智能医疗系统中 GA_SVM 特征选择和参数优化①

徐旭东, 王 群, 孔令韬

(北京工业大学 计算机学院, 北京 100124)

摘 要: 挂号是医疗过程最基本的单元, 通常患者不知道自己病情, 挂错科室的情况十分普遍, 智能医疗系统的挂号功能很好地解决了这一难题, 智能医疗系统利用医疗部门积累的海量病案文本进行训练和机器学习, 对患者的病例特征进行分析将其分类到正确的病种, 得出应挂的科室然后推荐给患者. 而影响传统的支持向量机(SVM)文本分类的效率和准确率主要是特征值的提取和核函数参数的优化问题, 由此提出了一种遗传算法(GA)和 SVM 相结合的文本分类方法, 即把文本特征值和核函数的参数看作遗传算法中的一个染色体(一个个体), 并进行二进制编码, 对每一个个体进行选择、交叉、变异的遗传操作, 得到最优的个体, 最后通过支持向量机利用最优特征和最优参数进行文本分类. 实验表明, 该模型提高了患者智能诊断挂号的正确率, 是一种较好的智能推荐诊断挂号算法.

关键词: 智能医疗系统; 特征选择; 核函数参数; 遗传算法; 支持向量机

Ga-Svm Based Feature Selection and Parameters Optimization in Intelligent Medical Systems

XU Xu-Dong, WANG Qun, KONG Ling-Tao

(College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract: Medical registration is the most basic unit of the medical profession. Generally patients don't understand the condition of their illness. So choosing the wrong department is completely common. Intelligent medical system solves this problem very well. Intelligent medical system makes use of the massive medical record texts which the medical department accumulates to train and carry on machine learning, and to analyze the characteristics of the registration patient's medical record and to classify to the right disease. The patient is recommended to the appropriate department according to getting the department to be registered. The influence on the efficiency and accuracy of the traditional support vector machine (SVM) text classification of the intelligent medical system is the feature extraction and kernel function parameters optimization. Therefore, the method of the Genetic Algorithm (GA) combining with SVM is proposed. The text feature values and kernel function parameters together is viewed as a chromosome of the genetic algorithm that is an individual to carry on the binary encoding. The optimal individual is obtained by the genetic manipulation of the selection, crossover and mutation. Finally, text classification is operated by support vector machine using the optimal features and optimal parameters. The test result shows that this model improves the accuracy of intelligent diagnosis and is a good intelligent diagnosis registration algorithm.

Key words: intelligent medical systems; feature selection; kernel function parameters; genetic algorithm; support vector machine

患者挂号就医是日常生活中的普遍现象, 然而患者选择挂号的科室时常是根据患者或其家人进行判断选择的, 这样主观性大, 判断准确率低, 给患者和家人造成了极大的不便, 因此提高挂号准确率一直是医疗界的一个热门研究领域. 随着电子病案应用的日益普及[1], 医疗部门积累了海量的病案文本, 通过对病案

① 收稿时间:2014-07-08;收到修改稿时间:2014-08-25

文本特征的分析, 可以给患者提供新型的挂号方式, 也就是智能医疗系统, 让患者对自己病情有一个初步判断, 这样使得患者能够在家通过网络、电话等方式进行准确挂号, 给患者带来了很大便利, 有利于提高医疗水平和治疗效率.

智能医疗系统在对病案进行文本分类过程中文本的特征通常高达几万, 而且特征之间存在大量冗余和不相关信息, 选择有效的特征值是文本分类的关键问题之一, 目前国内外学者对文本特征选择进行了深入的研究, 并提出了许多有效的方法. 如序列选择算法、关联规则选择算法、遗传算法、粒子群优化算法等[2]. 同时, 在分类领域中有很多的文本分类方法被引入, 并取得了很好的效果. 例如, KNN 文本分类、朴素贝叶斯分类、神经网络和 SVM 分类, 其中 SVM 是一种建立在统计学习理论基础上的结构风险最小化原则的分类方法. 在解决非线性及高维模式识别问题中表现出许多特有的优势, 成为文本分类中较好的方法. 而设置正确的 SVM 核函数参数是影响分类精度的关键问题,目前优化 SVM 参数常用“留一法”, 这种方法计算量大并且很难获得最优的参数. 同时, 目前的研究只对特征值选择和SVM参数分别进行选择和优化, 而特征值和 SVM 参数是相互影响智能医疗系统的分类功能的. 由此本文提出了将遗传算法(GA)和 SVM 相结合的文本特征值和核函数参数同时进行优化的方法, 即把文本特征值和核函数的参数看作遗传算法中的一个染色体(一个个体), 并进行二进制编码, 将支持向量机分类准确率作为遗传算法的适应度函数, 对每一个个体进行选择、交叉、变异的遗传操作, 得到最优的个体, 最后通过支持向量机利用最优特征和最优参数进行文本分类. 因此, 提出 GA_SVM 模型并将其应用于智能医疗系统, 无疑具有一定的创新意义和广阔的应用前景.

1 基于支持向量机的电子病案文本分类原理

在此过程中将病案样本分成训练样本和测试样本. 对电子病例的预处理是从文本中提取关键词来表示文本的处理过程, 主要任务是进行中文分词、去停用词等, 将文本表示成分类器可以处理的形式, 本文中使用比较普遍的 SVM 模型对文本进行表示, 假设特征词集合为 $T=\{t_1,t_2,t_3,t_4,\ldots,t_n\}$, 文本集合为 $D=\{d_1,d_2,d_3,d_4,\ldots,d_m\}$, 文档 d_j 用一个向量表示为 $d_j=(w_{j1},w_{j2},w_{j3},\ldots,w_{jn})$, 每一维对应特征词集合中的一个特征项, 其值通过权值计算公式(1)算出, 权值计算一般是特征词在文本集中出现频率的函数. 本文中特征值的权重采用 TF/IDF 权值法计算, 其中 TF 表示特征词在某文本中的出现频率, IDF 表示特征词在整个文本集中的出现频率. 公式为:

$$w_{jk}=f_{jk}*log\left(N/n_k\right) \tag{1}$$

式中 f_{jk} 表示词 k 在文档 j 中出现的频率, N 表示文本总数, n_k 为出现特征项 k 的文本总数.

然后利用分类算法构造分类器, 将每个病案文本作为分类器的输入, 用于病案本文的分类算法是 Vapnik 等人提出支持向量机算法(SVM), 根据 VC 维理论和结构风险最小原理, 对有限的样本信息进行学习, 找到模型的复杂性和学习能力最佳折中点, 得到最好的泛化推广性[1]. 基于病种的多样性, 问题最终归结为病种多分类, 而多分类问题是一个求解约束条件下的凸二次规划问题:

$$\begin{aligned}&Min\,Q(\alpha)=\frac{1}{2}\sum_{ij}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{n}\alpha_i\\&St\ \sum_{i}^{n}\alpha_i y_i=0;0\le\alpha_i\le C;i=1,2,\ldots,n;j=i=1,2,\ldots,n\end{aligned} \tag{2}$$

求解上述问题后得到的最优分类函数如下:

$$f(x)=\mathrm{sgn}[\sum_{i=1}^{n}\alpha_i y_i K(x_i\cdot x)+b];i=1,2,\ldots,n \tag{3}$$

式中, $K(x_i,x_j)$—核函数, 其中常用的核函数有以下几种:

1) 线性核函数: $K(x,x')=(x\cdot x')$;

2) 径向基核函数:

$K(x,x')=\exp(-||x-x'||\hat{}2)$;

3) 多项式核函数:

$K(x,x')=[(x\cdot x')+c]d$, 其中 $c\ge0$

由于径向基核函数(Radial Basis Function,RBF)对非线性和高维数据有较好的分析能力, 而且它需要的参数少(仅需要 C 和 γ 这两个参数)所以本文采用 RBF 径向基核函数作为支持向量机的核函数. 当使用 RBF 作为核函数时, 则由(2)式转化成为以下的最优化问题:

$$\begin{aligned}&Min\,Q(\alpha)=\frac{1}{2}\sum_{ij}^{n}\alpha_i\alpha_j y_i y_j\exp(-\gamma\|\chi_i-\chi_j\|^2)-\sum_{i=1}^{n}\alpha_i\\&St\ \sum_{i}^{n}\alpha_i y_i=0;0\le\alpha_i\le C;i=1,2,\ldots,n;j=i=1,2,\ldots,n\end{aligned} \tag{4}$$

这样, 求式(4)的最小化问题就取决于参数(C, γ)

的设置, 最优的参数可以使得支持向量机有更好地分类效果.

2　基于遗传算法电子病案特征值选择和SVM参数优化

2.1 遗传算法特征值选择和参数优化具体流程

特征选择也叫特征子集选择(FSS , Feature Subset Selection), 是指从已有的 M 个特征(Feature)中按照一定的规则选择 N 个特征使得系统的特定指标最优化. 病案文本分类中, 电子病历具有庞大的特征维数, 存在着大量的冗余和不相关的特征值, 在不降低分类正确率的前提下, 对其进行有效的降维, 可使智能医疗系统的诊断结果有所提高. 本文中我们使用的是径向基核函数, 所以要提高挂号正确率必须进行特征选择和核函数参数优化, 工程的具体流程如图 2 所示.

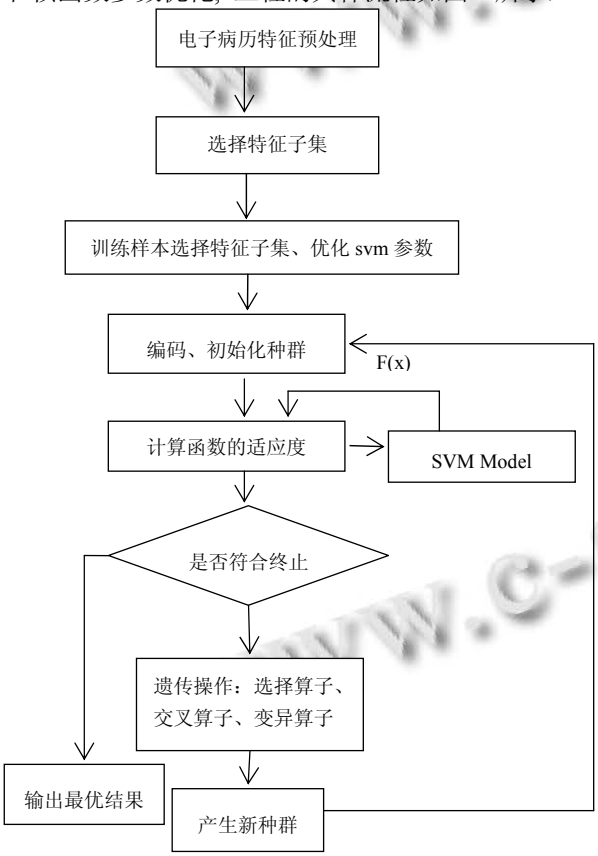

图2 GA-SVM特征值选择和核函数参数优化工作流程图

2.2 GA-SVM 详细设计

2.2.1 染色体编码

染色体编码是将对象抽象成由一定的符号按照一定的序列排列成串. 遗传算法具有隐含的并行性和全局强大的搜索能力, 能在短时间里搜索到全局最优解[5].

对于病案庞大的特征值, 利用遗传算法来进行特征值的选择, 需要保留有价值的病案特征值, 同时本文用的SVM核函数中的RBF也仅有C, 两个参数, 需利用遗传算法进行 SVM 分类模型的参数优化[6], 结合这两部分需求, 则染色体由两部分组成, 其中第一部分表示病案特征值, 第二部分是 SVM 核函数中 C 和 γ 两个参数, 如图 3 所示, 第一列是病案特征子集的二进制编码, 第二列 C 是 RBF 的第一个参数的二进制编码, 第三列是 RBF 的第二个参数的二进制编码. 在病案特征选择编码中, 1 代表该特征被选择, 0 代表该特征未被选择.

f				C				γ			
f_1	f_2	...	f_n	C_1	C_2	...	C_n	γ_1	γ_2	...	γ_n

图 3　GA_SVM 特征值选择和参数优化染色体的构成

在此为体现遗传算法选择病案特征值对挂号病种分类准确率的影响, 本文也进行了遗传算法仅对 SVM 核函数参数优化(GA_SVM1)的实验, 其挂号准确率结果用于与 GA_SVM 既病案特征值选择又 SVM 核函数参数优化(GA_SVM2)的挂号准确率结果比较, 由此 GA_SVM1 染色体的组成如图 4 所示.

C				γ			
C_1	C_2	...	C_n	γ_1	γ_2	...	γ_n

图 4　GA_SVM1 染色体的构成

2.2.2 适应度函数

适应度函数也称评价函数, 是根据目标函数确定的用于区分群体中个体好坏的标准,也是自然选择的唯一标准, 选择的好坏直接影响算法的优劣.

从定义上可以看得出, 适应度函数是影响分类结果的一个重要因素, 高的适应度值应该考虑到少的特征值和分类的准确率.

所以我们这里根据挂号患者病情分类到正确病种中的适应度函数定义为:$f(x)=f_1(x)-a\times f_2(x)$, $f_1(x)$函数代表病种分类的准确度, $f_2(x)$是病案特征值选择的数目. a 是一个调节的正参数,如果想让挂号病种分类的准确率高一些, 那么 a 值应该设置的大一些, 如果希望病案特征值选择的数目少一些,则 a 值设置的要小一些, 这里 $f_1(x)$挂号分类病种准确度越高, 适应度值

越大, $f_2(x)$病案特征值数目越少, 则适应度值越大. 这里 $a = 0.01$.

2.2.3 遗传算子

遗传算子是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型, 是一种通过模拟自然进化过程搜索最优解的方法. 其中对群体执行操作的遗传算子有选择、交叉、变异三种.

选择算子[7], 主要是从群体中选择出较适应环境的个体, 即适应度较高的个体. 本文选用轮盘赌法, 因为它既能保证高适应度对应的染色体有很高的概率被选中, 又能够让较低适应度对应的染色体有机会被选中进入下一代, 避免了算法陷入局部最优解.

交叉算子[8], 是在选中的用于繁殖下一代的个体中, 按交叉概率 Pc 对两两配对个体的相同的选中位置的基因进行交换, 目的在于产生新的基因组合, 也即产生新的个体. 在本文中交叉时, 在染色体的 3 段随机的选择 3 个交叉点, 然后与另一个染色体相应部分进行交叉. 本文采用单点交叉. 其中的交叉概率 Pc 计算如下公式(5).

变异算子[8], 根据生物遗传中基因变异的原理, 以变异概率 P_m 对某些个体的某些位执行变异. 在变异时, 对执行变异的串的对应位求反, 即把 1 变为 0, 把 0 变为 1. 其中的 P_m 计算如下公式(6).

$$P_c = \begin{cases} P_{c1} - \dfrac{(P_{c1} - P_{c2})(f' - f_{avg})}{f_{\max} - f_{avg}} & f \geq f_{avg} \\ P_{c1} & f < f_{avg} \end{cases} \tag{5}$$

$$P_m = \begin{cases} P_{m1} - \dfrac{(P_{m1} - P_{m2})(f - f_{avg})}{f_{\max} - f_{avg}} & f \geq f_{avg} \\ P_{m1} & f < f_{avg} \end{cases} \tag{6}$$

在(5)、(6)式中 f_{max} 是进化代中最大的适应度值, f_{avg} 是进化代中适应度值的平均值, f' 是两个交叉父母中的最大值的适应度值, f 是变异子代的适应值, P_{c1} 是当子代的平均值小于子代适应度值交叉概率, P_{c2} 是当子代的适应度值和平均值相等时的交叉概率, P_{m1} 是当子代的适应度值比平均值小时的变异概率, P_{m2} 是当子代的适应度值与平均值相等时的变异概率. 在本文中 P_{c1}=0.8,P_{c2}=0.5,P_{m1}=0.1,P_{m2}=0.01.

2.2.4 系统算法的基本步骤

文本分类采用具有全文检索功能和搜寻的开源程式库的 lucene 进行进行中文分词、去停用词, 并计算出每个特征值的权重值. 遗传算法采用 Matlab 的遗传算法工具箱, 初始种群数是 20, 进化代数 200. 支持向量机算法采用台湾林智仁教授的 Libsvm 工具箱.

Begin

Step1: Lucene 文本中文分词、去停用词, 计算病案文本的特征值权重;

Step2: T=0;

Step3: 初始化特征值种群 Initialpopulation(T);

Step4: 计算适应度函数的值 Fitness(T);

Step5: 若种群中适应度函数值已经达到足够大, 或者已经达到了终止代数, 则转到 Step8;

Step6: T=T+1;

Step7: 经过选择、交叉、变异遗传算子从 P(T)产生新的一代 P(T+1)代, 转到 Step3;

Step8: 将最优解进行反编码, 得到经过遗产算法优化的最优选择的特征值和 SVM 核函数参数.

Step9: 根据最优病案特征子集对病案训练样本处理, 并输入到最优参数 SVM 中进行建模, 并对测试样本进行检测, 根据检测结果对最优解性能进行分析.

End

而仅进行 GA_SVM 仅参数优化的实验步骤同上 Step1-Step9 所示.

3 系统验证及分析

本系统采用某医疗机构的数据, 选择了癫痫、脑栓塞和帕金森病这三种疾病各 200 例作为训练样本, 并从这三种病例中各抽取部分数据组成一个测试样本, 以测试系统新模型根据挂号患者特征分类到正确病种的准确率. 表 1 表示了该系统的部分病案训练样本.

表 1 模型训练样本(部分)

病例症状	意识丧失	头痛	意识不清	四肢抽搐	尿失禁
病例 1	0.13287117	0.16608897	0.076713204	0.10848885	0.076713204
病例 2	0.13287117	0.16608897	0.076713204	0.10848885	0.076713204
病例 3	0.13287117	0.16608897	0.076713204	0.10848885	0.076713204
病例 4	0.13287117	0.0	0.076713204	0.10848885	0.076713204
病例 5	0.0	0.19930676	0.0	0.16273327	0.11506981
病例 6	0.13287117	0.13287117	0.076713204	0.10848885	0.076713204
病例 7	0.16608897	0.0	0.095891505	0.13561106	0.095891505

为了使 GA-SVM 更有说服力和可比性, 这里选择 3 种对比模型, SVM(没有进行优化的 SVM), GA_SVM1(遗传算法仅对 SVM 核函数参数优化), GA-SVM2(遗传算法同时优化特征值和 SVM 核函数参数). 采用相同的病案测试样本对模型进行测试. 表 2 和表 3 分别给出了 SVM 模型分类和 GA_SVM1 分类的对比实验的结果; 以及 GA_SVM1 和 GA-SVM2 对比实验的结果.

表 2　SVM 模型分类和 GA-SVM 仅进行参数优化模型分类对比结果

参数|特征值个数	SVM	GA-SVM1
C	4.0	6.40888
γ	0.02	0.015831
特征值个数	204	204
挂号正确率(%)	62.0741	78.9474

表 3　GA-SVM1 和 GA-SVM2 的比较结果

参数|特征值个数	GA-SVM1	GA-SVM2
C	6.40888	4.3864
γ	0.015831	0.041294
特征值个数	204	103
挂号正确率(%)	78.9474	84.2105

图 5 是遗传算法选择特征值和优化参数时选择最优代过程的曲线, 图中可以看出在 20 代后曲线比较平稳; 从表 2 实验结果可以看出, 通过改进 SVM 参数 C 和 γ 两个参数, GA_SVM1 分类比 SVM 挂号准确率由 62.0741%提高到了 78.9474%, 说明了遗传算法对 SVM 核函数优化参数将挂号患者分类到正确病种是非常有效的. 从表 3 实验结果可以看出, 利用 GA_SVM2 分类, 对特征值选择的优化由最初的 204 个特征值缩减到 103 个特征值, 达到了去除冗余特征值的目的, 同时, 计算出了 SVM 分类的最优参数值, 从而使挂号准确率提高到了 84.2105%, 由此可以看出遗传算法进行特征值选择对分类的影响效果是明显的. 结合上述两个表的实验结果表明, GA_SVM2 与 SVM 模型挂号准确率相比以及与 GA_SVM1 模型挂号准确率结果相比, 取得了最好的挂号效果, 这有利于提高患者智能诊断挂号的正确率, 说明本文提出的 GA_SVM2 是可行的.

4　结论

智能医疗系统能为患者智能推荐挂号科室, 因此, 高效、正确地推荐挂号科室是此系统的最核心部分, 面临的主要问题就是特征值的选择和参数的优化, 为此, 本文提出了 GA-SVM 模型, 实验数据表明此模型比普通 SVM 算法在分类准确率上有较大的提升, 因而在医疗领域的应用具有广阔的前景.

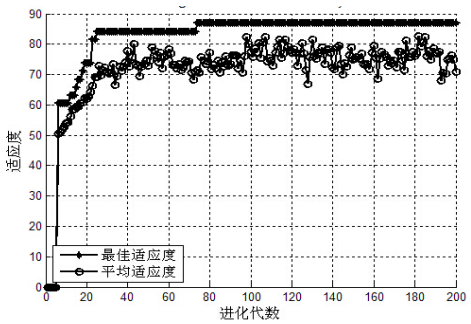

图 5　GA_SVM2 的选择过程图及实验结果显示

参考文献

1 杨孝光,李运明,张虎军,等.发达国家及地区电子病例发展现状与启示.西南军医,2013,5(3):261–282.
2 Heikkinen V, Korpela I, Tokola T, et al. An SVM classification of tree species radiometric signatures based on the Leica ADS40 sensor. IEEE Trans. Geosci Remote Sens, 2011, 49(11): 4539–4551.
3 Gao X, Yang S, Hu Y. Leakage forecasting for water supply network based on GA-SVM model. IEEE, 2010: 206–209.
4 Fu A, Sun G, Guo Z, et al. Forest cover classification with MODIS images in northeastern Asia. IEEE J Sel. Topics. Appl. Earth Observ. Remote Sens., 2010, 3(2): 178–189.
5 宋淑彩,庞慧,丁学钧.GA-SVM 算法在文本分类中的应用研究.计算机仿真,2011,28:222–225.
6 杜占龙,谭业双,甘彤.基于混沌遗传算法的 SVM 特征和参数优化.计算机工程,2012,38(5):163–166.
7 王俊年,刘云连,伍铁斌.改进的约束优化多目标遗传算法及工程应用.计算机工程与应用,2014,30:654–782.
8 王建玺,王刘涛.基于改进 GA 的 SVM 电力变压器过热诊断方法研究.计算机测量与控制,2014,22(2):456–651.
9 Wen Q, Zhang Z, Liu S, et al. Classification of grassland types by MODIS time-series images in Tibet, China. IEEE J Sel Topics Appl Earth Observ Remote Sens, 2010, 3(3): 404–409.
10 樊爱宛,时合生.基于特征选择和 SVM 参数同步优化的网络入侵检测.北京交通大学学报(自然科学版),2013,37(5): 58–61.