Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?

Michael E. Houle[1], Hans-Peter Kriegel[2], Peer Kröger[2], Erich Schubert[2], and Arthur Zimek[2]

[1] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
meh@nii.ac.jp
[2] Ludwig-Maximilians-Universität München
Oettingenstr. 67, 80538 München, Germany
{kriegel,kroegerp,schube,zimek}@dbs.ifi.lmu.de
http://www.dbs.ifi.lmu.de

Abstract. The performance of similarity measures for search, indexing, and data mining applications tends to degrade rapidly as the dimensionality of the data increases. The effects of the so-called 'curse of dimensionality' have been studied by researchers for data sets generated according to a single data distribution. In this paper, we study the effects of this phenomenon on different similarity measures for multiply-distributed data. In particular, we assess the performance of shared-neighbor similarity measures, which are secondary similarity measures based on the rankings of data objects induced by some primary distance measure. We find that rank-based similarity measures can result in more stable performance than their associated primary distance measures.

1 Introduction

Effective solutions for data indexing and data mining tasks often require that an appropriate measure of object-to-object similarity be provided. Operations such as the retrieval of objects similar to a query object are facilitated using a nearest-neighbor search with an appropriate distance measure. Any use of a similarity measure involves the implicit assumption that the data objects naturally form groups that can be regarded as arising from different generation mechanisms, and sharing common statistical characteristics. In the context of unsupervised learning, these groups can be clusters that follow some local 'natural' distribution. Sometimes, the learning process seeks to model the generation mechanism by fitting the data to known distributions; in other cases, only the groups themselves are sought. In outlier detection, the similarity measure is used to distinguish those objects that are conspicuously dissimilar from the majority of objects. In the context of classification, each class of the training set may be composed of one or more natural clusters, possibly together with outlier objects. In all contexts, we expect that a nearest-neighbor query based at an object from a particular natural grouping should rank objects from the same grouping ahead

M. Gertz and B. Ludäscher (Eds.): SSDBM 2010, LNCS 6187, pp. 482–500, 2010.

of other objects in the data set. In real-valued feature spaces, L_p norms or the cosine of the angle between the pair of vectors are commonly used to express similarities between vectors.

In general, similarity measures based on distances are sensitive to variations within a data distribution, or the dimensionality of a data space. These variations can limit the quality of the solution, the efficiency of the search, or both. For L_p norms in high dimensions, questions have been raised by several researchers, including Beyer et al. in [1], as to whether the concept of the nearest neighbor is meaningful. Intuitively, the key result of [1] states that if the ratio of the variance of the length of any point vector (denoted by $\|X_d\|$) with the length of the mean point vector (denoted by $E[\|X_d\|]$) converges to zero with increasing data dimensionality, then the proportional difference between the farthest-point distance D_{max} and the closest-point distance D_{min} (the *relative contrast*) vanishes:

$$\text{If } \lim_{d\to\infty} \text{var}\left(\frac{\|X_d\|}{E[\|X_d\|]}\right) = 0, \quad \text{then } \frac{D_{max} - D_{min}}{D_{min}} \to_p 0.$$

For a broad range of data distributions and distance measures, the relative contrast does diminish as the dimensionality increases. This *concentration effect* of the distance measure reduces the utility of the measure for discrimination. This phenomenon — recognized as one aspect of the *curse of dimensionality* — is quite general, and occurs for a broad range of data distributions and distance measures. In [2], the behavior of integer L_p norms in high dimensional spaces has been studied. The authors showed by means of an analytic argument that L_1 and L_2 are the only integer norms useful for higher dimensions. In addition, they studied the use of projections for discrimination, the effectiveness of which depended on localized dissimilarity measures that did not satisfy the symmetry and triangle inequality conditions of distance metrics. In [3], fractional L_p distance measures (with $0 < p < 1$) have been studied in a similar context. The authors provide evidence supporting the contention that smaller values of p offer better results in higher dimensional settings. These well-known studies generally assumed that the full data set followed a single data distribution, subject to certain restrictions. In fact, when the data follows a mixture of distributions, the concentration effect is not always observed; in such cases, distances between members of different distributions may not necessarily tend to the global mean as the dimensionality increases. As briefly noted in [1], if a data set is composed of many natural groupings or clusters, each following their own distribution, then the concentration effect will typically be less severe for queries based on points within a cluster of similar points generated according to the same mechanism, especially when the clusters are well-separated.

The fundamental differences between singly-distributed data and multiply-distributed data are discussed in detail in [4]. The authors demonstrate that nearest-neighbor queries are both theoretically and practically meaningful when the search is limited to objects from the same cluster (distribution) as the query point, and other clusters are well separated from the cluster in question. The key concept is that of *pairwise stability* of clusters, which is said to hold whenever the mean distance between points of different clusters dominates the mean

distance between points belonging to the same cluster. When the clusters are pairwise stable, for any point belonging to a given cluster, its nearest neighbors tend to belong to the same cluster. Here, a nearest-neighbor query of size on the order of the cluster size *can* be considered meaningful, whereas differentiation between nearest and farthest neighbors within the same cluster may still be meaningless. Note that for many common distributions these considerations may remain valid even as the dimension d tends to infinity: for example, two Gaussian distributions with widely separated means may find that their separability improves as the data dimension increases. However, it should also be noted that these arguments are based on the assumption that all dimensions bear information relevant to the different clusters, classes, or distributions. Depending on the ratio of relevant versus irrelevant attributes, and on the actual separation of sets of points belonging to different distributions, irrelevant attributes in a data set may impede the separability of different distributions and thus have the potential for rendering nearest neighbor query results less meaningful.

The observations of [4], and the important distinction between the effects of relevant and irrelevant attributes, both seem to have received little if any attention in the research literature. Despite the demonstrated deficiency of conventional L_p norms for high-dimensional data, a plethora of work based on the Euclidean distance has been dedicated to clustering strategies, which appear to be effective in practice to varying degrees for high-dimensional data [5]. Many heuristics have recently been proposed or evaluated for clustering [6, 7, 8, 9, 10, 11, 12, 13, 14], outlier detection [15, 16, 17, 18], and indexing or similarity search [6, 19, 20, 21, 22, 23] that seek to mitigate the effects of the curse of dimensionality. While some of these strategies, such as projected or subspace clustering, do recognize implicitly the effect of relevant versus irrelevant attributes for a cluster, all these papers (as well as others) abstain from discussing these effects, let alone studying them in detail. In particular, the concept of pairwise stability of clusters as introduced in [4] has not been taken into account in any of these papers. Although their underlying data models do generally assume (explicitly or otherwise) different underlying mechanisms for the formation of data groupings, they motivate their new approaches with only a passing reference to the curse of dimensionality. Indeed, it has been observed recently that many questions regarding these effects remain open [24]. Thus, a more detailed study of the effects of the curse of dimensionality on such heterogeneously distributed data sets in the presence of both relevant and irrelevant features is needed. One main objective of this paper is to attempt to address this need.

An interesting alternative to traditional similarity measurement is the definition of secondary measures based on the rankings induced by a specified primary similarity measure (such as an L_p norm, or the cosine measure). The simplest and most common of these methods involves the use of *shared nearest-neighbor* (SNN) information, in which the similarity value for an object pair (x, y) is a function of the number of data objects in the common intersection of fixed-sized neighborhoods centered at x and y, as determined by the primary measure. The primary similarity measure can be any function that determines a ranking of the

data objects relative to the query. It is not even necessary for the data objects to be represented as vectors.

The most basic form of shared nearest-neighbor similarity measure is that of the 'overlap'. Given a data set S consisting of $n = |S|$ objects and $s \in \mathbb{N}^+$, let $NN_s(x) \subseteq S$ be the set of s-nearest-neighbors of $x \in S$ as determined using some specified primary similarity measure. The overlap between objects x and y is defined to be the intersection size

$$SNN_s(x, y) = |NN_s(x) \cap NN_s(y)|. \tag{1}$$

Other similarity measures have been proposed based on the overlap, such as the *cosine measure*:

$$simcos_s(x, y) = \frac{SNN_s(x, y)}{s}, \tag{2}$$

so called as it is equivalent to the cosine of the angle between the zero-one set membership vectors for $NN_s(x)$ and $NN_s(y)$. This was used in [25,26] as a local density measure for clustering.

For computing the nearest neighbors in high dimensional data, SNN measures have been reported to be effective in practice, and supposedly less prone to the curse of dimensionality than conventional distance measures. SNN measures have found use in the design of merge criteria of agglomerative clustering algorithms [25,27,28], in approaches for clustering high-dimensional data sets [26,29], and in finding outliers in subspaces of high dimensional data [30]. However, in all of these studies, no systematic investigation has been made into the advantages of SNN measures over conventional distance measures for high-dimensional data.

The main contributions of this paper are as follows: (i) We present the first study of the effects of high data dimensionality for the more realistic scenario of data mixture models (as opposed to data following a single distribution), for a number of popular distance measures. (ii) We evaluate the performance of secondary similarity measures based on SNN information, as compared to the primary distances from which the rankings are derived. We demonstrate empirical evidence for the claim that SNN is more robust in higher dimensions than primary distances in widely varying settings of data set characteristics. (iii) We also provide interpretations for this observation: since the ranking of points is usually still meaningful in high dimensions, the overlap of the neighborhoods of two points in a common natural grouping can be expected to be substantially large, leading to a high SNN similarity value. The size of the overlap of neighborhoods of points from different groups is expected to be rather small, resulting in a low SNN similarity value. (iv) We derive several SNN-based secondary distance measures with the potential for good results for distance-based applications even when the curse of dimensionality limits the discrimination power of the underlying primary distance functions. In such situations, the distance-based implementation most likely performs worse than the SNN-based application for most choices of a primary distance function.

In the following section, we explore different aspects of the curse of dimensionality, and distinguish between the truth and myths surrounding this

phenomenon. We present the framework for our experimentation in Section 3. In Section 4, we evaluate how dimensionality affects the performance of SNN similarity, in contrast to that of the underlying primary similarity. In Section 5, we validate our findings on several real world data sets. This will motivate the formalization and discussion of possible distance measures based on the SNN dissimilarity, in Section 6. The results of the study are summarized in Section 7.

The data sets studied, the plots shown throughout this paper, as well as further information and experimental results and plots, are all available online via `http://www.dbs.ifi.lmu.de/research/SNN/`.

2 The Curse of Dimensionality Reconsidered

As mentioned earlier, previous studies of the effects of the curse of dimensionality on L_p norms mainly assume a common data distribution for all attributes of a given data set. Here, we investigate the effects of the curse in the presence of heterogeneous data distributions (for a more detailed discussion, see [5]):

Problem 1: Poor Discrimination of Distances

Concepts such as proximity, distance, or neighborhood become less meaningful with increasing dimensionality due to a loss of contrast of distances.

This is the fundamental problem studied in [1, 2, 3]. For any data mining, indexing, or similarity search application, this effect is a serious impediment to the successful treatment of high-dimensional data.

Problem 2: Presence of Irrelevant Attributes

Among the features of a high dimensional data set, for any given query object, many attributes can be expected to be irrelevant to that object. Irrelevant attributes can interfere with the performance of similarity queries for that object.

The relevance of any particular attribute may vary across different groups of objects within the same data set. Since natural clusters of the data are determined only by some subset of the available attributes, the presence of many irrelevant attributes may impede the efforts to identify these groups. The performance of distance measures may be seriously compromised even by a relatively small number of irrelevant attributes. As the total number of dimensions increases, one would expect more and more features to be irrelevant to a given query object. Many publications seem to confuse the problem of irrelevant attributes with that of Problem 1, but they are in fact different — it is not impossible to have poor discrimination of distances even when all attributes are relevant, and good discrimination even when many attributes are irrelevant.

Problem 3: Presence of Redundant Attributes

Similarly as with Problem 2, in a data set containing many attributes, there may be correlations or redundancies among subsets of attributes that also lead to special difficulties for data mining, indexing, or similarity search applications.

This issue relates to the concept of 'intrinsic dimensionality' of a data set. For spatial queries, the observation that the intrinsic dimensionality of a data set in

many cases is lower than the representational dimensionality (due to interdependencies among attributes) is often presented as a justification of strategies for obviating the curse of dimensionality [31,32,33,34]. It should be noted that there are scenarios where correlations among attributes do exist, but the problem of discrimination of distances still applies [1]. The correlations among attributes may be different within differing natural groups of a data set.

Since Problems 1 and 2 are often not well-differentiated in the literature, in our experimental studies, we will take care to demonstrate the differences in their natures and effects. In contrast with the earlier studies in [1,2,3], we limit our investigation here to mixtures of data distributions (as in [4]) as a realistic scenario for data mining or indexing or other similarity search applications.

3 Experimental Framework

3.1 Data Sets

To study the effects of the curse of dimensionality, we require a series of data sets that scale in dimensionality without introducing bias. After controlling for dimensionality, each of the sets in the series must be constructed so as to share common characteristics to the greatest degree possible. This is difficult to achieve with real world data, as the different attributes often vary in their scales and expressivity. When generating low-dimensional examples from a high-dimensional data set, it is not always clear how to select the projective dimensions fairly. In addition, well-defined ground truth sets necessary for assessing the expressiveness of query results are typically unavailable for large real data sets. The use of synthetic data allows us to study individual effects separately, while real data sets usually prevent the isolation of different influences. For these reasons we construct several series of artificial data sets using pseudo-random generators with largely fixed parameters, avoiding those parameter choices leading to data sets with groupings that are either too difficult or too easy to discriminate. Unless stated otherwise, the synthetic data sets were constructed with the following characteristics: $n = 10,000$ points grouped into $c = 100$ clusters in up to $d_{max} = 640$ dimensions. Cluster sizes are randomized with a mean of $\frac{n}{c} = 100$ and standard deviation $\frac{n}{10 \cdot c} = 10$, with the size of the last generated cluster adjusted so that the total number of points is n. When generating data sets for a series, those sets with dimensionality $d < d_{max}$ were generated so that their attributes coincided with the first d attributes of all other data sets having dimensionality greater than d.

For each object, attribute values were generated depending on whether the attribute is to be considered 'relevant' or 'irrelevant' for the formation of the cluster to which the object belongs. If the i-th attribute is deemed relevant to the j-th cluster, the value of this attribute for all members of c are normally distributed with a standard deviation in the range $\sigma_{j,i} \in [0.05 : 0.8]$, and a mean in the range $\mu_{j,i} \in [\frac{\sigma_{j,i}}{2} : 1 - \frac{\sigma_{j,i}}{2}]$. These ranges were chosen to avoid overly compact or overly wide distributions, as well as boundary effects, while still providing a wide variety of distributions and overlaps. No additional clipping or

normalization was applied. Any attributes irrelevant to the cluster were assigned noise values uniformly distributed in the interval $[0:1]$.

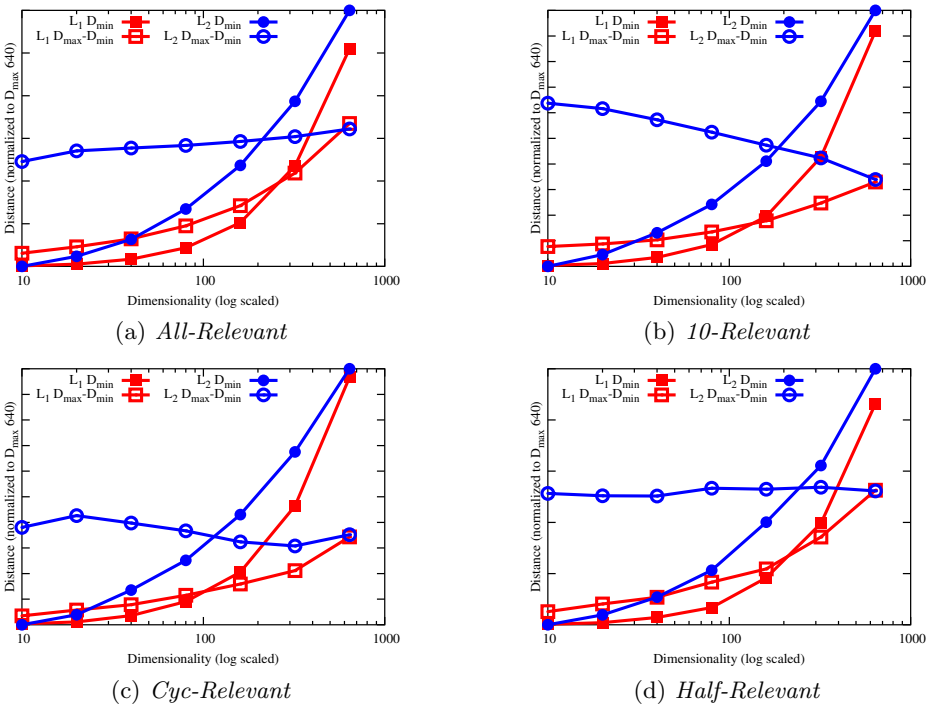
For the experimentation, 6 synthetic data series were created, each consisting of 7 sets of differing dimensionality $d = 10, 20, 40, 80, 160, 320, 640$: (i) *All-Relevant*: in this series, all attributes were generated so as to be relevant for all clusters. (ii) *10-Relevant*: in this series, the first 10 attributes are relevant for all clusters, the remaining attributes are irrelevant. (iii) *Cyc-Relevant*: in this series, the i-th attribute is relevant for the j-th cluster when $i \bmod c = j$; otherwise, the attribute is irrelevant. This series has $n = 1,000$ and $c = 10$. (iv) *Half-Relevant*: in this series, for each cluster, an attribute was chosen to be relevant with probability $\frac{1}{2}$, and irrelevant otherwise. The selection of attributes was consistent within a cluster, and performed independently of the selection for other clusters. (v) *All-Dependent*: this series is derived from *All-Relevant* introducing correlations among attributes. (vi) *10-Dependent*: this series is derived from *10-Relevant* introducing correlations among attributes.

For the correlated data sets *All-Dependent* and *10-Dependent*, the i-th attribute value X_i was generated by computing $X_i = Y_i$ for $1 \leq i \leq 10$, and $X_i = \frac{1}{2}(X_{i-10} + Y_i)$ for $i > 10$, where Y_i is the attribute of the corresponding uncorrelated data set *All-Relevant* or *10-Relevant*. This way of introducing correlations is inspired by Example 3 in [1].

These 6 series provide us with the means to study different aspects of the curse of dimensionality. Data series *All-Relevant* is the basic setting referred to in the statement of Problem 1. However, the sets differ from those considered in other studies [1,2,3], and conforms with [4] in that the data objects are partitioned into clusters (as are all our data sets). Data sets *10-Relevant* and *Cyc-Relevant* relate exclusively to Problem 2 in different settings. The clusters are further distinguished in the data set *Cyc-Relevant*, where every attribute is relevant for exactly one cluster. In the series *Half-Relevant*, we give up control of the number and choice of relevant attributes. Half the attributes are expected to be relevant to a given cluster, but the selection of relevant attributes varies (independently) from cluster to cluster.

Our synthetic data sets do not satisfy the IID (independent and identically-distributed) assumptions used in the proofs of [1], as the sets are composed of multiple clusters that overlap in some dimensions and are well-distinguished in others. However, the analysis of [24] applies when dimensional values are comparable in their extent and exhibit the same properties as for normalized data.

As intended, our synthetic data sets show the typical behavior ascribed to the curse of dimensionality. Figure 1 plots the numerator and denominator of the contrast formula $\frac{D_{max} - D_{min}}{D_{min}}$ for individual data sets, to demonstrate that D_{min} (the solid symbols) indeed grows much faster than the difference $D_{max} - D_{min}$ (the hollow symbols). The plots indicate that D_{min} grows exponentially faster than $D_{max} - D_{min}$. Plots for the correlated data series and other distance functions can be found on our web page, as well as plots showing the contrast directly.



(a) *All-Relevant* (b) *10-Relevant*

(c) *Cyc-Relevant* (d) *Half-Relevant*

Fig. 1. Curse of dimensionality: $D_{max} - D_{min}$ compared to D_{min}

In addition to the synthetic data sets described above, we also considered real-world data sets for our study. Real-world data sets suitable for this study are difficult to obtain, since they should have a reasonable size, number of classes, dimensionality, comparable dimensions and of course a solid ground truth. Results for real data can be difficult to interpret due to a lack of knowledge of the underlying data distributions, even when ground-truth class knowledge is available. Nevertheless, we report experimental results for 3 real data sets. The first real data set we used is the *Multiple Features* data set [35]. It consists of 2000 instances from 10 classes (corresponding to the digits 0 to 9). There are two variants, one with 649 dimensions (coming from multiple feature extraction algorithms and giving the data set its name), and another with 240 dimensions (the pixel averages features, which is the largest subset of directly comparable features). The second set considered is the *Optical Recognition of Handwritten Digits* data set [35]. It consists of 5620 instances from 10 classes (also corresponding to the digits 0 to 9) in 64 dimensions, in the form of an 8×8 grid of integer values in the range of 0 to 16 obtained by downsampling from a larger 32×32 grid. The third real data set comes from the *ALOI* image database [36], each image being described by 641 dense features based on color and texture histograms (for a detailed description of how the vectors were produced, see [37]). The full *ALOI* database consists of $110,250$ images of $1,000$ objects taken from different

orientations and in different lighting conditions, each object being treated as a class. We used only the first 22,050 instances of the data set, covering the first 205 objects, with an average class size of approximately 107 objects.

3.2 Distance Measures

As primary distance measures we considered for our experimental evaluation a range of different L_p distances, in particular the Manhattan (L_1) and Euclidean (L_2) distances, and the $p = 0.6$ and $p = 0.8$ fractional L_p distances. In addition, we used also the cosine distance, referred to here as *arccos*, as it is computed as the arc of the cosine similarity. All these distance measures can be used as the primary distance for the computation of a secondary similarity $simcos_s$, as defined in Equation 2. For our experiments, we use the distance measure $1 - simcos_s$, and compare the performance of this secondary distance measure with the corresponding primary distance measure to assess whether the accuracy is improved. There are other possibilities for constructing distance measures from similarity measures. The particular choice of method, however, does not affect the ranking of query results, although it may influence the contrast. We will discuss this further, in Section 6.

3.3 Evaluation Criteria

The purpose of a distance function is to facilitate the separation of data objects similar to the query from those objects which are not similar. The discriminative ability of a given distance function can best be evaluated by computing a nearest-neighbor ranking of all data points with respect to a given query point. Ideally, at the top positions of the ranking, we would find all objects drawn from the same natural cluster as the query object, followed by the objects from outside the cluster. To evaluate the discriminative ability of dissimilarity functions without referring to the actual values, we compute Receiver Operating Characteristic (ROC) curves that compare the true positive rate with the false positive rate. For each query, the objects are ranked according to their similarity to the query point. We can compute the matching ROC curve and the corresponding area under the curve (AUC) for each ranking result. An AUC of 1.0 indicates perfect discrimination — all relevant objects are ranked ahead of all other objects. An AUC of 0.5 indicates a total lack of discriminative ability, as this value is what would be expected with a uniform random permutation of the query result set. An AUC significantly less than 0.5 indicates a reversed ordering. The ROC curve and its AUC value provide a summary for a single ordering of points — that is, for a single query object. By generating a ROC curve and AUC value for each data object, the mean AUC value and standard deviation could then be used to rate the quality for a particular distance function. However, we expect points near the center of a cluster (the mean of the generating distribution) to discriminate well for many distance functions. On the other hand, for points near the border of a cluster or in the overlap of clusters, values of the dissimilarity measure will most likely perform less well. Therefore, at data generation time,

we assign to each point a centrality rank, based on both its deviation from the mean and the size of the cluster, so as to normalize across clusters of differing sizes. The point generated for cluster M that is closest to the mean of M is assigned a centrality of 1, and the point of M that is farthest from the mean of M is assigned a centrality of 0. To obtain readable graphs, the ROC AUC values then are aggregated into bins based on their centrality values. This allows us to plot the degradation of the distance function with respect to the centrality of a point within its distribution. For the plots shown in this paper, we will be using three bins for the central 20%, outer 20% and middle 60%. In the online material, 20 bins are used (each representing 5% of the data). In these plots we will also show the standard deviation along with the mean ROC AUC value.

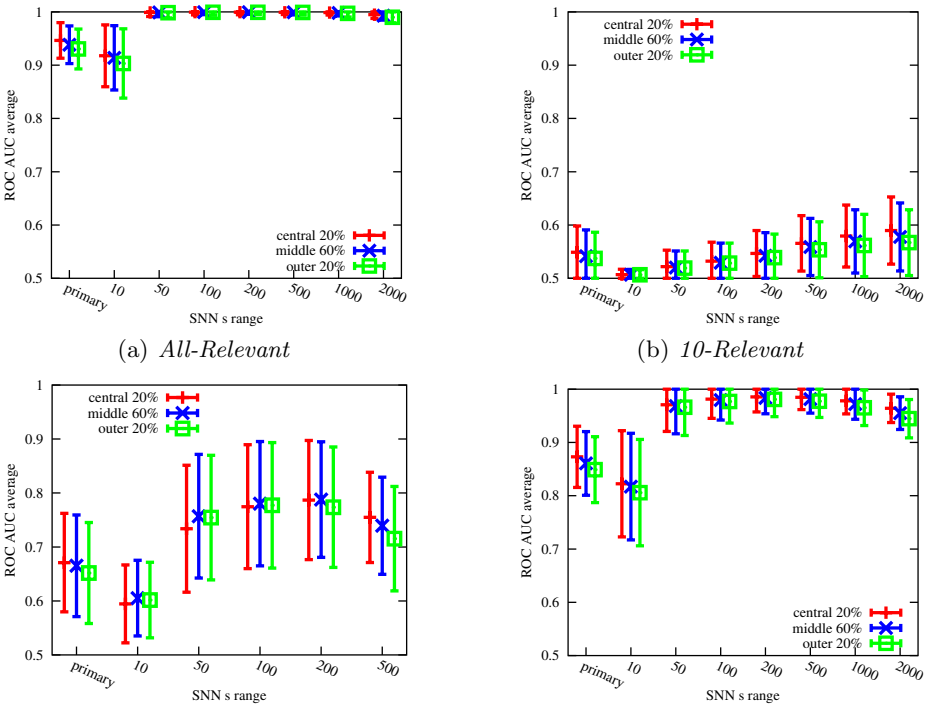
4 Effects of the Curse of Dimensionality

At first glance, the fact that our synthetic data sets exhibit the typical symptoms of the curse of dimensionality would seem to indicate that these data series are not amenable to indexing or mining. However, such a conclusion would be unnecessarily pessimistic. Especially for data sets with many relevant attributes (such as the *All-Relevant* series), any given number of clusters should become distinguishable when the number of relevant attributes becomes sufficiently large. This intuition is justified by examples such as the combination of kernels and support vector machines (SVM): the number of dimensions is increased in order to be able to separate classes linearly by hyperplanes. In fact, what is stated as a condition for the pairwise stability of clusters in [4], we would expect to hold for any two clusters where the number of discriminative attributes dominates. This is not an essentially original contribution of our study but confirms prior results. We provide, however, evidence for this effect in the online material.

One point that must be stressed is that while the curse of dimensionality tells us not to rely on the absolute values of distances, it is still viable to use distance values to derive a ranking of data objects. An ε-range query is dependent upon the choice of an appropriate value of ε, and thus suffers from the lack of contrast, whereas a k-nearest neighbor query will retrieve the top k neighbors independently of their absolute distance values. Hence, the computation of k-nearest neighbor queries and rankings has the potential to be viable in higher dimensions, whereas that of ε-range queries likely does not. Furthermore, although the curse of dimensionality contrast formula holds for all our data sets, the ranking results are not tied solely to the data dimensionality, and can in certain situations improve significantly with increasing dimensionality, as reported in [1]. The conclusion we draw is supported by the research literature as well as by our experiments on our synthetic data sets:

Conclusion 1: Relevant vs. Irrelevant Attributes

The quality of the ranking – and thus the separability of the different generating mechanisms – may not necessarily depend on the data dimensionality, but instead on the number of relevant attributes in the data set.



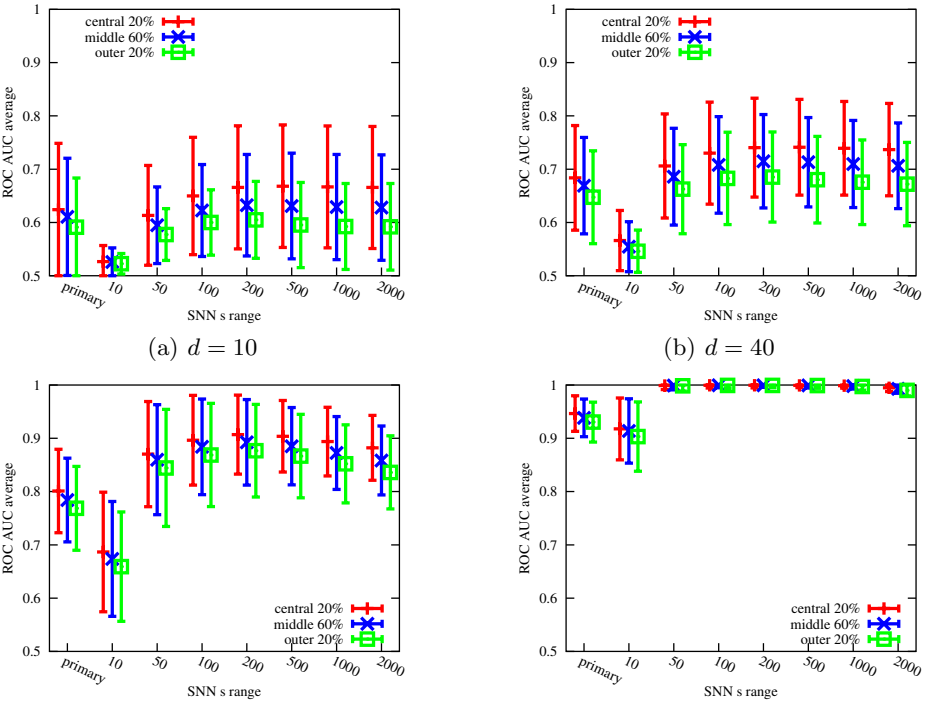
(a) *All-Relevant* (b) *10-Relevant*

(c) *Cyc-Relevant* (d) *Half-Relevant*

Fig. 2. Ranking quality with different SNN distances based on Euclidean distance at 640 dimensions

More specifically, there are two contrary effects of an increase in dimensionality when the number of relevant attributes is high: the relative contrast between points tends to decrease, but the separation among different generating mechanisms can increase. On the other hand, if the data dimensionality is high and the number of relevant dimensions is rather low, the curse of dimensionality fully applies, and hampers any analysis task. In retrospect, this is an important yet unsurprising conclusion to draw. Nevertheless, as mentioned in Section 1, it has not gained much recognition in the research literature to date.

As a further original contribution of this study, we evaluate the behavior of SNN as a secondary similarity measure. Motivated by the findings sketched above, an improved performance can be expected for a rank-based similarity measure such as SNN, whenever the ranking provided by the primary similarity measure is meaningful. Figure 2 compares results for the secondary distance measure with different SNN reference sizes s, based on Euclidean distance as the primary distance measure, for dimension $d = 640$. The performance of the corresponding primary distance is given on the left side of each diagram as a reference. Results for lower dimensionalities are comparable, and are shown in

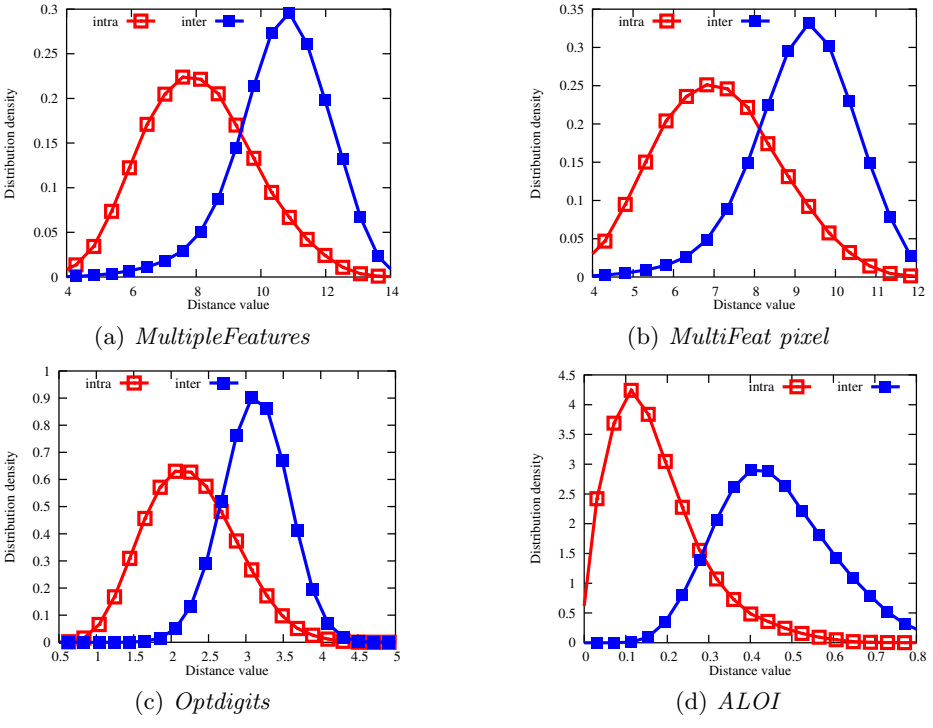


(a) $d = 10$ (b) $d = 40$

(c) $d = 160$ (d) $d = 640$

Fig. 3. Ranking quality for the *All-Relevant* set with different SNN distances based on L_2

the following figures. For easily separable data sets such as *All-Relevant*, most choices of s yield excellent results. On *Half-Relevant* and *Cyc-Relevant*, the best results are achieved for choices of s of the same order as the cluster size (100). This can also be seen for *All-Relevant* on lower dimensionality, where the contrast between the results is better. On the barely separable *10-Relevant* data set, even larger values of s seem to be needed, although the average ROC AUC score is not significant, being below 0.6. Figure 3 shows the same plots for different dimensionalities of the *All-Relevant* data set. It can be seen that by using an SNN distance, a considerable improvement can be achieved given that the data set is sufficiently separable, and that the parameter s is chosen roughly in the range of the cluster size. In particular, the secondary distance performs very well at high dimensionalities, and is reasonably robust with respect to the choice of s. The observations on the correlated data sets (given in the supplementary material) are quite similar. To summarize, we can draw the following conclusion from our experiments (see `http://www.dbs.ifi.lmu.de/research/SNN/` for the complete results with all distance functions on all data sets).



(a) *MultipleFeatures* (b) *MultiFeat pixel*

(c) *Optdigits* (d) *ALOI*

Fig. 4. Distributions of intra-class and inter-class distances (Euclidean distance)

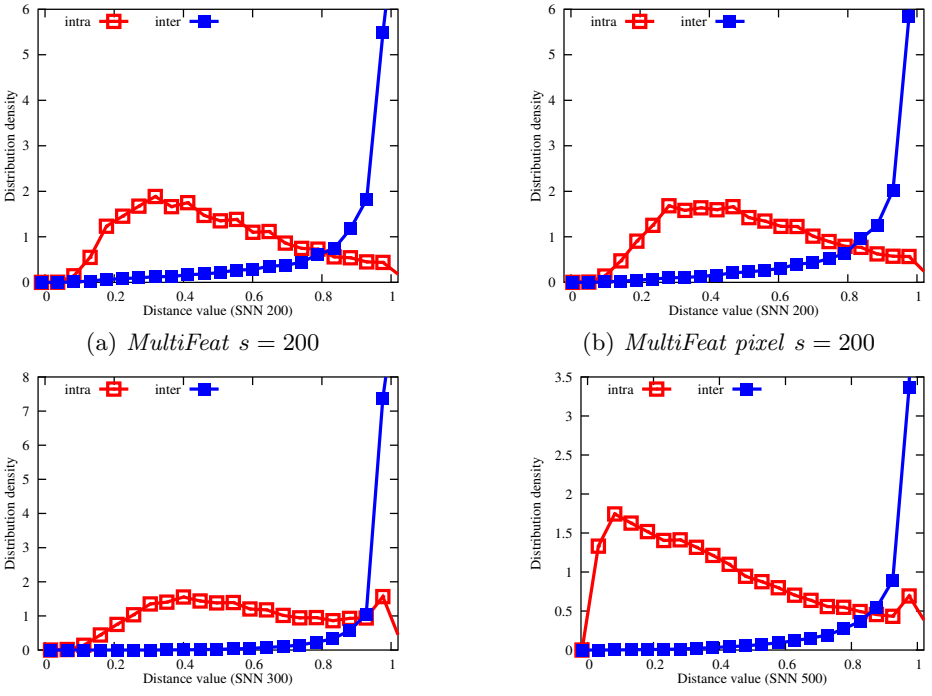
Conclusion 2: Ranking Quality Improvement

Our experiments suggest that the use of an SNN similarity measure can significantly boost the quality of a ranking compared to the use of the primary distance measure alone, provided that the primary distance already provides some degree of distinguishability of clusters.

The experimental results confirm that although the discrimination of primary distances worsens with increasing data dimensionality, the natural data groupings may still be separable and, if so, the neighborhoods of query points would contain many points from the same grouping. Clearly, for two points from a common data grouping, when increasing the value of s, the probability that their neighborhoods have significant overlap increases as well. On the other hand, if s is substantially larger than the size of the grouping, many objects from different groups are contained in the neighborhoods of the two points, and the performance of secondary distance measures become less predictable.

5 Experiments on Real Data

Experiments on artificial data allow more control over parameters such as the data dimension, and are more amenable to studying effects on the performance



(a) *MultiFeat* $s = 200$ (b) *MultiFeat pixel* $s = 200$

(c) *Optdigits* $s = 300$ (d) *ALOI* $s = 500$

Fig. 5. Distributions of intra-class and inter-class SNN distances (based on Euclidean distance)

of distance measures in isolation. Real-world data, on the other hand, is considerably more difficult to control in this way. Nevertheless, in this section we offer experimental results for real-world data sets in order to validate and confirm some of the effects observed for artificial data. As seen in Figure 4, on all the real data sets considered, the distance distributions are approximately Gaussian (which is to be expected in high dimensionalities for the L_p norms, due to the central limit theorem). It is also apparent that these data sets will be reasonably separable, as the overlap of the distance distributions is not very large. The results for other primary distances are comparable. Figure 5 shows the histogram results when using an SNN-based distance. The data set groupings have become very well separable. The effects of s on the results for real-data are as one would expect from the experiments on artificial data: Figure 6 displays the results for various sizes of s. Choosing s to match the class size gives reasonable results; however, the best performances are achieved with even larger values of s. Only when s approaches the full data set size does performance drop. The benefits of using SNN on the *ALOI* data set are minimal, as the groupings of that set are already very well separable for primary distances.

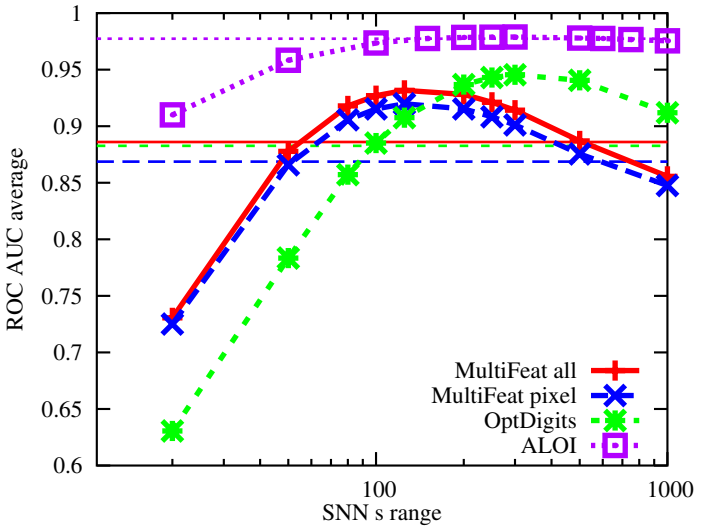


Fig. 6. Ranking quality with different SNN distances based on Euclidean distance (straight lines: ranking quality with primary distance)

6 Distance Measures Based on SNN

Let us recall, as described in Section 1, the observations reported by previous studies on the behavior of distance measures: (i) The relative contrast in Euclidean distances between nearest and farthest neighbor decreases with increasing dimensionality of the data [1]. (ii) This effect is stronger for L_p distances with higher values of p, while it remains weaker for the Manhattan distance L_1 [2]. (iii) Fractional distances — L_p distances with $p \in (0:1)$ — may even increase the relative contrast compared to L_p norms with $p \in \mathbb{N}^+$ [3]. The results of our experiments have not only confirmed these observations, but they also dispel some incorrectly held beliefs regarding the effects of dimensionality, through the investigation of data sets drawn from a mixture of distributions, each with varying relevance of attribute subsets.

Despite the performance limitations due to the presence of irrelevant attributes and due to the curse of dimensionality, our experimentation shows that traditional similarity measures can still serve as the basis of effective secondary similarity measures.

Conclusion 3: Stability of SNN

As an alternative to traditional distance measures such as L_p norms or the cosine distance, the performance of similarity search and its applications in data mining or indexing can be stabilized by using SNN secondary distance measures in preference to primary distances.

There are several common ways to convert a similarity measure into a dissimilarity measure. For the SNN similarity *simcos* (Equation 2) with a given number of neighbors s considered, we propose as possible distance measures:

$$dinv_s(x, y) = 1 - simcos_s(x, y) \tag{3}$$

$$dacos_s(x, y) = \arccos\left(simcos_s(x, y)\right) \tag{4}$$

$$dln_s(x, y) = -\ln simcos_s(x, y) \tag{5}$$

While *dinv*, which has been used throughout our experiments, is simply a linear inversion of the values, *dacos* penalizes slightly suboptimal similarities more strongly, whereas *dln* is more tolerant than *dinv* for a broad range of higher similarity values but approaches infinity for very low similarity values. In general, any function f that is monotonically decreasing on the interval $[0 : 1]$ with $f(1) = 0$ can be used to transform the SNN similarity measure into a dissimilarity measure. The functions only differ in their contrast at different ranges. All of these functions are symmetric (since *simcos* is symmetric) and maintain the same ranking. However, it should be noted that of the three, only *dacos* satisfies the triangle inequality. While most retrieval results (based simply on rankings) remain unaffected by different formulations of these secondary distances, the effects on indexing and clustering may vary from formulation to formulation. For example, the separation of clusters in terms of absolute distances depends on the concrete choice of the distance measure and on the secondary distance measure as well.

7 Conclusion

With the ever-increasing capabilities of automatic data generation, the demand is rising for analysis methods that can cope with high dimensional data. The notorious curse of dimensionality and its implications for similarity measurement have been the subject of several recent studies; however, these studies have evaluated only data sets generated according to a single distribution mechanism. Moreover, a number of myths surrounding the effects of the curse of dimensionality have been supported by too loose interpretations of these studies. Seemingly in contradiction to these studies, the SNN similarity measure has been reported to be able to alleviate the effects of the curse for clustering.

In light of these considerations, this paper has made the following contributions. We have presented the first study of the effects of high data dimensionality on a range of popular distance measures, for the more realistic scenario of data mixture models as opposed to data following a single distribution. We exposed some of the myths involving the curse of dimensionality, and partly confirmed previously reported truths. We demonstrated that although the contrast of pairwise distances diminishes with increasing dimensionality (severely hampering all distance-based algorithms), for realistic data sets with a mixture of local distributions, the discrimination power of a distance measure depends more strongly on the number of relevant dimensions, and can actually rise as the dimensionality increases. On the other hand, simultaneously increasing the data dimensionality and decreasing the number of relevant dimensions dramatically decreases the

separability of local distributions. In such a scenario, it seems to be more suitable to separate the groupings by means of projection into subspaces.

In addition, we evaluated the performance of secondary similarity measures based on SNN information, as compared to the primary distances from which the rankings are derived. We empirically confirmed that SNN is more robust in higher dimensions than primary distances in all settings. We also provided explanations for this observation: since the ranking of points is typically still meaningful in high dimensions, the overlap of the neighborhoods of two points in a common natural grouping can be expected to be substantially large, leading to a high SNN similarity value; the size of the overlap of neighborhoods of points from different groups is expected to be rather small, resulting in a low SNN similarity value.

Last but not least, we derived several SNN-based secondary distance measures with the potential for good results for distance-based applications even when the curse of dimensionality limits the discrimination power of the underlying primary distance functions. In such situations, the distance-based implementation most likely performs worse than the SNN-based application for most choices of a primary distance function.

In summary, for high dimensional applications, despite a deteriorating contrast in the chosen primary distance measure, we expect the incorporation of ranking information to enhance the quality of rankings and query results.

References

1. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: Beeri, C., Bruneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 217–235. Springer, Heidelberg (1998)
2. Hinneburg, A., Aggarwal, C.C., Keim, D.A.: What is the nearest neighbor in high dimensional spaces? In: Proc. VLDB (2000)
3. Aggarwal, C.C., Hinneburg, A., Keim, D.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, p. 420. Springer, Heidelberg (2000)
4. Bennett, K.P., Fayyad, U., Geiger, D.: Density-based indexing for approximate nearest-neighbor queries. In: Proc. KDD (1999)
5. Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM TKDD 3(1), 1–58 (2009)
6. Aggarwal, C.C.: Re-designing distance functions and distance-based applications for high dimensional data. SIGMOD Record 30(1), 13–18 (2001)
7. Domeniconi, C., Papadopoulos, D., Gunopulos, D., Ma, S.: Subspace clustering of high dimensional data. In: Proc. SDM (2004)
8. Woo, K.G., Lee, J.H., Kim, M.H., Lee, Y.J.: FINDIT: a fast and intelligent subspace clustering algorithm using dimension voting. Inform. Software Technol. 46(4), 255–271 (2004)

9. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: A review. SIGKDD Explorations 6(1), 90–105 (2004)
10. Yiu, M.L., Mamoulis, N.: Iterative projected clustering by subspace mining. IEEE TKDE 17(2), 176–189 (2005)
11. Liu, G., Li, J., Sim, K., Wong, L.: Distance based subspace clustering with flexible dimension partitioning. In: Proc. ICDE (2007)
12. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: A general framework for increasing the robustness of PCA-based correlation clustering algorithms. In: Ludäscher, B., Mamoulis, N. (eds.) SSDBM 2008. LNCS, vol. 5069, pp. 418–435. Springer, Heidelberg (2008)
13. Moise, G., Sander, J.: Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering. In: Proc. KDD (2008)
14. Achtert, E., Böhm, C., David, J., Kröger, P., Zimek, A.: Global correlation clustering based on the Hough transform. Stat. Anal. Data Min. 1(3), 111–127 (2008)
15. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. In: Proc. SIGMOD (2001)
16. Zhu, C., Kitagawa, H., Faloutsos, C.: Example-based robust outlier detection in high dimensional datasets. In: Proc. ICDM (2005)
17. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: Proc. KDD (2008)
18. Müller, E., Assent, I., Steinhausen, U., Seidl, T.: OutRank: ranking outliers in high dimensional data. In: Proc. ICDE Workshop DBRank (2008)
19. Katayama, N., Satoh, S.: Distinctiveness-sensitive nearest-neighbor search for efficient similarity retrieval of multimedia information. In: Proc. ICDE (2001)
20. Aggarwal, C.C., Yu, P.S.: Finding generalized projected clusters in high dimensional space. In: Proc. SIGMOD (2000)
21. Berchtold, S., Böhm, C., Jagadish, H.V., Kriegel, H.P., Sander, J.: Independent Quantization: An index compression technique for high-dimensional data spaces. In: Proc. ICDE (2000)
22. Jin, H., Ooi, B.C., Shen, H.T., Yu, C., Zhou, A.Y.: An adaptive and efficient dimensionality reduction algorithm for high-dimensional indexing. In: Proc. ICDE (2003)
23. Aggarwal, C.C., Yu, P.S.: On high dimensional indexing of uncertain data. In: Proc. ICDE (2008)
24. Francois, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. IEEE TKDE 19(7), 873–886 (2007)
25. Ertöz, L., Steinbach, M., Kumar, V.: Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: Proc. SDM (2003)
26. Houle, M.E.: Navigating massive data sets via local clustering. In: Proc. KDD (2003)
27. Guha, S., Rastogi, R., Shim, K.: CURE: An efficient clustering algorithm for large databases. In: Proc. SIGMOD, pp. 73–84 (1998)
28. Jarvis, R.A., Patrick, E.A.: Clustering using a similarity measure based on shared near neighbors. IEEE TC C-22(11), 1025–1034 (1973)
29. Houle, M.E.: The relevant-set correlation model for data clustering. Stat. Anal. Data Min. 1(3), 157–176 (2008)
30. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Outlier detection in axis-parallel subspaces of high dimensional data. In: Proc. PAKDD (2009)

31. Faloutsos, C., Kamel, I.: Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In: Proc. SIGMOD (1994)
32. Belussi, A., Faloutsos, C.: Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In: Proc. VLDB (1995)
33. Pagel, B.U., Korn, F., Faloutsos, C.: Deflating the dimensionality curse using multiple fractal dimensions. In: Proc. ICDE (2000)
34. Korn, F., Pagel, B.U., Falutsos, C.: On the "dimensionality curse" and the "self-similarity blessing". IEEE TKDE 13(1), 96–111 (2001)
35. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
36. Geusebroek, J.M., Burghouts, G.J., Smeulders, A.: The Amsterdam Library of Object Images. Int. J. Computer Vision 61(1), 103–112 (2005)
37. Boujemaa, N., Fauqueur, J., Ferecatu, M., Fleuret, F., Gouet, V., Saux, B.L., Sahbi, H.: IKONA: Interactive generic and specific image retrieval. In: Proc. MMCBIR, pp. 25–28 (2001)