Generalized NN Search

(projected nearest neighbor search)

KNN problems in high dimensional space

- performance
- **quality issue**
- insufficient low-dimensions algorithms
- what distance function to choose?

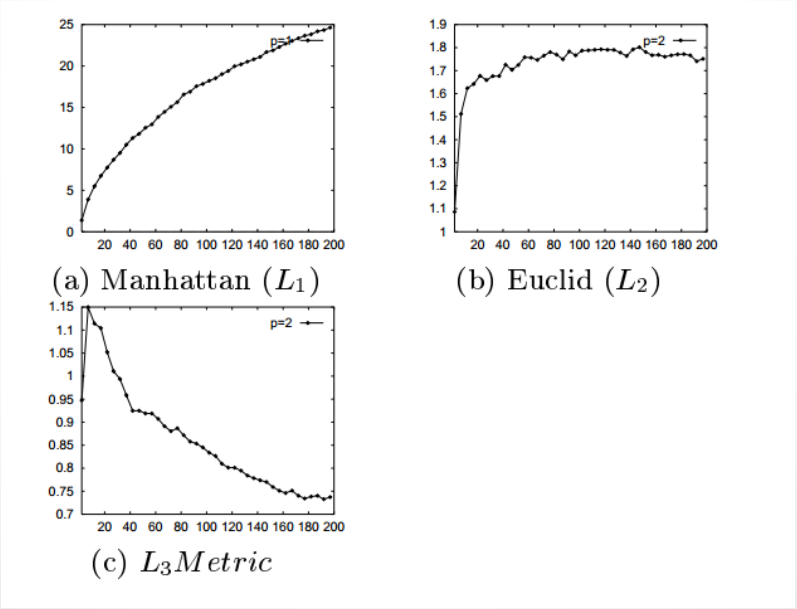
KNN problems in high dimensional space

- relation between max and min distance from query point

$$\frac{Dmax_d - Dmin_d}{Dmin_d} \rightarrow_p 0.$$

KNN problems in high dimensional space

- metrics does matter!



(a) Manhattan (L_1)

(b) Euclid (L_2)

(c) $L_3 Metric$

Generalized NN Search

- attributtes projection to smaller dimensions
- function of quality criterion (depends on query point, dataset, distance metrics)

Problem to solve

optimize criteria function over projections space

$$x_q \in \mathbb{R}^d \ is \ the \ point^3$$

$$x_{NN} = \{x' \in D | \forall x \in D, x \neq x' :$$

$$dist(p_{best}(x'), p_{best}(x_q)) \leq dist(p_{best}(x), p_{best}(x_q))\};$$

$$p_{best} = \{p \in P | \underset{p:\mathbb{R}^d \to \mathbb{R}^{d'}, d' \leq d}{MAX} \{ C(p, x_q, D, dist) \}\}.$$

Best projection search approaches

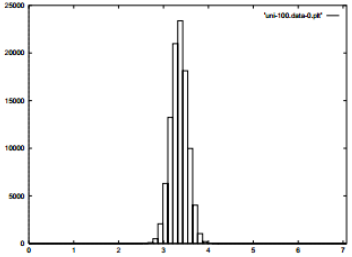
- random
- genetic
- greedy

Solution

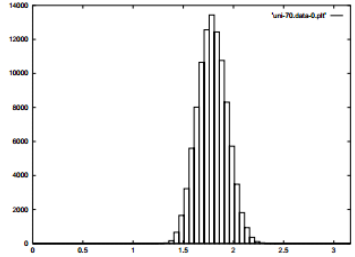
p_nn_search $(x_q, d_{tar}, D, C, dist)$
$d_{tmp} := 3$ to 5
$no\_iter := 10$ to 20
$p_{tmp} :=$ genetic_search$(\ x_q, d_{tmp}, D, C, dist, no_iter)$
$p_{best} :=$ greedy_search$(\ x_q, d_{tar}, D, C, dist, p_{tmp})$
$x_{NN} :=$ p_nn_search$(\ x_q, D, dist, p_{best})$
return $(\ x_{NN}\ )$

Observations

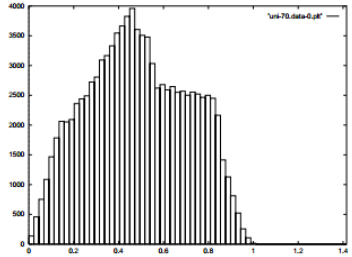
Distance distribution of uniform data



(a) 50 Dimensions



(b) 10 Dimensions



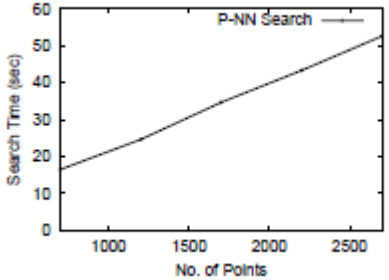
(c) 2 Dimensions

Observations

- best results for genetic algorithm
- linear complexity related to points count

Database	Class	NN	P-NN	Improv.
Ionosphere	0	0.52%	0.66%	27%
	1	0.95%	0.94%	0%
Spam	0	0.77%	0.85%	10%
	1	0.64%	0.79%	23%

Table 3: Generalized Nearest Neighbor Classification (Real Data)



Bibliografia

- A. Hinneburg, Charu C. Aggarwal, Daniel A. Keim: *What is the nearest neighbor in high dimensional spaces*
- http://stackoverflow.com/questions/5751114/nearest-neighbors-in-high-dimensional-data

Dziękujemy!

Radosław Chamot

Grzegorz Miejski