

基于杂交 BPSO-SVM 的网络故障特征选择

杨 婷, 孟相如, 徐 有, 温祥西

(空军工程大学 信息与导航学院, 陕西 西安 710077)

摘 要: 为提高网络故障诊断系统的诊断精度, 节约计算资源, 针对需要处理的含有大量无关或冗余特征的数据, 提出了一种基于杂交 BPSO-SVM 的网络故障特征选择算法. 该算法采用封装器模式, 以 SVM 的分类准确率和特征压缩比作为适应度函数来指导杂交 BPSO 进行特征选择, 将选择出的最优特征子集用于故障诊断. 运用 Kdd'99 数据集的实验结果表明, 杂交 BPSO-SVM 提高了诊断精度, 降低了特征维数, 可进一步提升网络故障诊断效果.

关键词: 网络故障诊断; 特征选择; 杂交二值粒子群; 支持向量机

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2014)01-0068-04

Network Fault Feature Selection Based on Breeding BPSO-SVM

YANG Ting, MENG Xiang-ru, XU You, WEN Xiang-xi

(School of Information and Navigation, Air Force Engineering University, Xi'an 710077, China)

Abstract: In allusion to deal with data which contains a lot of irrelevant and redundant features. A breeding binary particle swarm optimization - support vector machines (BPSO-SVM) algorithm was proposed for feature selection. In order to improve diagnosis accuracy and save computing resources of the network fault diagnosis system. The algorithm adopts wrapper mode, the classification accuracy of SVM and feature compression ratio as fitness function guide the breeding BPSO algorithm to search the feature space. Finally the best fitness subset was selected out. Experimental result on KDD'99 shows that the advanced algorithm improve the accuracy of diagnosis and reduce the feature dimension, and can further enhance network fault diagnosis effect.

Key words: network fault diagnosis; feature selection; breeding binary particle swarm optimization; support vector machines

1 引言

网络故障诊断本质上是一个模式识别过程, 实际获取的网络故障数据通常含有大量无关或冗余特征, 这些冗余不仅耗用大量的计算资源, 而且其中的冲突信息会干扰学习算法的正确决策, 影响最后的诊断精度. 因此对采集到的原始网络故障数据集进行特征选择可以删除无关、冗余的信息, 降低训练样本的维数和算法的复杂度, 从而提高模型的推广能力.

现有的特征选择算法根据是否依赖机器学习分为: 过滤器(Filter)模式和封装器(wrapper)模式. 其

中 Filter 模式利用数据本身的特性作为特征子集的评价指标, 其评价函数又可细分为距离测度、信息测度、相关性测度和一致性测度^[1]; wrapper 模式利用机器学习算法的正确率作为特征子集的评价指标, 选择效果一般较好. 文献[2]中提出了 BPSO-SVM 特征选择算法, 用粒子群(particle swarm optimization, PSO)算法进行特征寻优, 以 SVM 的分类结果作为适应度函数指导 PSO 进行寻优, 提高了分类器性能. 粒子群优化算法进行寻优, 简单容易实现, 对于复杂问题的求解在时间和求解精度上均有较好的效果^[3]. 但是 PSO 特征选择过程中的粒子仅仅通过当前搜索到的最优点进行搜索, 很大程度上是一

收稿日期: 2013-03-01; 修回日期: 2013-05-17

基金项目: 国家自然科学基金项目(61201209); 全军军事类研究生课题(2011JY002-524, 2012JY002-563)

种单项信息共享机制,全局搜索能力不高,易过早的陷入局部最优解。

针对上述方法的不足,本文将杂交 BPSO 算法用于特征寻优,通过杂交,使粒子之间互相共享信息,使整个种群比较均匀的向最优区域移动,提高全局搜索能力。

2 杂交 BPSO 算法

在 PSO 算法中,每个个体称为一个“粒子”,每个粒子位置代表一个潜在的解.所有粒子都有自己的位置、速度和一个由适应度函数决定的适应值来评价当前解的好坏.每个粒子根据自身记忆和群体知识搜索整个解空间,通过迭代方式寻优并收敛到最优解.迭代开始时,所有粒子的初始位置随机分布在解空间中,迭代过程中每个粒子根据自己发现过的最优位置 pbest 和整个群体已找到的最优位置 gbest 来更新当前飞行速度和位置。

速度和位置更新公式如下:

$$v_{i,j}(t+1) = \omega v_{i,j}(t) + c_1 r_1 [p_{i,j} - x_{i,j}(t)] + c_2 r_2 [p_{g,j} - x_{i,j}(t)] \quad (1)$$

$$x_{i,j}(t+1) = x_{i,j}(t) + v_{i,j}(t+1), j = 1, 2, \dots, d \quad (2)$$

式中, ω 为惯性权因子, c_1 和 c_2 为正的学习因子, r_1 和 r_2 为 0 到 1 之间均匀分布的随机数, $v_{i,j}(t)$ 和 $x_{i,j}(t)$ 为粒子当前的速度和位置, $v_{i,j}(t+1)$ 和 $x_{i,j}(t+1)$ 为粒子更新后的速度和位置, $p_{i,j}$ 为粒子的自身最好位置,即 p_{best} , $p_{g,j}$ 为粒子全局最好位置,即 g_{best} 。

为了解决离散组合优化问题, Kennedy 和 Eberhart 于 1997 年提出了一种二进制离散粒子群 (binary particle swarm optimization, BPSO) 优化算法^[4]。为了将粒子群算法离散化,算法由当前的状态变量决定粒子被判为 1 或 0 的概率,即由粒子速度决定一个范围在 $[0, 1]$ 之间的概率选择参数 s ; 若 s 接近于 1, 则粒子将更可能被选择为 1; 若 s 接近于 0, 则粒子更可能被选择为 0。使用 Sigmoid 函数求参数 s :

$$s = \text{Sigmoid}(v_i^k) = \frac{1}{1 + \exp(-v_i^k)} \quad (3)$$

式中, v_i^k 为粒子的第 k 次迭代时的速度。

修改后的离散粒子群优化算法与基本粒子群优化算法流程相类似,粒子速度和位置的更新公式为

$$v_i^{k+1} = \omega v_{i,d}^k + c_1 r_1 (x_{pbest}^k - x_i^k) + c_2 r_2 (x_{gbest}^k - x_i^k) \quad (4)$$

$$x_i^{k+1} = \begin{cases} 1, \rho < \text{Sigmoid}(v_i^{k+1}) \\ 0, \text{其他} \end{cases} \quad (5)$$

式中, ρ 是 $[0, 1]$ 之间的随机数, x_i^k 为粒子的第 k 次迭代时的位置, x_{gbest}^k 粒子的第 k 次迭代时的全局最优位置, x_{pbest}^k 粒子的第 k 次迭代时的自身最优位置。

采用线性变化的权重,让惯性权重从最大值 $\omega_{\max}$ 线性减小到最小值 $\omega_{\min}$ 。避免 PSO 算法的早熟以及在算法后期易在全局最优解附近产生震荡的现象, ω 随算法迭代次数的变化公式为

$$\omega = \omega_{\max} - \frac{t \times (\omega_{\max} - \omega_{\min})}{t_{\max}} \quad (6)$$

式中, $\omega_{\max}$ 、 $\omega_{\min}$ 分别表示 ω 的最大值和最小值, t 表示当前迭代步数, $t_{\max}$ 表示最大迭代步数, 通常取 $\omega_{\max} = 0.9$, $\omega_{\min} = 0.4$ 。

借鉴遗传算法中的杂交概念, Lovbjerg^[5] 提出了繁殖 (breeding) 粒子群算法, 粒子群中的粒子被赋予一个杂交概率, 这个杂交概率由用户定义, 与粒子的适应值无关。在每次迭代中, 根据杂交概率选取指定数量的粒子放入杂交池内, 池中的粒子随机两两杂交, 产生同样数目的子代粒子 (child), 并用子代粒子替换父代粒子 (parent)。子代位置和速度计算公式如下:

$$\begin{cases} \text{child}(x) = p \times \text{parent}_1(x) + (1-p) \times \text{parent}_2(x) \\ \text{child}(v) = \frac{\text{parent}_1(v) + \text{parent}_2(v)}{|\text{parent}_1(v) + \text{parent}_2(v)|} \times |\text{parent}_1(v)| \end{cases} \quad (7)$$

式中, $\text{parent}(x)$ 和 $\text{parent}(v)$ 代表所选父代的位置和速度, $\text{child}(x)$ 和 $\text{child}(v)$ 代表得到的子代的位置和速度, p 代表杂交概率, 选择 $[0, 1]$ 之间的随机数 (经验值约为 0.2), p_{best} 和 g_{best} 不变保持。

杂交 BPSO 算法执行的主要步骤如下:

Step1: 初始化, 确定种群规模 N 、杂交概率 p 和终止进化准则; 随机产生 N 个个体作为初始种群 $X(0)$, 进化代数计数器 $t \leftarrow 0$;

Step2: 个体评价. 计算 $X(t)$ 中各个个体的适应度, 得到 p_{best} 和 g_{best} ;

Step3: 更新, 用式 (4) 和 (5) 对粒子的位置和速度更新;

Step4: 根据杂交概率, 将粒子放入杂交池;

Step5: 将杂交池中的粒子进行杂交, 用式 (7) 进行位置和速度更新, 用子代粒子代替父代, 组成新一代种群 $X(t+1)$;

Step6: 如已满足终止准则, 则输出 $X(t+1)$ 中具有最大适应度的个体作为最优解, 终止计算; 否则

进入 Step2.

通过此方法产生的子代代替父代,选择父代没有基于适应值,防止了粒子群寻优过程中完全基于适应值易陷入局部最优的缺点.理论上讲,通过杂交增加了粒子间的交流,使粒子间更好的共享信息,从而使整个种群比较均匀的向最优区域逼近,两个在不同局部最优点的粒子经过杂交后,可以从局部最优逃离.

3 杂交 BPSO-SVM 特征选择算法

故障的特征组合空间为粒子的寻优空间,在杂交 BPSO 中,每个粒子用一个二进制的字符串表示,如第 i 个粒子的位置 x_i 可表示为 $x_{i1}, x_{i2}, \dots, x_{iD}$,其中 D 为粒子空间维数,由原始特征维数决定.空间中不同位置处的粒子,粒子中各位置分量取值对应特征组合中各特征分量的状态,特征选中状态为 1,丢弃状态则为 0.

特征选择的目的是使用少量的特征达到相同或更好的分类效果,适应度函数 F 指导粒子群搜索,因此适应度函数 F 的设计决定了粒子群最终的寻优结果.本文引入特征压缩比,即剔除的特征维数与原始特征维数之比.适应度函数 F 如式(8)所示.

$$F = \text{Accuracy} + \lambda \times \frac{D - \text{Feature_dim}}{D} \quad (8)$$

式中 Accuracy 为 SVM 的分类准确性. SVM 的分类结果受相关参数的影响,选择 3 折交叉验证进行 SVM 的参数寻优,Accuracy 为 3 折交叉验证的平均准确性. λ 为分类准确性和特征压缩比之间的折中参数,它的值越大,表明特征压缩情况越受重视,选择 λ 为 0.005.

图 1 中给出了该算法的过程框图.采集到的网络故障数据通过统一的数据预处理后,将故障数据分为两部分,一部分作为训练集,另一部分作为测试集.训练集送入杂交 BPSO 进行特征组合寻优,将寻优得到的特征组合送入 SVM 算法进行分类,用 SVM 的分类结果代入式(8)适应度函数动态指导杂交 BPSO 寻优,得到最优特征子集.最后将测试集经最优特征子集筛选后送入 SVM 分类器评估本文算法的有效性.

4 实验和分析

网络中的各种攻击事件和病毒会导致网络中出现大量的“软故障”,因此我们可以利用处于攻击下的网络状态模拟故障状态.本文采用美国国防部高

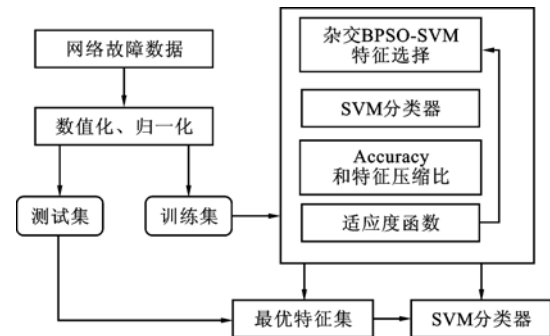


图 1 杂交 BPSO-SVM 网络故障特征选择过程框图

级研究计划局(DARPA)提供的 Kdd'99^[6]数据集.该数据集包括四类网络攻击,分别是 DOS、R2L、U2R 和 PROBE,将该数据集分为正常数据集和攻击状态下的“软故障”数据集.

SVM 网络故障诊断中碰到的数据都是高维异构数据,数据集中的特征有数值型、字符型、连续型、离散型.将数据集中的字符型特征数值化,按 kdd'99 数据集中对服务类型的说明和排序进行编码.取 1/-1 代表记录的类别标签正常/攻击.

对特征集进行归一化,避免大数值范围的属性支配那些相对较小数值范围的属性,同时也为避免计算中在数值上的困难,以提高故障诊断效率.通过式(9)将训练集和测试集中的特征数值 r 同时映射到 $[0, 1]$ 区间:

$$r_i = \frac{r_i - \min\{r_i\}}{\max\{r_i\} - \min\{r_i\}} \quad (9)$$

本文将杂交 BPSO-SVM 特征选择算法用于网络故障诊断.为了验证算法的分类性能,比较了不进行特征选择,使用 BPSO-SVM 和杂交 BPSO-SVM 算法进行特征选择对分类准确性及特征维数的影响.运用 C-SVM 分类器,核函数采用高斯径向基函数 $K(x, y) = \exp(-\beta \|x - y\|^2)$. 使用网格搜索 3 折交叉验证得到参数最优的分类器. BPSO 算法的参数设置如下:粒子数目 $N = 30$,问题维数 $D = 41$,学习因子 $c1 = 2, c2 = 2$,最大迭代次数 $M = 100$. 杂交概率为 0.5,杂交池的大小比例为 0.2,据 2.3 节的理论基础,采用了 1 000 个数据样本,其中 800 个训练集,200 个测试集.

图 2 中显示了 BPSO-SVM 算法和杂交 BPSO-SVM 算法的进化过程,从进化走势来看, BPSO-SVM 收敛速度比杂交 BPSO-SVM 快,但是陷入了局部最优.这是由于杂交 BPSO 在搜索过程中结合个体经验和群体智能,在选择父代时没有基于适应值,而是用杂交产生的子代代替父代,产生的子代需要重新寻找最优值,导致收敛速度变慢.同时,两个

在不同局部最优点的父代粒子经过杂交后,可以从局部最优逃离,寻找到全局最优解,避免了完全基于适应值的选择易陷入局部最优的缺点.

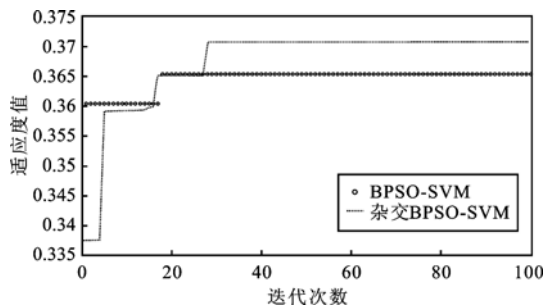


图 2 粒子群全局最优变化适应度曲线

表 1 中显示了三种算法的实验结果,可以看出,三种算法的分类准确性都比较高,这是因为 KDD'99 数据集相对其自身来说是一个比较完备的数据集,但是不影响其对本文算法有效性的验证. 基于 BPSO-SVM 算法有效降低了特征维数,降低了支持向量机在分类问题中的运算量,从而达到减少测试时间的目的. 杂交 BPSO-SVM 改进了 BPSO-SVM 基于适应度函数的寻优易陷入局部最优的问题,在 BPSO-SVM 特征选择的基础上进一步降低了特征维数,得到了更好的分类准确率.

表 1 实验结果

特征选择算法	维数	准确率/%
未经特征选择	41	94.4
BPSO-SVM	22	94
杂交 BPSO-SVM	16	95.2

5 结束语

针对网络故障诊断中的特征选择问题,本文提出了一种基于杂交 BPSO-SVM 网络故障特征选择算法. 该算法将样本特征选择归纳为组合优化问题进行研究,将整个特征空间看成一个解空间,用杂交 BPSO 算法在整个特征空间中寻找组合最优解,综

合考虑 SVM 分类器的分类准确性和特征维数压缩比作为适应度函数来评价杂交 BPSO 寻优结果,从而选择出最优分类特征子集. 将测试集特征选择后送入 SVM 分类器,验证了基于杂交 BPSO-SVM 特征选择算法的有效性. 实验结果表明,基于杂交 BPSO-SVM 的特征选择方法降低了特征维数,提高了分类器分类精度.

参考文献:

- [1] 陈友,程学旗,李洋,等. 基于特征选择的轻量级入侵检测系统[J]. 软件学报,2007,18(7):1639-1651.
- [2] 乔立岩,彭喜元,彭宇. 基于微粒群算法和支持向量机的特征子集选择方法[J]. 电子学报,2006,34(3):496-498.
- [3] 吕晓明,黄考利,连光耀. 基于 BPSO 的多故障最小候选集生成技术[J]. 系统工程与电子技术,2012,34(5):961-965.
- [4] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm[J]. IEEE Press,1997,97(1):4104-4108.
- [5] Lovjerg M, Rasmussen T K, Krink T. Hybrid particle swarm optimizer with breeding and subpopulation [C]//Proceedings of the Third Genetic and Evolutionary Computation Conference (GECCO-2001). San Francisco,2001:607-614.
- [6] 豆增发,高琳. 利用膜粒子群优化的条件随机域特征选择[J]. 西安电子科技大学学报,2012,39(5):138-146.

作者简介:

杨婷女,(1989—),硕士研究生. 研究方向为网络可生存性增强技术.

孟相如男,(1963—),教授,博士生导师. 研究方向为宽带通信网络技术.

徐有男,(1977—),硕士. 研究方向为宽带通信网络技术.

温祥西男,(1984—),博士研究生. 研究方向为人工智能与网络健康管理.