It has to be pointed out that smaller values suggest more discriminant data.

Let $\mathcal{N}_1^{=}(x_i)$ and $\mathcal{N}_1^{\neq}(x_i)$ be the intra-class nearest neighbor and the inter-class nearest neighbor of a given example (x_i, ω_i), respectively. Then, N2 can be computed as follows:

$$\mathrm{N2} = \frac{\sum_{i=1}^{n} \delta(\mathcal{N}_1^{=}(x_i), x_i)}{\sum_{i=1}^{n} \delta(\mathcal{N}_1^{\neq}(x_i), x_i)}. \tag{4}$$

2.4 Volume of local neighborhood (D2)

This measure represents the average volume occupied by the k nearest neighbors of each training instance. Since there generally exists an inverse relation between volumen and density, this measure can also be seen as a local estimate of density. Let $\mathcal{N}_k(x_i)$ be the set of the k nearest neighbors of a given example (x_i, ω_i), then the volume of this can be defined as follows:

$$\mathcal{V}_i = \prod_{h=1}^{d} (\max(f_h, \mathcal{N}_k(x_i)) - \min(f_h, \mathcal{N}_k(x_i))), \tag{5}$$

where $\max(f_h, \mathcal{N}_k(x_i))$ and $\min(f_h, \mathcal{N}_k(x_i))$ represent the maximum and minimum values of feature f_h among the k nearest neighbors of instance (x_i, ω_i).

From this, the volume of local neighborhood can be expressed as the average value of $\mathcal{V}_i$ for the n training instances.

$$\mathrm{D2} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{V}_i. \tag{6}$$

2.5 Class density in overlap region (D3)

The aim of this measure is to determine the *relative* density of each class in the overlap regions. In general, overlap regions contain the most critical cases for the classification task and accordingly give rise to most classifier errors. Taking this into account, in the present paper we propose a new measure of class density in overlap regions, namely D3, which is based on the well-known Wilson's editing algorithm [22].

D3 is a measure of the number of points lying in the region of some different class. To this end, we first find the k nearest neighbors of each example (x_i, ω_i). Then if a majority of these k neighbors belong to a class different from ω_i, it can be considered that (x_i, ω_i) lies in an overlap region. One can observe that the higher the value of D3 for a given class, the lower the number of examples from such a class in the overlap region (compared to the total number of instances).

3 Experiments and results

In the present section, we have run a number of experiments on several artificial databases whose characteristics can be fully controlled, allowing better interpretation of the results. Class overlapping, data dimensionality and class density experiments consist of estimating the k-NN classification accuracy (k was set to 1, 3, 5, 7, 9, 11, 13 and 15) and then computing the data complexity measures described in Sect. 2 (F1, F2, N2, D2 and D3), in order to analyze the relation between accuracy and each of the three situations: class overlapping, data dimensionality and class density. For each of these contexts, we have generated five different random data sets, which have then been divided into training and test partitions. Thus, the results here included correspond to the average over the five trials.

3.1 Experiments on class overlapping

To evaluate the effect of class overlapping on the *k*-NN classification accuracy, we generated 11 synthetic databases with different levels of overlapping. Each domain is described by two classes, randomly generated by two-dimensional Gaussian distributions whose means on axis X are: (1) $\mu_1 = 50.0$, $\mu_2 = 50.0$ (that is, absolutely overlapped); (2) $\mu_1 = 49.0$, $\mu_2 = 51.0$; (3) $\mu_1 = 48.0$, $\mu_2 = 52.0$; ...; (11) $\mu_1 = 40.0$, $\mu_2 = 60.0$ (that is, non-overlapped). Here the means on axis Y are irrelevant because, due to the way of defining the distributions, they do not affect overlapping. The covariance matrix for all databases is defined by $\sigma_{11} = 6.0$, $\sigma_{12} = \sigma_{21} = 20.0$, $\sigma_{22} = 36.0$. Each artificial database consists of 4,000 instances (half from class 1 and half from class 2). Figure 1 illustrates four different examples of these databases. Note that only 100 points from each class are represented here in order to make these figures clearer.

Figure 2 depicts the NN and best-*k*-NN classification accuracies for each of the 11 databases as compared to F1, F2, N2, D2 and D3 measures. For D2 and D3, the value of *k* has been set to 5. Labels on axis X represent the distance between the means of the two classes in each data set, going from the absolutely overlapped database ($\mu_2 - \mu_1 = 50 - 50 = 0$) to the non-overlapped set ($\mu_2 - \mu_1 = 60 - 40 = 20$). In the case of D3, we have plotted both the results corresponding to the overall NN and best-*k*-NN accuracies and also those for each class.

As expected, when the overlap between the two classes decreases, the NN and best-*k*-NN classification accuracies increase along with F1 (see Fig. 2a). Note that the lower the value of F1, the higher the overlapping between the two classes. When analyzing the classification results in function of F2 (see Fig. 2b), similar comments to those of F1 can be drawn here . Thus, the effect of class overlapping on the classifier performance can be fully explained by this measure: the lower the class overlapping, the lower will be the value of F2 and, correspondingly, the higher the classification accuracy.

Fig. 1 Examples of the class overlap databases

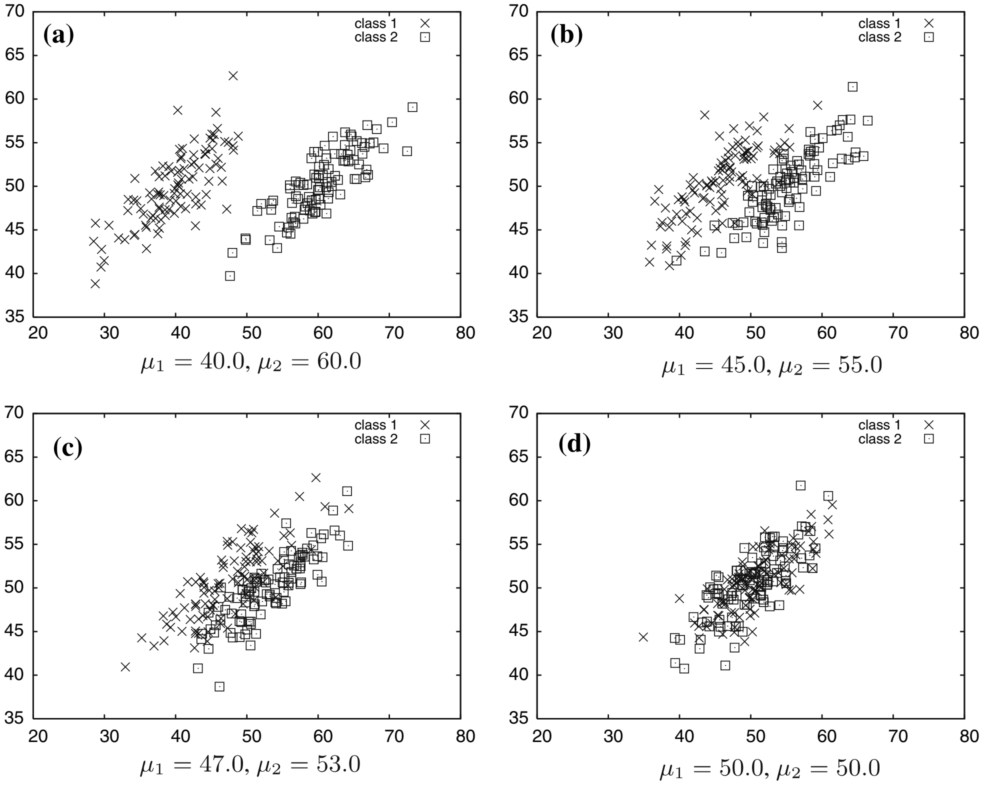
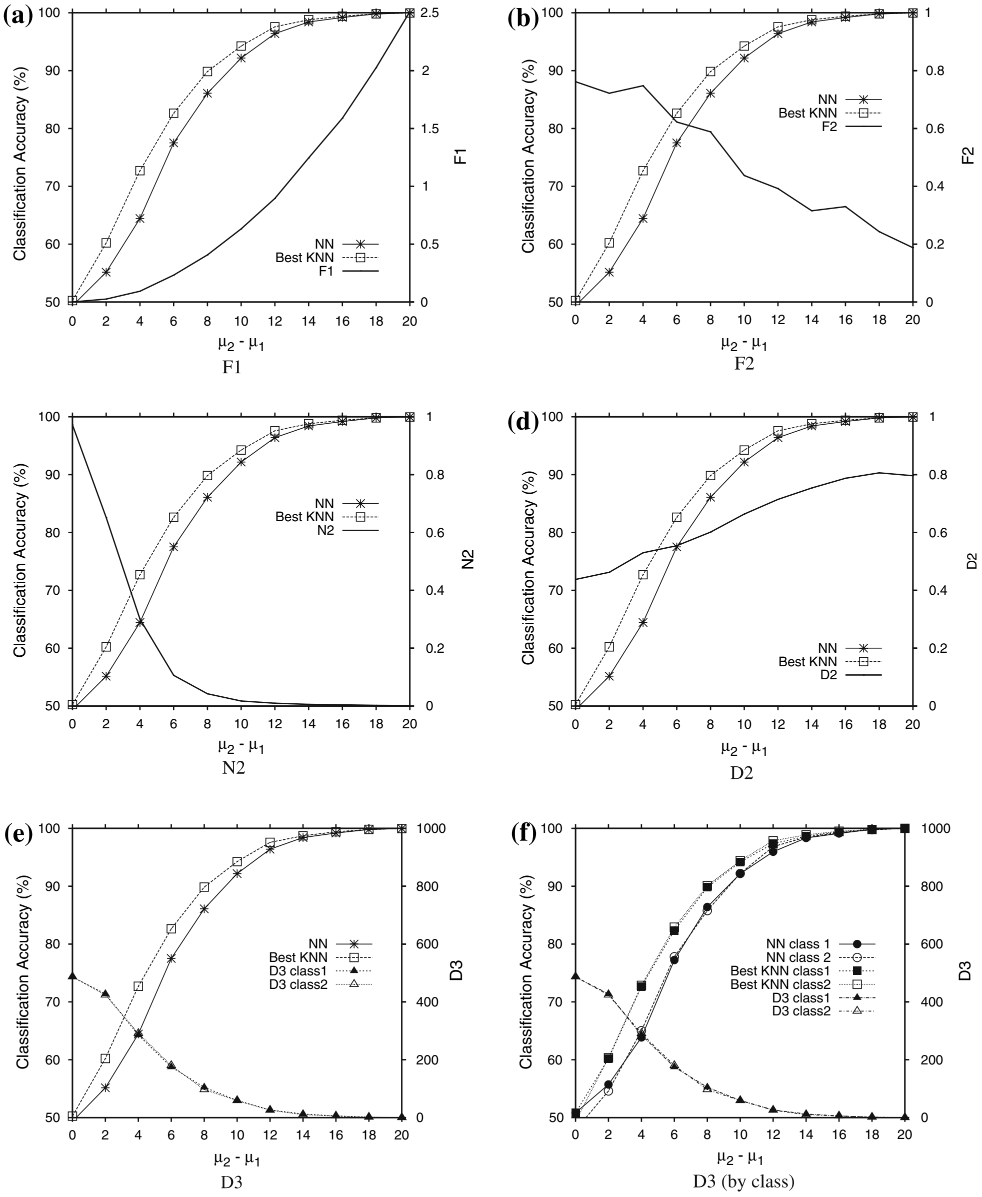


Fig. 2 Comparing NN and *k*-NN accuracies with F1, F2, N2, D2 and D3 in class overlap databases



F1

F2

N2

D2

D3

D3 (by class)

Analogously, the behavior of the classifiers could also be straightforwardly foreseen by analyzing the value of N2 (see Fig. 2c). As can be observed, N2 is high enough for the absolutely overlapped data set and then rapidly drops down while classification accuracies improve. On the other hand, D2 increases as the level of overlapping decreases. Since D2 reflects the average volume of the region comprising the *k* nearest neighbors of each training instance, it seems clear that such a region grows up (and correspondingly, density lessens) when the overlap is lower.

Focusing on measure D3, it can be observed from Fig. 2e, f that the classification accuracies increase as the number of examples appearing in the overlap regions diminishes. It must also be noted that the results for both classes are very similar, suggesting that there are not significant differences in the number of instances from each class present inside the corresponding overlap region. This could be expected from the definition of the present scenario, but it may result especially interesting in other domains where there does not exist a complete knowledge of the class distributions.

Taking into account that this first problem is primarily characterized by identical distributions and the same number of instances for both classes, but different levels of overlapping, the results suggest that here measures representing overlap and class separability (i.e., F1, F2 and N2) are those that better explain the behavior of the *k*-NN classifiers. Also, D3 constitutes an interesting measure for this problem, explaining the fact that the accuracies obtained over both classes are very similar.

3.2 Experiments on data dimensionality

To analyze how the feature space dimensionality affects the *k*-NN performance, we have employed a set of 19 synthetic databases corresponding to the same problem, but with dimensionality ranging from 2 to 20. They consist of two classes coming from Gaussian distributions with means equal to 0 and standard deviations 1 and 2 in all dimensions, respectively. It has to be noted that this represents a problem whose classes are strongly overlapped, albeit now the ultimate purpose is to characterize the effect of data dimensionality. There are a total of 2,500 instances belonging to each class. This domain constitutes an extension to the one employed in the ELENA European Project (http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA), in which seven databases with dimensionality ranging from 2 to 8 were generated.

Figure 5b illustrates the two-dimensional database. Due to the form of these Gaussians, the inner class matches the overlap region. Moreover, the density of the inner class (i.e., class 1) in this region is considerably higher than the one of the outer class (i.e., class 2). Finally, when dimensionality increases, the characteristics of the overlap region can be absolutely different from those appearing in the two-dimensional case, although it seems difficult to foresee the magnitude and consequences of these variations. Correspondingly, the experiments in this section are addressed to explain the behavior of the *k*-NN classifiers in this situation by using the measures F1, F2, N2, D2 and D3.

The results from these experiments plotted in Fig. 3 can be initially considered to be somehow surprising. Although the high class overlap as measured by F1 remains steady along all dimensions (F1$\approx$ 0, that is, close to maximum overlapping), it can be explained by the fact that F1 directly depends on the distance between the means of the two Gaussians, which is theoretically equal to 0 in all dimensions. On the other hand, the accuracies of both classifiers are not constant along all dimensions due to the differences in the underlying overlap. Therefore, this measure does not seem appropriate to describe the behavior of these classifiers in this particular problem.

In the case of F2 (see Fig. 3b), the value of this measure drastically drops down as dimensionality increases. According to the definition given in Sect. 2, F2 is the product of values in the range [0...1], which makes this measure strongly dependent on the number of features rather than the level of overlapping. Consequently, F2 is not especially useful to analyze situations with significant differences in the dimensionality.

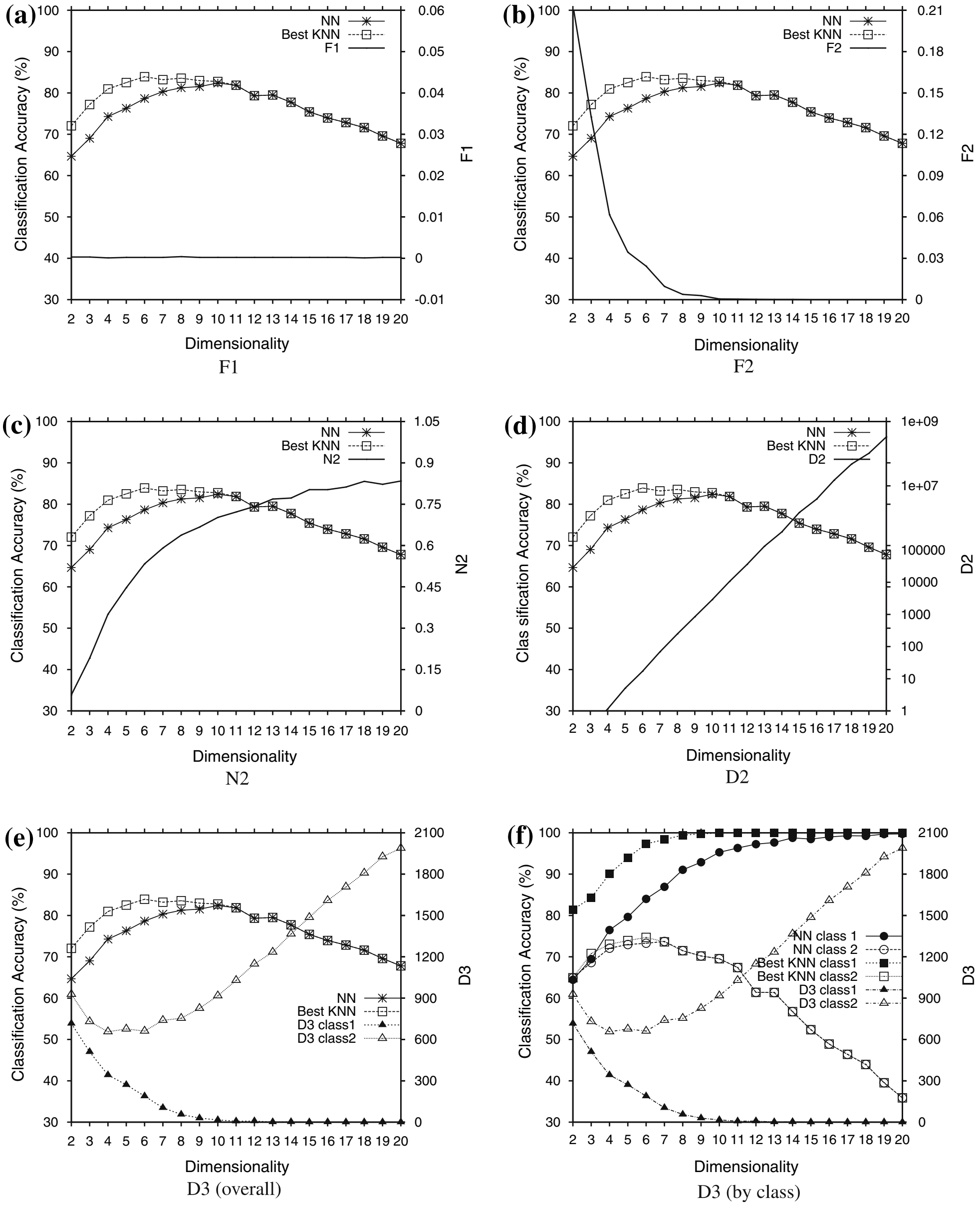
The behavior of the classifier when analyzed by means of N2 can be considered to be even more surprising. (see Fig. 3c). Theory dictates that greater values of N2 are related to lower classification results. Accordingly, by evaluating the tendency of N2, one would expect a degradation in classification performances. Nonetheless, the classification accuracies show a slight improvement up to dimensions 8–10, and then they decrease. A more exhaustive analysis reveals that the intra-class distances increase more rapidly than the inter-class distances, but their ratios still keep far from 1 in the lower dimensions (thus guaranteeing the correct classification of most instances), while approaching 1 in the higher dimensions.

On the other hand, values of D2 indicate that the average hypervolume defined by the *k* nearest neighbors increases as the dimensionality increases. This result indeed agrees with the behavior of N2. In fact, these two measures confirm that the data expands as dimensionality increases.

In Fig. 3e, we compare D3 with the overall accuracies produced by NN and *k*-NN classifiers. Firstly, it has to be remarked that while D3 in the inner class is close to 0 from dimension 9, D3 for the outer class clearly increases from dimension 8. This means that in the overlap region, the inner class becomes relatively denser and conversely the outer class results in becoming more sparse as the dimensionality increases. Note that overall classification accuracies also begin to drop down from dimension 8–10, suggesting that the degradation of accuracy can be mainly due to the variation of class distributions produced in the overlap region. Such a behavior can be better explained by examining the accuracies corresponding to each individual class (see Fig. 3f). Thus, one can see that classifiers over the inner class attain better results as dimensionality increases, but classification over the outer class suffers a very important degradation from dimension 8–9 (this is close to 30% in dimension 20). This results from increasing density in the inner class and diminishing intensity for the outer class.

In order to test how the TS size affects the present problem, we generated the same 19 databases but only with 500 instances. In Fig. 4, we have compared the NN and *k*-NN accuracies with D3 when using these reduced data sets. As can be seen, the results are very similar to those plotted in Fig. 3e, f corresponding to the databases with 2,500 instances. Nonetheless, note that now D3 computed on the outer class increases from dimension 4, while in the previous case it does from dimension 8. This suggests that the effect remarked above is now produced earlier because of the reduction in TS size. Moreover, the corresponding

Fig. 3 Results with varying data dimensionality



F1

F2

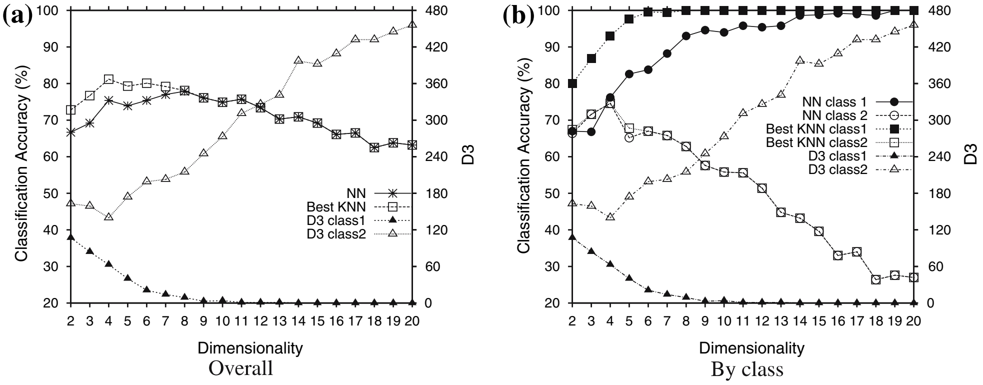
N2

D2

D3 (overall)

D3 (by class)

Fig. 4 Comparing NN and *k*-NN with D3 on dimensionality databases with 500 points



Overall

By class

classification accuracies are somewhat lower than those in the previous experiments, due to the lower density. No more differences are worthy of note.

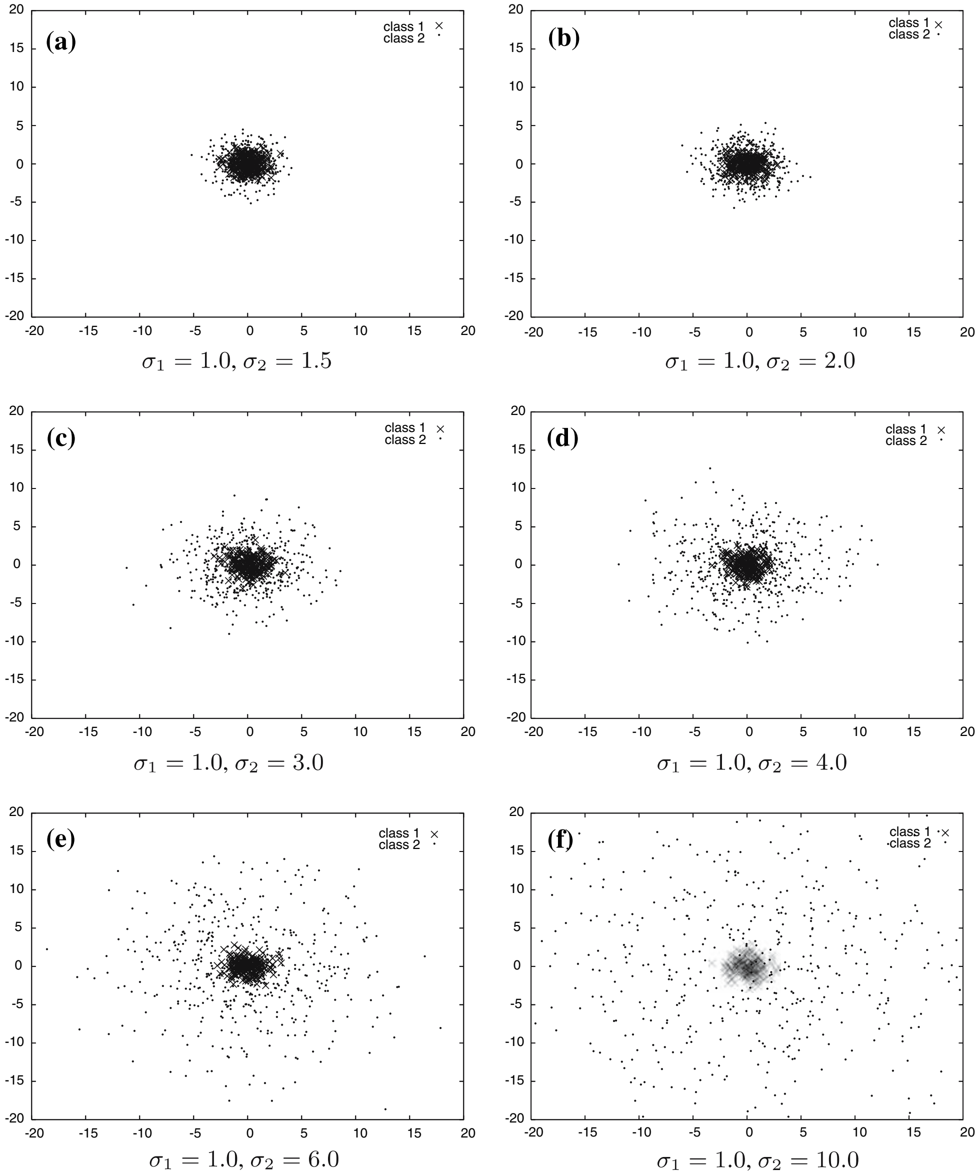
In this section, a strongly overlapped two-class data set has been characterized for a wide range of dimensions from 2 to 20. From the results provided here, it seems that the more useful measures to explain the behavior of NN and *k*-NN classifiers have been those that describe sparsity and density: N2, D2 and D3.

3.3 Experiments on class density

The aim of the third group of experiments is to study the effect of class density on the *k*-NN classifiers. To this end, we have generated six two-dimensional databases consisting of two classes. Both classes correspond to Gaussian distributions with zero mean ($\mu_1 = \mu_2 = 0$). Class 1 has standard deviation equal to 1 for all databases. Class 2 has standard deviation 1.5, 2.0, 3.0, 4.0, 6.0 and 10.0 for each of the six data sets, respectively. Note that the distinct standard deviations of class 2 describe the differences in class density within the overlap region (this matches the region of class 1), that is, the number of examples from class 2 lying in the overlap region diminishes as standard deviation increases. Figure 5 shows these synthetic databases.

These databases are similar to those employed in the experiments of Sect. 3.2. In fact, the data set in Fig. 5b ($\sigma_2 = 2.0$) corresponds to the two-dimensional case of the dimensionality databases. Therefore, some data characteristics coincide with those computed in the

Fig. 5 The density databases



$\sigma_1 = 1.0, \sigma_2 = 1.5$

$\sigma_1 = 1.0, \sigma_2 = 2.0$

$\sigma_1 = 1.0, \sigma_2 = 3.0$

$\sigma_1 = 1.0, \sigma_2 = 4.0$

$\sigma_1 = 1.0, \sigma_2 = 6.0$

$\sigma_1 = 1.0, \sigma_2 = 10.0$

previous experiments. Thus, the class overlap measured by F1 also remains steady, close to 0 along all databases, similar to the effect produced on the dimensionality experiments.

With respect to F2 (the volume of overlap), one can observe that values of F2 decrease and classification accuracies improve when increasing σ_2. The variation of σ_2 produces an expansion of the region of class 2 and consequently, a reduction in the relative significance of the overlap region. F2 describes this effect, which supports a certain improvement in classification performance.

In Fig. 6c, values of N2 indicate that classes become more separable as σ_2 increases: the average distance to inter-class nearest neighbor increases more rapidly than that to intra-class nearest neighbor, producing an increase in performance. On the other hand, D2 shows a very regular trend, except in the case of a more significant difference between the two classes ($\sigma_2 - \sigma_1 = 9.0$).

In the present experiments, both classes show identical behaviors in the tendency of D3. This is due to the reduction in the number of instances of class 2 lying in the overlap region, producing a situation in which class 1 is relatively denser than class 2. As a consequence, for each case ($\sigma_2 - \sigma_1$) the number of instances from class 2 removed is greater than that of class 1.

The analysis of D3, along with the one of F2, can fully explain the results of the classifiers in the present experiments. The increase in performance is mainly due to the relative reduction of the overlap region (described by F2) and also to the class imbalance generated in such a region (as measured by D3). It is worth noting that in the case of class 1, gains in *k*-NN classification accuracy are much more significant than those achieved using the NN classifier.

3.4 Experiments on real databases

After the analysis performed on synthetic databases, in the present section, we experiment with real databases in order to show the practical applicability of the data complexity measures in the prediction of NN and *k*-NN behaviors. Here, we have used three two-class databases taken from the UCI Machine Learning Database Repository (http://www.ics.uci.edu/mlearn/). The Cancer database consists of 683 instances with 9 attributes, the German database has 1,000 examples with 24 features, and the Phoneme database is a five-dimensional set of 5,404 instances. These data sets have been chosen to be quite different in the relation between the number of attributes and the number of instances.

Table 1 provides the data complexity measures (F1, F2, N2, D2, D3) and classification accuracies obtained over the three real databases. The results included here come from using the five-fold cross-validation estimate method. In the rows corresponding to D3, NN and best-*k*-NN, values in brackets refer to the results for class 1 and class 2, respectively. The meaning of D3* is the percentage of points belonging to both classes (the two values in D3) lying in the overlap region with respect to the total number of instances.

The results reported here seem to validate most of the conclusions suggested from the experiments with artificial databases. In the present situation, one does not know a priori the data characteristics for each particular problem, and the goal is to predict the behavior of the classifiers by means of the measures introduced in this paper. For example, a domain with high overlapping will be more difficult for the classifier than another with no overlap between classes. Similarly, the class density in the overlap region could indicate the difficulty of the classifier to correctly estimate the label of new samples from each class.

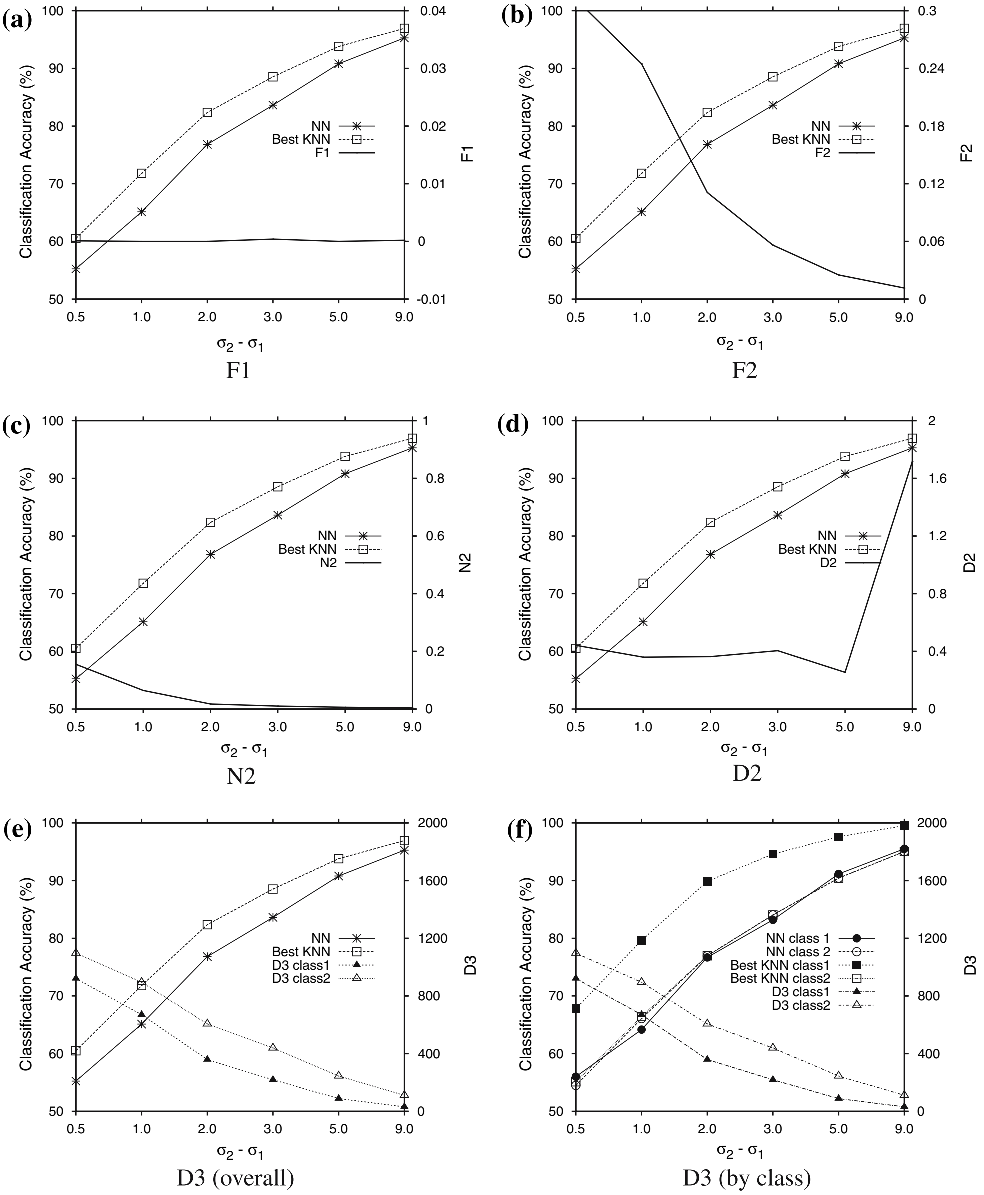
It has to be remarked that this analysis can become robust and accurate only if it is performed combining a set of measures: the employment of a unique measure could lead to erroneous conclusions. To this end, we will sort the three databases according to the desired values of each measure. In this sense, recall that higher classification accuracies are related to higher values of F1 and lower values of F2, N2, D2 and D3. Table 2 summarizes the ranks obtained for each measure.

According to the rankings reported in Table 2, it seems that the German database should be the problem with the worst classification performance. When comparing Cancer and Phoneme databases, the former wins in three out of five measures, which correspond to those that demonstrated a more important role in the previous analysis over synthetic problems. From this, it should be expected that the accuracies over the Cancer data set would be higher than those over the Phoneme data set. All these predictions are fully supported in the last two rows of Table 1.

Focusing on the results over the Cancer database (see Table 1), F1 is greater than 1, what in a two-class problem suggests low overlapping. This is strengthened by the fact that F2, N2 and D3* are close to 0. All these conditions indicate that this database constitutes a straightforward classification problem. D2 shows that there is a low density of patterns, which results from the relation between set size and data dimensionality.

The German database is an opposite example to the Cancer domain. The values in Table 1 describe a very high overlapping between both classes (F1 → 0,

Fig. 6 NN and *k*-NN accuracies compared to measures on density databases



F1

F2

N2

D2

D3 (overall)

D3 (by class)

Table 1 Data complexity measures and classification accuracies for the real databases

	Cancer	German	Phoneme
F1	1.38	0.03	0.08
F2	0.04	0.77	0.25
N2	0.22	0.82	0.09
D2	2.9×10^4	3.4×10^5	0.02
D3	(6) (5)	(76) (133)	(270) (194)
D3*	1.61	20.90	8.59
NN	96.10 (95.49) (97.22)	64.67 (70.48) (51.11)	88.71 (77.68) (93.28)
Best-*k*-NN	97.07 (96.24) (98.61)	70.67 (90.48) (51.11)	88.71 (77.68) (93.28)

F2 → 1, N2 → 1, D3* = 20.90%). The results of D3 indicate that class 2 is the most confusing one. This scenario produces low classification accuracies, especially with patterns from class 2.

The Phoneme database can be viewed as an intermediate case between the two previous domains. Analyzing F1 and F2, one can conclude that the means of both classes are close enough and there exists a

Table 2 Databases ranked according to each measure

	F1	F2	N2	D2	D3*
Cancer	1	1	2	2	1
German	3	3	3	3	3
Phoneme	2	2	1	1	2

significant overlap region. Nevertheless, N2 and D3 show that the amount of patterns in the overlap region is low, thus compensating the effect described by F1 and F2.

4 Concluding remarks

In this paper, several data complexity measures have been employed to analyze the effect of some data characteristics on the expected accuracy of the NN and *k*-NN classifiers. In particular, the present work has focused on studying how class overlap, data dimensionality and class density affect the practical accuracy of these classification algorithms.

From the empirical analysis carried out, a number of concluding remarks can be made. Firstly, the data complexity measures used here demonstrate that the *k*-NN classification performance is strongly sensitive to class overlap and class density and, at a lower level, also to data dimensionality. Secondly, one can see that those measures are domain-dependent, in the sense that their effectiveness hinges on the particular problem characteristics. Moreover, the experiments with real-problem data sets have shown the practical applicability of the data complexity measures to foresee the NN and *k*-NN behaviors.

The measures considered in this paper can be categorized into two groups: overlap and class separability measures, and class density measures. The first category includes a generalization of Fisher's ratio (F1), a relative measure on the volume of the overlap region (F2) and a non-parametric measure on separability of classes (N2). On the other hand, the second group involves the volume of local neighborhood (D2) and a measure on class density in the overlap region (D3). These two class density measures have been proposed in the present paper.

As expected, F1 has resulted to be especially worthwhile in the experiments on class overlap. In the other two synthetic experiments, where overlapping derives from variations in dimensionality and sparsity, F1 has not been suitable to explain the *k*-NN classifier performance. In the case of F2, this measure helps to understand the behavior of the classifiers in experiments where the dimensionality remains constant. In fact, F2 has resulted to be appropriate to compare problems with no significant differences in dimensionality. N2 has been useful in all experiments, basically as a complement to other more specific measures.

With respect to the class density measures introduced in this paper, D2 has played a role similar to N2. In this sense, it provides useful information, although some domain-driven measures are required to fully characterize the data distributions. Finally, D3 has been proven to be among the most explanatory measures. It provides an accurate measure on the density of each class in the overlap region. It is worth remarking that these densities in the overlap region mainly determine the *k*-NN classification performance. D3 is therefore an interesting measure to foresee the practical behavior of these classifiers.

This work should be viewed as a first step towards alternative ways of characterizing the behavior of the *k*-NN rules and therefore a number of extensions can be further investigated. First, some of the measures used here (e.g., N2 and D2) result from computing the average of individual values obtained for the training instances. In this case, a possible improvement could consist of analyzing the distributions of these individual values. Second, future research will address other problems present in the *k*-NN classifiers, such as imbalance in class distributions, presence of outliers and irrelevant attributes. Finally, the definition of new measures seems necessary to improve the analysis of the overlap scenario and to characterize the aforementioned situations.

5 Originality and contribution

This paper focuses on the analysis of three practical situations where the application of the *k*-NN classifiers can become less effective: class overlapping, high data dimensionality and class density. These constitute three crucial topics for analyzing the performance of most non-parametric classification systems, and especially in the case of distance-based classifiers. Briefly, we are interested in systematically characterizing the class overlapping, high feature dimensionality and class density effects on the *k*-NN classification results. To this end, we will make use of a number of data complexity measures.

This is not just another work to show how all these situations produce an important degradation in classifier performance. The aim of the present paper is to provide analytical measures of overlap, dimensionality and density, and relate them to the expected accuracy of the *k*-NN classifier. This could help to explain the practical behavior of this decision rule under the above-mentioned conditions.

Acknowledgments This work has been partially supported by grants DPI2006-15542 from the Spanish CICYT, GV04A-705 from Generalitat Valenciana and P1-1B2004-08 from Fundació Caixa Castelló-Bancaixa.

References

1. Bernadó E, Ho T-K (2004) On classifier domain of competence. In: Proceedings of 17th international conference on pattern recognition, Cambridge, pp 136–139
2. Beyer KS, Goldstein J, Ramakrishnan R, Shaft U (1999) When is ‘‘nearest neighbor’’ meaningful? In: Proceedings of 7th international conference on database theory, Jerusalem, pp 217–235
3. Cover TM (1968) Estimation by the nearest neighbor rule. IEEE Trans Inf Theory 14:50–55
4. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13:21–27
5. Dasarathy BV (1991) Nearest neighbor norms: NN pattern classification techniques. IEEE Computer Society Press, Los Alamos
6. Devijver PA, Kittler J (1992) Pattern recognition: a statistical approach. Prentice Hall, Englewood Cliffs
7. Friedman JH (1997) On bias, variance, 0/1-loss, and the curse of dimensionality. Data Mining Knowl Discov 1:55–77
8. Gama J, Brazdil P (1995) Characterization of classification algorithms. In: Progress in artificial intelligence. Springer, Heidelberg, pp 83–102
9. Hand DJ, Vinciotti V (2003) Choosing *k* for two-class nearest neighbour classifiers with unbalanced classes. Pattern Recognit Lett 24:1555–1562
10. Ho T-K, Basu M (2002) Complexity measures of supervised classification problems. IEEE Trans Pattern Anal Mach Intell 24:289–300
11. Hoekstra A, Duin RPW (1996) On the nonlinearity of pattern classifiers. In: Proceedings of 13th international conference on pattern recognition, Vienna, pp 271–275
12. Jain AK, Xu X, Ho T-K, Xiao F (2002) Uniformity testing using minimal spanning tree. In: Proceedings of 16th international conference on pattern recognition, Quebec City, pp 281–284
13. Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. Intell Data Anal 6:429–449
14. Kubat M, Chen W-K (1998) Weighted projection in nearest-neighbor classifiers. In: Proceedings of 1st Southern symposium on computing, Hattiesburg, pp 27–34
15. Little RJA, Rubin DB (2002) Statistical analysis with missing data. Wiley, New York
16. Mollineda RA, Sánchez JS, Sotoca JM (2005) Data characterization for effective prototype selection. In: Proceedings of 2nd Iberian conference on pattern recognition and image analysis, Estoril, pp 27–34
17. Okamoto S, Yugami N (2003) Effects of domain characteristics on instance-based learning algorithms. Theor Comput Sci 298:207–233
18. Sánchez JS, Barandela R, Marqués AI, Alejo R, Badenas J (2003) Analysis of new techniques to obtain quality training sets. Pattern Recognit Lett 24:1015–1022
19. Singh S (2003) PRISM—a novel framework for pattern recognition. Pattern Anal Appl 6:134–149
20. Sohn S-Y (1999) Meta analysis of classification algorithms for pattern recognition. IEEE Trans Pattern Anal Mach Intell 21:1137–1144
21. Zhang J, Mani I (2003) *k*NN approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of workshop on learning from imbalanced datasets, Washington DC
22. Wilson DL (1972) Asymptotic properties of nearest neighbor rules using edited data sets. IEEE Trans Syst Man Cybern 2:408–421

Author Biographies

J.S. Sánchez is an Associate Professor in the Department of Programming Languages and Information Systems at Universitat Jaume I (Castelló de la Plana, Spain) since 1992, and he is currently the head of the Pattern Analysis and Learning Group. He received a B.Sc. in Computer Science from the Universidad Politécnica de Valencia in 1990 and a Ph.D. in Computer Science Engineering from Universitat Jaume I in 1998. He is the author or co-author of more than 90 scientific publications, co-editor of two books and guest editor of several special issues in international journals. His current research interests lie in the areas of pattern recognition and machine learning, including nonparametric classification, feature and prototype selection, ensembles of classifiers, and clustering.

Ramón A. Mollineda Cárdenas is an Associate Professor in the Department of Programming Languages and Information Systems at Universitat Jaume I (Castelló de la Plana, Spain) since 2003, and currently belongs to the Pattern Analysis and Learning Group. He received a B.Sc. in Computer Science from the Universidad Central de Las Villas, Cuba, in 1995 and a Ph.D. in Computer Science from the Universidad Politécnica de Valencia in 2001. He is a member of IAPR and AERFAI (Spanish Association of Pattern Recognition and Image Analysis). He is author or co-author of more than 25 scientific publications, and he has been a member of review committees of several journals and conferences. His current research interests lie in the areas of pattern recognition and machine learning, including nonparametric classification, ensembles of classifiers, data analysis and string matching.

José Martínez Sotoca received a B.Sc. degree in Physics from the Universidad Nacional de Educación a Distancia (UNED, Spain) in 1996 and an M.Sc. degree in Physics from the University of Valencia, Spain, in 1999. He received a Ph.D. degree in Physics, also from the University of Valencia in 2001. Currently, he is an Assistant Lecturer in the Department of Programming Languages and Computer Systems at Universitat Jaume I (Castelló de la Plana, Spain). His main research interests lie in the area of pattern recognition and biomedical applications, including image recognition, hyperspectral data, structured light and feature extraction and selection. He has collaborated in different projects, most of them in medical application of computer science, and has published more than 35 scientific papers in national and international conferences, books and journals.