

使用粒子群算法进行特征选择及对支持向量机参数的优化

张俊才^{1,2}, 张 静^{1,2}

(1 许昌市耕新信息科学研究院, 河南 许昌 461000; 2 许昌职业技术学院 信息工程系, 河南 许昌 461000)

摘 要: 特征选取和参数设置是提升支持向量机分类器的效果的两个主要手段. 为了将两者结合起来, 实现同步优化, 以达到更好的分类效果, 设计了一种基于粒子群算法的分类器优化算法. 新算法对粒子采用 2 进制编码的, 设计适合的目标函数, 同步进行特征选择和支持向量机参数的优化. 经过对比验证, 新方法能够更加准确的得到待分类数据的特征子集跟支持向量机参数, 最终得到更优的处理结果.

关键词: 支持向量机; 参数优化; 粒子群算法; 2 进制编码

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2012)07-0138-04

Use PSO to Perform Feature Selection and Parameter Optimization of SVM

ZHANG Jun-cai^{1,2}, ZHANG Jing^{1,2}

(1 Xuchang Genxin Institute of Information Technology, Xuchang 461000, China;

2 Department of Information Engineering, Xuchang Vocational Technical College, Xuchang 461000, China)

Abstract: Feature selection as while as parameter optimization are two important way to improve SVM performance. In order to combine the two different way together to get a better classification effect, an new optimization algorithm based PSO is designed. The new algorithm uses Binary encode, and constructs proper target function to optimize the Feature selection and parameter optimization at the same time. Experiment s show that the new algorithm can quickly and exactly get a better suitable feature subset s and SVM parameters, and get a better classification effect.

Key words: SVM; parameter optimization; PSO; Binary encode

1 引言

数据分类是数据挖掘领域的主要问题之一, 其中主要涉及分类器模型的设计、分类空间的特征选取和分类器模型所用参数的取值等问题. 数据量越来越大, 分类空间的特征维数越来越高. 当待处理数据有上述特征时, 通过特征选择算法实现将降低维度, 把原来的分类空间转换为数据量较小的特征子空间, 减小问题时空复杂度^[1-3]. 此外, 适合的参数取值对提升分类器模型的分类精度及其重要, 常常会出现差之毫厘, 谬之千里的情况, 因此分类器参数优化也是数据挖掘领域很关键的研究问题. 一直以来, 特征选择与分类器参数优化这两个问题是独立考虑的, 但最近的研究表明, 把这两个方面融合起来

同步优化能够取得更好的分类效果. 文献[4]提到使用遗传优化算同步选择特征和支持向量机分类模型参数优化的算法, 以及文献[5]设计了基于应用遗传规划的方法, 均得到了较好的分类效果. 粒子群算法是著名优化算法, 其通过每个粒子记录自己得到的历史最优解以及粒子群全局的最优解综合考虑来进行优化^[6-7]. 最近几年, 粒子群优化在数据挖掘领域有了越来越多的应用^[8].

本文将 PSO 算法应用于解决分类问题, 即优化支持向量机的参数取值以及数据特征选择和支持向量机参数同步优化问题, 提出了一种使用 2 进制编码的 B-SVM-PSO. 经过对比试验, 与传统支持向量机相比, 新算法有较好的准确率, 与采用智能算法优化后的支持向量机相比, 新算法有较好的效率.

2 PSO 算法

粒子群优化算法是一种进化计算技术,由 Eberhart 和 Kennedy 根据对鸟群捕食观察,分析,得到灵感建立模型而创立的.算法在初始化阶段,在解空间中随机生成一组随机解,经过迭代搜索获取接近最优解的可接受解.粒子的搜索空间就是问题解的全集,其中所有粒子都代表空间中找到一个可能解,通过目标函数评价每个粒子所处于位置的优劣,粒子在解空间中移动的过程就是搜寻最优值的过程.所有的粒子都具有位置向量 $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$,代表目前在解集空间中所处于位置,根据粒子的位置向量值可由适应度函数计算出粒子适应度,以反映粒子所代表解的优劣程度.此外,每个粒子还有一个速度向量 $V_i = (v_{i1}, v_{i2}, \dots, v_{in})$,代表了粒子运动的方向和运动的快慢.在每次迭代过程中,粒子通过飞翔更新到新位置,并通过适应值函数计算出其在当前位置下的适应度,然后根据公式(1)更新其当前速度.

$$V_i = \omega V_i + c_1 \times \text{rand}() \times (Pbest_i - x_i) + c_2 \times \text{rand}() \times (Gbest_i - x_i) \quad (1)$$

接着按照公式(2)更新粒子位置 X_i .

$$X_i = X_i + V_i \quad (2)$$

其中, $Pbest_i$ 代表某个粒子 i 迭代个体最优值, $Gbest$ 代表整个粒子群搜索到的全局最优值. C_1 和 C_2 是两个参数,被称为学习因子. $\text{rand}()$ 代表 0 到 1 之间的随机数.另外, ω 是惯性因子, ω 值较大,全局寻优能力强,收敛较慢;反之局部寻优能力较强,收敛较快.

起初粒子群算法只被应用于连续问题的优化,为解决离散问题, Kennedy 和 Eberhart 随后有提出了 2 进制编码的粒子群算法.具体实现方式是将每个粒子采用 2 进制编码,使之成为一个 2 进制的向量.其中,粒子的速度表征了粒子所处位置取值为 1 的概率,则需要采用转换函数把速度转换到区间 $[0, 1]$. 本文使用 sigmoid 函数,则粒子更新公式如下:

$$p_{ij} = \begin{cases} 1, & \text{rand}() \leq S(v_{ij}) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

3 支持向量机

支持向量机(支持向量机)是一种重要的分类技术,具有良好的分类效果和较低的时间复杂度.将下支持向量机的原理介绍如下^[11]:

设有输入模式集 D , 其中样本维度为 n , 样本总数为 M . 需要将其划分 2 类, 分别标记为 1, 和 -1, 如式(4)所示:

$$D = \begin{cases} (x_i, y_i) | i \in \{1, 2, \dots, M\} \\ x_i \in R^n, y_i \in \{+1, -1\} \end{cases} \quad (4)$$

设存在超平面如式(5)所示:

$$(w \cdot x_i) - b = 0 \quad (5)$$

可将样本满足约定条件正确分类.

则可将此问题转换为式(6)所示.

$$\begin{cases} (w \cdot x_i) - b \geq 1 & \text{if } y_i = 1 \\ (w \cdot x_i) - b \leq -1 & \text{if } y_i = -1 \end{cases} \quad (6)$$

式(6)可改写为

$$y_i [(w \cdot x_i) - b] \geq 1 \quad i \in \{1, 2, \dots, M\} \quad (7)$$

由上式可得:满足式(4)的最优超平面解可表征为:

$$\text{Minimize } \Phi(w) = \|w\|^2 \quad (8)$$

同时该问题还可转化为如下形式,

$$f(x) = \text{sign}(\sum a_k y_k K(x_k, x) + b) \quad (9)$$

4 B-SVM-PSO 算法介绍

B-SVM-PSO 算法主要创新在于粒子设计以及目标函数的设计,着重介绍这两点以及算法流程.

4.1 粒子设计

粒子的代表采用 0 和 1 组成的 2 进制向量,粒子的构成需要根据所采用的核函数决定.此处使用 RBF 作为支持向量机的核函数, RBF 的数学表达式如下

$$K(x, y) = \exp(-\gamma \|x - y\|^2), \quad \gamma > 0 \quad (10)$$

对于使用 RBF 核函数的支持向量机,需要优化的参数有两个:核参数 γ 和惩罚参数 C ,除此之外,还需同时开始特征选取.所以,粒子必须含有三要素,即参数 C 、参数 γ 和特征掩码.图 1 为粒子的结构示意图.

P_C^l	P_C^i	P_C^r	P_γ^l	P_γ^i	P_γ^r	P_F^l	P_F^i	P_F^r
---------	---------	---------	--------------	--------------	--------------	---------	---------	---------

图 1 粒子的结构图

需要指出的是,粒子中,表征参数 C 和 γ 的编码需要根据式(11)变换成 10 进制的参数,以便在算法运行中使用.式(11)中 p 代表参数的 10 进制值, $\min_p$ 代表参数的最小值, $\max_p$ 代其最大值, l 代表参数的 2 进制位串长度, j 代表参其 10 进制的值.参数的准确度以及参数的值域均可依据需求而定.而针对特征掩码那部分,所有 2 进制位都映射特征

全集中某个特征位,其值等于 1,就代表对应的特征被选入特征子集,其值等于 0,就代表对应的特征不属于选出的特征子集的元素。

$$p = \min_p + (\max_p - \min_p) / (2^l - 1) \times d \quad (11)$$

4.2 目标函数设计

由前述可知,本文所研究的算法,其目的是同时优化特征子集和支持向量机的参数,从而获取更有分类准确率,与此同时,实现减少所处理数据的维度。即要是达到分类准确的高和得到的特征数目少的双重目的。本文的粒子适应值函数定义如下

$$\text{Fitness} = w_a \times \text{SVM} - \text{accuracy} + w_f \times (\sum f_i)^{-1} \quad (12)$$

其中, w_a 代表分类准确的所需的权值;支持向量机-accuracy 代表支持向量机的分类准确的; w_f 代表特征维度数量的权重; f_i 代表第 i 个特征维度在掩码中的相对偏移,1 代表特征被使用,0 代表相反; w_a 和 w_f 由实际需要决定。

4.3 算法流程

算法输入:待处理的样本集合 D ,学习因子 c_1 、 c_2 ,粒子群群的粒子个数 PN ,参数 C 、 γ 的值域 $(\min_c, \max_c)$ 、 $(\min_\gamma, \max_\gamma)$ 和 2 进制位串位数数量 n_c 和 n_γ ,特征 2 进制位串位数数量 n_f 、权值 w_a 、 w_f 、最多迭代轮数 N

输出结果:经过优化获取的特征子集、支持向量机的参数 C 、 γ 和对应的分类准确度

算法步骤:

- (1)初始化算法参数和整个粒子群的初始位置和速度;
- (2)把所有粒子的 2 进制表达转变成支持向量机的参数 C 、 γ 和对应的特征子集,再启动 LIBSVM 开始样本采集完成训练,在由式(12)运算得出粒子适应值;
- (3)依据粒子适应值,重新计算 P_i 及 P_g ;
- (4)由式(1),重新计算每个速度;
- (5)由据式(3),重新计算每个粒子的 2 进制位的值;
- (6)根据预先设计的最大迭代次数,评判是不是已经超出最大迭代次数 N ,若是,那就算法结束,并给出最后结果;若不是,那就跳转到步骤(2)继续运转。

5 实验结果

5.1 实验环境介绍

为了充分测试算法,减小误差,全面的对比 B-

SVM-PSO 效率,使用多个 UCI 样本集进行测试,样本集的相关信息如表 1 所示,均可从网站下载使用。

表 1 实验所用 UCI 样本集的相关信息

样本名称	类别数	样本数	特征数
Pima diabetes	2	768	8
finger disease	225	0	13
Flower	2	351	34
SuNOr	2	208	60
Iriss	3	150	4
channel	4	990	18

5.2 算法评判方法

把表 1 中所示的每个样本集划分成 n 个子集,其中 1 个为实验集,其余 $n-1$ 为样本集。对于每个样本集中的每个子集,分别作一次试验集,其他 $n-1$ 分别作一次样本集,即每个样本集上测试 n 次。计算 n 次结果的算术平均值作为最终结果。此外,作为对比试验,采用了普通向量机和文献[1]提出的基于遗传算法优化的支持向量机对相同的样本集进行对比试验。

定义了不同的准确率指标:正确率、错误率和命中率如下

$$\begin{cases} r_p = \text{num}_p / \text{Num}_p \\ r_n = \text{num}_n / \text{Num}_n \\ r_a = (\text{num}_p + \text{num}_n) / (\text{Num}_p + \text{Num}_n) \end{cases} \quad (13)$$

其中 r_p 为正确率; r_n 为错误率; r_a 为命中率; num_p 代表被 SVM 准确划分的正样本个数; num_n 代表被 SVM 准确划分的反样本数目; Num_p 代表正样本总量; Num_n 代表反例样本总量。针对分类目标数为 2 的样本集评判其分类效果,采用 $r_p * r_n$ 。针对多类别的样本集,那就采用命中率 r_a 描述其分类效果。

5.3 实验结果

表 2 是支持向量机算法、文献[1]提出的 GA-支持向量机算法和 B-SVM-PSO 在每个样本集上的测试结果。

表 3 是 GA-支持向量机方法和 B-SVM-PSO 方法在特征选择方面的比较,其中给出了两种算法在相同样本集上运行 10 次所选特征数目的平均值。第一列括号中的数字是该样本集原有的特征数目。

5.5 实验结果分析

表 2 展示的实验结果显示, B-SVM-PSO 的分类结果于普通支持向量机相比,总体上准确率有了显著的的提升。

表 2 三种不同算法在样本集上的实验结果对比

数据集	支持向量机			B-SVM-PSO			GA-支持向量机		
	准确率	错误率	命中率	准确率	错误率	命中率	准确率	错误率	命中率
Pima diabetes	0.731	0.744	0.739	0.743	0.792	0.774	0.783	0.870	0.815
finger disease	0.800	0.806	0.803	0.941	0.926	0.933	0.944	0.951	0.948
Flower	0.953	0.937	0.943	0.992	1.000	0.997	0.996	0.987	0.985
SuNOr	0.947	0.672	0.818	1.000	1.000	1.000	0.986	0.984	0.980
Iriss	NO	NO	0.966	NO	NO	0.980	NO	NO	1.000
channel	NO	NO	0.763	NO	NO	0.925	NO	NO	0.840

表 3 B-SVM-PSO 和 GA-支持向量机优化后特征维度个数

数据集	B-SVM-PSO	GA-支持向量机
Pima diabetes	1.4	3.7
finger disease	4.1	5.4
Flower	3.9	6.0
SuNOr	5.8	15.0
Iriss	1.0	1.0
channel	9.7	9.2

对类别数量较多的样本集 finger disease, 普通支持向量机方法的平均正确率为 0.754; 而 B-SVM-PSO 方法达到 0.936, 比前者提升了 21.2%。此外, 由 B-SVM-PSO 方法得到的分类结果相比结果优劣较为均衡, 即有着较好的稳定性。对 B-SVM-PSO 和遗传算法支持向量机的结果比较, 整体来说, 无论是算法准确程度抑或稳定性, 都比较相近。

从表 3 的结果展示了特征选择数, 总体来看, B-SVM-PSO 要少于对比算法, 相差也相对不小。由于上述两种算法的分类结果准确率相对较为接近, 综合来看, B-SVM-PSO 整体效果更优。

样本学习和处理待分类数据是分类问题的两个必要步骤。对样本学习阶段, 普通支持向量机方法速度必然是最迅速的。然而, 样本学习时阶段没有进行空间维度的选取, 势必会影响最终数据处理时的准确率。而对于同领域数据, 分类器可以一次学习, 多次使用, 因此, 样本学习阶段的效率相比就没有那么重要了。B-SVM-PSO 与之相比的劣势就可忽略了。

6 结束语

特征选取和参数设置是提升支持向量机分类器的效果的两个主要手段。为了将两者结合起来, 实现同步优化, 以达到更好的分类效果, 文中设计了一种基于粒子群算法的分类器优化算法。新算法对粒子采用 2 进制编码的, 设计适合的目标函数, 同步进行特征选择和支持向量机参数的优化。经过对比验证,

该方法能够更加准确的得到待分类数据的特征子集跟支持向量机参数, 得到更优的处理结果。

参考文献:

- [1] John G H, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem [C]//Proceedings of the Eleventh International Conference on Machine Learning. New Brunswick, NJ, USA: Morgan Kaufmann, 1994: 121-129.
- [2] Kohavi R, John G H. Wrappers for feature subset selection [J]. Artificial Intelligence, 2010, 97 (122): 273-324.
- [3] Huang C-L, Wang C-J. A GA-based feature selection and parameters optimization for support vector machines[J]. Expert Systems with Applications, 2011, 31 (2): 231-240.
- [4] Shi Y, Eberhart R. A modified particle swarm optimizer[C]//IEEE World Conf on Computational Intelligence. Piscataway: IEEE, 2005: 2675-2679.
- [5] Zhang L P, Yu H J, Hu S X. A new approach to improve particle swarm optimization[C]//Lecture Notes in Computer Science. Chicago: Springer Verlag, 2006: 1036-1043.
- [6] Kennedy J, Eberhart R C. Particle swarm optimization [C]//Proceedings of IEEE International Conference on Neural Networks. Piscataway, NJ: IEEE, 2010. 1942-1948.
- [7] 张雯, 杨春明. 改进的粒子群优化算法[J]. 微电子学与计算机, 2007: 627-630.
- [8] 任小康, 郝瑞芝, 孙正兴. 基于单纯形法的量子粒子群优化算法[J]. 微电子学与计算机, 2009: 986-989.

作者简介:

张俊才 男, (1968—), 硕士, 副教授。研究方向为机器学习、智能算法。

张 静 女, (1963—), 硕士, 副教授。研究方向为数据挖掘智能算法。