23th Benelux Meeting on Systems and Control
Helvoirt (The Netherlands), 17-19 March 2004, p. 112

Open questions about similarity search in high-dimensional spaces

Damien Francois(1), Vincent Wertz(1), Michel Verleysen(2)
Université catholique de Louvain
(1)CESAME Research Center
Av. G. Lemaitre, 4
B-1348 Louvain la Neuve
{francois,wertz}@auto.ucl.ac.be
(2)DICE - Machine Learning Group
Place du Levant, 3
B-1348 Louvain la Neuve
verleysen@dice.ucl.ac.be

1 Similarity search

Many data analysis methods (classification tools, clustering algorithms, ...) make use of a *similarity measure* on data. For example the well-known k-NN classifier determines the class of a new data element according to its 'similarity' with other elements for which the class label is known.

In many cases, those data are embedded (described) in a metric space (often a Euclidean space) and the similarity between two data elements is measured by the distance between their respective vector representations in the space.

A growing number of applications now involve complex data that require a high number of numerical components to be completely described. Those data have to be embedded in high-dimensional spaces (from tens to thousands dimensions). Examples are spectrophotometric data, gene expression data, texts, pictures, etc.

2 The concentration of measure phenomenon

It is well known that the Euclidean norm is subject to the *concentration phenomenon*, which expresses that the respective norms of two randomly chosen vectors in a high-dimensional space will be very similar, with a high probability. That leads to question the relevance of the Euclidean distance as a measure of similarity for complex data.

It can indeed be shown that, if $x = [x_1, \cdots, x_d]$ is a random variable in $\Re^d$,

$$\lim_{d\to\infty} \frac{\text{Std}\,(\|x\|)}{\text{E}\,(\|x\|)} = 0. \qquad (1)$$

The equation says that when dimensionality grows, the standard deviation of the norm (or Euclidean distance to origin) of a random vector gets small compared to the expected value of the norm. This means that the norm of a high dimensional vector becomes nearly a constant independant on the coordinates of the vector !

Furthermore, Beyer [1] have proved that for any random $x = [x_1, \cdots, x_d] \in \Re^d$ surrounded by other points $y_i \in \Re^d$,

$$\lim_{d\to\infty} \frac{\max_i(d(x, y_i)) - \min_i(d(x, y_i))}{\min_i(d(x, y_i))} = 0. \qquad (2)$$

In other words, the distances from a point to its nearest and farthest neighbours respectively, tend to be quite similar when dimension is high...

3 Practical considerations

The above-mentioned results were obtained from a theoretical viewpoint. We need to further investigate whether the intuition of irrelevance of the Euclidean norm as a similarity measure in high-dimensional spaces does have an impact in practical cases. The following questions are thus of interest.

Has the concentration phenomenon an impact on the stability of a nearest neighbour search ? Indeed we would like that if a data element x is similar to y, then y would be similar to x. In other words, if x is the nearest neighbour of y, y should be among the closest neighbours of x.

Will the use of some other metric less subject to concentration help ? It can be shown that Minkowski metrics

$$\|x\|_p = \left(\sum_i (x_i)^p\right)^{\frac{1}{p}}$$

have slower convergence rates to concentration when p is small. Will a nearest neighbour search be more stable if the value of p is well chosen ?

What impact has concentration on robustness to noise ? If two ideal elements are similar, we would like the corresponding observed data to be similar too. Are all Minkowski norms equally robust to noise ?

How influent is the intrinsic dimensionality of the data? Results (1) and (2) rely on the independance of the components x_i. Is concentration observed when there are dependences ?

Answering those questions will help improving the global performances of data analysis methods that rely on data similarity estimation.

References

[1] K.S. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft: "When Is 'Nearest Neighbor' Meaningful?", Proc. 7th International Conference on Database Theory (ICDT'99), pp.217-235