ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 181-186

Non-Euclidean Norms and Data Normalisation

K.A.J. Doherty, R.G. Adams and N. Davey

Department of Computer Science, University of Hertfordshire, College Lane,
Hatfield, Hertfordshire, AL10 9AB, United Kingdom

Abstract – In this paper, we empirically examine the use of a range of Minkowski norms for the clustering of real world data. We also investigate whether normalisation of the data prior to clustering affects the quality of the result. In a nearest neighbour search on raw real world data sets, fractional norms outperform the Euclidean and higher-order norms. However, when the data are normalised, the results of the nearest neighbour search with the fractional norms are very similar to the results obtained with the Euclidean norm. We show with the classic statistical technique, K-means clustering, and with the Neural Gas artificial neural network that on raw real world data the use of a fractional norm does not improve the recovery of cluster structure. However, the normalisation of the data results in improved recovery accuracy and minimises the effect of the differing norms.

Keywords – Clustering, Minkowski Metric, Normalisation

1.0 Introduction

The measurement of similarity or distance is fundamental in the cluster analysis process as most clustering techniques begin with the calculation of a matrix of distances (or dissimilarities). The use of non-Euclidean norms within the clustering framework has been the topic of recent research [1-3]. Previous results in [1] have shown that in the context of K-means partitioning on synthetic data and nearest neighbour search on real data, that fractional norms do give better results. We have extended this investigation in 2 ways: i) to see if normalisation of the data affects the results, and ii) to see if K-means and a good neural network classifier can repeat these results on real data sets (with and without normalisation).

2.0 The Minkowski Metric

A family of distance measures are the Minkowski metrics, where the distance between the d-dimensional entities i and j (denoted by M_{ij}) is given by:

$$M_{ij} = \left\{ \sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right|^{r} \right\}^{\frac{1}{r}} \tag{1}$$

where x_{ik} is the value of the kth variable for the ith entity, and x_{jk} is the value of the kth variable for the jth entity.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 181-186

The most familiar and common measure of distance 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. Fig. 1 shows unit length loci around the origin, plotted with a selection of L_r norms. The L_2 norm traces a circle, the fractional ($r < 1$) norms trace a hypoellipse, the L_1 norm traces a straight line and the higher order norms ($r > 2$) produce hyperelliptical traces.

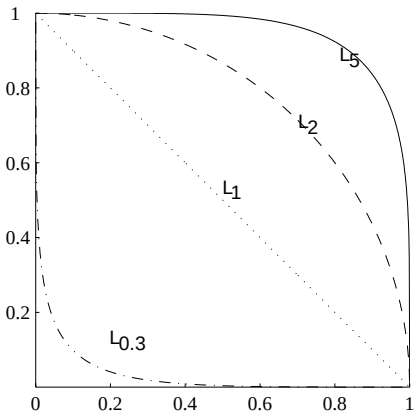

Fig 1. Unit length loci plotted with various L_r norms, from [1].

In a clustering 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 L_r norms (where $r > 1$) 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.

3.0 Fractional L_r norms and Data Normalisation

The clustering process aims to identify natural groupings within a data set. The notion of proximity is key in the identification of these natural groups, and the assumption is made that two entities that are in close proximity are likely to be members of the same group, or class.

3.1 Nearest Neighbour Search

Nearest neighbour (NN) search identifies entities in close proximity. The nearest neighbour problem is defined in [4] as: *Given a collection of data points and a query point in a d-dimensional metric space, find the data point that is closest to the query point*.

We repeated the empirical test of NN search using both fractional and higher-order L_r norms in [1] with the Ionosphere, Wisconsin Diagnostic Breast Cancer (WDBC), and Image Segmentation labelled data sets from the UCI Machine Learning Repository [5]. We report only a single representative result set. Table I shows the results of the nearest neighbour search on the raw Image Segmentation data set, and we confirmed that for this, and the two other data sets considered, the fractional norms generally

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 181-186

identified more nearest neighbours of the same class as the query point than the higher order norms.

Table I

Image Segmentation Training Data - Raw

210 Instances of 19 attributes plus the class label (7 classes)

n	$L_{0.1}$	$L_{0.5}$	L_1	L_2	L_4	L_{10}	L_∞
3	518	539	494	450	446	426	437
5	818	874	772	692	720	651	678
9	1345	1424	1249	1184	1167	1029	1066

For this experiment, the data were subjected to a NN search with no preprocessing. The larger the number of neighbours found belonging to the same class as the query point (when considering the nearest *n* neighbours), the better the performance of the NN search. In general, the fractional L_r norms identify more neighbours of the same class as the query point than the higher-order L_r norms.

The effect of 7 data standardisation methods on the recovery of class structure with a variety of perturbed data was examined in [6]. The results showed that normalization of each attribute for each datum, by the division of the range of each attribute, was beneficial to the cluster recovery accuracy for the synthetic data sources considered. The overall conclusion in [6] was the recovery of the underlying cluster structure improved when the data were normalised with either

$$x' = (x\text{-}X_{min})/(X_{max} - X_{min}) \quad (2)$$

or

$$x' = x/(X_{max} - X_{min}), \quad (3)$$

where x is the attribute value to be normalised, X_{max} is the maximum value of attribute x, and X_{min} is the minimum value of attribute x.

We repeated the nearest neighbour search with the Image Segmentation data normalised to the range [0,1] with (2). Table II details the results of the nearest neighbour search and shows that, for this normalised data set, the fractional norms generally identified more nearest neighbours of the same class as the query point than the higher-order norms.

Table II

Image Segmentation Training Data - Normalised

210 Instances of 19 attributes plus the class label (7 classes)

n	$L_{0.1}$	$L_{0.5}$	L_1	L_2	L_4	L_{10}	L_∞
3	506	536	507	515	502	493	464
5	819	872	821	821	809	799	736
9	1371	1489	1404	1364	1400	1362	1244

For this experiment, the data were normalised to the range [0,1] with (2) prior to the NN search. The larger the number of neighbours found belonging to the same class as the query point (for the *n* neighbours considered), the better the performance of the NN search. In general, the fractional L_r norms identify more neighbours of the same class as the query point than the higher-order L_r norms.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 181-186

However, the differentiation between the various norms (for a given *n*) is not as dramatic as the results obtained with the raw data shown in Table I. Although not shown here, it is interesting to note that the normalisation of the WDBC data set resulted in the L_2 norm identifying more nearest neighbours of the same class as the query point than the fractional norms considered.

The nearest neighbour search results show the lower-order norms outperforming the higher-order norms, and from this point forwards we concentrate our investigation principally on fractional norms.

3.2 K-means Class Recovery Accuracy

The improvement in the performance of K-means partitioning using a fractional distance norm was demonstrated in [1] on synthetic data. We extended this to see if this finding carried over to real world data sets. We empirically examined the effect of fractional norms on the performance of the K-means class recovery accuracy with a selection of data sets from [5], and report on a representative example (the Image Segmentation data set). We performed K-means partitioning with the number of codebook vectors equal to the number of classes in the data set, and the codebook vectors initialised with the attribute values of a datum drawn at random from the data set. We used either L_2 or $L_{0.3}$ as the distance measure for the duration of the K-means process. Once the codebook vectors were stabilised, the data were classified (allocated to the nearest codebook vector) with the L_2 and $L_{0.3}$ norms. We were looking for a correlation in the recovery between the training norm and the classification norm. No correlation was identified, but we include the results for interest.

Table III

		Raw		*Normalised*	
Training norm	Classification norm	Class Recovery	Likely Range	Class Recovery	Likely Range
L_2	L_2	59.2%	±0.7%	63.0%	±2.3%
L_2	$L_{0.3}$	58.4%	±0.8%	62.0%	±1.7%
$L_{0.3}$	L_2	54.5%	±3.6%	63.3%	±0.6%
$L_{0.3}$	$L_{0.3}$	54.2%	±3.4%	62.5%	±0.3%

K-means Class Recovery Accuracy for the Image Segmentation data set. Column 1 details the norm used to perform the K-means distance measurement. Column 2 details the norm used to classify the data to the nearest codebook vector when the K-means partitioning reached a quiescent state. The results suggest that normalisation of the data both improves class recovery accuracy and reduces the effects of the norm *r*-value.

Using labelled data allowed us to assess the class recovery accuracy with confusion matrices, using a count of the number of members from class *c* represented by the exemplar codebook vector for class *c*. The accuracy results are expressed as a percentage of the ideal partitioning scheme. For each set of parameters, we ran the K-

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 181-186

means partitioning 10 times, and show the precision of our estimates of class accuracy as 95% confidence limits. Table III shows the recovered class accuracy using the raw and normalised Image Segmentation data set. The most obvious result is the overall class recovery accuracy improved for all L_r norms when the data were normalised to the range [0,1]. Additionally, the normalisation of the data minimised the variation in class recovery obtained with the differing norms. In contrast to the results obtained with the synthetic data in [1], there was no improvement in the K-means class recovery accuracy obtained with the fractional norm when compared to the results obtained with the L_2 norm. The trends were similar in the other UCI data sets considered.

3.3 Neural Gas Class Recovery Accuracy

We examined the influence of the norm *r*-value on the performance of the Neural Gas algorithm [7] when clustering a selection of real world data sets.

Table IV shows the recovered class accuracy for the raw and normalised data sets. The results shown for the raw data are the optimum result set for a range of adaptation steps and temporal decay functions. When repeating the experiment with the data normalised to the range [0,1], the number of adaptation steps and temporal decay functions remained at these optimum settings. The results generally showed a marked improvement in the class accuracy recovery between the normalised and raw data. However, there was little, if any, significance in the class accuracy recovery between the L_r norms considered.

Table IV

			Raw		*Normalised*	
	Training norm	Classification norm	Class Recovery	Likely Range	Class Recovery	Likely Range
Image Segmentation	L_2	L_2	47.2%	±1.9%	61.4%	±3.4%
	L_2	$L_{0.3}$	46.0%	±0.1%	60.3%	±3.4%
	$L_{0.3}$	L_2	48.4%	±6.2%	57.5%	±4.2%
	$L_{0.3}$	$L_{0.3}$	52.1%	±2.6%	62.3%	±0.2%
WDBC Breast Cancer	L_2	L_2	72.9%	±8.1%	92.7%	±0.3%
	L_2	$L_{0.3}$	85.4%	±7.4%	91.1%	±0.2%
	$L_{0.3}$	L_2	82.0%	±8.9%	87.8%	±7.0%
	$L_{0.3}$	$L_{0.3}$	84.9%	±7.3%	89.2%	±1.2%
Wine	L_2	L_2	70.4%	±0.2%	95.3%	±0.4%
	L_2	$L_{0.3}$	77.0%	±1.0%	93.7%	±0.3%
	$L_{0.3}$	L_2	66.5%	±0.4%	95.4%	±1.3%
	$L_{0.3}$	$L_{0.3}$	69.8%	±0.3%	93.0%	±0.4%

Neural Gas Class Recovery Accuracy for a selection of UCI data sets. Column 1 lists the norm used to perform the Neural Gas distance measurements. Column 2 lists the norm used to classify the data to the nearest codebook vector. Again, the results suggest that normalisation of the data both improves class recovery accuracy and reduces the influence of norm *r*-value.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 181-186

4.0 Conclusions

In this paper, we demonstrated the important effect normalisation of the data has on the performance of nearest neighbour search, K-means, and Neural Gas clustering. Using the Neural Gas clustering algorithm, there was a significant and substantial improvement in the class recovery accuracy obtained by normalising the data.

In contrast to the results obtained with the synthetic data in [1], the K-means class recovery accuracy on real world data obtained with the fractional norm was not an improvement over the results obtained with the L_2 norm. The trends were similar in the other real world UCI data sets considered.

With nearest neighbour search, it is the case that the fractional norms, in general, do identify more neighbours of the same class as the query point with raw data as claimed in [1]. However, the results are not as convincing when the data are normalised, and we identified one data set for which normalisation resulted in the nearest neighbour search with the L_2 norm outperforming a nearest neighbour search with fractional norms.

Bibliography

[1] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," *Lecture Notes in Computer Science*, vol. 1973, pp. 420-434, 2001.

[2] D. L. Donoho, "High Dimensional Data Analysis: The Curses and Blessings of Dimensionality," presented at American Mathematics Society Conference: Math Challenges of the 21st Century, Los Angeles, USA, 2000.

[3] M. Verleysen, "Learning high dimensional data," presented at NATO Advanced Research Workshop on Limitations and Future Trends in Neural Computing, Siena (Italy), 2001.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "Nearest Neighbor" Meaningful?," presented at International Conference on Database Theory, 1999.

[5] C. L. Blake and C. J. Merz, "(UCI) Repository of machine learning databases," University of California, Department of Information and Computer Science., 1998.

[6] G. W. Milligan and M. C. Cooper, "A study of standardization of variables in cluster analysis," *Journal of Classification*, vol. 5, 1988.

[7] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten, "Neural Gas Network for Vector Quantization and its Application to Time-Series Prediction," *IEEE Transactions on Neural Networks*, vol. 4, pp. 558-569, 1993.