文章编号:1001 - 9081(2009)03 - 0833 - 03

基于支持向量机和 k-近邻分类器的多特征融合方法

陈 丽,陈 静
(中国农业大学 理学院,北京 100083)
(jing_quchen@163.com)

摘 要:针对传统分类方法只采用一种分类器而存在的片面性,分类精度不高,以及支持向量机分类超平面附近点易错分的问题,提出了基于支持向量机(SVM)和 k-近邻(KNN)的多特征融合方法。在该算法中,设样本集特征可分为 L 组,先用 SVM 算法根据训练集中每组特征数据构造分类超平面,共构造 L 个;其次用 SVM-KNN 方法对测试集进行测试,得到由 L 组后验概率构成的决策轮廓矩阵;最后将其进行多特征融合,输出最终的分类结果。用鸢尾属植物数据进行了数值实验,实验结果表明:采用基于 SVM-KNN 的多特征融合方法比单独使用一种 SVM 或 SVM-KNN 方法的平均预测精度分别提高了 28.7% 和 1.9%。

关键词:支持向量机;k-近邻;多特征融合;后验概率
中图分类号: TP311.13;TP181 **文献标志码**:A

Multi-feature fusion method based on support vector machine and k-nearest neighbor classifier

CHEN Li, CHEN Jing
(*College of Science, China Agricultural University, Beijing* 100083, *China*)

Abstract: The traditional classification methods only use one single classifier, which may lead to one-sidedness, low accuracy, and that the samples nearby the Support Vector Machine (SVM) hyperplanes are more easily misclassified. To solve these problems, the multi-feature fusion method based on SVM and K-Nearest Neighbor (KNN) classifiers was presented in this paper. Firstly, the features were divided into L groups and the SVM hyperplanes were constructed for each feature of training set. Secondly, the testing set was tested by SVM-KNN method, and the decision profile matrixes were obtained. Finally, these decision profile matrixes were implemented by multi-feature fusion method. The experimental results on Iris data show that the forecast accuracy of the multi-feature fusion method based on SVM-KNN classifiers increases by 28.7% and 1.9% than those of SVM and SVM-KNN methods respectively.

Key words: Support Vector Machine (SVM); K-Nearest Neighbor (KNN) algorithm; multi-feature fusion; inverse probability

0 引言

传统的分类方法大都采用单一的分类器对数据进行分类,或者通过增加单个分类器结构的复杂度来提高分类精度,结果往往令人不满意。而通过将多个结构较为简单的分类器进行融合的方法来提高整体的分类精度,不失为明智的选择。然而,常见的融合方法如线性组合法、投票法[1]、计数法[2]、模糊积分[3-4]等都没能利用样本被子分类器分类后的后验概率,忽略了子分类器的性能差异,分类结果都不理想。另外,还有一种融合方法:概率乘积准则[5],此方法虽利用了样本被每个子分类器分类后的后验概率,但由于概率乘积准则在融合过程中要将所有后验概率相乘,增加了融合后样本被错分的概率。这是因为假设存在某个子分类器将样本分为某类的概率为零,则融合过程中将所有后验概率相乘时,该样本被分为该类的概率将大大降低。而采用概率和准则[5]可有效避免此类情况的发生。

研究发现支持向量机[6](Support Vector Machine,SVM)与 k-近邻(K-Nearest Neighbor, KNN)相结合能提高支持向量机的分类精度[7]。因此,本文提出了基于支持向量机(SVM)和 k-近邻(KNN)的多特征融合分类方法,以期在不增加支持向量机算法时间复杂度的基础上,减少支持向量分类超平面附近样本的错分率,提高对数据的分类准确率。

1 SVM-KNN 分类器简介

支持向量机是由 Vapnik 等人于 1995 年提出的一种建立在统计学习理论基础上的分类方法,在解决小样本、非线性、高维模式识别中具有分类精度高、泛化能力强等优势。但是在使用 SVM 分类时,分界面附近的样本点比较容易被错分,而 SVM 分类器等价于每类只选一个代表点的 1-NN 分类器[7]。因而在对样本进行分类时,可以考虑根据空间的不同分布采用不同的分类方法:当样本距 SVM 最优超平面的距离大于一给定阈值 ε 时,即样本离分界面较远时,用 SVM 方法进行分类;反之,当样本和 SVM 最优超平面的距离小于 ε 时,用 KNN 方法对测试样本分类。

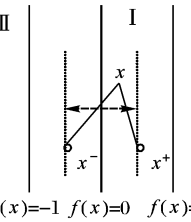

图 1 SVM-KNN 分类算法示意图

具体地,对于待识别样本 x,计算 x 与两类支持向量代表

收稿日期:2008 - 09 - 28;**修回日期**:2008 - 11 - 28。 **基金项目**:国家自然科学基金资助项目(10871022)。
作者简介:陈丽(1982 -),女,河南安阳人,硕士研究生,主要研究方向:数据挖掘、模式识别、支持向量机; 陈静(1964 -),女,河南南阳人,教授,主要研究方向:数据挖掘、数值模拟、最优化方法、支持向量机。

点 x^+ 和 x^- 的距离差，如果距离差大于 ε，如图1中区域Ⅱ，用SVM一般都可以正确分类；当距离差小于 ε，即落入区域Ⅰ，如果分类时仍用SVM，相当于只计算 x 与每类所取的一个代表点的距离，比较容易错分。所以这时采用KNN算法进行分类，将所有支持向量作为代表点，计算待识别样本和每个支持向量的距离，对待识别样本做出判断。

2 多特征融合分类方法

多特征融合[8]是指首先根据样本的每组特征分别对样本进行分类，然后将所有的分类结果进行融合，得到最终分类结果，其构架如图2所示。它能整合来自多信息源的信息，降低单信息源中存在的不确定性，从而提高系统的整体性能。

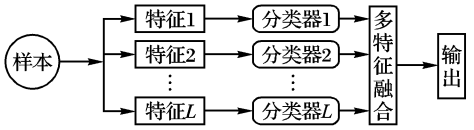

图2 多特征融合构架

图2中“样本”为测试集中一个样本，“特征 i”表示第 i 组参与融合的特征数据，$i = 1,2,\cdots,L$。设所有样本可分为 c 类。首先，根据训练样本集每组特征数据构造出 L 组基本分类器。然后，用这 L 组基本分类器分别对测试样本 x 进行分类。设经第 i 组分类器分类后得到的分类结果记为：$\{d_{i,1}(x),\cdots,d_{i,j}(x),\cdots,d_{i,c}(x)\}$，$d_{i,j}(x)$ 表示第 i 个分类器将测试样本分为第 j 类的概率（后验概率），其中 $i = 1,2,\cdots,L$，$j = 1,2,\cdots,c$。于是对每个测试样本 x，经 L 个分类器分类后可以得到一个决策轮廓矩阵：

$$\begin{bmatrix} d_{1,1}(x) & d_{1,2}(x) & \cdots & d_{1,j}(x) & \cdots & d_{1,c}(x) \\ d_{2,1}(x) & d_{2,2}(x) & \cdots & d_{2,j}(x) & \cdots & d_{2,c}(x) \\ \vdots & \vdots & & \vdots & & \vdots \\ d_{i,1}(x) & d_{i,2}(x) & \cdots & d_{i,j}(x) & \cdots & d_{i,c}(x) \\ \vdots & \vdots & & \vdots & & \vdots \\ d_{L,1}(x) & d_{L,2}(x) & \cdots & d_{L,j}(x) & \cdots & d_{L,c}(x) \end{bmatrix} \tag{1}$$

最后，将这些结果经多特征融合后输出最终分类结果。

3 基于SVM-KNN的多特征融合分类方法

多特征融合方法能避免只采用单一分类器分类方法存在的片面性和分类精度不高的问题。其中，多特征融合方法中的概率和准则充分利用了样本被分类后的后验概率，是一种比投票法、计数法等融合方法更有效的特征融合方法。同时，由于SVM-KNN可以减轻对核函数参数选择的敏感程度，一定程度上比使用SVM具有更好的分类性能[7]。因此，本文提出了基于SVM和KNN的多特征融合分类方法。新算法在不增加传统支持向量机方法时间复杂度的基础上，降低了其分类超平面附近样本点的错分率，考虑到了各个子分类器之间的差异，是一种比只采用单一的SVM或SVM-KNN分类器分类精度更高的一种方法，且通过不同的融合方案能够分析出在分类过程中占主要因素的特征。

3.1 概率和准则

概率和准则[5]的思想是：首先假设对于测试样本 x，经 L 个基本分类器分类后，得到了一个决策轮廓矩阵，如式(1)。

然后，计算测试样本 x 属于第 j 类的置信度 $\mu_j(x)$：

$$\mu_j(x) = (1-L)p(j) + \sum_{i=1}^{L} d_{i,j}(x);\ j = 1,2,\cdots,c \tag{2}$$

其中，$p(j) = N_j/N$，$j = 1,2,\cdots,c$，N_j 为训练样本集中属于第 j 类的样本数量，N 为训练样本的总数量。

则融合后样本 x 的判别结果为：

$$j^*(x) = \arg\max_{j=1}^{c}[\mu_j(x)] \tag{3}$$

3.2 SVM-KNN多特征融合分类算法

设样本集中共有 L 组特征。首先，根据训练样本集中的 L 组特征数据用SVM方法分别得到 L 组支持向量集 T_{sv}^k 和决策超平面 $g_k(x)$：

$$g_k(x) = \sum_{i=1}^{m} \alpha_i^* y_i K(x_{i,k}, x) + b^*;\ k = 1,2,\cdots,L \tag{4}$$

其中，$x_{i,k}$ 表示第 i 个训练样本的第 k 组特征，$K(\cdot,\cdot)$ 为支持向量机核函数。

对于一个两分类问题，基于SVM-KNN的多特征融合分类算法步骤如下。

输入：测试样本集 T 和训练样本集。

输出：测试集中每个样本的类别 $j^*(x)$。

方法：

1）取 $x \in T$，计算样本 x 的决策轮廓矩阵 $(d_{i,j}(x))_{L\times 2}$。

1.1）设初始 $k = 1$。根据式(4)计算 $g_k(x_k)$，其中 x_k 表示 x 的第 k 组特征数据。

1.1.1）若 $|g_k(x_k)| > \varepsilon$，计算 x_k 分别被分为正类和负类的概率 $\{d_{k,1}(x), d_{k,2}(x)\}$[9]：

$$d_{k,1}(x) = \frac{1}{1+\exp(-g_k(x_k))}, d_{k,2}(x) = 1 - d_{k,1}(x) \tag{5}$$

否则，转1.1.2）。

1.1.2）若 $|g_k(x_k)| < \varepsilon$，用KNN分类算法重新分类，训练集为第 k 个支持向量机分类器的支持向量集 T_{sv}^k，则：

$$d_{k,1}(x) = q/p,\quad d_{k,2}(x) = 1 - d_{k,1}(x) \tag{6}$$

其中，p 为KNN算法中所取近邻的个数，q 为所取近邻中属于正类的代表点的个数[10]。在KNN算法中计算 x_k 与每个支持向量的距离时采用下式：

$$d(x_k,x_i) = K(x_k,x_k) - 2K(x_k,x_i) + K(x_i,x_i);\quad x_i \in T_{sv}^k \tag{7}$$

1.2）$k \leftarrow k+1$，若 $k = L$，转2）；否则，转1.1）。

2）得到样本的决策轮廓矩阵 $\begin{bmatrix} d_{1,1}(x) & d_{1,2}(x) \\ \vdots & \vdots \\ d_{i,1}(x) & d_{i,2}(x) \\ \vdots & \vdots \\ d_{L,1}(x) & d_{L,2}(x) \end{bmatrix}$，代入到式(2~3)中，此时 $c = 2$，得到测试点 x 经多特征融合后的最终分类结果 $j^*(x)$ 并输出。

3）$T \leftarrow T - \{x\}$，若 $T = \varnothing$，停止；否则，转1）。

3.3 算法时间复杂度分析

设 n 为训练样本的个数，L 为将样本特征分成的组数，d 为每组特征包含的特征数目。对于一个测试样本，由于KNN法的时间复杂度是 $O(dn^2)$[11]，计算后验概率时只需要将分类后的结果代入式(5~6)中，因此其时间复杂度是 $O(1)$，而判断 $g_k(x_k)$ 与 ε 的大小的时间复杂度是一常数，步骤1.1.1）和1.1.2）共循环了 L 次，所以步骤1.1.1）的时间复杂度是 $O(L)$，步骤1.1.2）的时间复杂度是 $O(Ldn^2 + L)$。所以，步骤1）计算样本决策轮廓矩阵的总时间复杂度是 $O(Ldn^2 + L)$。步骤2）中多特征融合的计算相当于对决策轮廓矩阵的每一列元素求和，再分别加上一个常数后求每列的最大值，然后取其下标，所以步2）的时间复杂度是 $O(2L)$。又由于标准支持向量机的时间复杂度是 $O(n^3)$[12]，所以基于SVM-KNN的多特征融合分类算法总的时间复杂度是：

$$O(Ln^3) + O(Ldn^2 + L) + O(2L) =$$

$$O(Ln^3 + Ldn^2 + L + 2L) = O(n^3) \quad (8)$$

其中,L 是不依赖于 n 的指定正常数,$d < n$。

由上面可以看出,本文提出的基于 SVM-KNN 的多特征融合分类算法并没有增加支持向量机的时间复杂度。

4　实验与讨论

为了验证算法的有效性,采用鸢尾属植物数据集的 Iris-versicolor 和 Iris-virginca 两类数据进行实验,这两类数据共包含 100 个样本,每个样本包含四个特征,分别为萼片长度、萼片宽度、花瓣长度和花瓣宽度。实验采用五折交叉验证的方法,将 SVM、SVM-KNN、基于 SVM 的多特征融合算法与基于 SVM-KNN 的多特征融合算法四种算法的分类准确率作对比。实验中支持向量机核函数采用高斯径向基核函数 $K(x,y) = \exp\{-|x-y|^2/2\tau^2\}$,惩罚参数 C 取 1000。实验结果如图 3 和 4 所示。

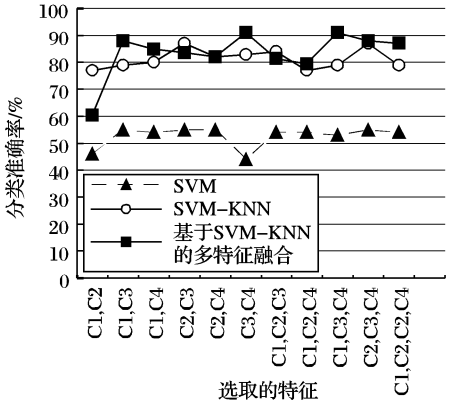

图 3　三种方法在不同特征选取方案下的分类准确率

图 3 中横轴表示各种不同的特征选取方案,共有 11 种选取方案,C1、C2、C3、C4 分别表示萼片长度、萼片宽度、花瓣长度、花瓣宽度四种特征。如方案 C1,C2 表示选取萼片长度,萼片宽度两个特征,分别用 SVM、SVM-KNN、基于 SVM-KNN 的多特征融合分类方法三种方法进行分类,纵轴表示平均分类准确率。通过图 3 可以看出:11 种特征选取方案中,基于 SVM-KNN 的多特征融合方法的分类正确率高于 SVM 的有 11 种,等于或高于 SVM-KNN 的有 8 种。并且 SVM-KNN 和 SVM 的所有融合方案的平均分类准确率分别只有 81.3% 和 52.6%,而基于 SVM-KNN 的多特征融合分类方法的平均准确率为 83.4%,分别比 SVM 和 SVM-KNN 方法提高了28.7% 和 1.9%。同时,从最后一个融合方案 C1,C2,C3,C4(选取样本的所有特征),可以看出只采用 SVM 或 SVM-KNN 方法对测试样本进行分类的平均准确率为 54% 和 79%,而基于 SVM-KNN 的多特征融合分类方法的分类准确率为 87%,比 SVM 和 SVM-KNN 准确率分别提高了 32% 和 8%。从以上分析可以看出,基于 SVM-KNN 的多特征融合分类方法与 SVM 或 SVM-KNN 分类方法相比有较高的准确率。

图 4 比较了基于 SVM 和 SVM-KNN 两种分类器的多特征融合方法的分类准确率,横轴表示参与融合的特征选取方案,纵轴表示分类准确率。从图中可以看出,基于 SVM-KNN 的多特征融合方法的 11 种融合方案的分类准确率均高于基于 SVM 的多特征融合方法,且基于 SVM 的多特征融合方法的平均准确率只有 70.8%,比基于 SVM-KNN 的多特征融合方法的平均准确率低 12.6%。这表明基于 SVM-KNN 的多特征融合方法是一种比基于 SVM 更有效的多特征融合方法。

与此同时,对比基于 SVM-KNN 的多特征融合分类方法中的 11 种融合方案,从 C1,C3 与 C1,C2,C3,C1,C4 与 C1,C2,C4,C1,C3,C4 与 C1,C2,C3,C4 的分类准确率可以看出各种特征包含的信息之间并非完全互补,而是存在着一定的冲突。同时,通过统计分类准确度在 85% 以上的融合方案中各种特征出现的次数,以及由 C3,C4 与 C1,C3,C4 的分类准确率相等,可以说明花瓣长度和花瓣宽度是鸢尾属的主要特征,其他特征是辅助特征。

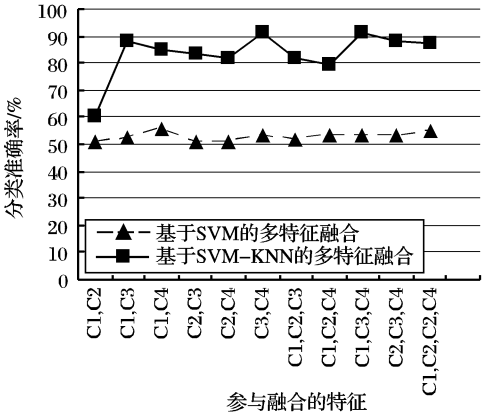

图 4　基于 SVM 和 SVM-KNN 的多特征融合分类方法准确率

5　结语

本文研究的基于 SVM-KNN 的多特征融合的方法,整合了多种特征信息,避免了单一分类器可能存在的片面性和在 SVM 分类中核参数选择问题,在不增加支持向量机算法时间复杂度的基础上,降低了 SVM 分类面附近样本点的错分率,提高了分类系统的总体性能,是一种有效的分类预测方法。同时也发现,参与融合的特征数据之间并非是完全互补的,而是存在着一定的冲突,如何选择合适的特征使它们间的信息互补达到最大化,是一个值得研究的问题。另外,如何将本文的算法从两分类推广到多分类,并将其运用到基因表达数据分类中是进一步要解决的问题。

参考文献:

[1] BATTITI R, COLLA M. Democracy in neural nets: Voting schemes for classification [J]. Neural Networks, 1994, 7(4): 691 – 707.

[2] HO T K, HULL J J, SRIHARI S N. Decision combination in multiple classifier systems[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(1): 66 – 75.

[3] CHO S B, JIN H K. Combining multiple neural networks by fuzzy integral for robust classification[J]. IEEE Transactions on Systems, Man and Cybernetics, 1995, 25(2): 380 – 384.

[4] CHO S B, JIN H K. Multiple network fusion using fuzzy logic[J]. IEEE Transactions on Neural Networks, 1995, 6(2): 497 – 501.

[5] KITTLER J, HATEF M, DUIN R P W, *et al*. On combining classifiers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 226 – 239.

[6] 邓乃扬, 田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京: 科学出版社, 2004: 45 – 76.

[7] LI R, YE S W, SHI Z Z. SVM-kNN classifier – A new method of improving the accuracy of SVM classifier[J]. Acta Electronica Sinica, 2002, 30(5): 745 – 748.

[8] 施建宇, 潘泉, 张绍武, 等. 基于多特征融合的蛋白质折叠子预测[J]. 北京生物医学工程, 2006, 25(5): 482 – 485.

[9] PLATT J C. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods [C]// Advances in Large Margin Classifiers. Cambridge, MA: MIT Press, 2000: 61 – 74.

[10] 娄震, 金忠, 杨静宇. 基于类条件置信变换的后验概率估计方法[J]. 计算机学报, 2005, 28(1): 19 – 24.

[11] RICHARD O D, PETER E H, DAVID G S. Pattern classification[M]. 李宏东, 姚天翔, 等译. 北京: 机械工业出版社, 2003: 151 – 158.

[12] 业宁, 王迪, 窦立君. 信息熵与支持向量的关系[J]. 广西师范大学学报: 自然科学版, 2006, 24(4): 127 – 130.