Unsupervised Learning with Normalised Data and Non-Euclidean Norms

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

[a] *University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, United Kingdom*
{*K.A.J.Doherty, R.G.Adams, N.Davey*}*@herts.ac.uk*

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.

Key words: 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 [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 examined in this work. In sections 5, 6 and 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:

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

where x_{ik} is the value of the kth variable for entity i, x_{jk} is 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

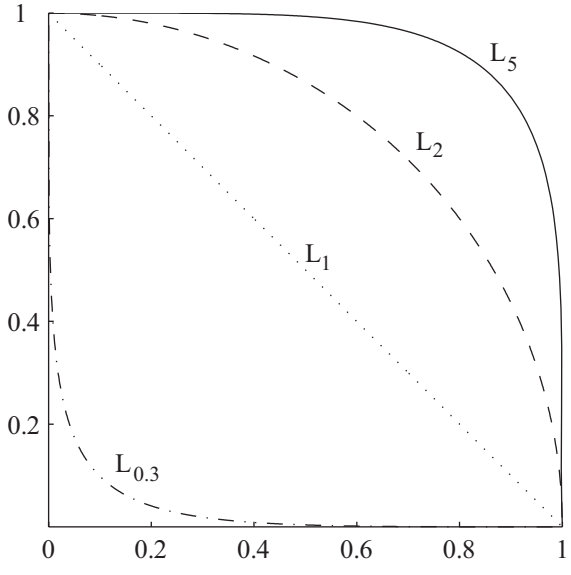

Fig. 1. First quadrant plot of unit length loci from the origin with various L_r norms

(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 1st quadrant.

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 $L_{\frac{1}{3}}$ norm, we now find $(\|ab\|_{\frac{1}{3}} = 2^3) > (\|bc\|_{\frac{1}{3}} = 6)$. 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 $L_{\frac{1}{3}}$ norms. The L_2 norm clearly emphasises the larger attributes. The $L_{\frac{1}{3}}$ 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 $\|x\|_{\frac{1}{3}}$ 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.

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

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

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 [4], where the influence of data pre-processing on the recovery

of class structure was evaluated. The results showed that normalisation was beneficial to the cluster recovery accuracy for the synthetic data sources considered. The accuracy of the recovered cluster structure improved when the data were normalised with either:

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

or

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

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.

4 DATA SETS

We performed our empirical tests of clustering accuracy using a selection of labelled data sets from the UCI Machine Learning Repository [5]. The data sets considered were the Ionosphere, Image Segmentation (training data), Wisconsin Diagnostic Breast Cancer (WDBC) and Wine data sets. These data sets were selected to show our approach on data with a range of classes, dimensionality and data distributions. The basic characteristics of each data set are shown in table 2.

Table 2
The basic characteristics of the UCI data sets examined in this paper, showing the dimensionality of the data, the number of instances in the data set and the total number of classes (C).

Name	Dimensionality	Instances	Classes
Ionosphere	34	351	2
WDBC	30	569	2
Image Segmentation	19	210	7
Wine	13	178	3

To repeat the experimental results presented in [1], we generated synthetic data following the description in the paper: six Gaussian sources (all of equal variance) in $\Re^{20}$, distributed randomly in $U[0, 100]$. From each source we drew 10000 elements, giving a total of 60000 data points. The results in [1] were presented as confusion matrices, in which the number of correct and incorrect elements classified are displayed. Confusion matrices are commonly used for measuring the performance of classification systems. We adjusted the variance of the Gaussian sources until the degree of cluster overlap (indicated by the number of incorrectly classified elements) was comparable to the overlap shown in the confusion matrices of [1].

5 NEAREST NEIGHBOUR SEARCH

Cluster analysis aims to identify natural groupings within a data set. The notion of proximity is key in the identification of these natural groups. Generally, the assumption is made that two entities in close proximity are likely to be members of the same group, or class. The nearest neighbour (NN) search identifies entities in close proximity and is defined in [6] 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". The natural extension to NN search is K-nearest neighbours (K-NN), where the nearest K neighbours to the query point are identified.

Table 3
K-NN search on a selection of *raw* UCI data sets. For a given K, the larger the number of neighbours found belonging to the same class as the query point, the better the K-NN search.

Ionosphere Data Set

K	$L_{0.1}$	$L_{0.5}$	L_1	L_2	L_4	L_∞
3	972	935	929	893	910	927
5	1613	1554	1526	1460	1515	1553
9	2830	2773	2713	2600	2706	2783

Wisconsin Diagnostic Breast Cancer

K	$L_{0.1}$	$L_{0.5}$	L_1	L_2	L_4	L_∞
3	1632	1637	1603	1564	1589	1573
5	2724	2708	2664	2594	2637	2627
9	4847	4844	4759	4639	4710	4714

Image Segmentation Training Data

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

Using the UCI Ionosphere, WDBC and Image Segmentation data sets we performed a K-NN search. For each member of the data set of class c, where $c \in C$ (q.v. Table 2), the K-NNs are identified and a count maintained of those neighbours whose class was also c. Table 3 shows our K-NN search results on

Table 4
K-NN search on a selection of *normalised* UCI data sets. Prior to the K-NN search, the data were normalised with equation 2. For a given K, the larger the number of neighbours found belonging to the same class as the query point, the better the K-NN search.

Ionosphere Data Set

K	$L_{0.1}$	$L_{0.5}$	L_1	L_2	L_4	L_∞
3	970	953	969	892	924	952
5	1609	1577	1596	1467	1530	1579
9	2835	2822	2810	2609	2712	2811

Wisconsin Diagnostic Breast Cancer

K	$L_{0.1}$	$L_{0.5}$	L_1	L_2	L_4	L_∞
3	1623	1625	1635	1638	1610	1587
5	2703	2691	2707	2713	2681	2653
9	4839	4820	4845	4848	4795	4722

Image Segmentation Training Data

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

data which repeat the trends identified by [1], in that the L_1 and fractional norms successfully identified more nearest neighbours of the same class than the L_2 norm. However, the argument for K-NN search with fractional norms is not that clear-cut, as for both the Ionosphere and WDBC data sets, the K-NN search with the L_4 and L_∞ norms were also more successful than than the search with L_2. We have presented results for 3 different values of K since in [1] the value of K was not specified, nor was it clear if the K-NN was performed on the training or test data set; so although our results do not match their results perfectly, they exhibit the same trend and are of the same order of magnitude. We assume the data in [1] are raw - that is, the data are subjected to the K-NN search without undergoing standardisation (to zero mean and unit variance) or normalisation (to range 0 to 1). Table 4 shows the results of repeating the K-NN search when the data are normalised with equation 2. The result of the effectiveness of nearest neighbour search with fractional norms when the data are normalised are not as convincing as the results obtained

with raw data. With normalised Image Segmentation data, the L_2 search for both 3 and 5 nearest neighbours outperformed the same search using $L_{0.1}$. Moreover, the results for the normalised WDBC data set show that nearest neighbour search with L_2 outperformed the search with all the other norms considered. These results suggest that the claimed improvement in nearest neighbour search brought about by the use of fractional norms is likely to be data dependent.

6 K-MEANS CLUSTERING

K-means [7] is a scalable partitioning process suitable for identifying data structures that are convex, compact and well separated. However, the number of codebook vectors must be prespecified, the partitioning process is susceptible to distortion by noise and outliers, and the final partitioning is sensitive to the initialisation of the codebook vectors.

For these experiments, the K-means algorithm was initialised with the number of codebook vectors equal to the number of classes in the data set, and the codebook vectors were initialised at a location drawn at random from the set of all data points, thus eliminating unused codebook vectors. For each data set, we performed the K-means training process using L_2 or $L_{0.3}$ for the distance calculations. Once the algorithm reached a quiescent state, the input data were classified based on the L_2 and $L_{0.3}$ distance from the nearest codebook vector. We assessed the accuracy of the recovered class structures with confusion matrices, which provided a valuable insight into the operation of the partitioning algorithm with the differing distance metrics.

6.1 Synthetic Data

We examined the claimed improvement in [1], where an improvement in the performance of K-means partitioning was reported when partitioning synthetic high dimensional data sets (q.v. section 4) using fractional norms. The reported improvement was an increase in class recovery accuracy from 89% with K-means partitioning performed with L_2 norm, to 99% using the $L_{0.3}$ norm. With very compact and well separated clusters, the accuracy of the recovered class structure was clearly related to the initialisation of the codebook vectors. Training K-means with L_2 or $L_{0.3}$ generally resulted in one of two results; either K-means placed one codebook vector per source and the reported class recovery rate was 98%+, or the reported class accuracy recovery dropped to 82% with one codebook vector classifying one and one half sources, with another codebook vector classifying the remaining half of the

split cluster. However with less compact clusters, we found a range of cluster variances where K-means with $L_{0.3}$ consistently resulted in the codebook vectors being placed one per source and returning a consistent class recovery accuracy of 98%+, but with the same data, the accuracy obtained with L_2 K-means remained dependent on the codebook initialisation.

We reproduced the claimed class recovery improvement in K-means clustering. However, this improvement in performance must be treated with care - the reported class accuracy improvement is not achievable with all data. To be effective, there must be close proximity (in the L_r norm) between the clusters. With very compact and well separated clusters there may be no improvement.

6.2 Real World Data

Table 5
K-means clustering on a selection of UCI data sets. Column 2 details the norm used to train K-means, and column 3 details the norm used to classify the data. In all cases, data normalisation improved the recovery of class structure, and the norm dependent variations in the accuracy were minimised.

			Class Recovery Accuracy (%)	
Data Set	Training Norm	Classification Norm	Raw Data	Normalised Data
Image Segmentation	L_2	L_2	59.2 ($\pm$0.7)	63.0 ($\pm$2.3)
	L_2	$L_{0.3}$	58.4 ($\pm$0.8)	62.0 ($\pm$1.7)
	$L_{0.3}$	L_2	54.5 ($\pm$3.6)	63.3 ($\pm$0.6)
	$L_{0.3}$	$L_{0.3}$	54.2 ($\pm$3.4)	62.5 ($\pm$0.3)
WDBC Breast Cancer	L_2	L_2	85.4 ($\pm$0.0)	92.8 ($\pm$0.0)
	L_2	$L_{0.3}$	85.1 ($\pm$0.0)	91.0 ($\pm$0.0)
	$L_{0.3}$	L_2	85.0 ($\pm$0.3)	92.1 ($\pm$0.0)
	$L_{0.3}$	$L_{0.3}$	84.5 ($\pm$0.4)	89.8 ($\pm$0.0)
Wine	L_2	L_2	70.0 ($\pm$0.4)	94.9 ($\pm$0.0)
	L_2	$L_{0.3}$	76.6 ($\pm$2.0)	93.6 ($\pm$0.4)
	$L_{0.3}$	L_2	71.6 ($\pm$0.8)	95.7 ($\pm$0.3)
	$L_{0.3}$	$L_{0.3}$	84.2 ($\pm$3.0)	92.9 ($\pm$0.3)

The improvement in the performance of K-means partitioning using fractional norms was demonstrated in [1] on synthetic data. We extended this work, and empirically examined the effect of fractional norms on the UCI Image Segmentation, WDBC Breast Cancer and Wine data sets.

We ran each K-means partitioning 10 times, and show the precision of our estimates of class accuracy as 95% confidence limits (which define the likely range of the true value in the population from which our results are drawn). Table 5 shows the recovered class accuracy. The results for the raw data show how the accuracy varies with the norm, but the results show no correlation between the norm and the accuracy achieved. Comparisons of the raw and normalised results suggest the normalisation of the data source with equation 2 minimises the effects produced by the use of a non-Euclidean norm. In addition and more significantly, the results suggest the recovery of class structure improves when the data are normalised, irrespective of the norm used.

7 UNSUPERVISED LEARNING

We examined the impact of normalisation and the norm, on the class recovery accuracy of three unsupervised competitive neural network algorithms [8,9]; the Neural Gas (NG) network [10], the Growing Neural Gas (GNG) network [11] and the Self-Organising Feature Map (SOM) [12]. The three networks use soft competition to distribute the network nodes, but the neighbourhood function varies between the algorithms. The networks were trained using the L_2 and $L_{0.3}$ norms for the distance measurements, and after training, the data were classified based on the L_2 and $L_{0.3}$ distances to the nearest node. Again, we ran each experiment 10 times and show the precision of our estimates of class accuracy as 95% confidence limits (which define the likely range of the true value in the population from which our results are drawn).

7.1 Neural Gas Class Recovery Accuracy

The NG algorithm is dependent on the number of adaptation steps, a neighbourhood function and a temporal decay function. The NG neighbourhood function is determined by the ordered ranking of the distance of the node to the current input vector. Our investigations suggest that the algorithm is not particularly sensitive to the parameter settings and for all tests we use the default parameters described in [10]. The training was performed with the number of nodes equal to the number of classes in the data set.

Table 6 shows the recovered class accuracy for the raw and normalised data sets. In general, for the raw data sets, there would appear to be no correlation between the norm and performance of the NG algorithm across all of the data sets. More significant is the class recovery accuracy obtained with normalised data, when compared to the equivalent raw data results. In all cases, the recovery of class structure improved with normalised data.

Table 6
Neural Gas clustering on a selection of UCI data sets. Column 2 details the norm used to train Neural Gas, and column 3 details the norm used to classify the data. In all cases, data normalisation improved the recovery of class structure, and the norm dependent variations in the accuracy were minimised.

			Class Recovery Accuracy (%)	
Data Set	Training Norm	Classification Norm	Raw Data	Normalised Data
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)

7.2 Self Organising Map Recovery Accuracy

The classification performance of the SOM algorithm is also dependent on the number of adaptation steps, a neighbourhood function and a temporal decay function. Determined by the topological layout of the SOM, the neighbourhood function is restricted to the direct topological neighbours of the winning node. Initially, the neighbourhood function must be large enough to ensure the map is ordered globally, but the neighbourhood extent is typically reduced over time. The algorithm is not particularly sensitive to parameter settings and we followed the suggestions for setting the parameters in [13]. We performed the clustering with a linear SOM. Configured with the number of network nodes equal to the number of classes in the data set, we are essentially mapping the data on to a line.

Table 7 shows the accuracy of the recovered class structure for the raw and normalised data sets. Once again, the results obtained with the raw data suggest there is no correlation between the training and classification norm, and the performance of the linear SOM across all the data sets. Again, in all

Table 7
SOM clustering on a selection of UCI data sets. Column 2 details the norm used to train the SOM, and column 3 details the norm used to classify the data. Again, data normalisation improved the recovery of class structure, and the norm dependent variations in the accuracy were minimised.

			Class Recovery Accuracy (%)	
Data Set	Training Norm	Classification Norm	Raw Data	Normalised Data
Image Segmentation	L_2	L_2	43.5 ($\pm$1.3)	62.1 ($\pm$1.5)
	L_2	$L_{0.3}$	43.5 ($\pm$0.7)	60.7 ($\pm$0.5)
	$L_{0.3}$	L_2	50.9 ($\pm$2.2)	63.2 ($\pm$0.1)
	$L_{0.3}$	$L_{0.3}$	51.8 ($\pm$1.9)	63.0 ($\pm$0.2)
WDBC Breast Cancer	L_2	L_2	85.1 ($\pm$1.8)	92.8 ($\pm$0.5)
	L_2	$L_{0.3}$	84.3 ($\pm$1.4)	91.4 ($\pm$0.6)
	$L_{0.3}$	L_2	84.2 ($\pm$1.1)	91.0 ($\pm$1.2)
	$L_{0.3}$	$L_{0.3}$	84.0 ($\pm$1.2)	88.6 ($\pm$1.6)
Wine	L_2	L_2	70.0 ($\pm$0.8)	94.5 ($\pm$0.9)
	L_2	$L_{0.3}$	77.8 ($\pm$1.7)	93.7 ($\pm$0.8)
	$L_{0.3}$	L_2	71.9 ($\pm$0.6)	95.1 ($\pm$0.7)
	$L_{0.3}$	$L_{0.3}$	84.5 ($\pm$2.4)	93.1 ($\pm$0.7)

the cases, the accuracy of the recovered class structure improved when the data are normalised, and the impact of the norm value is minimised.

7.3 Growing Neural Gas Recovery Accuracy

The GNG dynamically "grows" a structure until either a user defined performance criterion or network size is met. The topology representing network [14] generated by the Competitive Hebbian Learning not only determines the insertion point for a new node as the network grows, but also describes the neighbourhood of the winning node used for the soft competition update of the neighbouring nodes. Our own experiments, and the work of others [15], suggest that the network is reasonably insensitive to the network parameters, and for these experiments the parameters were set to the values in [11]. The maximum number of nodes was set equal to the number of classes in the data set.

Table 8 shows the accuracy of the recovered class structure for the raw and

Table 8
GNG clustering on a selection of UCI data sets. Column 2 details the norm used to train Growing Neural Gas, and column 3 details the classification norm. In all cases, data normalisation improved the recovery of class structure, and the norm dependent variations in the accuracy were minimised

			Class Recovery Accuracy (%)	
Data Set	Training Norm	Classification Norm	Raw Data	Normalised Data
Image Segmentation	L_2	L_2	46.0 ($\pm$1.4)	65.3 ($\pm$0.9)
	L_2	$L_{0.3}$	45.2 ($\pm$1.0)	62.8 ($\pm$1.3)
	$L_{0.3}$	L_2	51.3 ($\pm$2.0)	62.0 ($\pm$0.6)
	$L_{0.3}$	$L_{0.3}$	51.6 ($\pm$2.0)	61.6 ($\pm$0.5)
WDBC Breast Cancer	L_2	L_2	85.1 ($\pm$1.2)	92.3 ($\pm$1.2)
	L_2	$L_{0.3}$	85.2 ($\pm$1.2)	90.8 ($\pm$1.4)
	$L_{0.3}$	L_2	86.1 ($\pm$1.2)	90.8 ($\pm$1.7)
	$L_{0.3}$	$L_{0.3}$	86.1 ($\pm$1.5)	88.8 ($\pm$2.6)
Wine	L_2	L_2	70.1 ($\pm$1.2)	94.4 ($\pm$1.2)
	L_2	$L_{0.3}$	77.3 ($\pm$3.5)	93.1 ($\pm$1.2)
	$L_{0.3}$	L_2	71.6 ($\pm$1.0)	93.9 ($\pm$1.1)
	$L_{0.3}$	$L_{0.3}$	83.5 ($\pm$2.1)	92.7 ($\pm$1.0)

normalised data sets. Once again, the results for the raw data show no correlation between the performance of the GNG classifier and the norm. Again, in all the cases, the accuracy of the recovered class structure improved when the data are normalised.

8 CONCLUSION

Our results obtained with the three unsupervised neural clustering algorithms showed that with raw data, there was no consistent improvement in class recovery accuracy with the fractional norm. Indeed, no single training and classification norm pair produced consistent, good quality results. However, normalising the data to the range [0,1] consistently increased the accuracy of the recovered class structure. Using normalised data resulted in very similar accuracy levels for the K-means clustering and the three neural inspired models. These results are not surprising, as essentially the three neural algorithms are attempting to perform a minimisation of a sum squared error criterion, similar to K-means, but with the added constraint of the soft competition

update of the neighbourhood nodes.

When performing a K-NN search on the raw UCI data sets, fractional norms identified more nearest neighbours than the L_1 and Euclidean norms, repeating the findings of [1]. However, these results are not clear-cut in suggesting fractional norms outperform higher-order norms, as a K-NN search with L_4 and L_∞ on the raw Ionosphere and WDBC data sets out-performed a similar K-NN search with L_2. The results of our experiments show that any claimed improvement is likely to be data set dependent.

The application of fractional norms to the K-means partitioning algorithm can produce vast improvements in clustering accuracy. However, we have demonstrated the improvement is data dependent. With well separated clusters, there is no guarantee the performance improves, and the K-means algorithm remains susceptible to codebook vector initialisation and can remain trapped in a local minimum.

The results presented in this paper very clearly demonstrate the beneficial effect normalisation of the data has on the recovery of class structure with both K-NN search and squared error minimisation clustering. The improvement in the results achieved by normalising the data prior to analysis dwarfed the smaller, unpredictable influence of the norm to such an extent that our overall conclusion is that the data should always be normalised, and unless a strong argument can be made for the use of a specific measure of distance, the norm used for distance measurements might as well be the L_2 norm due to its familiarity.

References

[1] C. C. Aggarwal, A. Hinneburg, D. A. Keim, On the surprising behaviour of distance metrics in high dimensional space, in: J. Van den Bussche, V. Vianu (Eds.), Proc. 8th Int. Conf. on Database Theory, London UK, Vol. 1973 of Lecture Notes in Computer Science, Springer, 2001, pp. 420–434.

[2] K. Doherty, R. Adams, N. Davey, Non-Euclidean Norms and Data Normalisation, in: M. Verleysen (Ed.), Proc. 12th Euro. Symposium on Artificial Neural Networks, Brugges, Begium, d-side publications, Brugges, 2004, pp. 181–186.

[3] P. H. Sneath, R. R. Sokal, Numerical Taxonomy - The Priciples and Practice of Numerical Classification, W.H.Freeman and Company, San Francisco, 1973.

[4] G. W. Milligan, M. C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (2) (1985) 159–179.

[5] C. Blake, C. Merz, (UCI) Repository of Machine Learning Databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html (1998).

[6] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is "nearest neighbor" meaningful?, in: C. Beeri, P. Buneman (Eds.), Proc. 7th Int. Conf. on Database Theory, Jerusalem, Israel, Vol. 1540 of Lecture Notes in Computer Science, Springer, 1999, pp. 217–235.

[7] J. A. Hartigan, Clustering Algorithms, John Wiley and Sons, Inc., New York, 1975.

[8] J. Hertz, A. Krogh, R. Palmer, Introduction to the Theory of Neural Computation, Addison-Wesley, Redwood City, CA, 1991.

[9] B. Fritzke, Unsupervised Ontogenic Networks, in: E. Fiesler, R. Beale (Eds.), Handbook of Neural Computation, IOP Publishing Ltd and Oxford University Press, 1997, pp. C2.4:1–C2.4:16.

[10] T. M. Martinetz, S. G. Berkovich, K. J. Schulten, Neural Gas Network for Vector Quantization and its Application to Time-Series Prediction, IEEE Trans. on Neural Networks 4 (4) (1993) 558–569.

[11] B. Fritzke, A Growing Neural Gas Network Learns Topologies, in: G. Tesauro, D. Touretzky, T. Leen (Eds.), Advances in Neural Processing Systems 7 (NIPS'94), MIT Press, Cambridge, 1995, pp. 625–632.

[12] T. Kohonen, Self-Organizing Maps, 2nd Edition, Springer-Verlag, Berlin, 1997.

[13] T. Kohonen, The Self-Organizing Map, Proceedings of the IEEE 78 (9) (1990) 1464–1480.

[14] T. M. Martinetz, K. J. Schulten, Topology representing networks, Neural Networks 7 (3) (1994) 507–522.

[15] D. Heinke, F. H. Hamker, Comparing neural networks: A benchmark on growing neural gas, growing cell structures, and fuzzy artmap, IEEE Trans. Neural Networks 9 (6) (1998) 1279–1291.