基于 LSH 的 K 近邻搜索

赵巍 MG1433089 imagine4077@gmail.com 13383410557

(南京大学 计算机科学与技术系, 南京 210093)

1 实现细节

（一）BRUTE-FORCE 方法实现 KNN：

距离选择余弦距离。计算 $\cos\theta = \vec{u} \bullet \vec{v} \Big/ |u||v|$ ，函数值越大则相似度越高。

（二）LSH 实现：

1、选择随机投影法实现哈希。哈希函数取 $H(v) = (h_1(v)h_2(v))_2$ 。其中，$h_i(v)$ 的返回值为 0 或 1。

2、$h_i(v)$ 计算如下：随机选择一个维度等于向量 v 的向量 random_v ， 若 v 与 random_v 夹角为正，则返回 1；否则返回 0。

3、对训练集内所有数据，计算其哈希函数的值，并记录。

4、对待预测的数据，首先计算其哈希函数值，然后取出与此值相同的训练集内的数据，计算这些数据与待预测数据的余弦距离。

2 结果

2.1 实验设置

1、每类取 150 个关键词，各类间允许存在交集。

2、LSH 中，H（v）将训练集划分为 4 个桶。

3、距离取余弦距离。

2.2 试验结果

K	Precision of LSH (mean/std)	Time of LSH (mean/std)	Time of brute-force search (mean/std)
10	0.505389221557/ 0.322278097084	0.424676638163s/ 0.332231384697	1.98638322824s/ 1.4655180917
20	0.439820359281/ 0.272740522986	0.423646702738s/ 0.273219655305	1.97194011054s/ 1.46907154427
30	0.405189620758/ 0.267388644606	0.408616769813s/ 0.267410606773	1.9675928133s/ 1.48627901308
40	0.363922155689/ 0.243440976235	0.447107789045s/ 0.257261265074	2.01736526575s/ 1.55434754222
50	0.325988023952/ 0.211745129475	0.415305384619s/ 0.229812077083	2.01241916097s/ 1.56875096893