

QPSO 算法在朴素贝叶斯分类上的应用研究

李红彪, 张洪业

(东北电力大学 信息工程学院, 吉林省 吉林市 132012)

摘要: 朴素贝叶斯分类假定类条件独立,使得所选数据集的条件属性集在预处理时必须进行属性约简,如果处理不当,就会造成分类的不准确. 本文分别对在训练集上随机选取的属性子集组成粒子,构造适应度函数,从而构建了朴素贝叶斯分类器,并利用量子粒子群算法对分类效果进行择优操作. 实验证明,其分类效果优于传统的朴素贝叶斯分类方法.

关键词: 量子粒子群优化; 数据挖掘; 贝叶斯分类

中图分类号: TP311 **文献标志码:** A **文章编号:** 1007-2683(2010)04-0086-03

An Application Study of QPSO Algorithm on Naive Bayesian Classification

LI Hong-biao, ZHANG Hong-ye

(Institute of Information Engineering, Northeast Dianli University, Jilin 132012, China)

Abstract: The Bayesian classification approach assume that its attribute is independent, so the feature reduction of condition attribute set must be handled in data preprocess. In this process, data classification will not be accurate if it can not be handled properly. In this paper, random selected attributes of the training sets are used to form particles and a fitness function is set up at the same time, and therefore the naive Bayesian classifiers is constructed. At the same time, QPSO algorithm is used to optimize the effect of classifying. Experiment shows that this method has better performance than that of traditional naive Bayesian method.

Key words: quantum particle swarm optimization; data mining; Bayesian classification

0 引言

数据挖掘(Data Mining, DM)是从大量的、不完全的、有噪声的、模糊的、随机的数据中,通过设置一定的学习算法提取隐含的、事先未知的、潜在有用的信息处理过程. 分类作为数据挖掘的一个重要分支,其目的是提取描述重要数据类的模型及对未来数据的趋势作出预测或分类^[1-2].

朴素贝叶斯分类(Naive Bayesian Classification, NBC)是一种简单而高效的分类方法,其特点为速度快、准确率高. 但是由于在实际应用时,需要进行属

性约简,而这一过程可能会影响到分类效果. 同时,朴素贝叶斯分类模型所需估计的参数很少,对缺失数据不太敏感,算法也比较简单. 当属性相关性较小时, NBC 模型的性能最为良好.

本文将全局搜索的量子粒子群算法^[3]应用到朴素贝叶斯分类当中,最终取得了更佳分类效果.

1 PSO 算法^[4]

PSO(Particle Swarm Optimization)是在1995年由 Kennedy 和 Eberhart 提出的一种基于迭代的优化技术. 该算法与其他演化算法类似,也是基于群体的,根

据对环境的适应度将群体中的个体移动到好的区域,然而它不像其他演化算法那样对个体使用演化算子,而是将每个个体看做 D 维收缩空间中的一个没有体积的微粒,并在收索空间中以一定的速度飞行,并根据本身的飞行经验及同伴的飞行经验进行动态调整。^[6]但在实际应用中,必须调整好微粒自身的方向和速度,设置好相应的参数,合理地调整收索精确度,以实现粒子之间的信息共享与相互合作,否则微粒可能会飞过最优解或导致局部寻优而无法进行全局搜索.为改善传统 PSO 算法的性能,人们已经在原算法的基础上提出了各种各样的改进算法,但由于早已经有人证明该算法并不是一个全局收敛的算法,因此这种自身的缺陷是无法弥补的.

2 QPSO 算法^[6]

生物群体中个体的思维及运动非常复杂,有很大的不确定性,就像粒子有了量子行为.把量子理论应用于 PSO 算法就形成了一种改进的微粒群优化算法,即量子粒子群算法(Quantum Particle Swarm Optimization, QPSO).

QPSO 算法中的粒子按照下面三个公式更新位置信息,即

$$mbest = \frac{\sum_{i=1}^M P_i}{M} = \left(\frac{\sum_{i=1}^M P_{i1}}{M}, \frac{\sum_{i=1}^M P_{i2}}{M}, \dots, \frac{\sum_{i=1}^M P_{id}}{M} \right) \quad (1)$$

$$p = (\varphi_1 P_i + \varphi_2 p_g) / (\varphi_1 + \varphi_2) \quad (2)$$

$$X(t+1) = p \pm \beta \times |mbest - X(t)| \times \ln\left(\frac{1}{u}\right) \quad (3)$$

式中: $mbest$ 是粒子群中粒子的平均最好位置; M 是群体中粒子的数目; P_i 是粒子 i 的个体历史最优位置 $pbest$; p_g 是所有粒子中的全局最优位置 $gbest$; φ_1, φ_2, u 为均匀分布在区间 $[0, 1]$ 的随机数; β 为系数创造力,用于影响单个粒子的收敛速度和算法性能,是该算法中唯一的参数; t 为当前进化次数; $X(t)$ 为粒子的当前位置.

算法的基本步骤如下:

1) 初始化: 确定种群规模 M 和粒子维数 D , 初始化粒子群体、 $pbest$ 和 $gbest$;

2) 计算个体历史最优位置: 根据目标函数计算每一个粒子的适应度值,通过和个体历史最优值比较,如果当前值优于个体历史最优值,则把当前值替

换为个体最优值($pbest$),否则不替换;

3) 计算群体历史最优位置: 计算全部粒子的适应值,得到 $gbest$;

4) 计算 $mbest$;

5) 根据式(1)~(3),以一定概率取加或减,更新每个粒子的位置,从而生成新的粒子群体;

6) 判断粒子适应度是否满足收敛条件或者是否到达事先设定的最大进化代数,如果没有则返回到步骤 1) 继续运行,否则算法结束.

与传统粒子群算法相比,量子粒子群算法的最大优势在于它能够克服传统粒子群算法在收敛过程中经常收敛到一个局部最优点的缺陷.因此与传统的粒子群算法必须在一个有限的搜索范围内进行搜索求取最优值相比,其求得的最佳位置会具有更好的适应值,即会得出更优的结果;另外, QPSO 只用一个参数 β 决定粒子的收敛速度和位置,与传统 PSO 算法相比具有参数少、编程简单,易实现和收敛速度快等优点^[6].

3 基于量子粒子群算法的朴素贝叶斯分类

3.1 朴素贝叶斯分类^[4]

朴素贝叶斯分类的理论基础是贝叶斯定理和贝叶斯假设,使用概率规则来进行学习或实现某种推理过程,将学习或推理的结果表示为随机变量的概率分布.

朴素贝叶斯分类器将每个训练样本数据分解成一个 n 维特征向量 X 和决策类别变量 C ,并假定特征向量的各分量间相对于决策变量是相对独立的^[8-9].

设特征向量 $X = \{X_1, X_2, \dots, X_n\}$ 表示数据的 n 个属性($A_1, A_2, \dots, A_n$) 的具体取值,在这里表示数据集中各记录的对应属性.

假设决策类别变量 C 有 m 个不同的取值,即 $C = \{C_1, C_2, \dots, C_m\}$. 则有

$$P(X|C_k) = P(X_1, X_2, \dots, X_n|C_k), 1 \leq k \leq m \quad (4)$$

由贝叶斯定理可知:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)}, 1 \leq k \leq m \quad (5)$$

朴素贝叶斯分类将未知类别的变量 X 归属于类别 C_k , 当且仅当:

$$P(C_k|X) > P(C_j|X), 1 \leq j \leq m, j \neq k \quad (6)$$

3.2 编码方式

本文对粒子群中的粒子采用二进制编码方式,

即每个粒子由一组二进制位构成,其长度为数据库中随机属性的个数,每个二进制位分别与数据库中的某个属性相对应.若某个二进制位为1,则表示数据库中对应的属性参与构建朴素贝叶斯分类器.这样每个粒子事实上就对应着一个朴素贝叶斯分类器.

3.3 适应度函数

适应度函数用来评估个体的适应度,即区分群体中个体好坏的标准.衡量朴素贝叶斯分类器分类效果除了分类精确度要高之外,还应考虑分类误差在实例空间中的分布程度,即差异度^[4].本文适应度函数设为

$$F = R + \lambda D \quad (7)$$

式中: R 为朴素贝叶斯分类器在验证集上的分类精确度; D 为朴素贝叶斯分类器在验证集上的差异度; λ 为决定差异度影响的系数^[4].

3.4 分类过程

- 1) 采用分层随机取样方法将数据库分成训练集、验证集和测试集;
- 2) 需要随机生成 S 个随机属性子集参与朴素贝叶斯分类器的构造;
- 3) 利用随机生成的 S 个属性子集初始化粒子群中的各个粒子;
- 4) 采用上面介绍的量子粒子群算法进行择优操作;
- 5) 若得出全局最优解则分类过程结束;否则重复执行步骤3)和4)。

4 算法实验

本文实验所用的数据来自美国加州大学欧文分校(University of California, Irvine, UCI) 计算机科学与信息学院提供的机器学习数据库. 根据各数据集的特点,从中选择了 Mushroom, Tic - Tac - Toe Endgame, Breast - Cancer, MAGIC Gamma Telescope 及 Hill - Valley 5 个数据集,其中 Mushroom 数据集包含 2 480 个空属性值; Breast - Cancer 数据集包含 9 个空属性值. 实验首先对这些空属性进行数据补齐,然后对每个数据集,采用分层随机抽样,训练集数据占 50%,验证集和测试集各占 25%^[4]. 实验中, λ 取值为 0.5, 粒子群的种群规模为 100,进化的最大代数 100. 实验结果如表 1 和表 2 所示.

表 1 分类误差及对比

数据集	记录数	属性数	分类误差/%	
			朴素贝叶斯分类器	本文算法
Mushroom	8 124	22	91.76	92.53
Tic - Tac - Toe Endgame	958	9	66.74	68.28
Breast - Cancer	286	9	70.25	74.69
MAGIC Gamma Telescope	19 020	11	83.47	85.81
Hill - Valley	606	101	88.63	88.39
平均误差			80.32	81.94

表 2 训练时间及对比

数据集	记录数	属性数	训练时间/s	
			朴素贝叶斯分类器	本文算法
Mushroom	8 124	22	133.21	150.90
Tic - Tac - Toe Endgame	958	9	58.77	66.51
Breast - Cancer	286	9	15.25	20.73
MAGIC Gamma Telescope	19 020	11	816.35	1 005.87
Hill - Valley	606	101	33.16	46.98
平均训练时间			211.35	258.2

5 结 语

实验表明,与传统的朴素贝叶斯方法相比,在大多数情况下该方法具有更好的分类误差,本文所提出的方法在学习时间上要长于朴素贝叶斯分类方法,这是由于 QPSO 方法中粒子编码要浪费一定时间. 下一步的研究方向应是: 粒子群算法善于处理连续型变量. 因为描述各属性是否参与构造朴素贝叶斯分类器的变量是离散的 0 - 1 变量,因此本文中采用上下取整的方法,但该方法在一定程度上影响了分类效果. 因此有必要对量子粒子群算法进行改进,使之更好地适应各种情况下的分类. 另外 λ 的取值也会影响算法的分类能力,文中 λ 的取值是一个常量,如果在进化过程中根据实际情况适当地调整 λ 的取值,应该能够获得更好的分类效果.

参 考 文 献:

- [1] 龙海侠,须文波,孙 俊. 基于 QPSO 的数据聚类 [J]. 计算机应用研究,2006,23(12): 40 - 45.
- [2] HAN Jiawei, KAMBER M. 数据挖掘概念与技术 [M]. 范明,孟小峰,等译. 北京: 机械工业出版社,2005.
- [3] SUN J. A Global Search Strategy of Quantum-behaved Particle Swarm Optimization [C] // Proc. 2004 Congress on Cybernetics and Intelligent Systems, 2004.

(下转第 93 页)

穿场强增加. 可以通过图 1、2 解释, 在相同外施电场下, SF_6 的有效电离系数小于 N_2 , 因此, 前者比后者耐电强度高. $\text{SF}_6 - \text{N}_2$ 可按一定比例混合, 既能减少 SF_6 , 又能保证临界击穿场强 E_b .

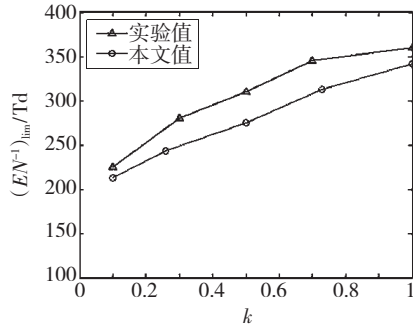


图6 混合气体临界击穿场强随 SF_6 分压比 k 的关系曲线

3 结 语

在 $EN^{-1} = 250 \sim 450 \text{ Td}$ 范围内对单一气体 SF_6 和 N_2 进行数值计算, 得出电子雪崩参数, 与实验结果比较, 有很好的一致性, 验证了蒙特卡罗法的可行性及碰撞截面的正确性. 采用蒙特卡罗法仿真了 SF_6 和 N_2 混合气体的电子雪崩运动, 仿真结果表明: 混合气体临界击穿场强 E_b 随着 SF_6 分压比的增加而提高. 随着计算机的发展, 计算精确度将会大大

增加. 文中得出混合气体的预放电参数可供实际工程应用参考.

参 考 文 献:

- [1] TAKUMA T. Gas Insulation and Greenhouse Effect [J]. J IEE Japan, 1999, 119: 232 - 235.
- [2] KÜÇÜKARPACI H N, LUCAS J. Simulation of Electron Swarm Parameters in Carbon Dioxide and Nitrogen for High E/N [J]. J Phy D: Appl Phys, 1979, 12: 2123 - 2137.
- [3] WU B T, XIAO D M. Predicting the Dielectric Strength of $\text{c-C}_4\text{F}_8$ and SF_6 Gas Mixtures by Monte Carlo Method [J]. Journal of Shanghai Jiaotong University, 2007, E - 12(1): 121 - 124.
- [4] DINCER M S, GOVINDA G R Raju. Monte Carlo Simulation of the Motion of Electrons in SF_6 in Uniform Electric Fields [J]. J. Appl. Phys, 1983, 54(11): 6311 - 6316.
- [5] DINCER M S, AYDIN T. Simulation of Limiting Field Behavior in Electron Swarms in $\text{SF}_6 - \text{N}_2$ Gas Mixtures [J]. IEEE Trans Elec Insul, 1994, 1(1): 139 - 145.
- [6] ITOH H, Miura Y, NAKAO Y, et al. Electron Swarm Development in SF_6 : I. Boltzmann Equation Analysis [J]. J. Phys. D.: Appl. Phys., 1988, 21: 922 - 930.
- [7] PHELPS A V, PITCFORD L C. Anisotropic Scattering of Electrons by N_2 and its Effect on Electron Transport [J]. Phys. Rev. A, 1985, 31(5): 2932 - 2949.
- [8] 裴鹿成, 张孝泽. 蒙特卡罗方法及其在粒子输运问题中得应用 [M]. 北京: 科学出版社, 1980: 500.

(编辑: 王 萍)

(上接第 88 页)

- [1] 张洪业, 曲朝阳, 王宇新. QPSO 算法在生产调度中的应用 [J]. 计算机工程与应用, 2009, 45(20): 208 - 210.
- [2] 谢小峰, 张文俊, 杨之康. 微粒群算法综述 [J]. 控制与决策, 2003, 18(2): 129 - 134.
- [3] SUM J, FENG B, XU W B. Particle Swarm Optimization with Particles Having Quantum Behavior [C] // Proceedings of the IEEE Congress on Evolutionary Computation. Portland, Oregon: IEEE Press, 2004: 325 - 331.
- [4] 张洪业, 金 刚, 王宇新. 微粒群算法在印染企业车间调度中的应用 [J]. 计算机工程与应用, 2009, 45(21): 245

- 248.

- [5] FIREDMAN N, GEIGER D. Bayesian Network Classifier [J]. Machine Learning, 1997, 29: 103 - 130.
- [6] 郑建军, 刘 炜, 刘琼听. 基于选择性的贝叶斯分类器集成方法 [J]. 北京理工大学学报, 2003, 23(6): 724 - 727.
- [7] 胡为成, 程转流, 王本年. 基于模拟退火遗传算法的贝叶斯分类 [J]. 计算机工程, 2007, 33(9): 219 - 221.

(编辑: 付长缨)