Non-Euclidean metrics for similarity search in noisy datasets

D. Francois[1], V. Wertz[1] and M. Verleysen[2] *

Université catholique de Louvain - Machine Learning Group
1- CESAME, Avenue George Lemaître, 4
2- DICE, Place du Levant, 3
B-1348 Louvain-la-Neuve, BELGIUM

Abstract. In the context of classification, the dissimilarity between data elements is often measured by a metric defined on the data space. Often, the choice of the metric is often disregarded and the Euclidean distance is used without further inquiries. This paper illustrates the fact that when other noise schemes than the white Gaussian noise are encountered, it can be interesting to use alternative metrics for similarity search.

1 Introduction

Many nonlinear tools for data analysis, among which many Artificial Neural Networks (ANN) models, rely on some similarity or dissimilarity measure between data elements. Examples include the k-Nearest Neighbours (k-NN) classifier, Kohonen maps (SOM), etc [1].

Many of those tools use the Euclidean distance to measure the similarity between data elements. The Euclidean distance might be a good choice *a priori*; however this paper arguments that the metric should be chosen according to the noise scheme that is assumed to affect the data.

The problem of nearest neighbour search is recalled in Section 2, while Section 3 will present Minkowski and fractional metrics. Section 4 will develop a general noise model, and Section 5 will give some insights on how to choose the right metric when a given noise scheme is assumed. Section 6 will describe some experiments and conclusions are drawn in Section 7.

2 Nearest Neighbour search for classification

The problem of nearest neighbour search consists in finding, among a dataset, the most similar data element to a given one, the latter is called *query point*. Mathematically, it is defined as follows : given $S = \{x_j\}_{j=1}^{N} \subset \Re^d$ a dataset of d-dimensional observations, x_q a query point, and $d(\cdot, \cdot)$ a *metric*, find $x_{nq} \in S$ the nearest neighbour of x_q among S such that $d(x_q, x_{nq}) \leq d(x_q, x_j) \forall x_j \neq x_{nq}$.

*M. Verleysen is a Senior Research Associate of the Belgian F.N.R.S. (National Fund For Scientific Research). The work of D. Francois is funded by a grant from the Belgian FRIA. Part of the work of D. Francois and of the work of V. Wertz is supported by the Interuniversity Attraction Pole (IAP), initiated by the Belgian Federal State, Ministry of Sciences, Technologies and Culture. The scientific responsibility rests with the authors.

The search for nearest neighbours is essential in many supervised and unsupervised classification techniques. For example the k-NN classifier determines the class label of a newly encountered data element according to the majority class among the nearest neighbours of the new data. Many unsupervised methods such as the Kohonen maps, as well as many methods for hierarchical clustering are based on the concept of nearest neighbour.

In this paper, we will focus on the choice of the metric for nearest neighbour search. Indeed, the choice of the metric is of high importance because the abovementioned methods and tools rely on the fact that similar data elements are close according to the chosen metric. If the metric does not reflect the 'right' notion of similarity, the methods will not perform well.

3 Minkowski and Fractional metrics

In this study, we will focus on the Minkowski metrics, based on the Minkowski norms. The Minkowski norms (also called L_p norms) are a family of norms parameterized by their exponent $1 \leq p \leq \infty$: For a $x_j = [x_{j1}, \ldots, x_{jd}] \in \Re^d$

$$\|x_j\|_p = \left(\sum_i |x_{ji}|^p \right)^{\frac{1}{p}} . \qquad (1)$$

When $p = 2$, we have the Euclidean norm. For $p = 1$, it induces the Manhattan metric. The limit for $p \to \infty$ induces the Chebychev metric.

Minkowski metrics have been successfully used in classification when the classes cannot be assumed to be hyper-spherical, or, equivalently, their populations to be Gaussian-distributed [2]. They have also been considered in the context of regression to build robust estimators. One interesting result is that the most robust estimator for regression are found by minimizing the L_p-norm of the residuals when the residuals are distributed as a generalized p-Gaussian [3]. Therefore, it has been proposed to choose the value of p according to the kurtosis of the noise distribution [4], according to the following relationship: $p = \frac{9}{\gamma^2} + 1$ where γ is the kurtosis of the noise.

Recently, fractional norms have been brought into light for high-dimensional data [5]. Those norms look like Minkowski norms except the value of the exponent p is still positive but less than one. Although those norms cannot be named norm in general because the triangle inequality is not ensured, they still can be used for nearest neighbour search.

4 A noise model

One of the most popular additive noise model is the *white Gaussian noise.* The term 'white' refers to the fact that the noise affects equally every component of the data. The term 'Gaussian' means that the perturbations are normal distributed : $x'_{ji} = x_{ji} + n_i \quad 1 \leq i \leq d$ where n_i is draw from a normally distributed random variable with mean zero and variance σ_n^2. However, in many real cases, the effective noise has the property to alter only a few components, but in a way that their values change drastically. We will call this scheme a highly coloured noise, meaning that few components are altered, in contrast to white noise. Examples of such noise scheme include the so called 'impulsive noise' or 'salt and pepper noise', or even 'burst noise' in the signal processing community [6, 7]. Coding errors and missing data markers can also be seen as highly coloured noise. The white noise model cannot fit such types of noise.

The noise model we propose to consider is the following :

$$x'_{ji} = \begin{cases} x_{ji} + n_i & \text{with probability } p_n \\ x_{ji} & \text{with probability } 1 - p_n \end{cases} \tag{2}$$

with n_i drawn from a random zero-mean variable. This model is general enough to represent both types of noises mentioned above : if $p_n = 1$, the model describes a white noise, while if p is low it describes an impulse noise.

5 The right metric for the right noise

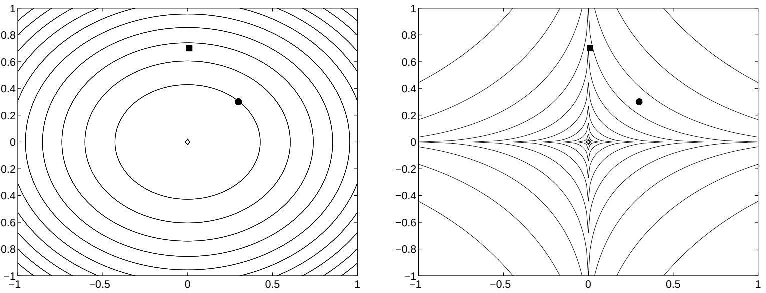
The role of a metric is to map a pair of points, or data elements x_1 and x_2, to a single value called the *distance* between those elements. If the components of x_1 and x_2 do not differ much, the distance will be small. In the case of absence of noise of any kind, virtually all metrics are equivalent. However this paper argues the fact that in the presence of noise, the metric must be carefully chosen.

If we observe a positive distance between x_1 and x_2, the natural question which arises is the following : is the distance due to real dissimilarity between x_1 and x_2, or is it due to noise in measurements ? If we suppose the noise is white, then very small componentwise differences between x_1 and x_2 will indicate that the distance is most probably due to noise, and not to real dissimilarity. In contrast, if many components are very close but some others are completely different, the distance should be interpreted as resulting from real dissimilarity. On the other side, if the noise is assumed to be coloured, opposite conclusions must be drawn. In any way, the metric should map differences due to noise to small values of the distance, and differences due to real dissimilarities to larger values of the distance.

The thesis of this paper is that fractional norms are a better dissimilarity estimator than the Euclidean norm when a highly coloured noise is present. Looking at the shapes of the isocurves can help us understand why fractional norms will better handle coloured noise than the Euclidean norm would. We can see in Figure 1 that, in dimension 2, provided one component is very similar, the other can be altered by a significant level of noise while still being a small distance to the center.

6 Experiments

This section will present the results of experiments carried out on both synthetic and real datasets. The performances of Minkowski and fractional metrics at the



(a) Euclidean norm (L_2) (b) Fractional norm ($L_{1/2}$)

Fig. 1: Isocurves from center for two different metrics. Depending on the metric, the nearest neighbour of the center is the square or the circle.

task of recovering a data element from its noisy version are evaluated. Given a dataset, we alter one by one the data elements according to a given noise scheme. Then, the nearest neighbour of the altered data element is searched in the hope that the nearest neighbour found is the original point. The score associated to a metric is the proportion of searches that recovered the original data element.

6.1 Synthetic dataset

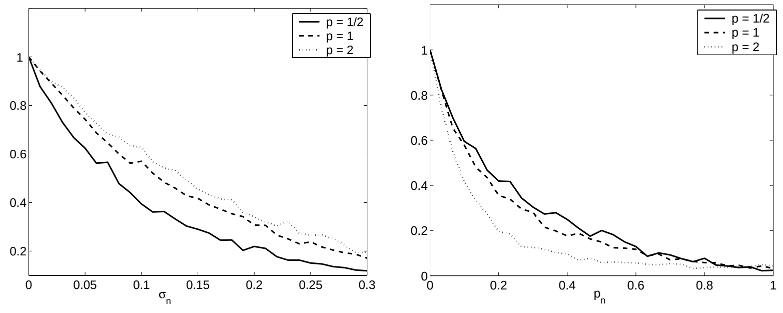
This dataset consists in 100 points uniformly distributed in $[0, 1]^{20}$. In a first experiment, a white Gaussian noise with standard deviation σ_n ranging from 0 (no noise) to 0.3. In the second experiment, a more coloured noise is added, with p_n ranging from 0 to 1 ($\sigma_n = 1$). Results are averaged over 10 trials and presented on Figure 2. The leftmost graph refers to a white noise, while the rightmost graph presents the scores of the same metrics for a coloured noise. We can see that, whatever noise level is considered, the Euclidean norm performs better at the white noise experiment, while the fractional norm outperforms the euclidean norm when the noise is coloured.

6.2 Chemometrics data

This section presents experiments conducted on databases of spectra of meat[1] and of orange juice [2]. As all measurements, spectra are subject to white noise, but averaging techniques exist to deal with it. Another type of noise is often encountered for such data. Sometimes, the spectra are shifted i.e. labels or indices of components do not match. Such shifts give rise to high differences between spectra in the sense of the Euclidean distance, but we can hope that fractional distances will handle them in a better way. Indeed, a shift of coordinates gives rise to very low componentwise differences in the flat regions of the spectrum, but results in large differences near the peaks of the spectra. An example from

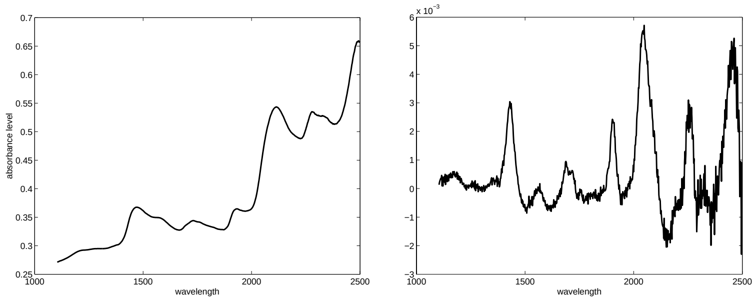
[1] Tecator Meat sample dataset, `http://lib.stat.cmu.edu/`, 215 100-dimensional spectra

[2] Orange juice dataset, `http://www.ucl.ac.be/mlg`, 216 700-dimensional spectra



(a) White noise (b) Coloured noise

Fig. 2: Scores (see text for definition) of several metrics for experiments with (a) white noise and (b) coloured noise; the value of p identifies the metric (See Eq (1)).



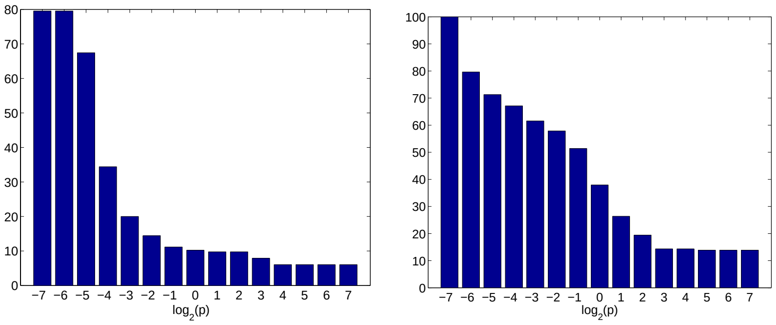
(a) sample spectrum (b) componentwise differences

Fig. 3: Orange juice dataset : (a) sample spectrum and (b) componentwise differences between this spectrum and itself shifted one component to the right. Some differences are very small while others are much larger, contrasting from Gaussian noise.

the orange juice dataset is presented in Figure 3 along with the componentwise differences between this spectrum and the same spectrum shifted from one component to the right. Figure 4, presents the scores of retrieval of the right spectrum from a shifted version of it. Minkowski and fractional metrics were used with values of p from 2^{-7} to 2^7. As expected, fractional norms perform significantly better than the Euclidean norm, which in Fig. 4 is related to the bars of value $\log_2(p) = 1$.

7 Conclusions

Since the notion of metric is crucial in many classification methods, it is important to choose the right metric for the right problem. This paper suggests to choose the metric according to the shape of the noise that is assumed on the data. While it is known that the Euclidean metric is optimal in presence of white Gaussian noise, it is shown that other type of noise require other metric.



(a) TECATOR dataset

(b) Orange Juice dataset

Fig. 4: Scores of several metrics in presence of coloured noise induced by spectrum shifts on Tecator and Orange Juice datasets.

Many high-dimensional data are prone to impulse or burst noise, that is a noise which affects only a minority of the components of the data elements, but in a significant way. Such noises are encountered in many signal processing applications. The experiments conducted on both synthetic and real datasets show that fractional norms are preferable when such noise scheme is encountered. The Euclidean norm, although heavily used, fails at measuring dissimilarity correctly in those cases.

Based on the idea developped in this paper and confirmed by experiments on artificial and real datasets, further work will consist in choosing in a more quantitative way the metric that should be used when a specific type of coloured noise in assumed on high-dimenisonal data.

References

[1] Michael A. Arbib. *The Handbook of Brain Theory and Neural Networks.* MIT Press, 1995.

[2] N. B. Karayiannis and M. M.Randolph-Gips. Non-euclidean c-means clustering algorithms. *Intelligent Data Analysis-An International Journal*, 7(5):405–425, 2003.

[3] J. M. Chen B. S. Chen and S. C. Chen. An ARMA robust system identification using a generalized lp norm estimation algorithm. *IEEE Trans. Signal Processing*, 42:1063–1073, 1994.

[4] T. T. Pham and R. J. P. deFigueiredo. Maximum likelihood estimation of a class of non-gaussian densities with application to lp deconvolution. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(1):73–82, 1978.

[5] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science*, 1973:420–434, 2001.

[6] A. Bovik. *Handbook of Image and Video Processing.* Academic Press, 2000.

[7] E. O. Elliot. Estimates of error rates for codes on burst–noise channels. *Bell Systems Technical Journal*, 42, 1977.