基于余弦相似度和实例加权改进的贝叶斯算法①

王行甫, 付欢欢, 王 琳

(中国科学技术大学 计算机学院, 合肥 230027)

摘 要: 面对大量样本特征时很多分类器无法取得较好的分类效果, 样本数有限导致贝叶斯算法无法获得精确的联合概率分布估计, 在样本局部构建高质量分类器需要有效的样本相似性度量指标. 针对以上问题, 提出了一种基于余弦相似度进行实例加权改进的朴素贝叶斯分类算法. 算法考虑特征对分类的决策权重不同, 使用余弦相似度度量样本的相似性, 选出最优训练样本子集, 用相似度值作为训练样本的权值来训练修正后的贝叶斯模型进行分类. 基于 UCI 数据集的对比实验结果表明, 提出的改进算法易于实现且具有更高的平均分类准确率.

关键词: 实例加权; 朴素贝叶斯; 余弦相似度; 逆文本频率; 文本分类

Improved Naïve Bayes Algorithm Based on Weighted Instance with Cosine Similarity

WANG Xing-Fu, FU Huan-Huan, WANG Lin

(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: Many classifiers cannot get good results facing numerous sample features, bayes algorithms get poor estimate of joint probability distribution with limited samples, effective similarity measure is needed to build a local classifier. Considering these problems, an improved multinomial Naïve Bayes algorithm based on weighted instances (IWIMNB) with cosine similarity is proposed. Taking different attributes contributing differently to the classification decision weight into account this algorithm uses cosine similarity as a metric of the similarity between training and validation instances. Selected training instances weighted by cosine similarity are used to train modified Naïve Bayes model. Results of final experiments show that the average classification accuracy of the proposed IWIMNB algorithm gets significant improvement.

Key words: weighted instance; Naïve Bayes; cosine similarity; inverse document frequency; text categorization

文本分类的主要任务是预先提供一定量已标记类别的训练文本, 根据测试文本的内容判定其真实类别, 常见的分类器有支持向量机、人工神经网络、遗传算法等. 贝叶斯算法假定待分类样本遵循某种概率分布, 根据已观测样本数据对测试样本的类别进行概率计算从而做出最优的分类决策. 最早用于分类任务的贝叶斯模型是基于比较后验概率的朴素贝叶斯算法(NB), 计算过程简单高效易于实现, 在大量文本数据集上分类效果甚至超过 SVM、神经网络等分类算法[1]. K 近邻算法是一种非常有效的基于实例归纳推理方法, 用 K 个邻居中最普通的类来预测未知样本的类别.

为了提高贝叶斯算法的分类效果, 文献[2,3,4]中主要提出了三类改进方法: 1) 结构拓展法, 用有向边表示属性间依赖关系; 2) 特征选择法, 在样本的特征空间中搜索一个特征子集; 3) 局部学习法, 在验证样本的局部构建分类器. 其中局部学习法的具体思路是在整个训练样本空间找到验证样本的局部, 然后在这个局部训练分类器, 该方法为不同的验证样本建立不同的分类模型.

很多分类算法面对大量的样本特征往往无法取得较好的分类效果, 通过互信息、信息增益等方法进行特征选择可适当减少特征值, 但同时会丢失用于分类

① 基金项目:国家科技重大专项(2012ZX10004-301-609);国家自然科学基金(61472382,61272472,61232018)
收稿时间:2015-12-19;收到修改稿时间:2016-01-28 [doi:10.15888/j.cnki.csa.005306]

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的关键特征信息. 文献[5]表明朴素贝叶斯分类器在大数据集上的分类性能并不会得到改善, 可以考虑适当提高训练样本的质量, 同时降低可能存在的特征间依赖给分类结果造成的负面影响. 常见的相似性度量方法如欧式距离法[4]等并未考虑到文本样本的特性, 选择的训练样本子集对分类结果改进效果有限. 不同特征对分类决策权重的作用不同.

为了进一步提高分类准确率, 本文将使用余弦相似度度量训练样本和验证样本的距离, 选择最优训练样本子集, 以相似度值作为权值(WI)来训练修正后的朴素贝叶斯分类器(MNB), 同时考虑不同特征对分类具有不同的决策权重进行改进(I), 最终提出一种基于余弦相似度进行实例加权改进的朴素贝叶斯分类算法(IWIMNB). 在若干 UCI 数据集上的对比实验表明, 本文提出的 IWIMNB 算法能有效提高训练样本的质量、弱化特征间依赖并考虑逆文本频率, 操作性较强并有效提高了算法平均分类准确率.

1 朴素贝叶斯算法

算法假定给定样本所有特征相互条件独立, 使用贝叶斯公式通过类别的先验概率和特征的条件概率来估计样本属于每个类别的后验概率, 以最大后验概率为标准确定样本的类别. 假设训练样本集合分为 k 类, 类别集合记为 $C=\{\mathrm{c}_1,\mathrm{c}_2,...,\mathrm{c}_i,...,\mathrm{c}_k\}$, 类别 c 的先验概率为 $\Pr(\mathrm{c})$, 测试样本 t_i 属于类别 c 的类条件概率为 $\Pr(\mathrm{t}_i|c)$, 其先验概率为 $\Pr(\mathrm{t}_i)$, 根据贝叶斯公式[1]其后验概率 $\Pr(\mathrm{c}|t_i)$ 可表示为式(1):

$$\Pr(\mathrm{c}|t_i)=\frac{\Pr(\mathrm{c})*\Pr(\mathrm{t}_i|c)}{\Pr(\mathrm{t}_i)},\mathrm{c}\in C \tag{1}$$

式(1)中样本的先验概率 $\Pr(\mathrm{t}_i)$ 恒定, 测试样本 t_i 的类别取最大后验概率值对应的类别, 即式(2):

$$\mathrm{C}(\mathrm{t}_i)=\operatorname{argmax}\{\Pr(\mathrm{c}_i)*\Pr(\mathrm{t}_i|c_i)\},\mathrm{c}_i\in C \tag{2}$$

多项式模型(MNB)和二项独立模型是用朴素贝叶斯算法进行文本分类的两种主要实现模型[6], 前者考虑了单词在文本中的词频信息, 这一附加信息使得朴素贝叶斯算法的分类效果更好. 对于多项式模型, 式(1)中类先验概率 $\Pr(\mathrm{c})$ 可表示为类别为 c 的训练样本数除以所有训练样本数, 测试文本 t_i 属于类别 c 的类条件概率 $\Pr(\mathrm{t}_i|c)$ 表示在类别为 c 的样本中获得文本 t_i 的概率. 另外 $\Pr(\mathrm{t}_i|c)$ 可表示为式(3)[6]:

$$\Pr(\mathrm{t}_i|c)=(\sum\nolimits_n f_{ni})!\prod\nolimits_n\frac{\Pr(\mathrm{w}_n|c)^{f_{ni}}}{f_{ni}!},\quad \Pr(\mathrm{w}_n|c)=\frac{1+F_{nc}}{N+\sum_{x=1}^{N}F_{xc}} \tag{3}$$

式(3)中 f_{ni} 表示文本 t_i 中单词 w_n 的词频, $\Pr(\mathrm{w}_n|c)$ 表示类别 c 中单词 w_n 的条件概率, F_{xc} 表示类别为 c 的训练样本中单词 w_x 个数, N 表示训练样本中所有不重复单词的个数. 当类别 c 中不包含单词 w_x 时 F_{xc} 为 0, 对应条件概率 $\Pr(\mathrm{w}_n|c)$ 为0, 为保证后验概率的有效性式(3)进行了laplace平滑[7]处理.

2 局部学习法研究现状

实例加权方法的基本思路是在整个训练样本空间找到验证样本的局部[4], 然后在这个局部学习分类器, 该方法为不同的验证样本建立不同的分类模型, 从而有效提高分类准确率. 设计一个高效的相似性度量算法是基于实例加权学习贝叶斯分类器的关键. K 近邻算法是一种非常有效的基于实例推理算法, 算法计算出验证样本与每个训练样本的距离, 然后找到距离验证样本最近的 K 个邻居, 最后用这 K 个邻居中最普通的类来预测验证样本的类别.

文献[4,8,9,10,13]中讨论了很多距离度量方法. Frank 等综述了距离函数、平滑参数、加权函数以及局部学习模型的结构等, 使用欧式距离度量相似性实现局部加权学习并应用. Huang 等提出 K-mode 算法处理 K-means 算法无法处理的类别特征值, 聚类过程中为了最小化代价函数使用基于频率的方法更新 mode, 从而实现对类别特征的聚类. Han 等提出一种迭代算法以验证样本的分类准确率为目标函数确定训练样本权值. Li 等介绍了处理离散特征值的值差分度量(VDM), 用来定义不同特征的距离. Yamada 等综述了既能处理离散特征值又能处理连续特征值的距离函数的改进版本(HOEM、HVDM、IVDM、DVDM、WVDM). 以上距离度量方法根据特征的类型为连续值、线性离散值等分别进行定义, 在不同类型数据集上对分类准确率影响差异较大.

3 改进的贝叶斯算法

3.1 余弦相似度和实例加权修正的贝叶斯公式

针对以上问题及文本样本的特性本文提出使用余弦相似度度量样本的相似性[11]. 余弦相似度通过测量

© 中国科学院软件研究所 http://www.c-s-a.org.cn

两个向量内积空间的夹角的余弦值来度量相似性. 在信息检索中, 文本由一个有权值的特征向量表示, 权值的计算取决于词条在文本中出现的频率, 余弦相似度可判断两个文本主题方面的相似性, 任意两文本 A,B 的余弦相似度可表示为式(4):

$$sim_{A,B}=\frac{\sum_{i=1}^{n}A_i*B_i}{\sqrt{\sum_{i=1}^{n}(\mathrm{A}_i)^2}*\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{4}$$

式(4)中 n 表示样本空间单词总数, A_i, B_i 表示第 i 个单词在文本 A, B 中对应频数. 对每个验证样本计算所有训练样本与其余弦相似度, 选择值最大的 k 个训练样本, 对相似度值进行标准化使其和为 k. 考虑相似度加权, 式(1)中 Pr(c) 可修正表示为式(5):

$$\Pr(\mathrm{c})=\frac{1+\sum_{i=1}^{k}I(c,c_i)sim_i'}{k+\sum_{i=1}^{k}sim_i'},I(c,c_i)=\begin{cases}1,c_i=c\\0,c_i\neq c\end{cases} \tag{5}$$

式(5)中 sim_i' 表示第 i 个训练样本与验证样本的相似度标准化值, 为保证结果的有效性式(5)进行了 laplace 平滑[7]处理. 考虑相似度加权, 式(3)中用于计算类条件概率的 $\Pr(\mathrm{w}_n|c)$ 可修正表示为式(6):

$$\Pr(\mathrm{w}_n|c)=\frac{1+\sum_{i=1}^{k}I(c,c_i)*F_{ni}*sim_i'}{N+\sum_{i=1}^{k}\sum_{j=1}^{N}I(c,c_i)*F_{ji}*sim_i'} \tag{6}$$

式(6)中 F_{ji} 表示类别为 i 的训练样本中单词 w_j 的个数, N 表示类别为c的训练样本中所有不重复单词的个数, 为保证有效性式(6)进行了laplace平滑[7]处理. 式(6)中分母表示加权后 c 类样本中所有单词词频, 分子表示加权后 c 类样本中单词 w_n 的词频.

3.2 逆文本频率

常见词出现在大量的训练样本中, 即使它们在不同类别中权重差异很小仍会在样本的分类决策计算中起重要作用. 给与分类无关的特征和其他特征赋予相同的权重将会降低分类的准确率. 在信息检索文本处理等领域, 逆文本频率[12]通常用来衡量一个词是否为常见词, 如果某个词比较少见但是它在某文本中多次出现, 那么它很可能就反映了该文本的特性, 因此可以对词频 f_{ni} 按式(7)进行转换:

$$f_{ni}^{'}=\frac{f_{ni}}{\sum_{j=1}^{k}f_{nj}} \tag{7}$$

式(7)中 f_{ni} 表示文本 t_i 中单词 w_n 的词频, k 表示文本数. 式(7)能降低常见词的分类权重, 提高非常见词的分类权重, 从而提高模型的分类准确率.

3.3 改进算法的具体流程

综合 3.1 和 3.2, 改进算法的应用步骤总结如下:

Step1: 对所有样本进行预处理(词根转换、去掉常见停用词), 提取样本的词频形成词频向量, 为消除文本长度差异进行标准化处理生成相对词频向量;

Step2: 考虑不同特征对分类的决策权重不同, 根据公式(7)进一步转换词频向量;

Step3: 计算验证样本与每个训练样本的余弦相似度, 根据公式(4)选出相似度值最大的 k 个训练样本;

Step4: 根据修正后的贝叶斯公式进行分类, 根据公式(2)(3)(5)(6)取最大后验概率值对应的类别为验证样本类别;

Step5: 改变 k 值重复 Step4, 进行交叉验证, 最终选择具有最高平均准确率的分类模型.

4 实验结果及分析

为了验证上述改进算法是否有效, 本文进行了对比实验并分析了实验结果. 实验从 UCI 数据库中挑选数据集, 结合数据挖掘工具 Weka 进行. 实验平台为: 硬件环境 CPU2.26GHz、RAM 2GB; 软件环境 win7、Weka、Matlab.

文献[10,13]综合对比分析了 MNB 算法、MNB-LWL-Euclidean 算法、MNB-LWL-HVDM 算法、MNB-LWL-IVDM 算法、MNB-LWL-WVDM 算法、SVM 算法, 并对从 UCI 数据库中选出的 10 个数据集进行分类, 分类结果见表 1.

从所选的 10 个数据集来看, 基于 MNB 改进模型的平均分类准确率比 MNB 要高 3%~7%, 而 SVM 平均分类准确率比基于 MNB 改进模型的分类准确率高. 与 Euclidean 相比, HVDM、IVDM、WVDM 等针对不同类型特征值进行相应改进的距离度量方法的分类准确率都有所提高, 其中 IVDM 度量方法具有最高的平均分类准确率 84.04%, 比 Euclidean 度量方法平均分类准确率提高了约 2.2%, 表明改进后的度量方法在所选数据集上对类别特征的处理具有较好的健壮性. 另外 HVDM 度量方法的平均分类准确率比 WVDM 方法高 0.5%, 这两种方法的分类效果都稍差于 IVDM 度量方法.

针对本文提出的改进算法, 选择在文献[14]中频繁使用的 20 Newsgroups、WebKB、Industry Sector 以

© 中国科学院软件研究所 http://www.c-s-a.org.cn

及 Reuters-21578 数据集进行实验. Reuters-21578 数据集中每个样本被标记为多个类别, 其余数据集中每个样本被标记为单个类别. 20 Newsgroups 数据集被去重后只剩 18828 个样本, 分别属于 20 个不同类别. Industry Sector 数据集总共包括 9558 个样本, 根据类型分别属于 105 个不同类别. WebKB 数据集中只使用 4391 个样本, 分别属于 students、faculty、course 和 project 类别. Reuters-21578 数据集被处理后剩下 10789 个样本, 分别属于 90 个不同类别.

使用 10 重交叉验证, 最终分类准确率取历次结果的平均值, 每次实验随机选取数据集的 60%作为训练样本, 20%作为验证样本, 剩余 20%为测试样本. 参考 3.3 中的改进算法的具体流程, 分别使用 MNB、IVDM、WIMNB、KNN、KNNDW(knn with distance weight)、IWIMNB 等算法对验证样本进行分类. 当使用 KNN、KNNDW、IWIMNB 算法进行分类时, 发现随着 k 值从 1 不断增大, IWIMNB 的分类准确率先缓慢增加后基本保持不变, KNN、KNNDW 的分类准确率缓慢降低并一直低于 IWIMNB 的分类准确率. 分析可知, 改进算法在 k 值较小时(k<20)可能出现过拟合现象, k 值超过一定范围后(k>80)改进算法对 k 的取值并不敏感, 最终 k 取 50. 作为对照, 使用 Weka 中基于序列最小优化算法[15]实现的 SVM 对数据集进行分类. 各算法的分类结果见表 2.

表 1 不同距离度量算法对应的分类准确率

数据集	实例数	实数型	整数型	类别型	MNB	Euclidean	HVDM	IVDM	WVDM	SVM
Annealing	798	6	3	29	92.34	94.99	95.61	96.11	95.87	97.23
Bridges	108	1	3	7	55.43	58.64	59.64	60.55	58.14	64.32
Glass	214	9	0	0	70.26	72.36	74.37	74.54	73.99	74.24
Hepatitis	155	6	0	13	74.50	77.50	80.28	82.58	78.88	84.76
Iris	150	4	0	0	92.32	94.67	95.68	94.89	96.01	98.43
LED	1000	0	0	7	54.53	57.20	60.28	60.28	60.27	62.85
Vehicle	846	18	0	0	64.23	70.93	70.93	75.27	71.37	76.43
Vowel	528	10	0	0	94.32	99.24	99.24	99.53	96.21	99.78
Wine	178	13	0	0	93.65	95.46	95.46	97.78	96.48	98.56
Zoo	90	0	0	16	95.78	97.78	98.89	98.89	98.89	98.84
平均	-	-	-	-	78.74	81.88	83.02	84.04	82.60	85.55

表 2 MNB、IVDM、WIMNB、KNN、IWIMNB 和 SVM 算法分类准确率

数据集	实例数	类别	MNB	IVDM	WIMNB	KNN	IWIMNB	SVM
20Newsgroups	18828	20	88.32	89.01	89.67	89.40	91.03	93.52
WebKB	4391	4	80.08	79.10	80.09	82.67	86.87	91.13
WebKB-NoHTMLTagsOrStoplist	4382	4	85.95	87.21	87.68	87.08	92.88	93.42
IndustrySector	9558	105	54.20	55.32	58.89	57.68	65.36	70.56
Industry-SectorNoHTMLTags	9549	105	64.21	66.01	67.48	64.10	74.43	88.06
Reuters-21578	10789	90	53.21	54.45	55.04	52.89	64.40	75.43

对几种数据集的分类准确率求平均值, 统计各种改进算法对应的平均分类准确率, 结果见表 3.

表 3 算法平均分类准确率对比

算法	平均分类准确率
MNB	70.99
MNB-LWL-IVDM	71.85
WIMNB	73.14
KNN	72.30
IWIMNB	79.16
SVM	85.33

从表 2、表 3 可以看出, 与 MNB 算法相比除个别数据集外(如 20Newsgroups), 改进算法的分类准确率都有一定程度的提高, 其中 IWIMNB 算法的平均分类准确率增幅最高约 8%, IVDM 算法增幅最低不到 1%. 基于余弦相似度进行的实例加权法与文献[13]中提到改进效果最好的 IVDM 算法相比, 平均分类准确率有显著提升, 这表明改进算法能有效获得高质量的训练样本子集并进行了特征权重改进.

综上所述, 改进算法尽管仍无法达到 SVM 的分类

© 中国科学院软件研究所 http://www.c-s-a.org.cn

效果, 但可操作性强且能显著提高分类准确率.

5 结语

本文提出了一种改进的基于余弦相似度进行实例加权的朴素贝叶斯分类算法(IWIMNB), 算法考虑不同特征具有不同的决策权重, 使用余弦相似度度量训练样本与验证样本之间的相似性, 选择最佳训练样本子集, 用标准化的余弦相似度数值作为训练样本的权值对修正后的朴素贝叶斯模型进行参数训练, 有效提高了训练样本的质量并弱化了特征条件独立假设. 对一系列 UCI 数据集进行分类的对比实验结果表明, 本文提出的改进算法可操作性强并具有更好的分类效果, 能有效提高算法平均分类准确率.

值得注意的是, 改进算法没有考虑结构拓展以及特征选择对分类准确率的影响, 当文本特征量级不同时需要对余弦相似度的计算进行修正, 在原有计算基础上需要多次计算每个训练样本与验证样本的余弦相似度并排序时间复杂度较高, 针对不同数据集如何确定 k 的全局最优值. 下一步将从结构拓展和特征选择等角度入手, 尝试解决上述问题, 重点研究基于结构拓展和特征选择算法的改进方法, 尽量多包含与分类相关的信息使算法的结构更加合理, 同时进一步减少算法时间复杂度并提高算法平均性能.

参考文献

1 Kulkarni AR, Tokekar V, Kulkarni P. Identifying context of text documents using Naïve Bayes classification and Apriori association rule mining. Software Engineering (CONSEG), 2012 CSI Sixth International Conference. IEEE. 2012. 1–4.

2 Song W, Yu JX, Cheng H, et al. Bayesian network structure learning from attribute uncertain data. Web-Age Information Management. Springer Berlin Heidelberg, 2012: 314–321.

3 Hall MA. Correlation-Based Feature Selection for Machine Learning. The University of Waikato, 1999.

4 Frank E, Hall M, Pfahringer B. Locally weighted naive bayes. Proc. of the Nineteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc. 2002. 249–256.

5 Kohavi R. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. KDD. 1996: 202–207.

6 Aggarwal CC, Zhai CX. A survey of text classification algorithms. Mining Text Data. Springer US, 2012: 163–222.

7 McCallum A, Nigam K. A comparison of event models for naive bayes text classification. AAAI-98 Workshop on Learning for Text Categorization. 1998, 752. 41–48.

8 Huang Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. DMKD, 1997.

9 Han EHS, Karypis G, Kumar V. Text Categorization Using Weight Adjusted K-nearest Neighbor Classification. Springer Berlin Heidelberg, 2001.

10 Li C, Jiang L, Li H, et al. Attribute weighted value difference metric. 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE. 2013. 575–580.

11 Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations. HLT-NAACL. 2013. 746–751.

12 Baena-García M, Carmona-Cejudo JM, Castillo G, et al. TF-SIDF: Term frequency, sketched inverse document frequency. 2011 11th International Conference on Intelligent Systems Design and Applications (ISDA). IEEE. 2011. 1044–1049

13 Yamada T, Yamashita K, Ishii N, et al. Text classification by combining different distance functions withweight. Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2006). IEEE. 2006. 85–90.

14 Hu Y, Milios EE, Blustein J. Enhancing semi-supervised document clustering with feature supervision. Proc. of the 27th Annual ACM Symposium on Applied Computing. ACM. 2012. 929–936.

15 Yang XS, Deb S, Fong S. Accelerated particle swarm optimization and support vector machine for business optimization and applications. Networked Digital Technologies. Springer Berlin Heidelberg, 2011: 53–66.

© 中国科学院软件研究所 http://www.c-s-a.org.cn