一种改进的线性支持向量机的特征筛选算法*

张 阳[1], 刘永革[2], 景 旭[1]
(1 西北农林科技大学 信息工程学院, 陕西 杨凌 712100;
2 安阳师范学院 计算机科学系, 河南 安阳 455100)

[摘 要] 针对SVM 法线特征筛选算法仅考虑法线对特征筛选的贡献, 而忽略了特征分布对特征筛选的贡献的不足, 在对SVM 法线算法进行分析的基础上, 基于特征在正、负例中出现概率的不同提出了加权SVM 法线算法, 该算法考虑到了法线和特征的分布。通过试验可以看出, 在使用较小的特征空间时, 与SVM 法线算法和信息增益算法相比, 加权SVM 法线算法具有更好的特征筛选性能。

[关键词] 特征筛选; 支持向量机; 加权SVM 法线算法; 文本分类
[中图分类号] TP31 [文献标识码] A [文章编号] 1671-9387(2006)10-0210-03

在文本分类研究中, 特征空间规模对分类器的分类准确度和效率影响很大。线性支持向量机(Support Vector Machine, SVM)是一种优秀的文本分类器[1]。针对线性SVM 的特征筛选, 是目前文本分类研究的热点之一。Joachims[1]用信息增益(Information Gain, IG)进行特征筛选, 用TDIDF 权重表示文档; Dumais 等[2]用互信息(Mutual Information, MI)进行特征筛选, 用布尔权重表示文档; Yang 等[3-4]用χ^2 进行特征筛选, 用TDIDF 权重表示文档; Brank 等[5]针对朴素贝叶斯、SVM 等分类器, 提出了SVM 法线算法, 通过对IG、SVM 法线等多种特征筛选算法的比较, 指出SVM 法线算法具有最好的性能。

但SVM 法线算法仅仅考虑了法线对特征筛选的影响, 即线性SVM 通过学习给每个特征赋予的权重仅与法线有关, 而未考虑特征的分布情况。针对此项不足, 为了进一步提高SVM 法线算法在特征筛选上的性能, 本文提出了加权SVM 法线特征筛选算法, 并在算法中同时体现法线和特征的分布对特征筛选的贡献, 以期为文本分类研究提供更为可靠的特征筛选算法。

1 SVM 法线特征筛选算法

SVM 法线特征筛选算法[5]中, 对于由d 个样品构成的训练数据集$\{\vec{X}_i y_i\}$, $i=1, 2, \cdots, d$, $\vec{X}_i \in R^n$ 表示1 个样品, $y_i \in \{-1, +1\}$表示该样品所属的类别。线性SVM 通过在输入空间中构造最优超平面来分隔正例、负例样品数据。假设$\vec{W} \in R^n$, $b \in R$, 用"$<\ ,\ >$"表示向量之间的内积, 则该分类超平面可以表示为:

$$<\vec{W}, \vec{X}> + b = 0 \qquad (1)$$

SVM 分类器的分类函数可以用分类超平面表示为:

$$f(\vec{X}) = \mathrm{sgn}(<\vec{W}, \vec{X}> + b) \qquad (2)$$

式中, sgn()为符号函数。

根据式(2)有: $\frac{\partial f}{\partial \vec{X}_i} = \vec{W}_i$。也就是说, 特征$A_i$ 对分类做出的贡献可以用$|\vec{W}_i|$来衡量。SVM 法线特征筛选算法中, 特征权重定义为:

$$SN(A_i) = |\vec{W}_i| \qquad (3)$$

经过线性SVM 学习超平面后得到的向量$\vec{W}$ 中, $|\vec{W}_i|$越大, 其对应特征A_i 就越重要; 反之, 就越不重要。若$\vec{W}_i > 0$, 特征A_i 表示样品倾向于正例; $\vec{W}_i < 0$, 特征A_i 表示样品倾向于负例。

可以看出, 式(3)中仅考虑了SVM 法线对特征筛选起到的贡献, 而忽视了特征分布对特征筛选的影响。

2 加权SVM 法线特征筛选算法

在SVM 法线特征筛选算法中, 如果在样品中观察到特征A_i 的概率很低, 即使SVM 学习的结果中$|\vec{W}_i|$很大, 但针对整个测试数据集来说, 特征A_i 对分

* [收稿日期] 2006-04-12
[基金项目] 西北农林科技大学人才基金项目(01140401; 01140402)
[作者简介] 张 阳(1975-), 男, 江苏扬州人, 教授, 博士, 主要从事数据挖掘方面的研究。E-mail: zhangyang@nwsuaf.edu.cn

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

类函数$f(\vec{X})$贡献低的概率较大; 相反, 如果在样品中观察到特征A_i的概率很高, 即使$|\vec{W}_i|$很小, 但对整个测试数据集来说, 特征A_i对分类函数$f(\vec{X})$贡献高的概率较大。可以看出, 在特征筛选时, 应该考虑到特征的分布情况。为此, 通过对SVM 法线特征算法的分析, 在算法中引入特征的分布情况, 现提出加权SVM 法线特征筛选算法。

用$\vec{X}_{ji}$表示向量$\vec{X}_j$在第i个维度上的分量, 在加权SVM 法线算法中, 特征的权重定义如下:

$$WSN(A_i)=\left|\left(\frac{\sum_{y_j=+1}\vec{X}_{ji}}{|y_j=+1|}-\frac{\sum_{y_j=-1}\vec{X}_{ji}}{|y_j=-1|}\right)\vec{W}_i\right| \quad (4)$$

式中, $\left|\left(\frac{\sum_{y_j=+1}\vec{X}_{ji}}{|y_j=+1|}-\frac{\sum_{y_j=-1}\vec{X}_{ji}}{|y_j=-1|}\right)\vec{W}_i\right|$表示特征$A_i$在正、负例样品中出现的概率之差。该值大于0 时, 表示特征A_i倾向于在正例样品中出现; 该值小于0 时, 表示特征A_i倾向于在负例样品中出现。其绝对值越大, 表示特征A_i在正、负例样品中出现的概率相差越大, 越能对分类做出贡献; 该项的绝对值越小, 表示特征A_i在正、负例样品中出现的概率越接近, 对分类做出的贡献越小。

可以看出, 加权SVM 法线特征筛选算法中同时考虑了特征的分布和SVM 学习的结果, 能进一步提高特征筛选的性能。

3　实例验证

以文本分类的标准实验文档集Reuters21578[6]作为实验样本集, 选用其中最大的10 个文档类别进行试验, 使用常用的*ModApte* 分割将其分割为训练文档集和测试文档集。

试验步骤如下:

(1) 对数据集进行预处理, 包括终结词[7](Stop Word) 处理和词干化[8](Stemming) 处理;

(2) 用SVM 法线算法(SN)、加权SVM 法线算法(WSN)、信息增益算法(IG) 等3 种算法分别对选用的样本集进行特征筛选;

(3) 用特征筛选后的特征集来表示原始文档;

(4) 在训练数据集上用线性SVM 分类算法学习分类模式;

(5) 用该分类模式对测试数据集进行分类分析。

通常使用F1 和BEP 来衡量文本分类器的性能, 对这两种指标的计算又有宏观值和微观值之分。文本分类研究中经常使用微观F1 和微观BEP 来评价文本分类的结果, 这里也采用这两个指标(其计算方法见文献[3, 4])。分类结果如表1 和表2 所示, 其中第1 列给出了特征集的大小, 第2, 3, 4 列分别给出了用信息增益算法、SVM 法线算法和加权SVM 法线算法进行特征筛选的分类结果。

表1　用微观F1 衡量的分类结果

Table 1　Classification result measured by micro F1

特征集 Feature set	IG	SN	WSN
100	0.915 751	0.928 623	0.931 636
200	0.928 976	0.930 182	0.930 758
300	0.928 990	0.932 293	0.933 430

表2　用微观BEP 衡量的分类结果

Table 2　Classification result measured by micro BEP

特征集 Feature set	IG	SN	WSN
100	0.916 150	0.928 831	0.931 805
200	0.929 267	0.930 350	0.930 987
300	0.929 366	0.932 422	0.933 609

由表1, 2 可以看出, 在信息增益算法、SVM 法线算法及加权SVM 法线算法中, 信息增益算法的性能最差。当特征集规模较小(如不超过300 个特征) 时, 由于加权SVM 法线算法同时考虑了法线和特征分布对特征筛选的贡献, 故加权SVM 法线算法的性能优于SVM 法线算法。

在文本分类技术的应用中, 往往希望使用较小的特征空间, 从而提高学习和分类的效率。此时, 相比信息增益算法和SVM 法线算法, 使用加权SVM 法线算法进行特征筛选将会使文本分类器具有更好的分类性能。

本文提出的特征筛选算法, 在DRC-BK 分类器[9]的规则筛选中亦取得了良好的效果。

4　结　论

在文本分类研究中, 较好的特征筛选算法有助于提高文本分类的性能和效率。SVM 法线算法虽是一种优秀的特征筛选算法, 但该算法仅仅考虑了法线对特征筛选的贡献, 而忽略了特征分布的影响。本文提出的加权SVM 法线算法, 同时考虑了法线和特征分布对特征筛选的影响。实例验证结果表明, 针对较小的特征空间, 加权SVM 法线算法较信息增量算法和SVM 法线算法具有更好的分类性能。但今后还拟进一步研究针对线性SVM 的特征筛选算法, 以提高加权SVM 法线算法的特征筛选性能。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[参考文献]

[1] Joachims T. Text categorization with support vector machines learning with many relevant features[C]//In Proceedings of the 10th European Conference on Machine Learning. Berlin: Spring-Verlag, 1998: 250-265.

[2] Dumas S, Platt J, Heckerman D, et al. Inductive learning algorithms and representations for text categorization[C]//In Proceedings of the 7th International ACM Conference on Information and Knowledge Management. Berlin: Spring-Verlag, 1998: 350-362.

[3] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]//In Proceedings of ICML-97, 14th International Conference on Machine Learning. Berlin: Spring-Verlag, 1997: 412-420.

[4] Yang Y, Liu X. A re-examination of text categorization methods[C]//In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research And Development in Information Retrieval (SIGIR '99). New York: ACM, 1999: 158-166.

[5] Brank J, Grobelnik M, Milic'-Frayling N, et al. Interaction of feature selection methods and linear classification models[C]//In Proceedings of the ICML-02 Workshop on Text Learning. Sydney: AU, 2002: 1100-1112.

[6] [2006-04-01]. http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[7] [2006-04-01]. ftp://sunsite.dcc.uchile.cl/pub/users/rbaeza/irbook/stopper/.

[8] Martin Porter. The porter stemming algorithm[EB/OL]. [2006-04-01]. http://www.tartarus.org/~martin/PorterStemmer/.

[9] Zhang Yang, Li Zhan-huai, Cui Ke-bin. DRC-BK: mining classification rules by using boolean kernels[C]//In the Proceedings of the International Conference on Computational Science and its Applications. Berlin: Spring-Verlag, 2005: 850-865.

An improved feature selection algorithm for linear SVM classifiers

ZHANG Yang[1], **LIU Yong-ge**[2], **JING Xu**[1]

(1 *College of Information Engineering, Northwest A & F University, Yangling, Shaanxi* 712100, *China*;
2 *Deptartment Computer Science, Anyang Normal University, Anyang, He'nan* 455100, *China*)

Abstract: SVM normal feature selection algorithm only considers the contribution made to feature selection by the normal, ignoring the contribution made by the distribution of features totally. In this paper, based on the analysis of SVM normal algorithm, and the difference between the probability that attributes occurs in positive and negative samples, a weighed SVM normal algorithm, which considers both the distribution and the normal was presented. The experiment results show that, compared with SVM normal algorithm and information gain algorithm, when a small feature space is applied, the weighed SVM normal algorithm has better feature selection performance than SVM normal algorithm and information gain algorithm.

Key words: feature selection; SVM; weighed SVM normal algorithm; text categorization

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net