Available online at www.sciencedirect.com

ScienceDirect

Applied Soft Computing 7 (2007) 203–210

Applied Soft Computing

www.elsevier.com/locate/asoc

Unsupervised learning with normalised data and non-Euclidean norms

K.A.J. Doherty [*], R.G. Adams, N. Davey

University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK

Received 14 October 2004; received in revised form 16 February 2005; accepted 19 May 2005

Abstract

The measurement of distance is one of the key steps in the unsupervised learning process, as it is through these distance measurements that patterns and correlations are discovered. We examined the characteristics of both non-Euclidean norms and data normalisation within the unsupervised learning environment. We empirically assessed the performance of the *K*-means, neural gas, growing neural gas and self-organising map algorithms with a range of real-world data sets and concluded that data normalisation is both beneficial in learning class structure and in reducing the unpredictable influence of the norm.

Keywords: Distance measures; Data normalisation; Unsupervised learning; Neural gas; Growing neural gas; Self-organising map; *K*-means

1. Introduction

The measurement of distance is fundamental in the unsupervised learning process as most learning techniques require the calculation of a measure of similarity (respectively, dissimilarity) between training examples. Within the artificial neural network unsupervised learning community, the choice of distance measure often seems quite arbitrary. Inspired by a claimed improvement in nearest neighbour search and *K*-means class recovery accuracy when using fractional norms [1], we empirically examined the characteristics of non-Euclidean norms within the unsupervised learning framework. The claimed improvement arising from the use of fractional norms was therefore the motivation for this work.

Within the data driven sciences, the benefits of data pre-processing, such as normalisation or standardisation, are well known. However, in many fields of research these benefits are often overlooked and our work reported in this paper examines the consequences of combining data normalisation and non-Euclidean norms. The results presented here are an extension of our work initially reported in Ref. [2]. The remainder of this paper is organised as follows: In Section 2, we recapitulate the Minkowski metric. Section 3 describes data normalisation. Section 4 describes the synthetic and real-world data sets

[*] Corresponding author. Tel.: +44 1707 284 326; fax: +44 1707 284 185.

E-mail addresses: K.A.J.Doherty@herts.ac.uk (K.A.J. Doherty), R.G.Adams@herts.ac.uk (R.G. Adams), N.Davey@herts.ac.uk (N. Davey).

1568-4946/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2005.05.005

examined in this work. In Sections 5–7, we describe the results of nearest neighbour search, K-means clustering and clustering using three neural-inspired clustering algorithms, and finally, Section 8 presents our conclusions.

2. The Minkowski metric

A family of distance measures are the Minkowski metrics [3], where the distance between the d-dimensional entities i and j (denoted by $||ij||_r$) is given by:

$$||ij||_r = \left\{ \sum_{k=1}^{d} |x_{ik} - x_{jk}|^r \right\}^{1/r} \qquad (1)$$

where x_{ik} is the value of the kth variable for entity i, x_{jk} the value of the kth variable for entity j and $r > 0$.

The most familiar and common distance measure is the Euclidean or L_2 norm—a special case of the Minkowski metric where $r = 2$. Human understanding and experience makes us familiar with the results when applying L_2 measurements (to a problem space on a Euclidean plane), but the application of non-L_2 norms can lead to some counter-intuitive results. Consider the unit length loci from a point when plotted in the Euclidean plane with an L_r norm. In this Euclidean 2-space, the L_2 norm traces a circle, the fractional ($r < 1$) norms trace a hypoellipse, the L_1 norm trace a straight line and the higher order norms ($r > 2$) produce hyperelliptical traces. See Fig. 1 for a plot of these loci in the first quadrant.

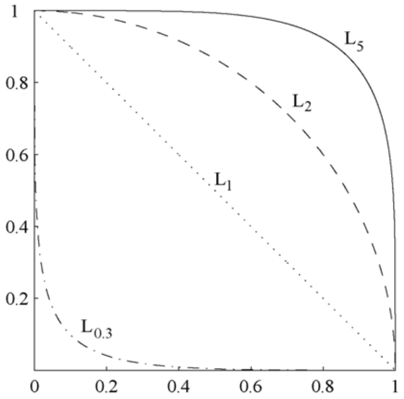

Fig. 1. First quadrant plot of unit length loci from the origin with various L_r norms.

Table 1
The distance of vectors a and b from the origin measured with L_2 and $L_{1/3}$

Norm	$\|\|a\|\|_r$	$\|\|b\|\|_r$	$\|\|a\|\|_r/\|\|b\|\|_r$
L_2	40.18	60.12	1.5
$L_{1/3}$	361.27	441.94	1.2

The ratio of the $L_{1/3}$ distance between the two vectors is less than the ratio of the L_2 distance, demonstrating how the fractional norm reduces the effect of the large feature vector attribute differences.

Consider the three feature vectors $a = (0, 1)$, $b = (1, 0)$ and $c = (7, 0)$. Let $||xy||_r$ be the L_r distance between vectors x and y. Generating a measure of dissimilarity with the L_2 norm, we find $(||ab||_2 = \sqrt{2}) < (||bc||_2 = 6)$. However, if we generate a measure of dissimilarity with the $L_{1/3}$ norm, we now find $(||ab||_{1/3} = 2^3) > (||bc||_{1/3} = 6)$. In a learning context when measuring dissimilarities between two entities, the use of a fractional norm reduces the impact of extreme individual attribute differences when compared to the equivalent Euclidean measurements. Conversely, the higher-order norms emphasise the larger attribute dissimilarities between the two entities and taken to the limit, L_∞ reports the distance based on the single attribute with the maximum dissimilarity. To further illustrate these points, consider the following feature vectors $a = (3, 2, 1, 40)$ and $b = (3, 2, 1, 60)$, and let $||x||_r$ be the L_r distance between vector x and the origin. Table 1 shows the distances of vectors a and b from the origin measured with the L_2 and $L_{1/3}$ norms. The L_2 norm clearly emphasises the larger attributes. The $L_{1/3}$ norm reports the relative distance from the origin to the vectors a and b in line with intuition—that is, b is further from the origin than a. However, the ratio of the $||x||_{1/3}$ distances is less than the ratio of the equivalent $||x||_2$ distances demonstrating how the fractional norm can reduce the effect of large differences in individual attributes.

3. Normalisation

Data *normalisation* (or *ranging*) is the linear transformation of data to within the range [0, 1] [3]. Normalisation was one of seven data pre-processing methods examined in Ref. [4], where the influence of data pre-processing on the recovery of class structure