用于大数据分类的 KNN 算法研究*

耿丽娟，李星毅
（江苏大学 计算机科学与通信工程学院，江苏 镇江 212013）

摘 要： 针对 KNN 算法在处理大数据时的两个不足对其进行了研究，提出多层差分 KNN 算法。算法对已知样本根据类域进行分层，既避免了传统改进算法中剪辑样本带来的判别误差，又大大降低了无效的计算量；同时在最后一层采用差分的方法进行决策，而不是直接根据最近邻进行分类，大大提高了分类的准确性。实验结果表明，该算法在对样本容量大、涉及邻域多的大数据样本进行分类时能取得较好的分类效果。

关键词： 大数据；KNN；差分多层
中图分类号： TP391；TP301.6 **文献标志码：** A **文章编号：** 1001-3695(2014)05-1342-03
doi:10.3969/j.issn.1001-3695.2014.05.013

Improvements of KNN algorithm for big data classification

GENG Li-juan, LI Xing-yi
(*School of Computer Science & Telecommunication Engineering, Jiangsu University, Zhenjiang Jiangsu* 212013, *China*)

Abstract: When dealing with big data there are two disadvantages of KNN algorithm, this paper proposed multi-differential KNN algorithm, that was, according to class field using the algorithm to stratify the known samples, it avoided errors by traditional algorithm in the sample clipping, and also greatly reduced the invalid amount of computation. While in the last layer adopting the method of differential decisions, rather than classifying directly from the nearest neighbor, which greatly improved the classification accuracy. Experimental results show that on a large sample size, the algorithm can achieve better classification effects when involving more large neighbor fields to classify data samples.

Key words: big date; KNN; multi-differential

随着信息技术的快速发展，大数据时代已经到来，人们迫切需要研究出更加方便有效的工具对收集到的海量信息进行快速准确的分类，以便从中提取符合需要的、简洁的、精炼的、可理解的知识。目前关于这方面的研究已经取得了很大的进步。现有的分类算法有很多种，比较常用的有 KNN(K-nearest neighbor)[1]、Native Bayes[2]、Neural Net[3]、SVM(support vector machine)[4]、LLSF(linear least square fit)[5] 等方法。

针对这些算法处理大规模数据时存在的问题，国内外已经进行了很多相关方面的研究。文献[6]针对传统支持向量机方法处理大规模数据时时间复杂度和空间复杂度随数据量的增加直线上升的缺点，提出了核向量机(core vector machine，CVM)方法，大大减小了算法的时间和空间复杂度；文献[7]对向量机方法进行了进一步研究，提高了核向量机的分类速度和泛化能力，但是其分类精度依然没有得到改善；文献[8]提出了一种聚簇消减大规模数据的支持向量分类算法，提高了传统算法处理大规模数据时的速度，同时降低了算法的时间复杂度，但是精度也只有在阈值选择适当时才有可能达到既减少训练时间又提高精度的双赢目的。

KNN 作为一种经典的统计模式识别方法，也是效果最好的分类方法之一，而且 KNN 方法主要靠周围有限的邻近样本，而不是靠判别类域的方法来确定所属类别，因此对于类域的交叉或重叠较多的大数据来说，KNN 方法较其他方法更为适合，但 KNN 在分类时主要的不足是该算法只计算最近的邻居样本，某一类的样本数量很大，容易出现误判。现在主要采用权值的方法(与该样本距离小的邻居权值大)来改进，但是权值的设置针对不同的领域又要有不同的要求，实用性不是很高。该方法的另一个不足之处是计算量较大，因为对每一个待分类的文本都要计算它到全体已知样本的距离，才能求得它的 K 个最近邻点。目前常用的解决方法是事先对已知样本点进行剪辑，但样本的剪辑经常带来后续的判别误差。本文针对大数据和 KNN 算法的特点，提出了 DM-KNN 算法，有效地解决了 KNN 算法对大数据的分类问题。

1 KNN 算法及其存在问题

1.1 KNN 文本分类算法

KNN 法由 Cover 和 Hart 于 1968 年提出，是一个理论上比较成熟的方法。该算法的基本思想是：根据传统的向量空间模型，文本内容被形式化为特征空间中的加权特征向量。对于一个测试文本，计算它与训练样本集中每个文本的相似度，找出 K 个最相似的文本，根据加权距离和判断测试文本所属的类别。具体算法步骤[9]如下：

a) 对于一个测试文本，根据特征词形成测试文本向量。

b) 计算该测试文本与训练集中每个文本的文本相似度，按照文本相似度，在训练文本集中选出与测试文本最相似的 k

收稿日期：2013-06-17；修回日期：2013-08-13 基金项目：国家“十一五”科技支撑计划资助项目(2006BAG01A0)；国家自然科学基金资助项目(10972027)；江苏大学校基金资助项目(11JDG064)

作者简介：耿丽娟(1988-)，女，河南安阳人，硕士研究生，主要研究方向为数据挖掘(13253338824@126.com)；李星毅(1969-)，男，江苏镇江人，教授，硕导，主要研究方向为人工智能、智能交通、复杂系统智能分析.

个文本。

c) 在测试文本的 k 个近邻中,依次计算每类的权重。

d) 比较类的权重,将文本分到权重最大的那个类别中。

1.2　KNN 算法处理大数据时存在的问题

KNN 算法稳定性好、准确率高、简单易用,针对大数据的分类问题,它存在着如下缺点: a) 对每一个待分类的文本都要计算它到全体已知样本的距离,才能求得它的 K 个最近邻点,而大数据的典型特点就是数据信息海量、价值密度低,这就显然出现了很大的无效计算量[10,11]; b) 在决定测试样本的类别时,该算法只计算最近邻的样本[12,13],而大数据的另一个显著特点是涉及领域繁多、类别界限不明显,对于此类文本容易使判决结果产生偏差; c) 随着信息爆炸时代的到来,各种新的事物层出不穷,出现新的类别的概率极大,而 KNN 算法的邻居都是已知的类别样本,也就导致了对新样本的无知或者误判。

2　改进的 KNN 算法

2.1　分层模型的应用

分层模型的基本思想是根据所属类别的不同对已知样本进行分层,第一层包含的类别数最少,最后一层包含的类别数最多,然后依层对未知样本进行分类。图 1 以社区民情民意信息的分层为例,图中共分了三层。

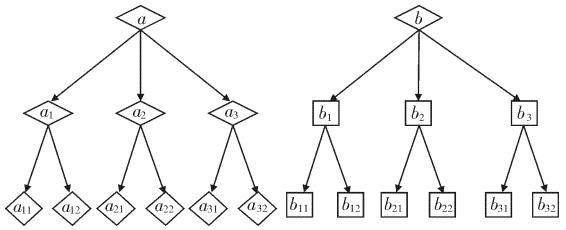

图 1　分层模型

第一层只有 a 和 b 两个类别,如果判断出来未知样本属于 a 类,那么在第二层时只需在 a_1、a_2、a_3 类中进行比较,不需要在 b 类的其他文本进行比较。在第二层判断时,如果判断出来属于 a_1 类,那么在第三层进行比较时只需要在 a_{11}、a_{12} 类中进行比较,依此类推即可。

图 1 中菱形部分为分层模型需要比较的类别数,而传统的方法是需要对所有的数据进行比较。从图 1 中可见分层模型可以大大减少无效计算量。

2.2　差分模型的应用

本文用图示的方法来解释差分模型的思想。图 2 中,x 是未知样本,a、b、c、d 为已知类别,其中 x 到 a、b、c、d 的距离分别为 Δa、Δb、Δc、Δd,如图 2 所示。

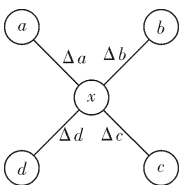

图 2　差分模型 1

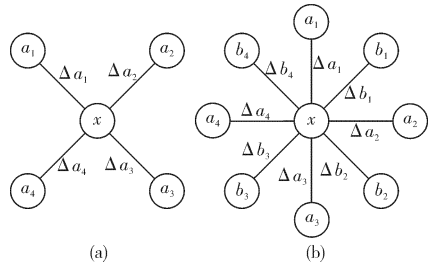

图 3　差分模型 2

如果 $\Delta a = \max\{\Delta a, \Delta b, \Delta c, \Delta d\}$,$\Delta b = \max\{\Delta b, \Delta c, \Delta d\}$,那么按照传统 KNN 算法的思想,只需要把未知样本分配到 a 类中,根据分层思想,此时只需要在把未知样本 x 与 a 类中的子类 a_1、a_2、a_3、a_4 再次利用 KNN 算法进行分类即可; 但是如果利用差分模型,当且仅当 $|\Delta a| - |\Delta b| > m$ 时,才能将 x 判别到 a 类中,否则将 x 判别到 a 和 b 类中,然后对 a 和 b 类的子类再次进行 KNN 算法,如图 3(b) 所示。

2.3　改进的 KNN 算法——差分多层 KNN(DM-KNN) 算法

针对大数据的自身特点以及 KNN 算法的缺点,算法主要在以下几个方面进行了改进: a) 构建树状分层结构,针对 KNN 算法计算量比较大的缺点[14],本文改进后的算法采用构建树状分层结构首先对高层进行比较,然后依据高层比较结果的不同,再依次对下一层次进行比较,相比直接对所有文本进行距离计算,计算量明显减少,同时提高了运算速度; b) 差分比较,由于大数据具有类域交叉性的特点,该算法不是在权重比较结束后直接进行判断,而是又针对大数据的类域交叉性进行了一次差分比较,可以有效地防止最近邻和次近邻误判的情况; c) 动态增加类别,由于大数据中信息的不可预知性,该算法针对最终比较结果不能判断隶属于哪个类别的情况,在算法最后可以动态增加新类别。具体算法步骤如下:

a) 对于一个测试文本,根据特征词形成测试文本向量。

b) 对于训练文本集,利用专业领域知识,通过文本数据的分析定义出分层类别,将其构建成 n 层树状形式。

c) 依次计算该测试文本与第 $1 \sim n$ 层训练集中每个文本的文本相似度(以下以第 1 层为例) :

(a) 文本相似度计算式为

$$\mathrm{sim}_1(d_i, d_{1j}) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{1jk}}{\sqrt{\sum_{k=1}^{M} W_{ik}^2}\sqrt{\sum_{k=1}^{M} W_{1jk}^2}}$$

式中: d_i 为测试文本的特征向量; d_{1j} 为第 1 层第 j 类的中心向量; M 为特征向量的维数; W_k 为向量的第 k 维; K 值的确定一般先采用一个初始值,然后根据实验测试的结果调整 K 值。

(b) 按照文本相似度,在训练文本集中选出与测试文本最相似的 K 个文本。

(c) 在测试文本的 K 个近邻中,依次计算每类的权重,计算式为 $P_{1J} = \sum_{d_i \in kNN} \mathrm{sim}(x, d_i)\, y(d_i, C_j)$。式中: x 为测试文本的特征向量; $\mathrm{sim}(x, d_i)$ 为相似度计算公式; 而 $y(d_i, C_j)$ 的取值为 1 或者 0,如果 d_i 属于 C_j,则函数 y 值为 1,否则为 0。

(d) 对计算的权重进行排序:

$$P_{11} \geqslant P_{12} \geqslant P_{13} \geqslant \cdots \geqslant P_{1J} \geqslant \cdots$$

(e) 对排序后的权重进行差分比较:

①$D_{12} = P_{11} - P_{12}$。如果 $D_{12} \geqslant D_0$(D_0 为阈值,有待于优化选择),则测试文本属于第 1 类,在对第二层进行相似度比较的时候,只需要比较第二层中第 1 类的子类; 如果 $D_{12} \leqslant D_0$ 则继续进行判断。

②$D_{k(k+1)} = P_{1k} - P_{1(k+1)}$。如果 $D_{k(k+1)} \geqslant D_0$,则测试文本属于第 $1 \sim k$ 类中的其中一类,在对第二层进行比较时,只需要比较第二层中第 1 类中第 k 类的子类; 如果 $D_{23} \leqslant D_0$ 则继续进行判断。

d) 第 n 层,对于权重的差分比较结果,若比较结果不是单一的,采用动态增加类别的方法,在文本第 n 层增加一个类; 若比较结果单一,就将文本分到权重最大的那个类别中。

2.4　时间复杂度分析

由算法可知,KNN 算法在时间上的代价主要在于测试样

本与训练样本库中样本之间的相似度计算。按照传统的 KNN 算法中直接计算测试样本与训练集中每个样本的相似度的方法，其时间复杂度为 $O(n^2)$；而 DM-KNN 算法中构建树状分层结构，计算相似度时不需要对 n 个样本集都进行计算，只需要与分层之后相似层下的样本进行相似度计算，算法的复杂度降为 $O(n \log n)$。也就是说改进后的算法时间复杂度远小于 KNN 算法或者类似于 KNN 算法的时间复杂度。

3 实验结果与分析

3.1 实验数据

本文对上述方法的分类效果进行了实验。实验数据为某公安机关提供的 8 990 篇社区民情民意文本，其文本符合大数据的数据量大、数据种类多、类域交叉等多个特点，本实验将数据分为三层、39 个类别。文本类别与文本如图 4 ~ 6 所示。

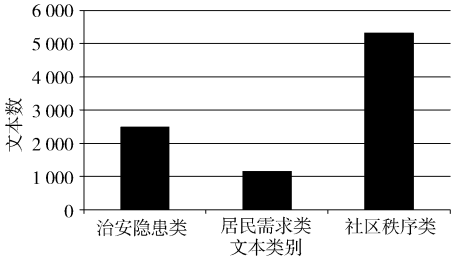

图 4 第一层

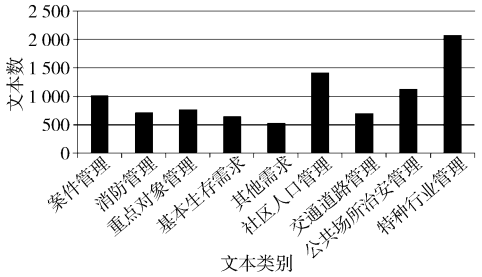

图 5 第二层

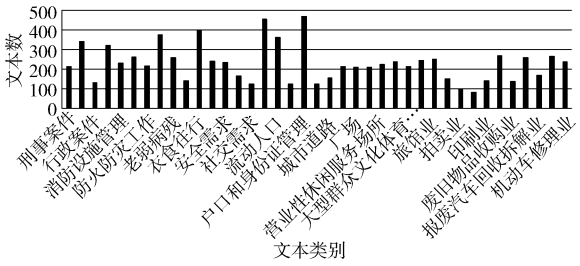

图 6 第三层

3.2 实验设置与结果分析

实验 1 本文就改进算法进行了实验。本实验使用了中科院提供的 ICTCLAS 免费中文分词系统。首先从公安机关提供的社区民情民意文本中随机抽出 500 份，由行业专家进行人工分类，然后对此 300 份文本用文中的算法进行测试，循环测试三次，取平均值为测试结果。经过反复实验，最终确定 KNN 算法中的参数 $k = 17$。分类效果评估指标使用常用的查准率、查全率以及 F_1 测试值(对应方法的值以下标进行标注)。

$$\text{查准率} = \text{分类的正确文本数} / \text{实际分类文本数}$$

$$\text{查全率} = \text{分类的正确文本数} / \text{应有文本数}$$

$$F_1 = \frac{\text{查准率} \times \text{查全率} \times 2}{\text{查准率} + \text{查全率}}$$

将传统 KNN 分类器分类效果与之比较，结果如图 7 ~ 9 所示。

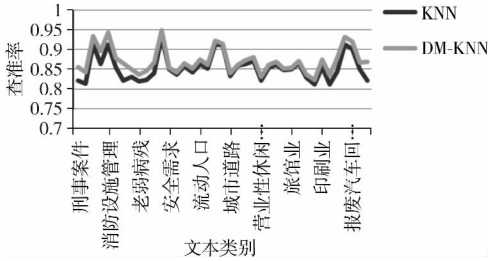

图 7 KNN 与 DM-KNN 查准率比较

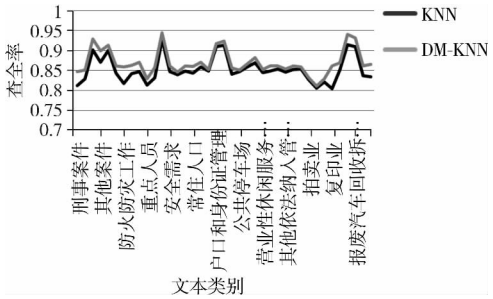

图 8 KNN 与 DM-KNN 查全率比较

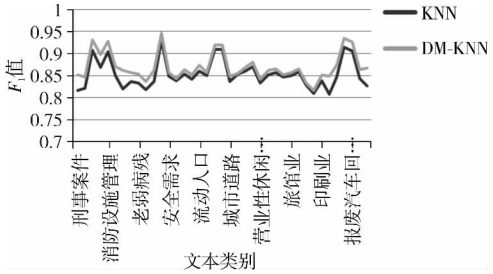

图 9 KNN 与 DM-KNN F_1 值比较

从以上实验可以发现，从查准率来看，采用 DM-KNN 算法的查准率高于传统的 KNN 算法，但是从查全率来看，改进后的算法虽然整体上有所提高，但是提高得不是很明显，而且针对一些稀有样本查全率反而降低，因此采用了综合测试因子 F_1 来评估两种算法的精度。由图 7 可以看出，DM-KNN 算法的精度相比 KNN 算法有了显著提高，证明改进后的效果是令人满意的。

实验 2 本实验把公安机关提供的 8 990 篇社区民情民意文本分为三份数据，分别为 Database1、Database2、Database3。其中 Database1 和 Database2 分别包含 3 000 篇文本，Database3 包含 2 990 篇文本。然后分别采用 KNN 和 DM-KNN 算法对其进行分类，记录其分类用时并进行比较，实验结果如表 1 所示。

表 1 KNN 和 DM-KNN 分类时间比较

	KNN/s	DM-KNN/s
Data1 用时	2 617.9	1 582.3
Data2 用时	2 493.4	1 415.8
Data3 用时	2 521	1 423.9

由表 1 可以看出，DM-KNN 算法在时间上明显优于传统的 KNN 算法，是一种对于社区民情文本不仅准确率高且分类速度快的文本分类算法。

4 结束语

随着信息化的深入发展、大数据时代的到来，如何有效地将网络技术和计算机技术融入到现实生活中，使其发挥作用，成为待研究的课题。在此背景下，本文提出了用于大数据分类的改进 KNN 算法，以在合理的时间内从海量数据(下转第 1373 页)

示每组数据使用两种方法分别得到的聚类精度。从图 4 中可以看到，DTW 方法得到的聚类精度的取值在 0.65 ~ 0.9，而 HDTW 算法得到的聚类精度的取值在 0.75 ~ 0.9，并且除了第 10 组数据之外，HDTW 算法得到的聚类精度总是高于 DTW 算法的。图 5 展示了 10 组数据的总平均聚类精度和运算时间的对比，从图 5 中可以看到 HDTW 算法的结果是远远优于 DTW 算法的，通过 HDTW 算法进行相似性度量得到的聚类精度比 DTW 算法要高出近 10%。另外，从运算时间来看，HDTW 算法与 DTW 算法相比大约减少了 30%。

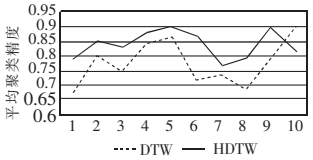

图 4　两种算法在数据集 1 中的分组聚类精度对比

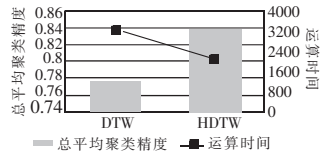

图 5　两种算法在数据集 1 中的总平均聚类精度和运算时间的对比

2.2.3　两种算法在数据集 2 上的结果对比

图 6 展示了 DTW 算法和 HDTW 算法在每组数据上得到的聚类精度的对比，横轴表示数据集 2 中的 10 组数据，纵轴表示每组数据使用两种方法分别得到的聚类精度。从图 6 中可以看到，DTW 方法得到的聚类精度的取值在 0.65 ~ 0.95，并且不同的数据组得到的聚类精度的波动很大。HDTW 算法得到的聚类精度的取值在 0.8 ~ 0.9，并且不同的数据组得到的聚类精度值波动较小，除了第 7 组数据之外，HDTW 算法得到的聚类精度总是高于 DTW 算法。图 7 展示了 10 组数据的总平均聚类精度和运算时间的对比，同样，从图 7 中可以看到 HDTW 算法的结果优于 DTW 算法。

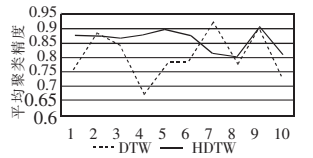

图 6　两种算法在数据集 2 中的分组聚类精度对比

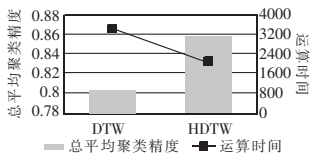

图 7　两种算法在数据集 2 中的总平均聚类精度和运算时间的对比

3　结束语

本文提出一种基于分层动态时间弯曲的序列相似性度量方法，对较长时间序列进行多层次的分段后再进行相似性度量，并在算法中充分考虑了时间序列的形态特征和趋势。实验结果证明，该方法具有较好的效果。

参考文献：

[1] 李海林，郭崇慧. 时间序列数据挖掘中特征表示与相似性度量研究综述 [J]. 计算机应用研究，2013，30(5)：1285-1291.

[2] KEOGH E. Data mining and machine learning in time series databases [C] //Proc of the 4th IEEE International Conference on Data Mining. 2004.

[3] BERNDT D，CLIFFORD J. Using dynamic time warping to find patterns in time series [C] //Proc of AAAI Workshop on Knowledge Discovery in Databases. 1994：359-371.

[4] KEOGH E，PAZZANI M. Derivative dynamic time warping [C] //Proc of the 1st SIAM International Conference on Data Mining. 2001：1-11.

[5] 陈胜利，李俊奎，刘小东. 基于提前终止的加速时间序列弯曲算法 [J]. 计算机应用，2010，30(4)：1068-1071.

[6] 尚福华，孙达辰，吕海霞. 提高 DTW 运算效率的改进算法 [J]. 计算机工程与设计，2010，31(15)：3518-3520.

[7] LEMIRE D. Faster retrieval with a two-pass dynamic time warping lower bound [J]. Pattern Recognition，2009，42(9)：2169-2180.

[8] MARTEAU P F. Time warp edit distance with stiffness adjustment for time series matching [J]. IEEE Trans on Pattern Analysis and Machine Intelligence Archive，2009，31(2)：306-318.

[9] 陈乾，胡谷雨. 一种新的 DTW 最佳弯曲窗口学习方法 [J]. 计算机科学，2012，39(8)：191-195.

[10] 吴学雁，黄道平，莫赞. 基于极值点特征的时间序列相似性查询方法 [J]. 计算机应用研究，2010，27(6)：2068-2070.

[11] GAVRILOV M，ANGUELOV D，INDYK P，*et al*. Mining the stock market：which measure is best [C] //Proc of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston：ACM Press，2000：487-496.

(上接第 1344 页) 中提取对企业或者社会有用的信息。而且实验通过对社区民情民意文本进行分类，证明了该方法在对大数据分类时效果是令人满意的。

参考文献：

[1] YANG Yi-ming，PEDERSEN J O. A comparative study on feature selection in text categorization [C] //Proc of the 14th International Conference on Machine Learning. 1997：412-420.

[2] CHAKRABARTI S，DOM B，AGRAEAL R，*et al*. Using taxonomy，discriminants，and signature for navigating in text databases [C] //Proc of the 23rd VLDB Conference. 1997：446-455.

[3] NG H T，GOH W B，LOW K L. Feature selection，perceptron learning，and a usability case study for text categorizaion [C] //Proc of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1997：67-73.

[4] JOACHIMS T. Text categorization with support vector machines：learning with many relevant features [C] //Proc of the 10th European Conference on ML. 1998：137-152.

[5] YANG Yi-ming，CHUTE C G. An example-based mapping method for text categorization and retrieval [J]. ACM Trans on Information Systems，1994，12(3)：252-277.

[6] TSANG I W，KWOK J T，CHEUNG P M. Core vector machines fast SVM training on very large data sets [J]. Journal of Machine Learning Research，2005，6：363-392.

[7] 蔡磊，程国建，潘华贤，等. 分类大规模数据的核向量机方法研究 [J]. 西安石油大学学报：自然科学版，2009，24(5)：89-92.

[8] 陈光喜，徐建，成彦. 一种聚簇消减大规模数据的支持向量分类算法 [J]. 计算机科学，2009，36(3)：184-188.

[9] YANG Yi-ming，LIU Xin. A re-examination of text categorization methods [C] //Proc of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley：ACM Press，1999：42-49.

[10] 江涛，陈小莉，张玉芳，等. 基于聚类算法的 KNN 文本分类算法研究 [J]. 计算机工程与应用，2009，45(7)：61-65.

[11] 鲁婷，王浩，姚宏亮. 一种基于中心文档的 KNN 中文文本分类算法 [J]. 计算机工程与应用，2011，47(2)：41-44.

[12] TAN Song-bo. Neighbor-weighted K-nearest neighbor for unbalanced text corpus [J]. Expert Systems with Applications，2005，28(4)：667-671.

[13] 郝秀兰，陶晓鹏，徐和祥，等. KNN 文本分类器类偏斜问题的一种处理对策 [J]. 计算机研究与发展，2009，46(1)：52-61.

[14] 李荣陆，胡运发. 基于密度的 KNN 文本分类器训练样本裁剪方法 [J]. 计算机研究与发展，2004，41(4)：539-545.