文章编号:1000-5862(2015)03-0297-07

基于样本重要性原理的 KNN 文本分类算法

万韩永,左家莉,万剑怡*,王明文
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:KNN 是重要数据挖掘算法之一,具有良好的文本分类性能. 传统的 KNN 方法对所有样本权重看作相同,而忽略了不同样本对于分类贡献的不同. 为了解决该个问题,提出了一种样本重要性原理,并在此基础上构造 KNN 分类器. 应用随机游走算法识别类边界点,并计算出每个样本点的边界值,生成每个样本点的重要性得分,将样本重要性与 KNN 方法融合形成一种新的分类模型——SI-KNN. 在中英文文本语料上的实验表明:改进的 SI-KNN 分类模型相比于传统的 KNN 方法有一定的提高.

关键词:文本分类;KNN;样本重要性原理;SI-KNN
中图分类号:TP 391 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2015.03.14

0 引言

当前,网络信息呈现爆炸性增长,人们利用互联网查找信息的难度随之增大,信息检索有效地缓解了这一矛盾. 文本分类技术在信息检索中占有不可或缺的地位,成为处理互联网上大量文本的关键技术. 现有的文本分类方法主要包括:决策树分类[1]、最近邻分类(K-Nearest Neighbor,KNN)[2]、朴素贝叶斯分类[3]、支持向量机分类[4]以及集成分类[5]等. 与其他方法相比,最近邻分类具有简单、无参数等特点,在实际应用中被广泛使用.

对于分类而言,主要考虑的问题是分类的边界. 在训练集中,样本应该以一个较高的置信度被分配到正确的类别,学习使得到的分类器对于未知样本的分类性能更加稳健. 支持向量机(Support vector machine,SVM)[6-7]是一种通过最大化分类边界的距离来维持全局分类准确性的分类方法. 该方法通过找寻训练样本中的重要样本点,即类边界的样本点来确定分类边界,利用分类边界确定最优的分类超平面来使得分类间隔最大化,这些重要样本点也就是支持向量. SVM 作为最有效的文本分类方法之一,有效地证明了识别重要样本点对于提高分类准确率具有重要作用.

在大间隔分类方法中主要分为 2 种典型的支持向量机和 Boosting 分类器. 在分类过程中,它们将不同样本点分类贡献区别对待[8-10]. Boosting 分类方法[11]作为一种集成分类方法,在迭代过程中通过提升错分样本点的权重值来改进分类器的性能. 该方法在每次迭代过程中,通过提高上一轮迭代时错分样本的权重值使得分类器更加关注被上一轮分类器错分的样本点,从而使得最终产生的组合分类器的性能最优. Boosting 分类方法表明在训练分类器的过程中,不同的样本点对于分类器的贡献是不同的,即不同样本点的权重应该不一样.

KNN 作为一种被广泛使用的分类方法,其有着简单、无参数等特点. 但在 KNN 分类过程中,所有邻近样本点都具有相同的重要性,忽视了不同样本点对于分类贡献的不同.

本文在整个训练集上通过随机游走算法识别类边界点,并计算出每个样本点的边界值,生成每个样本点的重要性得分,将得到的样本重要性得分与 KNN 方法融合形成一种新的分类模型.

1 KNN 相关工作

KNN 方法的基本思想是:对于给定的一组具有类标签的训练样本,通过相似性度量函数来找到与待分类样本最相似的 K 个最近邻,根据 K 个最近邻的类标签按照多数投票的原则决策出待分类样本的类标签. 影响 KNN 算法分类性能的主要因素包括 3 个方面:(i) 度量与待分类样本相似程度的相似性度

收稿日期:2014-12-28
基金项目:国家自然科学基金(61272212,61163006,61203313,61365002,61462045)资助项目.
通信作者:万剑怡(1974-),女,江西南昌人,教授,博士,主要从事信息检索、并行计算和智能信息处理的研究.

量函数;(ii) 最近邻个数 K 值的选取;(iii) 从 K 最近邻中决策出待分类样本的类标签的决策规则[12]. 针对这 3 个方面的因素,研究人员开展了相关的研究工作,并取得了一定的成果.

文献[13]提出了一种 Mahanalobis 距离度量规则,通过缩小同类样本之间的距离,放大不同类样本间的距离使得 KNN 算法中的近邻点都属于同一类,而不同类的点用一个尽可能大的间隔隔开. 文献[14]利用强度值自适应计算出每个待分类样本最优的 K 值,而不是在整个数据集上寻找最优的 K 值. 文献[15]提出了基于区域划分的 KNN 文本快速分类方法,将训练集划分成多个不同区域,测试样本的 K 个最近邻一定落在测试样本的最近邻区域中,所以只需把最近邻区域看成该测试样本的训练集,计算出 K 个最近邻,有效减少了 KNN 算法的计算量. 文献[16]提出了一种改进型的 KNN 决策规则 SWF,根据 K 个最近邻与待分类样本的相似度来加权每个近邻对分类的贡献,即相似度越大,贡献越大. 最终,以相似度之和作为待分类样本的隶属度的判决规则. 该方法对传统 KNN 方法的多数投票规则进行改进,考虑了相似度不同的 K 个最近邻样本对于待分类样本的贡献程度不同的问题. 文献[17]提出了 GAK-KNN 算法,首先给训练集合的各类分配相同的权重;然后根据各个样本点对于所隶属类别的代表能力分配给该样本相应的权重,代表能力越大,分配的权重也就越大. 该方法利用每个类别的各自样本点的代表能力作为权重分配原则. 针对 KNN 方法会偏向不均衡类的大类问题,文献[18]分析了 over-sampling 和 under-sampling 采样法,通过增加少数类样本和减少多数类的样本,以达到类与类的样本平衡,从而提高少数类的分类性能. 但是,over-sampling 方法的时间复杂度太高,under-sampling 则可能会丢失一些有用的信息.

虽然上述研究工作从不同的方面有效地提高了 KNN 的分类准确性,但注意到,KNN 在确定了待分类样本的 K 个近邻后,按照多数投票的原则决策出待分类样本的类标签,而这种简单的多数投票原则通常忽略了不同样本点的重要性问题.

2 样本重要性原理

SVM 分类方法采用类边界点,即支持向量,来确定最优的分类超平面. 如图 1 所示,样本点 a,b,c,d 对于分类超平面的确定没有起到任何作用. 这表明了对于分类问题而言,不同样本点在分类过程中的重要性是有所不同的.

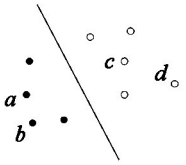

图 1 最优的分类超平面图

由于不同样本点对于分类的作用是不同的,在分类过程中,要考虑样本点的重要性程度,即对所有样本点赋予不同的权重,必须根据训练数据集的原始分布. 通常认为,一个类别的中心点可以更好地表征该类别,而越靠近边界的样本点对于类别的表示能力越差. 尽管没有任何一种模型可以完整地表示原始文本的类别分布情况. 但是,通过向量空间模型表示之后,靠近类边界的样本点的所有近邻样本中的异类样本点的数量应该大于类中心样本点的近邻样本中的异类样本点的数量. 所以,通过在异类样本点之间进行随机游走可以识别出边界样本点,而越靠近类边界的样本点的重要性越低,即样本点的权重更小.

不同于 GAK-KNN 算法和改进型的 KNN 决策规则 SWF 对样本点赋予不同权重的方法,本文在整个训练集上通过随机游走算法识别类边界点,并计算出每个样本点的边界值. 在理论上,靠近分类边界的样本点被错分的代价应该小于远离边界的样本被错分的代价,从而可以从样本点的边界值中生成样本的重要性得分. 最后,计算出 K 个近邻样本点中各个类别的样本重要性得分,将样本重要性得分最大的类标签作为待分类样本的类别.

如某个训练集合分布如图 2 所示,首先,在不同类别的样本点之间建立一个加权无向图,权重值与样本点之间的距离成反比关系,样本点 a 与 c,d,f,g 有权重值,与 b,e,m,n 没有权重值,其他样本点相同. 然后,在建立的无向图上生成马尔科夫转移矩阵,如 $P(a,c)$ 表示样本点 a 到样本点 c 的转移概率. 最后,在整个无向图上进行随机游走算法,由于类边界样本点的近邻样本点具有更多的异类样本点,所以,算法最终必将收敛于类边界样本点上,即图 1 中 a,b,c,d. 然而,靠近分类边界的样本点对于该类的表示能力要小于远离类边界的样本点,如样本点 a 的边界值大于样本点 e 的边界值,因此,样本点 a 的重要性得分应该小于样本点 e 的重要性得分,即 a 的权重要小于 e 的权重. 计算出每个样本点的重要性得分之后,对于测试样本 j 而言,计算出 K 个最近邻样本点后,根据这 K 个最近邻样本的重要

性得分最大的类标签来决策出 j 的目标类别.

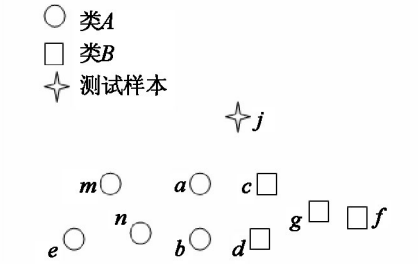

图 2　训练集合分布图

对于给定的一组具有类标签的文本训练集合,(i) 在训练集上建立一个不同类样本点之间的加权无向图;(ii) 生成样本点之间的概率转移矩阵;(iii) 通过随机游走计算出每个样本点的边界值,并利用边界值计算出样本重要性.

本文作如下几个假设:(i) 样本的边界值是非负的,样本的边界值反映出样本的重要性得分,取非负数;(ii) 靠近边界的样本比远离边界的样本的边界值更大,边界值反映出样本距离类边界的距离,离类边界越近边界值也越大;(iii) 当一个样本被同类样本包围时,可以确定它一定不是边界点,它的边界值为 0;(iv) 当一个样本被不同类标签的样本包围时,它被认定为离群点,边界值为 0.

2.1　样本边界值的计算

$X = \{x_1, x_2, \cdots, x_m\}$ 表示 m 个已知类标签的样本,$Y = \{y_1, y_2, \cdots, y_m\}$ 表示相应的类标签,其中 $y_i \in \{-1, 1\}$, $i \in \{1, 2, \cdots, m\}$.

(i) 生成一个加权无向图

无向图的顶点为 m 个已知标签的样本,边连接的 2 个顶点是 2 个类标签不同的样本,边的权重计算方法为

$$r(j,i) = \exp\{-\lambda dist(j,i)\},$$

其中 $y_i \neq y_j$, $dist(j,i)$ 为样本 x_i 和 y_j 之间的欧几里德距离,λ 为衰减程度,通常取大于 1 的整数,样本间距离越大权重值越小.

(ii) 生成马尔科夫转移矩阵

$$P(j,i) = \begin{cases} 0, & \text{当} \sum_{k \in N_j} r(j,k)\, I(j,k) = 0, \\ \dfrac{r(j,i)\, I(j,i)}{\sum_{k \in N_j} r(j,k)\, I(j,k)}, & \text{其他,} \end{cases}$$

其中

$$I(j,i) = \begin{cases} 1, & \text{当}((y_i \neq y_j) \text{ 且 } (x_i) \in N_j), \\ 0, & \text{其他,} \end{cases}$$

N_j 为样本 x_j 的 K 个最近邻. 对于一个样本,它的边界值是由最近邻中不同类标签样本的边界值决定. 因此,$P(j,i)$ 可以看作样本 x_j 到 x_i 的转移概率.

(iii) 生成马尔科夫链的平稳分布

$$\pi(j) = \sum_{i \in X} \pi(i)\, P(i,j),$$

其中 $\pi(j)$ 为样本 x_j 的边界值. 下面给出计算样本边界值的伪代码.

算法 1　计算边界值.

$m_1 = 0$, $m_{-1} = 0$, $l = 0$

for $\forall x_i \in X$, do

N_i = the K nearest neighbors for x_i (if $k \geqslant m$, $N_i = X - X_i$)

N_i^+ = the subset of N_i from the same class as x_i

N_i^- = the subset of N_i from the different class as x_i

β is the threshold of the ratio of $|N_i^-|$ to $|N_i^+|$

if $|N_i^-| = 0$ OR $|N_i^-| / |N_i| \geqslant \beta$ then $\alpha_i^{(0)} = 0$

else $\alpha_i^{(0)} = 1$, $m_{y_i} = m_{y_i} + 1$

end if

end for

for $\forall x_i \in X$ do $\alpha_i^{(0)} = \alpha_i^{(0)} / m_{y_i}$

end for

repeat

for $\forall x_i \in X$ do $\alpha_i^{l+1} = \sum_{j \in N_i} P(j,i)\, \alpha_j^l$

end for

normalize α_i^{l+1}

so that $\sum_{i \in +} \alpha_i^{k+1} = 1$, $\sum_{i \in -} \alpha_i^{k+1} = 1$, $l = l + 1$

until $l < \max(iteration)$ AND $\| \alpha_i^{(l)} - \alpha_i^{(l-1)} \| \leqslant \varepsilon$

output $\alpha_i^{(i)}$.

算法收敛: $\boldsymbol{P} = P(j,i)$ 为随机转移矩阵,为了保证不可约条件,对矩阵 $\boldsymbol{P}$ 做调整: $\boldsymbol{P} = \alpha \times \boldsymbol{P} + (1 - \alpha) \times \boldsymbol{U}/m$,其中 $0 \leqslant \alpha \leqslant 1$, $\boldsymbol{U}$ 为单位矩阵, m 为样本集的大小. 该算法是基于随机游走方法,根据随机游走理论,该算法最终能够收敛.

2.2　样本重要性的计算

根据样本的边界值可以计算出样本的重要性得分,其计算方法为

$$SI_i = 1 - \beta' \alpha_i / \max(\alpha_i),$$

其中 $0 \leqslant \beta' \leqslant 1$, α_i 为样本 x_i 的边界值得分, SI_i 为样本 x_i 的重要性得分. 两者呈反比关系, α_i 越大,表示样本 x_i 越靠近分类边界,其重要性得分 SI_i 也越小.

3　SI-KNN 算法

对于训练集,计算出每个样本的重要性得分 (SI_i 和相应的类标签). 在 KNN 分类算法中,根据近

邻样本点决策出测试样本的类别标签的过程中,将近邻样本的重要性得分考虑进来,而不是等权重的对待每个近邻样本点,从而形成一种加权的 KNN 分类算法(即 SI-KNN). 下面给出 SI-KNN 的伪代码.

算法 2 SI-KNN.

for each $x_{test} \in X_{test}$ do

N_{test_i} = the k nearest neighbors in the training samples for x_{test_i}

$\text{label}_{test_i} = \text{sign}\left(\sum_{j \in N_{test_i}} (SI_j Y_j)\right)$

end for

output $label_{test_i}$

根据算法 2 可知,在结合样本重要性得分(SI)后的 KNN 分类方法,不仅考虑了测试样本(X_{test})的最近邻样本的类别标签,而且考虑了这些近邻样本对于所属类别的表示能力,即样本的重要性得分. 采用基于样本重要性得分的加权 KNN 分类方法,有效地融合了不同样本点对于分类贡献度各不相同的问题.

4 实验设计与分析

4.1 实验数据

为了测试样本重要性理论对于 KNN 分类方法的影响,分别从复旦大学中文语料库中选取 10 类共 2 815 篇中文文本,20 newsgroup 中 20 类共 18 595 篇英文文本,Reuters-21578 的 ModeApte split 版本中选取 10 个常用类别进行实验. 见表 1 ~ 表 3.

表 1 复旦大学中文语料库 10 个类文档统计

编号	1	2	3	4	5	6	7	8	9	10
类别	交通	体育	军事	医疗	政治	教育	环境	经济	艺术	计算机
文本数	214	450	249	204	505	220	200	325	248	200

表 2 20 Newsgroup 语料库各类别文档统计

编号	类别	文本数	编号	类别	文本数
1	alt	799	11	hockey	999
2	graphic	973	12	crypt	991
3	misc	985	13	electronics	984
4	pc. hardware	982	14	med	990
5	mac. hardware	963	15	space	987
6	windows. x	988	16	christian	997
7	forsale	975	17	guns	910
8	autors	990	18	mideast	940
9	motorcycles	996	19	politics. misc	775
10	baseball	994	20	religion. misc	377

表 3 Reuters-21578 语料库各类别文档统计

类别	earn	acq	money-fx	grain	crude	trade	interest	wheat	ship	corn
训练文档数	2 877	1 650	538	433	389	369	347	212	197	182
测试文档数	1 087	719	179	149	189	119	131	71	89	56

4.2 实验分析

实验过程采用中科院分词系统,LTC 权重方法,复旦大学中文语料库和 20 Newsgroup 语料库文档频率(DF)进行特征降维,并进行十折交叉验证. 复旦中文语料库过滤 $DF < 25$ 的特征,20Newsgroup 过滤 $DF < 91$ 的特征. Reuters-21578 采用卡方检验选取 3 000 个特征词. 评价指标采用 F1 值,宏平均 F1 和微平均 F1. 复旦中文语料库的实验结果如表 4 所示,其中近邻数 K 值为 15.

表 4　复旦大学中文语料库文本分类性能

编号	1	2	3	4	5	6	7	8	9	10	MacF1	MicF1
KNN	78.79	92.80	90.17	91.16	84.76	83.87	84.10	84.37	83.33	83.75	85.71	83.75
SI-KNN	78.79	92.80	90.17	91.16	86.01	85.01	85.43	85.83	84.66	85.20	86.51	85.20
SVM	93.33	98.33	95.95	96.33	95.71	96.09	96.46	96.28	96.50	96.75	96.17	96.75

由表4可知,在复旦大学中文语料库上,SVM的分类性能确实要优于 KNN 方法. 但是,考虑训练集合中每个样本的重要性得分后的 SI-KNN 分类方法相比于按多数投票原则的 KNN 方法在性能上有一定的提高. 注意到,一些类别中的文本分类性能并没有得到提高,可能是实验过程中对中文进行预处理后仍然存留许多噪音,并且所使用的中文语料库的文本数目不大,才会导致在总体宏平均 F1 和微平均 F1 值提高的情况下,某些类别的 F1 值却没有得到提升.

20 Newsgroup 的实验结果如表 5 所示,其中近邻数 K 值为 10.

表 5　20Newsgroup 语料库文本分类性能

	1	2	3	4	5	6	7	8	9	10	11
	87.18	72.63	72.48	71.01	69.29	69.81	70.55	72.21	73.9	75.2	76.7
KNN	12	13	14	15	16	17	18	19	20	MacF1	MicF1
	78.51	77.8	77.67	78.16	78.53	79	79.86	79.76	79.52	75.99	79.52
	1	2	3	4	5	6	7	8	9	10	11
	87.18	73.68	73.89	72.06	72.28	72.32	72.24	73.53	75.07	76.07	77.59
SI-KNN	12	13	14	15	16	17	18	19	20	MacF1	MicF1
	79.4	78.54	78.32	78.82	79.15	79.63	80.43	80.40	80.16	77.05	80.16
	1	2	3	4	5	6	7	8	9	10	11
	89.93	82.57	83.96	83.50	85.22	85.74	85.91	86.61	87.40	88.48	89.12
SVM	12	13	14	15	16	17	18	19	20	MacF1	MicF1
	90.10	89.55	89.80	90.22	90.33	90.69	91.07	91.09	90.81	88.10	90.81

由表5可知,在20 Newsgroup语料库上,SVM方法的分类性能最优,但SI-KNN分类方法依旧要优于传统的 KNN 方法. 而且,随着语料库中文本数目的增大,在 20 Newsgroup 中分类性能没有得到提升的类别数量得到明显的改进.

Reuters-21578 语料库的实验结果如表 6 所示,其中近邻数 K 值取 10.

表 6　Reuters-21578 语料库的文本分类性能

编号	earn	acq	money-fx	grain	crude	trade	interest	wheat	ship	corn	MacF1	MicF1
KNN	98.34	95.36	82.47	87.37	83.78	77.52	78.54	75.36	82.41	78.63	83.98	91.78
SI-KNN	98.34	95.36	83.17	88.58	85.38	80.01	78.95	77.56	84.37	82.15	85.39	93.06
SVM	98.62	96.56	83.24	92.68	88.32	77.53	77.64	86.76	84.66	89.52	87.56	93.39

上述实验所使用的复旦中文文本语料库和 20 Newsgroup 语料库中,各个类别的文本数量几乎没有太大差异,最小类和最大类的文本数量也就在 1∶2 左右. 而表 6 的实验中所使用的 Reuters-21578 的最小类与最大类之间的文本数量超过 1∶10. 可见,通过样本重要性原理对样本点进行加权,不仅改进了 KNN 方法中忽略样本权重的多数投票原则,而且对于 KNN 方法在不均衡类中偏向大类问题也有一定的改进,并且使得 KNN 与 SVM 的分类性能接近.

4.2.1　参数敏感性分析　样本重要性理论中所涉及的参数较多,其中算法 1 中的 β 表示异类样本所占比例的阈值,即当某个样本的近邻样本点中没有任何异类近邻点或者异类近邻点所占比例非常小时,直接给该样本点边界值赋为 0. ε 表示随机游走过程中前后 2 次迭代的边界值得分小于该阈值时算法立即收敛. 实验中,$\beta = 0.01$,$\varepsilon = 0.001$.

转移概率矩阵的调整参数 α 取值 0.95,λ 表示衰减程度,通常取大于 1 的整数. λ 太小,不能体现出样本间距离与权重的反比关系; λ 太大,会使得权重值衰减过快. 算法 2 中的 β' 表示边界值与样本重

要性得分的关系系数,一般而言,越靠近类边界的样本点对于该类别的表示性越差,即重要性得分应该更低. 图 3 列出了参数 λ 和 β' 在复旦和 20 新闻组上的实验分析,图 3(a) 可以看出参数 λ 稳定性非常好. 图 3(b) 中显示在复旦数据上 β' 大于 0.5 时会有一定提高,但是在 20 新闻组却有一定的下降. 所以,本文提出的算法中 β' 取 0.5,而 λ 取 2.

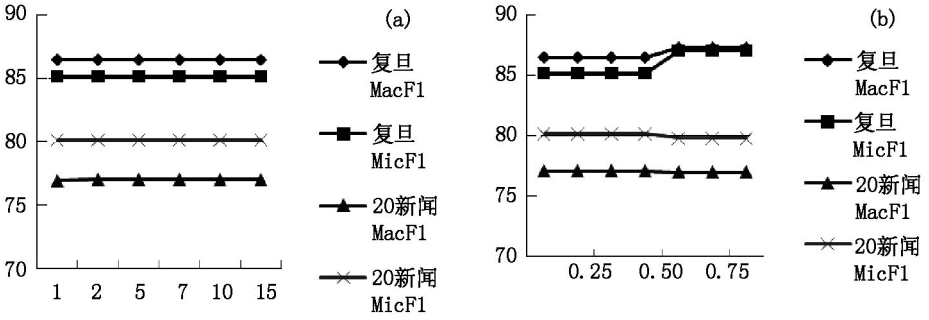

图 3 参数 λ 和 β' 的实验分析

4.2.2 近邻数 K 值的分析 实验过程中,为了找出各实验数据集下的最优 K 值,分别测试了不同 K 值下 KNN 的分类性能. 其中复旦大学中文语料库的实验结果如表 7 所示,20 Newsgroup 语料库的实验结果如表 8 所示. Reuters-21578 语料库的实验结果如表 9 所示.

表 7 复旦大学中文语料库的 K 值分析

K	1	2	3	4	5	6	7	8	9	10	MacF1	MicF1
10	72.22	91.34	89.89	90.91	85.98	85.56	85.37	85.65	84.66	85.20	85.68	85.20
15	78.79	92.80	90.17	91.16	84.76	83.87	84.10	84.37	83.33	83.75	85.71	83.75
20	83.87	93.55	89.41	90.57	83.03	82.67	82.73	83.16	82.17	82.67	85.38	82.67
25	90.32	95.16	88.62	89.95	81.32	81.17	81.36	82.55	81.62	81.95	85.40	81.95

表 8 20 Newsgroup 语料库的 K 值分析

K	5	10	15	20
1	85.71	87.18	87.58	89.93
2	72.63	72.63	72.98	72.36
3	70.15	72.48	72.01	73.00
4	69.04	71.01	70.00	71.06
5	66.73	69.29	68.51	69.49
6	67.45	69.81	69.24	70.63
7	67.17	70.55	70.06	70.97
8	69.24	72.21	71.70	72.27
9	71.35	73.90	73.45	74.22
10	73.10	75.20	74.87	75.33
11	74.68	76.70	76.50	76.85
12	76.49	78.51	78.20	78.53
13	76.11	77.80	77.55	77.51
14	76.17	77.67	77.36	77.36
15	76.79	78.16	77.84	77.92
16	77.29	78.53	78.08	78.21
17	77.71	79.00	78.38	78.52
18	78.68	79.86	79.26	79.40
19	78.77	79.76	79.16	79.27
20	78.66	79.52	78.82	78.92
MacF1	74.20	75.99	75.58	76.09
MicF1	78.66	79.52	78.82	78.92

表 9　Reuter-21578 语料库的 K 值分析

K	earn	acq	money-fx	grain	crude	trade	interest	wheat	ship	corn	MacF1	MicF1
5	96.62	92.41	83.58	85.49	83.20	78.18	75.81	74.84	83.17	80.07	83.34	91.25
10	98.34	95.36	82.47	87.37	83.78	77.52	78.54	75.36	82.41	78.63	83.98	91.78
15	98.58	93.46	81.12	85.48	84.28	80.16	75.09	75.12	80.68	77.38	83.14	91.61
20	96.69	92.14	80.36	85.83	80.84	79.03	77.34	74.42	81.73	77.38	82.58	90.58

由表 7 ~ 表 9 可知，本文所使用的复旦大学中文语料库的最适合的 K 值取 15，20 Newsgroup 语料库的最适合的 K 值取 10，Reuters-21578 语料库最合适 K 值为 10.

5　总结与展望

KNN 方法具有简单、无参数的特点，被广泛应用于各种领域. 但是，该方法对 K 个近邻样本点采用多数投票的原则，忽视了不同样本点对于分类贡献不同的问题. 针对这个问题，本文提出了样本重要性原理来对所有的样本点进行加权. 在分类过程中，通过考虑样本点本身的权重，形成了一种基于样本重要性原理的加权 KNN 分类算法.

通过在整个训练样本集中利用随机游走算法识别类边界点，并计算出每个样本点的边界值，生成所有样本点的重要性得分，将样本重要性与 KNN 方法结合形成了 SI-KNN 分类模型. 实验结果表明，在中英文文本语料库中，基于样本重要性原理的 SI-KNN 分类方法相比于传统的 KNN 分类方法有一定的提高，并且对于 KNN 方法偏向大类的问题也有一定的改进.

注意到，本文提出的样本重要性原理会在一定程度上加大 KNN 方法的时间复杂度，但是提高的时间复杂度是在分类器的训练阶段，而对于未知样本的预测阶段的时间复杂度却没有任何影响.

下一步的工作应该集中在本文实验过程中遇到的 2 个问题：1) 实验中所使用的复旦中文文本语料库的某些类别的分类性能没有得到改进，是否由于中文预处理和文本数量过小造成的；2) 虽然在不均衡数据集 Reuters-21578 的常用 10 个类别中的实验结果表明，基于样本重要性原理的 KNN 改进方法可以改善 KNN 方法偏向大类的问题，但是应该使用 Reuters-21578 中更多的小类来组成不均衡数据集来进行实验.

6　参考文献

[1] Rutkowski L，Jaworski M，Pietruczuk L，et al. The CART decision tree for mining data streams [J]. Information Sciences，2014，266：1-15.

[2] Jiang Liangxiao，Cai Zhihua，Wang Dianhong，et al. Bayesian citation-KNN with distance weighting [J]. International Journal of Machine Learning and Cybernetics，2014，5(2)：193-199.

[3] Bollen K A，Harden J J，Ray S，et al. BIC and alternative Bayesian information criteria in the selection of structural equation models [J]. Structural Equation Modeling：A Multidisciplinary Journal，2014，21(1)：1-19.

[4] Rebentrost P，Mohseni M，Lloyd S. Quantum support vector machine for big data classification [J]. Physical Review Letters，2014，113(13)：130503.

[5] Utkin L V，Zhuk Y A. Robust boosting classification models with local sets of probability distributions [J]. Knowledge-Based Systems，2014，61：59-75.

[6] Vapnik V N，Vapnik V. Statistical learning theory [M]. New York：Wiley，1998.

[7] Hastie T，Tibshirani R，Friedman J，et al. The elements of statistical learning [M]. New York：Springer，2009.

[8] Bermejo S，Cabestany J. Large margin nearest neighbor classifiers [M]. Springer Berlin Heidelberg，2001，84：669-676.

[9] Domeniconi C，Gunopulos D，Peng J. Large margin nearest neighbor classifiers [J]. Neural Networks，IEEE Transactions on，2005，16(4)：899-909.

[10] Chai Jing，Liu Hongwei，Chen Bo，et al. Large margin nearest local mean classifier [J]. Signal Processing，2010，90(1)：236-248.

[11] Schapire R E，Freund Y，Bartlett P，et al. Boosting the margin：A new explanation for the effectiveness of voting methods [J]. Annals of statistics，1998，26(5)：1651-1686.

[12] Nguyen N，Guo Y. Metric learning：A support vector approach [M]. Berlin：Springer Berlin Heidelberg，2008：125-136.

[13] Weinberger K Q，Saul L K. Distance metric learning for large margin nearest neighbor classification [J]. The Journal of Machine Learning Research，2009，10：207-244.

[14] 杨柳，于剑，景丽萍. 一种自适应的大间隔近邻分类算法 [J]. 计算机研究与发展，2013(11)：2269-2277.

(下转第 314 页)

cameras. A classification strategy based on camera coverage quality levels of video frames is provided and a data set of quality levels of camera video frames is labeled. A multi-dimension label assignment method is designed for utilizing deep convolution neural network to learn a robust video frame indication, and furthermore, to learn a video quality regression function based on Support Vector Regression (SVR), thus a robust evaluation on video coverage quality is performed. The experiment result shows that the algorithm of the article can perform an automatic evaluation on the surveillance camera coverage quality precisely, and effectively monitors the real-time change of camera surveillance quality.

Key words: video surveillance camera; coverage quality; deep convolution neural network; support vector regression

(责任编辑: 冉小晓)

(上接第 303 页)

[15] 胡元,石冰. 基于区域划分的 kNN 文本快速分类算法研究 [J]. 计算机科学,2012,39(10):182-186.

[16] 周奇. 基于指纹识别特征选择的改进加权 KNN 算法 [J]. 现代计算机: 专业版,2014(2):27-29.

[17] 王超学,潘正茂,马春森,等. 改进型加权 KNN 算法的不平衡数据集分类 [J]. 计算机工程,2012,38(20):160-163.

[18] Jindaluang W, Chouvatut V, Kantabutra S. Under-sampling by algorithm with performance guaranteed for class-imbalance problem [C]. Computer Science and Engineering Conference, 2014: 215-221.

The KNN Text Classification Based on Sample Importance Principals

WAN Hanyong, ZUO Jiali, WAN Jianyi*, WANG Mingwen

(School of Computer Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: As one of the top ten data mining algorithms, KNN has good performance of text classification. All samples are treated as the same as its weight in the traditional KNN method, but the question that the different sample has the different contribution to the classification has been ignored. To solve the problem, a sample importance principals and KNN classifier constructed on the basis of this principle has been presented. Using the random walk algorithm to identify these samples near the class boundary, and calculate the boundary value of each sample. To generate the score of sample importance of each sample from the boundary value, combined sample importance with KNN method to form a new classification model. Experimental results show that the new SI-KNN classifier has some improvement compared to the traditional KNN method on the Chinese and English text corpus.

Key words: text classification; KNN; sample importance principals; SI-KNN

(责任编辑: 冉小晓)