Features of Neighbors Spaces*

Marcel Jirina[1], Marcel Jirina, jr.[2]

[1] Institute of Computer Science AS CR, Pod vodarenskou vezi 2,
182 07 Prague 8 – Liben, Czech Republic
marcel@cs.cas.cz
http://www.cs.cas.cz/~jirina

[2] Center of Applied Cybernetics, Faculty of Electrical Engineering,
Czech Technical University in Prague, Technicka 2,
166 27 Prague 6 – Dejvice, Czech Republic
jirina@fel.cvut.cz

Contents

Abstract. Distances of the nearest neighbor or several nearest neighbors are essential in probability density estimate by the method of *k* nearest neighbors or in problems of searching in large databases. A typical task of the probability density estimate using several nearest neighbors is the Bayes's classifier. The task of searching in large databases is looking for other nearest neighbor queries. In this paper it is shown that for a uniform distribution of points in an *n*-dimensional Euclidean space the distribution of the distance of the i-th nearest neighbor to the *n*-power has Erlang distribution. The power approximation of the newly introduced probability distribution mapping function of distances of nearest neighbors in the form of suitable power of the distance is presented. A way to state distribution mapping exponent *q* for a probability density estimation including boundary effect in high dimensions is shown.

* Final version published: SOFSEM 2004: Theory and Practice of Computer Science (Peter Van Emde Boas, Jaroslav Pokorný, Mária Bieliková, Eds.), Berlin : Springer Verlag, Lecture Notes in Computer Science, vol. 2932, 2004. pp. 241-248. ISBN 3-540-20779-1. Held: SOFSEM 2004. Conference on Current Trends in Theory and Practice of Computer Science /30./, Měřín, 24.01.2004-30.01.2004, CZ.

1 Introduction

In a probability density estimate by the method of *k* nearest neighbors [5], [7] or in problems of searching in large databases [1], [2], [3], distances of the nearest neighbor or several nearest neighbors are essential.
The rather strange behavior of the nearest neighbors in high dimensional spaces was accoutered. For the problem of searching the nearest neighbor in large databases the boundary phenomenon was studied in [1] using approximation by so called bucketing algorithm and l_{max} metrics. In [10] the problem of finding *k* nearest neighbors was studied in general metric spaces and in [11] the so-called concentration phenomenon was described. In [2] and in [11] it was found that as dimensionality increases, the distance to the nearest data point approaches the distance to the farthest data point of the learning set.

For probability density estimation by the *k*-nearest-neighbor method, the best value of *k* must be carefully tuned to find optimal results. Let there be a ball with its center in *x* and containing *k* points. Let the volume of the ball be V_k and total number of points m_T. Then for the probability density estimate in point *x* (a query point [3]) it holds [5]

$$p_k(x) = \frac{k / m_T}{V_k} . \qquad (1)$$

It will be shown that starting from some *k* the value of $p_k(x)$ is not constant for larger *k*, as it should be, but lessens. It is caused by the ”boundary effect“.

The goal of this study is to analyze the distances of the nearest neighbors from the query point *x* and the distances between two of these neighbors, the *i*-th and (*i*-1)-st in the space of randomly distributed points without and with boundary effect consideration. We introduce the probability distribution mapping function, and the distribution density mapping function which maps probability density distribution of points in E_n to a similar distribution in the space of distances, which is one-dimensional, i.e. E_1. The power approximation of the probability distribution mapping function in the form of (distance)^q is introduced and a way to choose distribution mapping exponent *q* for a probability density estimation including the boundary effect in high dimensions is shown.

2 Probability Density Estimate Based on Powers of Distances

The nearest-neighbor-based methods usually use (1) for a probability density estimate and are based on the distances of neighbors from a given (unknown) point, i.e. on a simple transformation $E_n \rightarrow E_1$.
The idea of most nearest-neighbors-based methods as well as kernel methods [7] does not reflect the boundary effects. That means that for any point *x*, the statistical distribution of the data points x_i surrounding it is supposed to be independent of the location of the neighbor points and their distances x_i from point *x*. This assumption is often not met, especially for small data sets and higher dimensions.

To illustrate this, let us consider points uniformly distributed in a cube and a ball inserted tightly into the cube. The higher space dimension the smaller amount of the cube is occupied by the ball. In other words, the majority of points lie outside the ball somewhere “in the corner” of the cube (the boundary effect [1]). It seems that in farther places from the origin, the space is less dense than near the origin.

Let us look at function $f(i) = r_i^n$, where r_i is the mean distance of the *i*-th neighbor from point *x*. The function should grow linearly with index *i* in the case of uniform distribution without the boundary effect mentioned. In the other case this function grows faster than linearly and therefore we suggest choosing function $f(i)=r_i^q$, where $q \leq n$ is a suitable power discussed later.

3 Uniform Distribution without Boundary Effects

Let us assume random and uniform distribution of points in some subspace S of E_n. Further suppose that point x is inside S in the following sense: For each neighbor y considered the ball with its center at x and radius equal to $||x\text{-}y||$ lies inside S. This is the case where the boundary effects do not take place.

In this Chapter one-dimensional case is studied, the multidimensional case is subject of the next Chapter.

Points spread on a line randomly and uniformly. In this case the distance Δ between two neighbor points is a random variable with exponential distribution function $P(\Delta)=1-e^{-\lambda\Delta}$ and probability density $p(\Delta)=e^{-\lambda\Delta}$ [4], [6]. For this distribution the mean is $E\{\Delta\}= 1/\lambda = d$ and it is the mean distance between two neighbor points.

Randomly chosen point on a line with randomly and uniformly spread points. It can be simply derived that the distance Δ between this point and the nearest of two its neighbor points is a random variable with exponential distribution function which differs from the previous case by $\lambda = 2/d$.

The second, third, etc. nearest neighbor. Let us sort consecutive nearest neighbors from point x in an ascending order. The mean distance between two successive points is $d/2$. It is the same situation as in the previous case. From it it follows that the distance between two successive points has the exponential distribution with $\lambda = 2/d$. $1/\lambda$ corresponds to one half of the mean distance of the neighbor points on the line.

Composed distribution

Let the true distance of the i-th neighbor be denoted x_i. Let the difference in the distances of two successive neighbors, the (k-1)-st and the k-th be denoted $x_{k-1,\,k}$. Then it holds

$$x_i = x_1 + \sum_{k=2}^{i} x_{k-1,k}$$

because the distance is simply a sum of all successive differences and it is a sum of independent exponentially distributed random variables. The probability density of the sum of independent random variables is given by convolution of the probability densities of these variables. Assuming exponential distribution of individual random variables then the composed distribution has gamma or Erlang distribution Erl(i, λ).

For statistical distribution of distances of the first, the second, the third, ..., the i-th nearest point from point x with $\lambda = 2/d$ it holds Erl(1, λ) = Exp(λ), Erl(2, λ) = $\lambda^2 x e^{-\lambda x}$, etc.

4 Probability distribution mapping function

To study a probability distribution of points in the neighborhood of a query point x let us construct individual balls around point x embedded one into another like peels of onion. Radii of individual balls can be expressed by formula $r_i = const.\sqrt[n]{V_i}$. A mapping between the mean density in an i-th peel ρ_i and its radius r_i is $\rho_i = p(r_i)$. $p(r_i)$ is the mean probability density in the i-th ball peel with radius r_i. The probability distribution of points in the neighborhood of a query point x is thus simplified to a function of a scalar variable. We call this function a probability distribution mapping function $D(x, r)$ and its partial derivation according to r the distribution density mapping function $d(x, r)$. Functions $D(x, r)$ and $d(x, r)$ for x fixed are one-dimensional analogs to the probability distribution function and the probability density, respectively.

4.1 Nearest Neighbors in E_n

A number of points in a ball neighborhood with a center in the query point x and the probability distribution mapping function $D(x, r)$ grow with the n-th power of distance from the query point. Let us denote this n-th power of distance from the query point $d_{(n)}$. Let a, b be distances of two points from a query point x in E_n. Then it holds $d_{(n)} = \left|a^n - b^n\right|$. Using $d_{(n)}$ instead of r, both the number of points and the $D(x, r)$ grow linearly with r^n. The distribution density mapping function $d(x, r^n)$ taken as $\frac{\partial}{\partial(r^n)}D(x, r^n)$ is constant.

It can be easily seen that $d_{(n)}$ of successive neighbors is a random variable with exponential distribution function. The $d_{(n)}$ of the i-th nearest neighbor from the query point is given by the sum of $d_{(n)}$ between the successive neighbors. Then, it is a random variable with Erlang distribution Erl(i, λ), $\lambda = 1/\bar{d}_{(n)}$, where $\bar{d}_{(n)}$ is mean $d_{(n)}$ between the successive neighbors. The only difference is that instead of distance r the n-th power of distance is used in an n-dimensional Euclidean space and then $d_{(n)} = r_i^n - r_{i-1}^n$, $r_0^n = 0$.

Example of a Uniform Ball. Let us suppose a ball in an n-dimensional space containing uniformly distributed points over its volume. Let us divide the ball on concentric "peels" of the same volume. Using the formula $r_i = S(n)/2^n.\sqrt[n]{V_i}$ we obtain a quite interesting succession of radii corresponding to the individual volumes. The symbol $S(n)$ denotes the volume of a ball with unit radius in E_n; note $S(3) = {}^4/_3\pi$. The higher space dimension is considered the more similar values of the radii approaching the outer ball radius are obtained. The inner part of the ball is thus nearly empty. The $d_{(n)}$ or radius to the n-th power correspond much better to the probability distribution mapping function $D(x, r) = const.r^n$. The radii of balls to the n-th power grow thus linearly. Differences are then constant, which well corresponds to uniformity of the distribution.

4.2 Influence of a dimensionality and a total number of points

Distances, as well as $d_{(n)}$, of several nearest neighbors depend on probability distribution $p(z)$ of points in the neighborhood of a query point x and also on true density of points in this neighborhood.

Example of normal distribution. Let us consider a distribution of m_T samples with distribution density mapping function $d(x, r^n) = 2N(0,1)\big|_{r^N \geq 0} = 2\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}r^{2n}}$ in E_n. Coefficient 2 was introduced to get $D(x,\ r_n) \rightarrow 1$ for $r_n \rightarrow \infty$. Thus $D(x,\ r_n)$ can be considered as a probability distribution function. For the mean r_i^n of r^n for the i-th nearest neighbor of the query point x it holds that $r_i^n = N^{-1}(0.5 + 0.5i / m_T)$.

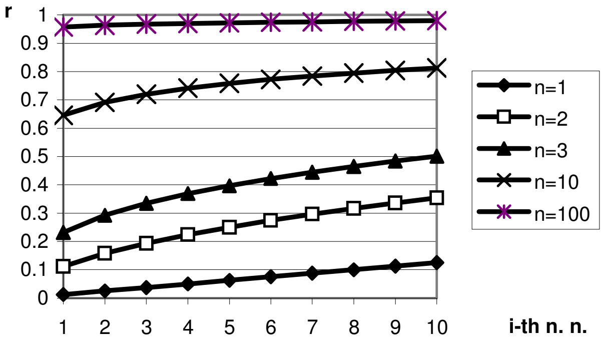

Fig. 1. Distances r_i^n of the first several nearest neighbors for different space dimensions n

Fig. 1 shows distances r_i^n of the first several nearest neighbors for different space dimensions *n*. Fig. 2 shows distances r_i^n for different numbers of points (samples in the set) for different space dimensions *n*.

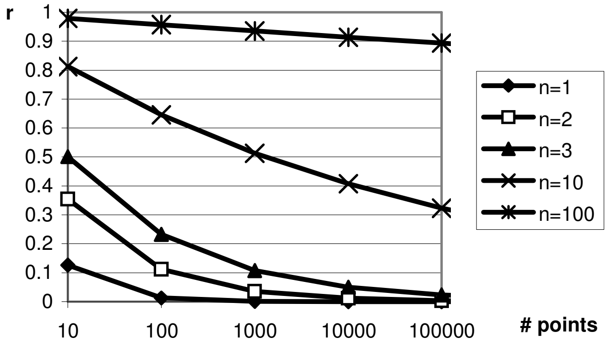

Fig. 2. Distances r_i^n for different numbers of points for different space dimensions *n*

5. Influence of boundary effects

The problem of boundary effects was studied in [1] in l_{max} metric and for different problems of searching the nearest neighbor in large databases. Taking the boundary effects into account, the estimation of the searching time was lesser than if no boundary effect is considered and is closer to reality.

5.1 Boundary effect phenomenon

Let the boundary effect be understood as a phenomenon such that within a given spherical neighborhood with radius *r* the probability density is a (not strictly) decreasing function and decreases starting from some point. Moreover, this fact influences the mean distances of neighbors of the query point so that these distances do not correspond to the uniform distribution. The boundary effect is demonstrated on the example of a uniform cube in Fig. 3.

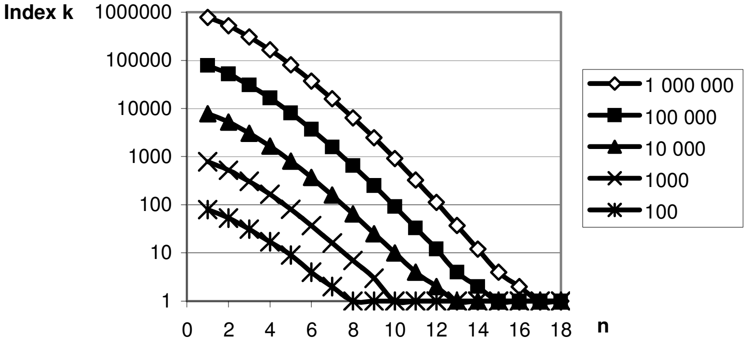

Fig. 3. Index k of the k-th nearest neighbor for which the boundary effect arises for different dimensionalities n and for m_T = 100 till 1,000,000 points (a unit cube with uniform distribution and the query point x in the origin)

5.2 Power approximation of the probability distribution mapping function

A course of the probability distribution mapping function is often not known and it is not easy to derive it analytically. Therefore, we suggest using power approximation r_q of the probability distribution mapping function $D(x, r_n)$ such that $\frac{D(x, r^n)}{r^q} \to const$ for $r \to 0+$.

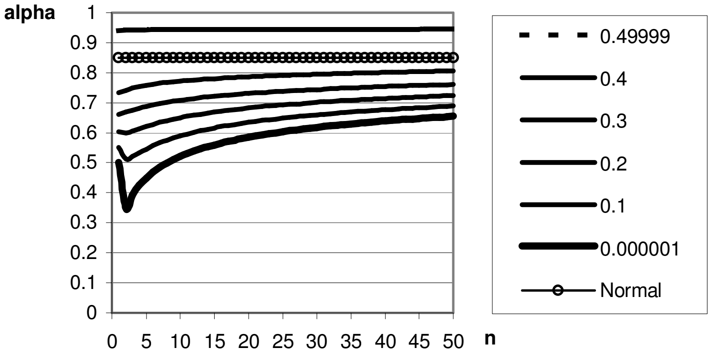

Fig. 4. Distribution mapping ratio α as a function of dimension n for normal distribution and for the uniform unit cube, with the query point in the distance l ($0 < l < 0.5$, see legend) from the center of the cube to the center of one of its walls.

The exponent q is the distribution mapping exponent. The variable $\alpha = q/n$ we call the distribution mapping ratio. Using the approximation of the probability distribution mapping function by $D(x, r^n) = const.(r^n)^\alpha$, the distribution mapping exponent is $q = n\alpha$.

An example of values of the distribution mapping exponent is shown in Fig. 4.

6. Conclusion

By using a notion of distance, i.e. a simple transformation En → E1, the problems with dimensionality are easily eliminated at a loss of information on the true distribution of points in the neighborhood of the query point. It is known [1], [3] that for larger dimensions something like local approximation of real distribution by uniform distribution does not exist. But the assumption of at least local uniformity in the neighborhood of a query point is usually inherent in methods based on the distances of neighbors.

This problem is solved by introduction of a power approximation of the probability distribution mapping function here. An essential variable of this approximation is the distribution mapping exponent. By using this exponent the real distribution is transformed to be uniform. It is possible to do it either locally or globally.

In essence, there are two ways in estimating the distribution mapping exponent. One of them is to estimate this exponent globally for the whole data set and rely on not too large local differences. The other way is to estimate the distribution mapping exponent locally, i.e. for each query point anew. A disadvantage of this approach is a large possible error in the distribution mapping exponent when a small number of neighbor points is used. Processing of a larger number of points, on the other hand, makes estimation closer to global estimation especially for small data sets.

References

[1] Arya, S., Mount, D.M., Narayan, O.: Accounting for Boundary Effects in Nearest Neighbor Searching. Discrete and Computational Geometry, Vol. 16 (1996) 155-176

[2] Beyer, K. et al.: When is “Nearest Neighbor” Meaningful? Proc. of the 7th International Conference on Database Theory. Jerusalem, Israel (1999) 217-235

[3] Burton, B.G.: The Poisson distribution and the Poisson process. http://www.zoo.cam.ac.uk/zoostaff/laughlin/brian/minireviews/poisson/poisson.pdf

[4] Demaret, J.C., Gareet, A.: Sum of Exponential Random Variables. AEÜ, Vol. 31, No. 11 (1977) 445-448

[5] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Second Edition. John Wiley and Sons, Inc., New York (2000)

[6] Eadie, Wt.T. et al.: Statistical Methods in Experimental Physics. North-Holland (1982)

[7] Hinnenburg, A., Aggarwal, C.C., Keim, D.A.: What is the nearest neighbor in high dimensional spaces? Proc. of the 26th VLDB Conf., Cairo, Egypt (2000) 506-515

[8] Kleinrock, L.: Queueing Systems, Volume I: Theory. John Wiley & Sons, New York (1975)

[9] Silverman, B. W.: Density Estimation for Statistics and data Analysis. Chapman and Hall, London (1986)

[10] Chávez, E., Figueroa, K., Navarro, G.: A Fast Algorithm for the All k Nearest Neighbors Problem in General Metric Spaces. http://citeseer.nj.nec.com/correct/462760

[11] Pestov, V.: On the geometry of similarity search: Dimensionality curse and concentration of measure. Information Processing Letters, Vol. 73, No. 1-2, 31 January 2000, 47-51.