基于随机子空间的多分类器集成

叶云龙, 杨 明

(南京师范大学 数学与计算机科学学院, 江苏 南京 210097)

[摘要] 提出了一种基于随机子空间的多分类器集成算法 RFSEn 首先选择一个合适的子空间大小, 然后随机选择特征子集并投影, 并得到子空间上的基分类器, 从而通过基分类器构成集成分类器, 并由集成分类器来进行文本的分类. 将该算法与单一分类器和基于重抽样技术的 bagging算法进行了比较, 在标准数据集上进行了实验. 结果表明, 该方法不仅优于单一分类器的分类性能, 而且一定程度上优于 bagging算法.

[关键词] 随机子空间, 分类器集成, 重抽样

[中图分类号] TP 391. 4 [文献标识码] A [文章编号] 1672-1292(2008) 04-0087-04

Multi-Classifier Ensemble Based on Random Feature Subspace

Ye Yunlong Yang Ming

(School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China)

Abstract In this paper, we propose an ensemble algorithm called RFSEn which is based on random feature subspace. First, an appropriate feature subset size is selected, then subsets of features are randomly and projected on the training set, and the primary classifiers of subspace are obtained, and thus ensembled classifiers are formed with these primary classifiers. At last, we use the ensembled classifier to classify the text. We compare the algorithm with bagging algorithm which is based on re-sampling techniques and single classifier on the standard datasets. The results show that RFSEn algorithm is not only superior to single classifier in performance, but better than bagging algorithm in some degree.

Key words random sub-space, classifier ensemble, re-sampling

近年来, 集成学习已成为模式识别研究的热点问题, 并已在模式识别的多个应用方面, 如字符识别、目标识别、文本分类等领域, 获得了较好的应用效果. 集成学习的研究被 Dietterich 认为是当前机器学习的四大研究方向之首[1].

按对训练数据进行处理得到个体学习器方式的不同, 可以将集成学习大致分为两类. 一类是以 Freund 和 Schapire[2]提出的 AdaBoost为代表, 在这类方法中, 前一级分类器为后一级分类器提供分类信息, 指导下一级分类器的训练和分类过程. 另一类是以 Breiman[3]提出的 Bagging算法为代表, 该类方法中各分类器是独立设计的, 相互不受干扰, 其核心是找到一种合适的组合准则来将各分类器的输出综合起来形成最终的结果. 目前, 已经有很多研究者投入到集成学习的研究中. Zhou等[4]提出了人工神经网络集成, 同时提出了“选择性集成”理论, 并证明通过选择部分个体学习器来构建集成可能要优于使用所有个体学习器构建的集成. 20世纪 90年代末, 集成学习技术开始被引入文本分类领域, 并成为该领域的一个研究热点. 例如, Weiss等[5]用决策树的集成进行文本分类, 并且成功地用于 email的过滤. Schapire等[6]将决策树桩(decision stump) 的集成用于文本分类系统 Boostexter中, 也取得了较好的效果.

理论和实验表明[7], 如果用于集成的分类器之间相互独立, 将可以得到最佳的分类精度. 而获得这种独立性的最有效方法是从特征子集得到集成成员分类器[7]. 换句话说, 在集成学习中, 特征划分将获得比数据划分更优的性能.

目前, 随机子空间法常用于增强集成分类器之间的独立性, Wang等[8]将其引入人脸识别中, 取得了不错的效果. Bay S D[9]将随机子空间法引入 kNN分类器集成, 取得了比 Bagging算法更高的分类精度.

收稿日期: 2008-06-18

通讯联系人: 杨 明, 教授, 博士, 研究方向: 数据挖掘、机器学习和粗糙集理论及应用. E-mail: myang@njnu.edu.cn

Robert等[10]将随机子空间法运用到手写体识别中,并取得了很好的效果.

Internet中存储了海量的文本信息,如何有效地发现、处理、过滤和管理这些信息资源是一个亟需解决的问题.文本分类是指在给定的分类体系下,根据文本的内容自动判别文本类别的过程.近年来,文本分类技术已经逐渐与搜索引擎、信息推送、信息过滤等信息处理技术相结合,正在各个领域得到广泛的应用,有效地提高了信息服务的效率和质量,而集成学习能够得到更高的分类准确率.由于文本样本的复杂性,经过预处理后,特征空间的规模仍然十分庞大,单一分类器的分类精度有限.为此,本文提出了基于随机子空间的多分类器集成算法(Random Feature Subspace Ensemble, RFSEn),并与 Bagging算法以及单一分类器进行了比较.实验结果表明,该方法是有效可行的.

1 相关概念

1.1 向量空间模型

目前,文本的表示主要采用向量空间模型(VSM).向量空间模型将文本的内容形式化为多维空间中的一个点,并以向量的形式来描述,向量的每一个分量表示特征项在该文本中的权重.根据实验结果,普遍认为选取词作为特征项要优于字和词组.因此,要将文本表示为向量空间中的一个向量,就首先要将文本分词,由这些词作为向量的维数来表示文本,用词频来表示每个词在文本中的权重.词频分为绝对词频和相对词频.绝对词频,即使用词在文本中出现的频率表示文本.相对词频为归一化的词频,其计算方法主要运用 TF-IDF公式.目前存在多种 TF-IDF公式,本系统中采用了一种比较普遍的 TF-IDF公式:

$$W(t_j, d_i) = \frac{tf(t_j, d_i) \times \log\left(\frac{N}{df(t_j)} + 0.01\right)}{\sqrt{\sum_{k=1}^{M}\left[tf(t_k, d_i) \times \log\left(\frac{N}{df(t_k)} + 0.01\right)\right]^2}}. \quad (1)$$

式中,$tf(t_j, d_i)$表示特征项 t_j在文本 d_i中出现的频率,$df(t_j)$表示特征项的文本频,N表示总的文本数,M表示特征总数,$W(t_j, d_i)$表示了第 i个样本矢量的第 j个分量值.

1.2 特征选择

特征选择一般是构造一个评价函数,对特征集中的所有特征进行分别评估,这样每个特征项都得到一个评估分值,然后对全部的特征按照其分值的大小进行排序,一般选取前 N 个最佳特征作为结果.其中N初始由人为设定,然后由实验来确定.常用的特征选择方法有文档频率、信息增益、互信息、X^2 统计量、期望交叉熵、文本证据权和几率比等,本文选用了 X^2 统计量的方法.在文本分类中,X^2 统计量表示如下:

$$CHI(T, C_i) = \frac{n[P(T, C_i) \times P(\overline{T}, \overline{C}_i) - P(\overline{T}, C_i) \times P(T, \overline{C}_i)]^2}{P(T) \times P(C) \times P(\overline{T}) \times P(\overline{C}_i)}. \quad (2)$$

特征 T的全局 CHI值如式(3)所示:

$$CHI(T) = \sum_{i=1}^{m} CHI(T, C_i). \quad (3)$$

CHI度量了特征 T与类别 C_i的相关程度,CHI值越大,表示 T与 C_i越相关,越依赖于 C_i进行特征选择时,选择 CHI值大的特征.

2 基于随机子空间的多分类器集成算法 *RFSEn*

由于文本样本的复杂性,经过特征项的提取后,特征维数仍然可以达到上千维,直接在原始空间上再进行降维,可能会丢失某些重要信息,因此本文提出基于随机子空间的分类集成算法 RFSEn.本文中子空间的选择是根据均匀分布 U随机抽取 m 个不同的子集 $A = \{d_1, d_2, \cdots, d_m\}$,每个子集的大小(即子空间的维数)为 r,每个子空间都定义一个映射 $P_A: F^n \rightarrow F^m$,在此基础上得到每个训练子集 $D_i = \{(P_A(x_j), y_j) \mid 1 \leqslant j \leqslant N\}$.再由分类算法 L得到待检样本的决策 h_i,重复 m 次,最后利用择多投票法得到最终决策.其中,子空间维数 r和基分类器的个数 m 可自动确定,为简便采用事先设定的方法.本文实验中,分别采用了在文本分类中常用的 SVM、C4.5、NaiveBayes、RBFNetwork 4种分类器作为基分类器,进行同态集成.

依据上述分析,基于随机子空间的分类集成算法(RFSEn)的具体步骤描述如下:

输入: (1) 训练集 $D = \{(x_j, y_j) \mid 1 \leqslant j \leqslant N\}$, $x_j \in X \subset \mathbf{R}^n$, $y_j \in \mathbf{C} = \{1, \cdots, k\}$;

(2) 学习算法 (训练基分类器) L;

(3) 子空间维数 $r < n$;

(4) 基分类器个数 m;

输出: 集成分类器 H.

步骤:

(1) for($i = 1$; $i <= m$; $i++$)

{随机产生 r维向量 $\boldsymbol{select}_i$, 保存该子空间在原始空间中的位置;

}

(2) 由 (1) 得到的所有特征子集为 $\boldsymbol{select}_1, \cdots, \boldsymbol{select}_m$, 将 $\boldsymbol{select}_1, \cdots, \boldsymbol{select}_m$ 在训练集 D 上进行投影得到不同子空间上的训练数据集 $D_1, \cdots, D_m$;

(3) 由分类算法 L 和各个子空间上的训练数据 $D_1, \cdots, D_m$ 来训练分类器 $\text{classifier}_1, \text{classifier}_2, \cdots, \text{classifier}_m$;

(4) 对于待检样本 X_i, 根据特征子集 $\boldsymbol{select}_1, \cdots, \boldsymbol{select}_m$ 得到该样本对应的子样本 $X_{i1}, \cdots, X_{im}$;

(5) $H(X_i) = \underset{y \in C}{\arg\max} \sum_{j: \text{Classifier}_j(X_{ij}) = y} 1$.

基于随机子空间的分类集成算法 (RFSEn)降低了特征空间的维数, 并且由于训练数据使用的是不同的特征空间投影得到的子数据集, 从而降低了分类器之间的相关性, 同时也避免了特直接在原始空间上进行特征选择带来的信息丢失问题, 提高了集成的效果.

3 实验结果及分析

为了验证 RFSEn算法的有效性, 本文在 20Newsgroup、WebKB、Reuters 3个数据集上做了实验.

3.1 数据集

20Newsgroups数据集包含从新闻组收集到的 19 997篇文章, 本文使用了其中的 5类共 5 000个文档; WebKB数据集包含着从 4个学校获得的 8 282个 Web网页, 本文采用了其中的 4 000个网页; Reuters数据集包含 21 578篇取自路透社新闻专线的文章, 实验采用的是“ModApte”版本的 Reuters-21578文集, 在样本数最多的 5个类别上进行了实验.

3.2 分类性能的评估

本文实验中使用的性能评估指标为常用的 F1-value 为了更好地介绍查全率和查准率以及 F1-value的含义, 建立如表 1所示的混合矩阵. 混合矩阵是针对某一个类而言的, 它统计了所有测试文本与这个待定类之间的分类情况.

表 1 混合矩阵

Table 1 Confusion matrix

	实际属于该类的文本数	实际不属于该类的文本数
分类器认为属于该类	***TP***	***FP***
分类器认为不属于该类	***FN***	***TN***

则查全率 (Recall)和查准率 (Precision)以及 F1-value 定义如下:

$$\text{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}, \tag{4}$$

$$\text{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}, \tag{5}$$

$$\text{F1-value} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{6}$$

对于多类别的问题, 一般采用平均的方法即微平均 (micro-average)和宏平均 (macro-average).

3.3 实验结果

本文以 SVM、RBFNetwork、NavieBayes和 C4.5作为基本分类器, 在上述 3个数据集上做了实验. 实验中, 原始空间的维数为 900, 子空间的维数取为 500, 集成的大小为 15. 表 2给出了 RFSEn算法、Single算法 (原始空间上单一分类器)以及 Bagging算法在上述 3个数据集上的实验结果.

表 2 分类性能比较

Table 2 Classification performance comparison

		SVM		RBFNetwork		NavieBayes		C4.5	
		mac-f1	mic-f1	mac-f1	mic-f1	mac-f1	mic-f1	mac-f1	mic-f1
Newsgroup	RFSEn	0.9240	0.9249	0.8659	0.8677	0.8639	0.8576	0.8132	0.8111
	Single	0.9144	0.9154	0.8526	0.8565	0.8629	0.8520	0.7796	0.7757
	Bagging	0.9158	0.9170	0.8589	0.8627	0.8695	0.8632	0.8363	0.8363
Reuters	RFSEn	0.8505	0.8618	0.7237	0.7422	0.7634	0.7764	0.8203	0.8354
	Single	0.8478	0.8602	0.7080	0.7236	0.7684	0.7810	0.7513	0.7655
	Bagging	0.8486	0.8618	0.7103	0.7314	0.7822	0.7950	0.7856	0.7981
Webkb	RFSEn	0.8834	0.8967	0.7556	0.7710	0.7849	0.7929	0.8567	0.8723
	Single	0.8755	0.8884	0.7337	0.7530	0.7933	0.8012	0.7766	0.8041
	Bagging	0.8754	0.8874	0.7344	0.7569	0.8058	0.8159	0.8121	0.8329

3.4 实验结果分析

从表 2可以看出,除了 NavieBayes外,只有以 C4.5为集成的基分类器,在 Newsgoup数据集上子空间集成不如 bagging算法.多数情况下,子空间集成算法都优于 Bagging,并且明显好于原始空间上单一分类器的分类效果.NavieBayes作为基分类器时,子空间集成失效的原因可能是 NavieBayes假设特征之间相互独立,而多数情况下,这种假设是不成立的[11].

4 结语

本文提出了基于随机子空间的多分类器集成算法 RFSEn,并在标准数据集上与单一分类器及 Bagging算法进行了实验对比.实验结果表明,RFSEn算法能有效增强文本分类器的推广性,同时给文本分类提供了一种新的途径.由于本文在对分类结果进行集成时采用的是最简单的择多投票法,未考虑个体分类器的多样性,这可能也是导致 RFSEn算法在某些数据集上效果不尽人意的原因.下一步工作将研究各种加权投票方法,充分考虑分类器的多样性,进一步提高文本分类的性能.

[参考文献] (References)

[1] Dietterich T G. Machine learning research: four current directions [J]. AIMagazine, 1997, 18(4): 97-136.

[2] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C] // Proceedings of the 13th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1996: 148-156.

[3] Breiman L. Bagging predictors[J]. Maching Learning, 1996, 24(2): 123-140.

[4] Zhou Z H, Wu J, Tang W. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002, 137(1/2): 239-263.

[5] Weiss S M, Apte C, Damerau F J. Maximizing text-mining performance[J]. IEEE Intelligent Systems, 1999, 14(4): 63-69.

[6] Schapire R E, Singer Y. Boostexter: a boosting-based system for text categorization [J]. Machine Learning, 2000, 39(223): 135-168.

[7] Tumer K, Ghosh J. Classifier combining: analytical results and implications[C] // Proceeding of the AAAI-96 Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learing Algorithms. Portland: AAAI Press, 1996.

[8] Wang Xiaogang, Tang Xiaoou. Using random subspace to combine multiple features for face recognition[C] // Proceeding of the 6th IEEE International Conference on Automatic Face and Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2004: 284-289.

[9] Bay S D. Combining nearest neighbor classifiers through multiple feature subsets[C] // Proceeding of the Proceedings of the 17th International Conference on Machine Learning. Madison, WI: Morgan Kaufmann, 1998: 37-45.

[10] Robert Bryll a. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets[J]. Pattern Recognition, 2003, 36(6): 1291-1302.

[11] Lewis D D. Naïve (Bayes) at forty: the independence assumption in information retrieva[C] // Proceedings of 10th European Conference on Machine Learning. Chemnitz, DE: Springer Verlag, 1998: 4-15.

[责任编辑: 严海琳]