基于特征选择的 Bagging 分类算法研究

姚明海,赵连朋,刘维学
(渤海大学 信息科学与技术学院,辽宁 锦州 121013)

摘 要:为了提高数据的分类性能,提出了一种基于特征选择的 Bagging 分类算法。通过 Fisher 准则和互信息的方法给定一种能够直接评价特征区分度和与类别相关性的评价方法,重新构造了计算特征区分度和与类别相关性的计算公式。并将该方法应用到 Bagging 分类算法当中。实现了算法迭代过程中的特征选择,使得每个基分类器都是由不同的特征子集训练所得,保证了基分类器的独立性,降低了训练误差。通过理论分析和大量的实验,对文中的方法与经典特征选择方法进行了比较,实验结果显示文中的方法能够得到更高的预测精准度。

关键词:数据挖掘;特征选择;集成学习;互信息;Bagging;分类器
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)04-0103-04
doi:10.3969/j.issn.1673-629X.2014.04.026

Research on Bagging Classification Algorithm Based on Feature Selection

YAO Ming-hai,ZHAO Lian-peng,LIU Wei-xue
(College of Information Science and Technology,Bohai University,Jinzhou 121013,China)

Abstract:In order to improve the classification performance of data,a Bagging classification algorithm based on feature selection is proposed in this paper. An evaluation method is proposed for full account of the discrimination and class information of each feature by the Fisher criterion and mutual information,built on the formula about discrimination and class information. The feature selection algorithm is applied to the Bagging classification algorithm. The feature selection is implemented in the iterative process of algorithm,so that each base classifier is trained by different feature subsets,which ensures the independence of each base classifier,reducing the training error. Compared the method with several classical feature selection methods by theoretical analysis and extensive experiments,the results show that the method can achieve higher predictive accuracy.

Key words:data mining;feature selection;ensemble learning;mutual information;Bagging;classifier

0 引 言

数据挖掘[1](Data Mining)是近年来随着数据库技术和人工智能技术的发展而出现的。目的在于从海量的数据中发现内在的、隐藏的有价值的知识和信息。它主要采用机器学习算法或统计方法进行知识学习,数据分类是数据挖掘领域的一个重要分支,主要通过分析训练数据样本,产生关于类别的精确描述。目前的分类方法有很多,如决策树、神经网络、贝叶斯、关联规则等。数据分类的目的在于构造一个分类模型,该模型能把数据库中的数据项映射到给定的类别中的某一个。分类技术解决问题的关键是构造分类器[2]。不同的分类器有其不同的特点。对于不同的数据样本,各种分类器表现的性能各不相同。为了充分发挥各分类器的特长,多分类器组合技术应运而生。大量的理论和实验结果表明,通过多分类器组合不但可以提高分类的准确率,而且能够提高模式识别系统的效率和鲁棒性。Bagging 是采用集成学习思想组合多个弱分类器的算法,通过对多个弱分类器的融合,有效提高最终强分类器的泛化性能,被认为是性能较好的分类方法之一。但传统的多分类器融合算法的特征选择和分类器训练不是同时进行的,其迭代过程中没有进行特征选择,其固有不变的特征集合无法降低弱分类器的训练误差,同时特征选择的原则也缺乏对数据样本的针对性,这样也就无法有效地提高分类器的泛化性能。文献[3-4]中提出的算法在迭代过程中加入过滤式和 Wrapper[5] 的嵌入式的特征选择方法,但是这些特征选

收稿日期:2013-06-07 **修回日期**:2013-09-16 **网络出版时间**:2014-01-28
基金项目:吉林省科技发展计划项目青年科研基金(201201070);辽宁省社科联项目(2010lslktjyx-03)
作者简介:姚明海(1980-),男,博士研究生,研究方向为数据挖掘、机器学习;赵连朋,副教授,博士,研究方向为数值分析;刘维学,讲师,研究方向为数据挖掘。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140128.1149.041.html

择方法和传统的特征选择方法都存在单一的考虑特征自身特点或特征间的冗余程度。依然无法将训练误差最小化。

文中在传统特征选择方法的基础上提出了一种新的特征选择方法,既考虑了特征自身的区分度,又考虑了特征与类别的相关性。并将其与采用 Bagging 算法的多分类器组合的分类方法相结合。在该方法的迭代过程中特征选择与分类器训练同步进行,在迭代过程中进行特征选择。保证了训练不同的弱分类器采用不同的特征集合,充分利用了特征集所蕴含的丰富信息,从而降低了弱分类器的训练误差。

1 分类器设计

90 年代初,Schapire 在概率近似正确框架内证明了强可学习与弱可学习问题是等价的,从此在机器学习领域集成学习成为了研究的热点问题。目前集成学习的思想主要体现在三个方面:一是基于样本的随机采样策略;二是基于多分类器的集成策略;三是基于样本特征空间的集成策略。Bagging 算法是较为常用的一种集成学习算法,其思想是:对训练样本进行有放回的随机采样,形成多个规模相近的训练子集,通过这些训练子集训练多个基分类器。通过多个基分类器的分类结果确定样本的分类属性。由于 Bagging 的样本采样方式使得各训练子集间是相互独立的,从而保证了基分类器的多样性,能够提高集成模型的泛化性能。Bagging 的具体过程如下:给定包含 n 个样本的训练样本集 D,从 D 中独立随机地抽取 $n'(n' < n)$ 个样本构成训练子集 D_1,通过样本子集 D_1 训练基分类器 C_1。重复以上过程 T 次后得到 T 个基分类器 $C_1, C_2, \cdots, C_T$。用这些基分类器对样本集进行预测,然后按照多数投票的规则得到样本集的预测结果。

2 基于特征选择的分类算法

2.1 特征选择

所谓特征选择,就是从数据样本集中按照某种准则选取一组有效的特征以降低特征空间的维数。同时,由于特征选择去除了特征空间中的一些冗余信息避免了这些信息对分类预测的影响,很大程度上提高了分类算法的预测准确率,也提高了算法的计算效率。特征选择方法按特征子集评价策略主要分为两大类:一类是过滤式的特征选择方法,这是一种计算效率较高的方法,它独立于后续的学习算法,采用统计的方法来评价特征的预测能力。具有代表性的方法有信息熵法[5]、Fisher score[6]、T 检验[7]等等。过滤式方法仅使用数据的内在属性,根据评价准则判断特征的预测能力。忽略了特征之间及特征与类别之间的相互关系。另一类是封装式的方法,这种方法在特征选择时依赖于具体的机器学习算法,根据分类器的预测性能来评价特征子集的优劣。封装式的方法相对于过滤式的方法计算量会有很大的提高。文中主要针对过滤式的方法进行了改进,在特征选择过程中综合考虑了特征的内在属性和特征与类别的相关性。具体流程如图 1 所示。

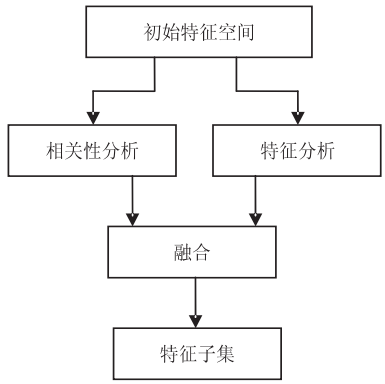

图 1 特征选择流程图

2.2 特征区分度评价

在特征分析上文中主要采用的是 Fisher 准则。费舍尔得分是使用最广泛的有监督特征选择方法之一,它根据特征的内在属性为特征进行打分,定义如下:

$$F^j = \frac{\sum_{k=1}^{c} n_k (\overline{x_j^k} - \overline{x_j})^2}{\sum_{k=1}^{c} n_k (\sigma_j^k)^2} \tag{1}$$

其中,k 为类标签;c 表示类的数量;n_k 表示第 k 类样本的数量;$\overline{x_j^k}$ 表示第 k 类中第 j 个特征的均值;$\overline{x_j}$ 表示所有样本中第 j 个特征的均值;σ_j^k 表示第 k 类中第 j 个特征的方差;F^j 越大表示该特征的区分度越好。

费舍尔得分分别考虑了类内散度和类间散度。类内散度越小越好,类间散度越大越好。但是,费舍尔得分没有考虑特征与类别的相关性。当特征的区分度非常好,但是与类别的相关性非常低时,对于分类预测来说这不是一个好的特征。针对这一问题,文中采用了互信息的方法来判别特征与类别的相关性。

2.3 类别相关性分析

在信息论中互信息[8]是一种信息的度量方法。它代表了两个事件之间的相关性,互信息定义为

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2}$$

其中,$I(X,Y)$ 表示事件 X 和事件 Y 的互信息;$H(X)$ 和 $H(Y)$ 表示事件 X 和事件 Y 的熵;$H(X,Y)$ 为联合熵。

熵的定义如下:

$$H(X) = \sum_{i=1}^{c} -p(x_i)\log_2 p(x_i) \tag{3}$$

联合熵的定义如公式(4)所示:

$$H(X,Y) = -\sum_{j=1}^{h} p(y_j)\left(\sum_{i=1}^{c} - p(x_i \mid y_j)\log_2 p(x_i \mid y_j)\right) \quad (4)$$

公式(3)和(4)中的 i,j 表示类别; h,c 表示类的数量; p 为事件发生的概率。

通过以上的分析相关性可以定义为:

$$D(X,c) = \frac{1}{|X|}\sum_{x_i \in X} I(x_i,c) \quad (5)$$

其中, X 表示特征子集; c 表示类别标签; $I(x_i,c)$ 为第 i 个特征与类标签 c 的互信息; D 越大表示特征与类别的相关度越高。

考虑到不同数据集的特性,给 F 和 D 分配不同的权重系数。当 F 起到主导作用时增加 F 的权重。相反增加 D 的权重。融合公式(1)和公式(5)得到目标评价函数如公式(6)所示:

$$S_i = \alpha \times F_i + (1-\alpha) \times D_i \quad \alpha \in [0,1] \quad (6)$$

当 S_i 的值越大表示第 i 个特征有更好的区分度和类别相关性。

2.4　算法实现

输入:训练样本集 X,训练样本类标签 Y,分类器算法 C,权重 α。

设 $X = \{x_1, x_2, \cdots, x_n\}$ 为训练样本集,共有 n 个样本,其中 x_i 为训练样本,每个样本由 d 个特征表示, $A = \{a_1, a_2, \cdots, a_d\}$。$Y = \{y_1, y_2, \cdots, y_n\}$ 为训练集的类标签集, y_i 为 x_i 的类标签。

输出:分类结果 G。

步骤 1:给定 j 和 T (j 为每次选取的特征数量, T 是构建基分类器的数量);

步骤 2: for i = 1 to T do

步骤 2.1:从训练样本集 X 中随机抽取 $m(m < n)$ 个训练样本构成样本子集 D_i ;

步骤 2.2:用公式(1)和公式(5)分别计算样本子集 D_i 中样本特征的区分度和类别相关度;

步骤 2.3:用公式(6)计算 D_i 中样本特征的得分,选择得分最高的 j 个特征构成训练子集 D_{ij} ;

步骤 2.4:用 D_{ij} 训练基分类器 C_i。

步骤 3:用 T 个基分类器对测试样本进行分类预测;

步骤 4:根据 T 个基分类器的预测结果,采用投票的机制预测分类样本的类别;

步骤 5:返回分类结果 G。

3　仿真实验

为了验证算法的有效性和优越性,文中的算法与经典特征选择算法 t-test, Information Gain (InfoGain)[9], Laplacian Score (LS)[10], ChiSquare (CS)[11], KruskalWallis(KW)[12], Mrmr[8] 进行了比较。

3.1　实验环境及数据

实验环境:Matlab2009b 开发平台,Windows7 操作系统。实验数据选用了 8 个标准机器学习数据库 UCI[13] 中的标准数据集,这些数据集的选取是从不同方面考虑的。从数据属性看,8 个数据集都是分类属性数据。从数据集大小考虑,较大的数据集有 magic 和 spambase,较小的数据集有 wine 和 glass;从数据集的特征数量来考虑,数量较多的有 ionosphere 和 spambase,数量较少的有 yeast 和 breast;从分类数量来看,分类数量较多的有 yeast 和 glass,分类数量较少的有 housing。具体数据集组成如表 1 所示。

表 1　实验数据集

数据集	特征维度	样本数量	分类数量
breast	9	699	2
glass	9	214	6
housing	13	506	2
ionosphere	34	51	2
magic	10	19 020	2
spambase	57	4 601	2
wine	13	173	3
yeast	8	1 484	8

以上 8 个数据集的数据都是筛选后的,一些不适合分类的特征如“ID”等已经被删除。

3.2　实验结果及分析

实验中的基分类器采用了台湾大学林智仁博士通过对 SVM 深入研究开发设计的多重分类支持向量机 LibSVM[14]。通过对不同数据集的统计,为每组数据集设置了一个权重值 α,α 的取值总体在 0.2 到 0.4 之间。为了降低样本偏差,实验采用了 4 折交叉验证,每组数据集分成 4 份,轮流选取其中的 3 份做训练,1 份做测试。4 次的均值作为分类准确率的估计。为了提高精确率,每次折交叉验证训练 9 个基分类器对测试集进行预测,这样的过程重复 30 次取均值。部分实验结果如图 2、图 3 和表 2 所示。

从实验结果可以看出在选择相同的特征维度时文中提出的方法基本都高于其他特征选择方法,只有在选择特征维度极少和选择的特征出现冗余时,预测准确率略低于其他个别方法。这就表明,充分考虑特征的区分度和特征与类别的相关性对于分类问题的特征选择是至关重要的,这些区分度较差和与类别相关性较低的特征不仅不能提高预测精准度,反而会降低计算效率,甚至影响分类效果。

4　结束语

文中提出了一种既考虑了特征自身的区分度,又

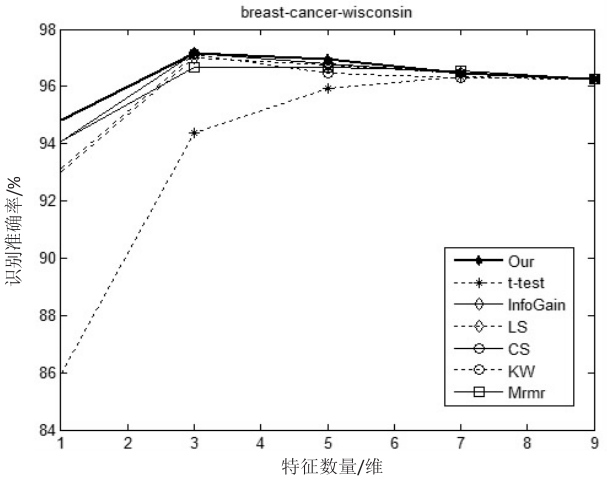

图 2 不同特征维度下的分类准确率(breast 数据集)

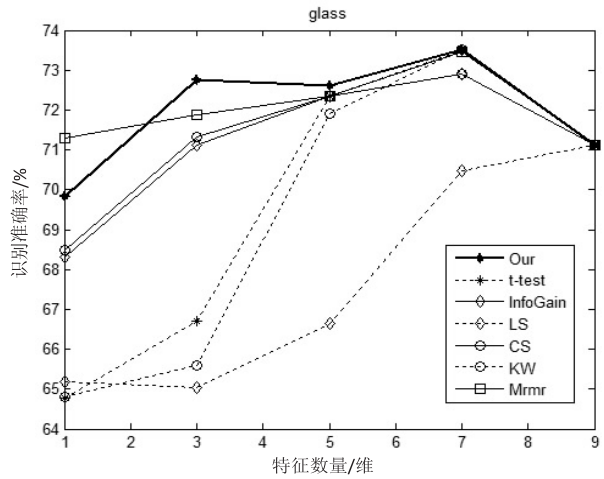

图 3 不同特征维度下的分类准确率(glass 数据集)

表 2 不同特征选择方法最高识别率统计(* 括弧内为特征维度)

	文中方法	t-test	InfoGain	LS	CS	KW	Mrmr
breast	97.18(2)	96.25(4)	96.81(2)	96.75(2)	96.79(2)	96.48(2)	96.65(3)
glass	73.51(7)	72.51(7)	72.68(7)	70.49(7)	72.91(7)	72.95(7)	72.91(7)
housing	100(1)	100(1)	99.9(1)	100(1)	99.9(1)	97.73(3)	99.98(1)
ionosphere	94.35(25)	93.23(34)	93.64(25)	93.45(30)	94.12(25)	93.85(30)	94.08(25)
magic	77.41(3)	71.92(9)	77.25(5)	76.97(10)	77.26(5)	76.97(10)	77.14(7)
spambase	88.66(10)	87.28(10)	87.74(10)	72.73(50)	87.9(5)	90.5(55)	72.63(50)
wine	97.82(5)	97.6(13)	97.6(13)	97.6(13)	97.6(13)	97.6(13)	97.6(13)
yeast	89.2(3)	87.38(3)	88.96(3)	87.73(5)	87.7(5)	87.38(3)	86.93(7)

考虑了特征与类别的相关性的特征选择方法。并将其与采用集成学习思想的多分类器组合的分类方法相结合,将特征选择过程融入到弱分类器的迭代过程中,有效降低了弱分类器的训练误差,最终改善了分类器泛化能力。实验结果表明文中的方法在分类性能上要高于其他方法。但是,文中的方法只能对特征进行评价,还不能实现自适应的选择特征维度。如何自适应的选择特征维度将是下一步的研究方向。

参考文献:

[1] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范 明, 孟小峰, 译. 北京:机械工业出版社, 2007.

[2] 张伟松, 高智英. 快速多分类器集成算法研究[J]. 计算机工程, 2012, 38(2):178-180.

[3] Sun Yijun, Li Jian. Adaptive learning approach to landmine detection[J]. IEEE transactions on aerospace and electronic systems, 2005, 41(3):973-985.

[4] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proc of CVPR 2001. Hawaii, USA:[s. n.], 2001:511-518.

[5] 许智伟, 胡 珉, 尹建新. 特征选择算法综述[J]. 电子设计工程, 2011, 19(9):46-51.

[6] Duda R O, Hart P E, Stork D G. Pattern classification[M]. New York:Wiley, 2001.

[7] Press W H, Teukolsky S A, Vetterling W T, et al. Numerical recipes in C[M]. 2nd ed. New York:Cambridge University Press, 1992.

[8] Ding J R, Huang J H J, Liu F, et al. Elastogram features selection and classification based on mRMR and SVM[J]. Journal of Harbin institute of technology, 2012, 44:81-85.

[9] Li X M, Li H R, Xue L, et al. TFIDF algorithm based on information gain and information entropy[J]. Computer engineering, 2012, 38(8):37-40.

[10] He X, Cai D, Niyogi P. Laplacian score for feature selection[M]//Advances in neural information processing system. Cambridge:MIT Press, 2005.

[11] Liu H, Setiono R. Chi2:Feature selection and discretization of numeric attributes[C]//Proc of 7th international conference on tools artificial intelligence. [s. l.]:[s. n.], 1995:388-391.

[12] Hollander M, Wolfe D A. Nonparametric statistical methods[M]. Hoboken, NJ:John Wiley & Sons, Inc., 1999.

[13] Blake C, Merz C. UCI repository of machine learning database[EB/OL]. 2005-09. http://www.ics.uci.edu/~mlearn/MLRepository.

[14] Hsu C V, Chang C C, Lin C J. LIBSVM:A library for support vector machines[EB/OL]. 2013-04. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

基于特征选择的Bagging分类算法研究

万方数据 WANFANG DATA 文献链接

作者： 姚明海， 赵连朋， 刘维学， YAO Ming-hai， ZHAO Lian-peng， LIU Wei-xue
作者单位： 渤海大学 信息科学与技术学院,辽宁 锦州,121013

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(4)

本文链接：http://d.wanfangdata.com.cn/Periodical_wjfz201404026.aspx