第35卷 第14期 计 算 机 工 程 2009年7月
Vol.35 No.14 Computer Engineering July 2009
·工程应用技术与实现· 文章编号：1000—3428(2009)14—0200—03 文献标识码：A 中图分类号：TP301.6

基于自适应遗传算法和SVM的特征选择

计智伟[1]，吴耿锋[1]，胡 珉[2]
(1. 上海大学计算机工程与科学学院，上海 200072；2. 上海大学悉尼工商学院，上海 200072)

摘 要：针对传统风险辨识方法无法实现盾构隧道施工过程中的风险状态实时识别的问题，提出一种自适应遗传算法和支持向量机结合的特征选择方法(AGASVM)，筛选出与施工质量风险关系最为密切的关键特征集。实验结果表明，用AGASVM所获得的关键特征集用于施工风险状态实时识别的分类准确率较高。其特征集规模比原始特征集有明显缩减，而且绝大部分关键特征与领域专家的意见是吻合的。
关键词：风险；特征选择；遗传算法；支持向量机

Feature Selection Based on Adaptive Genetic Algorithm and SVM

JI Zhi-wei[1], WU Geng-feng[1], HU Min[2]
(1. School of Computer Engineering & Science, Shanghai University, Shanghai 200072;
2. Sydney Institute of Language & Commerce, Shanghai University, Shanghai 200072)

【Abstract】 Aiming at the question that the traditional method for discerning risk can not come true the real-time recognition of risk statue in the shield tunneling constructing process, this paper proposes a feature selection method which combines Adaptive Genetic Algorithm with Support Vector Machine(AGASVM). It is used to filter a pivotal feature subset which is super correlative with risk of constructing quality. Experimental result shows that the pivotal feature subset selected by AGASVM can make the classification accuracy higher when it is used in the real-time recognition of risk statue. The dimension of pivotal feature subset is obviously smaller than the one of original factors set, and the most of pivotal features are the same as the ideas of domain experts.
【Key words】 risk; feature selection; genetic algorithm; Support Vector Machine(SVM)

1 概述

随着城市土地资源的日趋紧张，地铁隧道的建设已经成为解决城市交通问题的重要手段。由于地下工程的施工复杂性以及周边环境因素的影响，在工程的各个环节都存在一定的风险，对人员安全、经济以及环境构成较大的威胁。然而，我国在地铁建设方面的发展历史较短，经验不足，安全防范和风险管理相对落后，导致地下工程事故时有发生。尤其在大城市中，施工地点往往处于繁华地段，一旦发生危险，将会对周边建筑物造成严重影响，损失更是难以估量[1]。因此，研究有效的风险辨识和评估方法将有助于施工全过程的风险控制。

风险管理的方法直到20世纪70年代才由美国学者Einstein.H.H引入到隧道工程的风险控制中，随后一些著名学者提出了多个隧道工程风险辨识模型。目前应用较广的方法主要有故障树、层次分析法、蒙特卡罗模拟法等。上述方法都属于静态风险分析法，较多地依赖专家和工程人员的经验，主要用于设计阶段和工程初期的方案选择[2]。但是在实际施工过程中，风险是时时存在的，若能在施工过程中及时发现不安全的施工状态，并迅速采取有效的风险防范和控制措施，就可以将风险的影响降低到最小。然而目前的隧道施工实时监测数据维度很大，且具有很大的不规则性(数据含有噪声和缺失值，特征间量纲差异很大)，给实时风险识别带来了极大的困难。

为了达到对实时施工状态快速而准确的识别，有必要在诸多的风险因素中筛选出与施工质量关系最密切的关键特征集。为此，本文提出了一种采用自适应遗传算法(AGA)和支持向量机(SVM)结合的风险因素特征选择方法(Adaptive Genetic Algorithm with Support Vector Machine, AGASVM)，实验表明该方法是十分有效的。

2 盾构施工风险因素的特征选择

特征选择是模式识别中的一个关键技术，它的任务是从N个特征中按一定的选择标准选出一组由$n(n<N)$个特征组成的特征子集。其目的是从原始特征集中剔除冗余特征，选取出对分类有贡献的特征组合，从而使得分类效果更好。

要完成特征选择的任务需要解决2个问题：一是在有限的时间内快速找出最优特征集。二是特征选择的标准，通常以分类器的最大分类准确率作为特征选择评价准则。

盾构施工中的质量风险因素(原始特征)有200多个，涉及盾构机各部件状态、地面沉降、土质参数以及成形隧道变化等因素。因此，本文采用自适应遗传算法进行特征子集快速寻优，用SVM分类器的分类准确率来评价特征集的优劣，以便从高维的盾构风险特征集中选取关键特征集。

3 AGASVM方法

AGASVM方法属于Wappers(缠绕)法，即将SVM分类器嵌入到特征选择过程中，以分类器的最大分类准确率作为特征选择的评价指标。AGASVM的流程如图1所示。

基金项目：国家自然科学基金资助项目(50778109)；上海市重点学科基金资助项目(J50103)
作者简介：计智伟(1980－)，男，讲师、硕士研究生，主研方向：人工智能与模式识别；吴耿锋，教授、博士生导师；胡 珉，副教授、博士
收稿日期：2008-02-20 **E-mail：**jzw18@hotmail.com

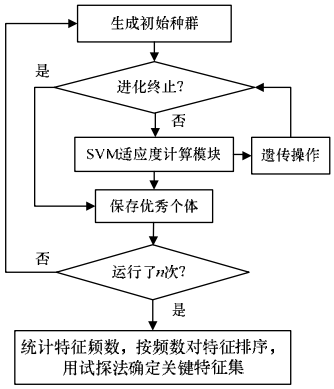

图 1　AGASVM 流程

从图 1 可见，在遗传算法的迭代过程中，每个个体的适应度计算均需要依赖 SVM 对当前特征子集分类性能的评价结果。

AGASVM 的具体步骤如下：

(1)染色体编码及初始种群的产生

设含有 N 个特征的原始特征集 $F=\{f_1,f_2,\cdots,f_N\}$，对每个特征子集进行二进制编码，得到长度为 N 的二进制码串：$X=x_1x_2\cdots x_N$。若 $x_i=1$，表示特征子集包含第 i 个特征 f_i，否则表示不包含。因此，$H=\{x_1x_2\cdots x_N \mid x_i\in(0,1),\ i\in[1,N]\}$ 可表示为所有特征子集的集合，也即个体空间。初始种群个体一般是随机生成。

(2)用 SVM 适应度计算模块计算个体适应度值

SVM 作为近年来提出的一种新型的机器学习方法，它建立在统计学习理论的基础上，具有较好的泛化能力[3]。支持向量机用于二分类时目标函数如下：

$$L=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_jK(x_i,x_j)-\frac{1}{2C}(\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}) \tag{1}$$

其中，C 用来控制对训练误差的评估；$\boldsymbol{\alpha}$ 是拉格朗日矢量乘子；$K(X_i, X_j)$是支持向量机用来解决非线性问题的核函数。选择径向基函数(RBF)作为核函数：

$$K(x,\ z)=\exp(-\|x-z\|^2/\sigma^2) \tag{2}$$

其中，x,z 是输入样本；σ 是半径。支持向量机的分类性能受惩罚参数 C 和核函数的影响。

SVM 适应度计算模块的核心是 SVM 分类机的训练和分类预测，其结构如图 2 所示。

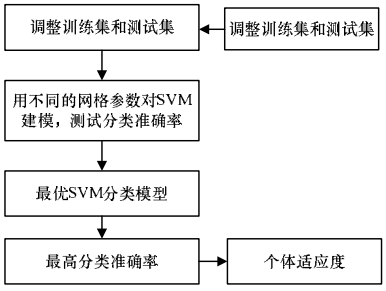

图 2　SVM 适应度计算模块

从图 2 可见，在计算每个个体的适应度之前，首先需要根据个体编码串调整训练集和测试集，然后用二维网格法对 SVM 进行多次建模和预测，直到 SVM 对当前测试集的分类准确率达到最大为止。得到最高分类准确率后，再通过下式计算适应度值。

遗传个体的适应度函数定义如下：

$$f(X)=f(x_1x_2\cdots x_N)=acc(X)-\lambda\frac{\sum_{i}^{N}x_i}{P} \tag{3}$$

其中，X 为个体；x_i 为第 i 个基因位；$acc(X)$为个体 X 在分类器上的最高分类准确率；$\lambda\in(0,1)$；$\sum_{i}^{N}x_i$ 表示特征子集所含特征个数(特征子集规模)；P 表示原始特征总数。

分类准确率 $acc(X)$的计算方法为

$$acc(X)=ano/tno \tag{4}$$

其中，ano 为被分类器正确分类的测试样本个数；tno 为测试样本总数。从式(3)可以看出，适应度较高的个体不但要具有较高分类准确率而且特征集规模应尽可能小。

由于 GA 种群中的不同个体代表不同的特征子集，因此在 SVM 计算特征集分类准确率之前，需要根据对应个体所表示的特征子集调整分类器的训练集和测试集。方法是：设当前个体为 $X=x_1x_2\cdots x_N$，对所有满足 $x_i=1$ 的基因位，其对应特征的全部观测值保存到新的训练样本和测试样本集中，删除 $x_i=0$ 对应特征的所有观测值。

另外，由于 SVM 的分类性能受到惩罚参数 C 和核函数参数 σ 的影响，因此，本文用二维网格法寻找最优参数对(C_i, σ_i)，使得分类器性能达到最佳[4]。二维网格寻优的具体步骤如下：

1)确定参数 C 和 σ 的取值范围。

2)构建参数对(C_i, σ_i)二维网格平面。例如 2 个参数各选取 10 个数值，则构成的网格平面有 100 个(C_i, σ_i)参数对。

3)将每对参数对输入 C-SVC 中，测试当前 SVM 分类器的分类准确率，取最小误差对应的节点值为最优参数对。

(3)遗传操作

遗传算法(GA)是模拟生物进化机制的一种自适应寻优算法，具有较强的并行模式空间搜索能力，可以较快地接近全局最优解，是解决大规模组合优化问题的常见方法。然而简单遗传算法自身固有的缺陷(如固定的交叉和变异概率、单点交叉操作等)，使得算法在寻优时稳定性不高，局部搜索能力不强，收敛速度较慢，且容易产生早熟现象[5]。因此，本文在对盾构施工风险特征集寻优的过程中采用了自适应遗传算法，使种群稳定地进化，避免陷入局部最优。

自适应的交叉概率公式如下：

$$p_c=\begin{cases}p_{c1}-\dfrac{(p_{c1}-p_{c2})(f'-f_{\text{avg}})}{f_{\max}-f_{\text{avg}}} & f'\ge f_{\text{avg}}\\ p_{c0}-(p_{c0}-p_{c1})\dfrac{f'}{f_{\text{avg}}} & f'<f_{\text{avg}}\end{cases} \tag{5}$$

自适应的变异概率公式如下：

$$p_m=\begin{cases}p_{m1}-\dfrac{(p_{m1}-p_{m2})(f_{\max}-f)}{f_{\max}-f_{\text{avg}}} & f\ge f_{\text{avg}}\\ p_{m0}-(p_{m0}-p_{m1})\dfrac{f}{f_{\text{avg}}} & f<f_{\text{avg}}\end{cases} \tag{6}$$

式(5)中 f' 为 2 个交叉个体中较大的适应度值 ;式(6)中 f 表示变异个体的适应度值。$f_{\max}$, f_{avg} 分别表示本代个体的最高适应度和平均适应度。

从式(5)和式(6)可以看出，改进的交叉和变异概率随种群的进化而不断地自适应调整。在进化初期，交叉概率和变异概率较大，促使优良个体不断地生长。到进化后期，交叉和变异逐渐减小，确保种群稳定，不使好的个体过早被破坏。

交叉操作采用两点交叉法代替传统的单点交叉，以便产

生更多的新模式。2 个交叉位置随机生成。

选择概率采用按比例的适应度分配方法。若个体 X_i 的适应度为 $f(X_i)$，则它被选中的概率为

$$p_i = f(X_i) / \sum_{i=1}^{N} f(X_i) \tag{7}$$

采用轮盘赌方式进行交叉个体的选择。每个个体按照自适应变异概率进行变异。为了保证每一代的优良个体不被破坏，采用精英选择策略，从父代个体中选择一小部分优良个体替代下一代中较差的个体。

当种群进化到一定代数或者种群中出现了满足要求的优秀个体后停止进化。

(4)根据特征排序，用试探法确定关键特征集

利用遗传算法本身的随机性，将遗传算法重复运行 n 次，每次按 10%~15%的比例选取末代优良个体 y_i 个。其中，i=1, 2,⋯,n。n 次后总共得到 $Y=\sum_{i=1}^{n} y_i$ 个优良个体。统计每个特征在这 Y 个较优特征子集中出现的频度，并按照频度值对特征进行降序排列，得到排序后的特征序列 $G=\{g_1,g_2,\cdots,g_N\}$。

最后，采用试探法得到关键特征集 R。方法是先假设关键特征集 R_0 为空，按照特征的排序进行试探。第 i 次操作是将特征 g_i 加入已测试特征集 $R_{i-1}=\{g_1,g_2,\cdots,g_{i-1}\}$，观察当前特征集 R_i 对于测试样本的分类能力，在达到较高准确率后即得到较优的特征集。

4 实验及结果分析

4.1 盾构隧道施工实时监测数据预处理

本文实验采用的原始数据为南京某地铁施工现场监测数据，先进行数据预处理：

(1)以环号为单位对施工参数进行合并。

(2)根据地面沉降每次对各个测点的沉降监测数据，计算出每个沉降监测时刻盾构机切口前方段、通过段、盾尾段 3 个区间内的土体表面最大沉降值以及坐标位置。

(3)对沿线地质资料、设计轴线描述、土质信息等数据以环号为单位进行插值处理。

(4)由于地面沉降数据的采集时间粒度最粗，因此选择地面沉降监测时刻对应的环号将以上各部分数据进行合并，得到了反映施工状况的实时状态向量。

(5)剔除值为 NULL 和常数的特征，得到原始特征共 85 个。其中，施工参数 47 个；盾构姿态参数 6 个；管片姿态参数 4 个；地面沉降参数 9 个；轴线信息及土质参数 19 个。

(6)将施工状态分为正常或不正常两类，然后根据地面沉降及管片偏差等工程验收指标对每个施工状态向量标定类别。

隧道在施工过程中对地面沉降的监测次数为 172 次。根据数据预处理方法，共获得 172 个实时施工状态向量，其中，正负两类样本个数分别为 87 和 85。本文实验选择 100 个样本作为训练集，30 个作为测试集，其中正负两类样本的比例均为 50%。

4.2 施工风险特征选择

本文实验将 SVM 以及 K-近邻算法(KNN)2 种分类器分别嵌入 AGA 中比较特征选择效果。前者即为 AGASVM，后者称为 AGAKNN。

采用 Matlab2007 编制遗传算法程序，种群大小为 40，进化 12 代，式(5)和式(6)中的参数 P_{c0}=0.83, P_{c1}=0.78, P_{c2}=0.6, P_{m0}=0.12, P_{m1}=0.1, P_{m2}=0.001。同时，使用 libsvm2.82 软件包作为 SVM 工具，支持向量分类机为 C-SVC，选用 RBF 为核函数，并用二维网格法寻优惩罚参数 C 和核半径 σ。网格法参数寻优的具体措施为：设惩罚参数 $C=10^{\beta}$，β 在[0,6]范围内变动，间隔为 0.5；设核半径 $\sigma=10^{\gamma}$，γ 在[-1,2]范围内变动，间隔 0.25。KNN 算法中对 K 的取值在[2,10]之间进行寻优，保证 KNN 的分类性能。

实验中将 AGASVM 和 AGAKNN 重复运行 10 次，在每次算法终止后，均从末代种群中选取非重复优良个体若干个，10 次共选取优良个体 70 个。表 1 显示了 AGA 结合 2 种分类器分别获取的较优特征子集的平均数。其中，AGASVM 和 AGAKNN 获取的特征子集平均规模接近，但 AGASVM 得到的特征子集分类准确率更高。

表 1 AGASVM 与 AGAKNN 寻优特征集比较

比较内容	AGASVM	AGAKNN
平均每个特征集中包含的特征数	34.085 7	33.953
特征子集平均分类准确率/(%)	88.19	84.61

下面通过实验比较在相同分类准确率下 AGASVM 和 AGAKNN 获得的关键特征集的规模。先对 AGASVM 运算得到的 70 个较优特征集中每个特征出现的频度进行统计，并按频度值从大到小对特征排序，频度排名前 50 的特征号依次为：6, 45, 58, 53, ⋯, 18, 27, 34 等。

根据前文所述的试探法，依次计算测试特征集不断加入新特征的分类准确率。最初，关键特征集 R_0 为空，将特征 6 加入 R_0 后得 $R_1=\{g_1\}=\{f_6\}$，R_1 分类准确率为 0.63。再加入特征 45，得到 $R_2=\{g_1, g_2\}=\{f_6, f_{45}\}$，分类准确率达到 0.76。依次类推，发现从特征 65 加入 R_{25} 后得到 R_{26} 开始，分类准确率达到 0.9，此后到 R_{39}，分类准确率一直保持 0.9。从 R_{40} 开始，分类准确率缓慢下降，如图 3 所示。

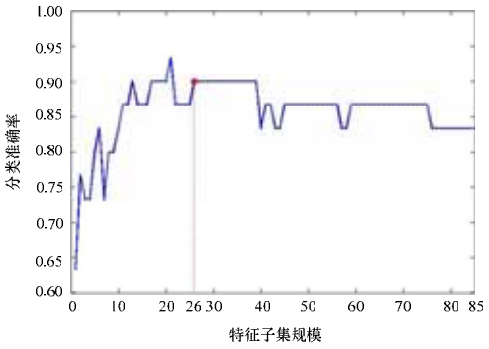

图 3 AGASVM 中特征集规模与分类准确率的关系

图 3 反映了 AGASVM 中特征子集规模与分类准确率的变化关系。曲线的变化呈现先上升后平缓的趋势。当特征子集规模小于 26 时，分类准确率值总体呈现上升趋势，曲线局部有所波动。在特征集规模大于 26 后，曲线变化开始趋于平稳，之后再加入剩余特征，分类准确率值开始缓慢下降。说明 R_{26} 较完整地包含了关键特征，因此，可选择频度排序的前 26 个特征构成关键特征集。

此外，对 AGAKNN 获取的 70 个较优特征集采用同样方法，取分类准确率阈值 0.9。发现从 R_{31} 到 R_{42}，特征集分类准确率维持在 0.9，之后再加入特征，分类准确率开始下降。因此，可认为 AGAKNN 所求关键特征集至少含有 31 个特征。实验结果概括如表 2 所示，即在相同分类准确率下 AGASVM 获得的关键特征集的规模明显比 AGAKNN 获得的少。

(下转第 226 页)