Fast k–neighbour classification of human vertebrae levels*

Francisco Moreno-Seco, José Manuel Iñesta, Luisa Micó, José Oncina

Departamento de Lenguajes y Sistemas Informáticos,
Universidad de Alicante,
E-03071 Alicante, Spain.
e-mail: {paco,inesta,mico,oncina}@dlsi.ua.es

Abstract

Recently, a fast approximate k–NN search algorithm has been proposed for metric spaces. This algorithm is a version of an exact 1–NN search algorithm (LAESA) which reduces its time complexity and improves its error rate. In this paper we test the performance of this new algorithm in real–data classification problems. Also, we compare the behaviour of the algorithm with other classification rules.

Keywords: k–nearest neighbour, real–data classification.

1 Introduction

Nonparametric classifiers are widely used in pattern recognition due mainly to its simplicity and good behaviour [1]. Among these classifiers, the most popular is probably the nearest neighbour (NN) classifier. Most of NN–based classifiers need a vectorial space to represent each pattern, and thus the Euclidean distance is almost always used. However, more general classifiers have been proposed (AESA [2], LAESA [3]) which do not require a vectorial space of representation; they simply require some way to measure the similarity between patterns (a distance) which has the properties of a metric. These algorithms find the NN of an unknown sample (in a metric space) calculating a very low number of distances. When the distance used is very time consuming (e.g. the Levenshtein distance), these algorithms become very fast classifiers.

The AESA and LAESA algorithms use the 1–NN rule for classification. Some extensions to these algorithms have been developed to use more than one neighbour

*Thanks to the CICYT for partial support of this work through project TIC2000–1703–CO3–02.

for classification [4, 5, 6]. In particular, the Ak–LAESA algorithm [5] is a variant of LAESA which uses at most k neighbours[1] to classify the sample, and obtains error rates very similar to those of an exact k–NN classifier.

Up to now, the Ak–LAESA algorithm has proved its good performance in experiments that use only synthetic data. In this paper we will apply this algorithm to a problem with real data: human vertebral level classification. Ak–LAESA uses a classification rule which is explained in the next section. We have made some experiments comparing this new kind of classification rule with other well–known and efficient classification rules, such as the NN rule, the nearest–centroid–neighbour rule (NCN [7]) and other non–parametric classification paradigms, including a human expert.

2 The Ak–LAESA algorithm

The LAESA algorithm [3] has two phases: preprocessing and classification. The Ak–LAESA [5, 6] is a small but important modification made in the classification phase. During the preprocessing phase, these algorithms select from the training set a number n_b of *base prototypes*; this number n_{bp} depends on the dimensionality of the data, and the base prototypes are selected so that they allow pruning of the training set during the classification phase. Also, during this phase all the distances between each base prototype and all other prototypes in the training set are calculated and stored for further use.

The classification phase has two stages: first, base prototypes are selected as candidates to NN, and its distance to the unknown sample is calculated and used (along with the distances computed in the preprocessing phase) to obtain a lower bound of the distance between each prototype and the sample. This lower bound of the distance is then used to prune the training set. When all base prototypes have been used, the second stage starts selecting new candidates to NN. While the LAESA algorithm goes on pruning and selecting prototypes until it finds the nearest neighbour, the Ak–LAESA stops this search when the number of remaining prototypes is less or equal than k, and then it classifies the sample by voting.

The behaviour of the Ak–LAESA leads it to classify using *at most* k neighbours, which are not necessarily the k nearest ones, although some of these prototypes are among them (see [6] for more details). However, the error rates of the Ak–LAESA algorithm are very similar to those of a k-NN classifier [5, 6], while at the same time the number of distance calculations is very low and does not depend on the training set size (this feature is derived from AESA and LAESA, see section 1).

[1] Some of these neighbours are among the k nearest neighbours, but most of them are not.

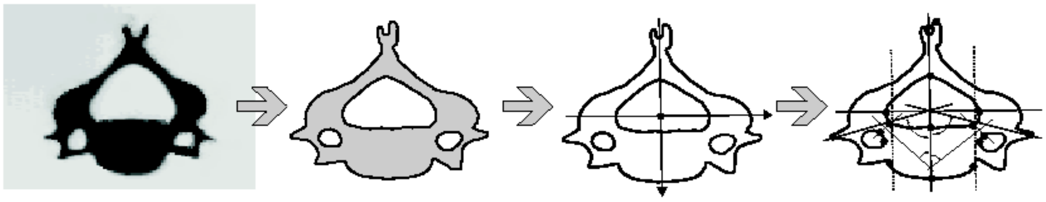

Figure 1: Main stages of the feature extraction procedure. From left to right: 1. vertebra image, 2. segmentation, 3. location and orientation, 4. geometric analysis, landmark detection and feature extraction.

3 Feature extraction and selection of vertebrae data

A set of 318 images of human vertebrae were digitised and calibrated for giving morphometric meaning to the measurements. After a careful segmentation that preserve the morphology, edges were extracted and coded. The calculated global descriptors were perimeters, geometric moments and derived features, and Fourier descriptors of the contours. The local features are more anatomical oriented measurements like widths, depths, thicknesses, angles, and ratios. Some anatomical landmarks were automatically located by geometric algorithms, translating anatomical knowledge into mathematical loci for their design [8]. See figure 1 for an outline of the process.

In this case of study, the high intra-class variability of the biological shapes makes the discrimination between classes a difficult task. A lot of features are extracted (50), but the size of the training set is reduced, constrained by the availability of samples, so for designing the classifier it is necessary to select the ones that discriminate the best. To do this dimension reduction, a linear discriminant analysis has been followed. A variable has a high discrimination power between different classes if the variance among the elements belonging to each of them is small, and if the variance between means of the classes is large. In the stepwise discriminant analysis a subset is built with a number of features, and its discrimination power is tested. This stepwise procedure is guided by the Fisher test, F. The value of F is computed as the ratio of the total variance in the data over the average within-class variance. F_i indicates how much a variable x_i contributes to the discrimination capability of the constructed subset. Moreover, a F_i can be calculated for each variable as the ratio of F for the subsets of selected variables both including x_i ($F_{i,in}$) and not ($F_{i,out}$). The higher F_i is, the higher the discrimination power of x_i will be. Thus, stepwise discriminant analysis proceeds starting with the variable of a higher F_i. Then, a second one x_j is sought with the highest $F_{j,in}$ (so we have the most discriminative combination). After this, the same is performed with three

Set Name	No. of prototypes	Dimensionality	No. of classes
CERVICAL	64	10	5
THORACIC	186	14	12
LUMBAR	68	14	5

Table 1: Characteristics of the datasets used in experiments.

variables, etc. During this procedure some variables can be excluded of the subset if a linear combination of the ones in the subset can explain the values of it in some degree (detected by a high F_{out}). The process ends when the descriptors out of the subset do not add discrimination capability to the subset ($F_i < 1$ for all of them).

4 Experiments

Three datasets have been used to develop all these experiments. In table 1 the main characteristics of these datasets are shown. Every prototype is a d–dimensional vector of selected features, thus the Euclidean distance has been used with all the algorithms tested.

We have developed two different experiments with these datasets: first, we wanted to obtain the number of *base prototypes* optimal for Ak–LAESA algorithm. In order to obtain these values (one for each dataset), exhaustive experiments have been developed: we have tested Ak–LAESA with all possible (and reasonable) values of n_b and with an increasing value of k. Then we compared the error rates for each value of n_b, and we determined that the best values were 8, 24 and 24 for the datasets CERVICAL, THORACIC and LUMBAR.

In our second experiment, we compared the error rates of Ak–LAESA, k–NN and k–NCN [7] classifiers with each dataset. Also, other classification paradigms have been considered as a reference. All three algorithms were run in a leave–one–out scheme with an increasing value of k. The results for each dataset are plotted in figure 2, where can also be noted that the behaviour of Ak–LAESA does not become worse as the value of k increases, as happens with k–NN and k–NCN. The classification made by a multilayer perceptron and anatomists has been also considered as a reference. The best error rates for each algorithm (with the value of k between parentheses) are shown on table 2. Note that it is a very hard problem even for a human expert, and NN rules perform comparably to it (neural networks classify slightly better). Also, we shall remark that Ak–LAESA error rates were

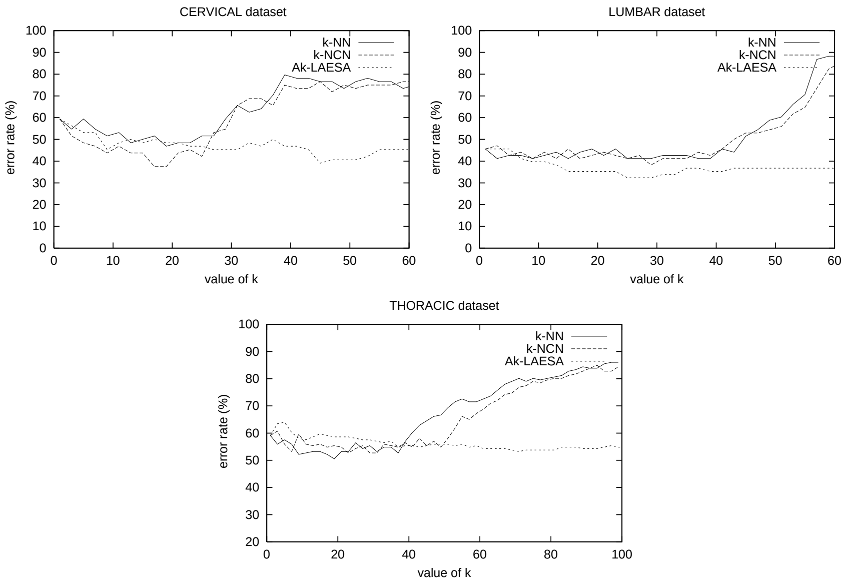

Figure 2: Error rates comparison of Ak–LAESA, k–NN and k–NCN.

obtained with a very low number of distance calculations: 9.2, 25.6 and 25.4 for the CERVICAL, THORACIC and LUMBAR datasets, while the distance calculations of k–NN and k–NCN were 63, 185 and 67.

5 Conclusions and future work

As shown in table 2, the Ak–LAESA obtains results comparable to those of other classification rules with real–data. It also obtains good results on synthetic data [5, 6]. The behaviour of this algorithm is somewhat defining a new kind of classification rule (based mainly on the triangle inequality) which obtains error rates comparable to those of other rules, but with a low number of distance calculations. As for the future, we plan to test this algorithm with some other real and synthetic classification tasks, requiring (or not) a vectorial space of representation.

References

[1] R. Duda, P. Hart. “Pattern Classification and Scene Analysis”. Wiley (1973).

Classifier	CERVICAL	THORACIC	LUMBAR
k-NN	46.875 (19)	**50.53** (19)	41.17 (3)
k-NCN	**37.5** (17)	52.68 (23)	38.23 (29)
Ak-LAESA	39.06 (45)	53.22 (71)	**32.35** (29)
MLP	35.9	37.7	28.5
Human expert	41.9	64.4	55.2

Table 2: Best error rates comparison of Ak–LAESA, k–NN, k–NCN, multilayer perceptron (MLP) and human expert. For the three k–NN classifiers, the value of k for the best rate is displayed between parentheses. Note that the results for MLP were obtained with a leave–k–out scheme, not with a leave–one–out.

[2] E. Vidal, "New formulation and improvements of the Nearest-Neighbour Approximating and Eliminating Search Algorithm (AESA)". Pattern Recognition Letters, (1994) **15**, pp. 1–7.

[3] L. Micó, J. Oncina, E. Vidal. "A new version of the nearest neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing-time and memory requirements". Pattern Recognition Letters, (1994) **15**, pp. 9–17.

[4] P. Aibar, A. Juan, E. Vidal. "Extensions to the approximating and eliminating search algorithm (AESA) for finding k-nearest-neighbours". New Advances and Trends in Speech Recognition and Coding, (1993), pp. 23–28.

[5] Moreno-Seco, F., Oncina, J., Micó, L. "Improving the LAESA algorithm error rates". Proceedings of the VIII Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes, Bilbao (1999), pp. 413–419.

[6] Moreno-Seco, F., Micó, L., Oncina, J. "A fast approximately k-nearest-neighbour search algorithm for classification tasks". Advances in Pattern Recognition. Lecture Notes in Computer Science, F.J. Ferri et al (Eds.) vol. 1876, Springer-Verlag, Berlin (2000), pp. 823–831.

[7] Sánchez, J.S., Pla, F. and Ferri, F.J. "Surrounding neighbourhood techniques for nearest–neighbour based classification and prototype selection", Recent Research Developments in Pattern Recognition, Vol. 1, Transworld Research Network, pp.63–76, 2000.

[8] Iñesta, J.M., Sanz, P.J., Buendía, M., Sarti, M.A. "Geometric algorithms for feature labelling in vertebrae images", Proc. of V International Symposium on Biomedical Engineering, pp. 99–100, 1994.