基于粒子群优化算法和相关性分析的特征子集选择*)

郭文忠[1,2] 陈国龙[2] 陈庆良[2] 余 轮[1]

(福州大学物理与信息工程学院 福州 350002)[1] (福州大学数学与计算机科学学院 福州 350002)[2]

摘 要 特征选择是模式识别与数据挖掘等领域的重要问题之一。针对此问题,提出了基于离散粒子群和相关性分析的特征子集选择算法,算法中采用过滤模式的特征选择方法,通过分析网络入侵数据中所有特征之间的相关性,利用离散粒子群算法在所有特征的空间里优化搜索,自动选择有效的特征子集以降低数据维度。1999 KDD Cup Data中IDS数据集的实验结果表明了提出算法的有效性。

关键词 数据挖掘,入侵检测,粒子群优化,相关性,特征子集选择

Feature Subset Selection Based on Particle Swarm Optimization Algorithm and Relevance Analysis

GUO Wen-Zhong[1,2] CHEN Guo-Long[2] CHEN Qing-Liang[2] YU Lun[1]

(Institute of Physics and Information Engineering, Fuzhou University, Fuzhou 350002)[1]

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350002)[2]

Abstract Feature selection is one of the important problems in the pattern recognition and data mining areas. The new feature subset selection method based on discrete binary version of particle swarm optimization (PSO) algorithm and relevance analysis is proposed. This new method employs the filter mode feature selection algorithm, which focuses on the correlation among the features of the network traffic data and employs the discrete particle swarm algorithm to find an optimized feature set. Experiments in 1999 KDD Cup Data confirm the effectiveness of the proposed strategy.

Keywords Data mining, Intrusion detection, Particle swarm optimization, Relevance, Feature subset selection

1 引言

Lee[1]把数据挖掘应用于入侵检测后,各种机器学习的方法也相继引入。但是,在网络入侵检测中,探测器收集到的数据量庞大且提取出来的特征繁多,其中有些特征是无关的,这些特征一方面可能降低分类或聚类的精度,另一方面会大大增加学习及训练的时间和空间复杂度,影响算法运行效率。因此从众多的特征(属性)中抽取有效的特征子集,不仅能改善学习算法的性能,还能提高算法训练速度。为此,设计一个适用的特征选择算法为入侵检测系统后续的检测工作提供一个精简、准确的特征描述子集是至关重要的。

2 过滤模式的特征选择

定义1 特征子集是将输入属性集合中不相关和冗余属性删除后得到的新属性集合。

定义2 特征选择(也称为属性选择)指的是识别和选取一个有效的属性子集,用于描述一个相对较大的通常含有冗余和不相关属性的数据集中的有效模式[2]。

特征子集选择过程中,算法一般根据以下原则选取有效的、最小规模的属性集合:(1)保证分类精度不能明显下降;(2)特征选择前后,类的分布尽可能一致。目前,根据是否依赖于机器学习算法,特征选择有两种模式:封装(Wrapper)模式和过滤(Filter)模式[3]。在封装模式中,选择方法直接优化某一特定的预测器(算法),这可以是通过评价预测器对每一步选择的特征子集的泛化能力来实现的(例如通过交叉验证的方法),封装方法把特征选择看作在所有可能的特征空间的搜索问题。在过滤模式中,选择方法独立于归纳学习算法,作为数据挖掘中预处理部分预先完成特征子集选择,然后进行归纳学习,过滤方法是根据数据的普遍性做出独立评估。一般说来,过滤法比封装法计算效率更高并且不依赖最终选择的分类算法,但由于使用的评估函数有可能给出不正确的偏爱导向,使得特征的选择偏离了最终目标。封装法将学习算法封装于特征选择算法中,最终选出的子集分类精度较高,但是当处理大数据量、多属性的数据时,其计算规模和资源占用十分庞大。

从优化的角度出发,特征选择问题实际上是一个组合优化问题[3,4],因此可以采用一些启发式搜索算法来选择次优特征子集,这样可以减少计算的工作量,在产生特征子集之后,用一个评价函数来计算子集的好坏,并且把结果和以前最好的比较经过若干次迭代之后,将最佳适应度的子集作为最终的选择结果。由于网络入侵数据量庞大,采用耗时的封装法显然是不合适的。为此,本文采用过滤形式的特征选择,利用收敛速度较快的粒子群算法在特征空间里搜索,同时引入相关性分析以指导算法的搜索。

3 基于PSO和相关性分析的特征选择

3.1 粒子群算法

粒子群算法最初是由Kennedy和Eberhart提出的[5],并

*)基金项目:国家自然科学基金项目(60673161);教育部科技重点项目(206073);福建省自然科学基金项目(A0610012)。**郭文忠** 讲师,博士生,研究方向为计算智能、计算机网络等;**陈国龙** 博士后,教授,CCF高级会员,研究方向为人工智能、网络信息安全等;**陈庆良** 硕士,研究方向为网络信息安全;**余 轮** 教授,博导,CSIG副理事长,研究方向为计算机网络与通信,计算机图像与图形等。

应用于连续空间的优化，在连续空间坐标系中，粒子群算法的数学描述如下：

假设在一个 D 维的目标搜索空间中，由 N 个粒子组成一个群体，其中每个粒子可表示为一个 D 维向量 $x_i=(x_{i1},x_{i2},\cdots,x_{id}\cdots,x_{iD})$，粒子 $i=(i=1,2,\cdots,N)$ 的速度定义为每次迭代中粒子移动的距离，用 $v_i=(v_{i1},v_{i2},\cdots,v_{id}\cdots,v_{iD})$ 表示。于是，粒子 $i(i=1,2,\cdots,N)$ 在第 $d(d=1,2,\cdots,N)$ 维子空间中的飞行速度 v_{id} 以及位置 x_{id} 可以根据下式进行调整：

$$v_{id}=w*v_{id}+c_1*r_1(p_{id}-x_{id})+c_2*r_2(p_{gd}-x_{id}) \quad (1)$$

$$x_{id}=v_{id}+v_{id} \quad (2)$$

其中 w 称为惯性权值，c_1，c_2 是两个正常数，称为加速因子，r_1 和 r_2 是两个 0～1 之间的随机数。通常使用一个常量 V_{max} 来限制粒子的速度，改善搜索结果。

1997 年 Kennedy 和 Eberhart 又提出了 PSO 算法的离散二进制版本（BPSO）[6]，使该算法进入组合优化领域。BPSO 采用二进制编码形式，将每一维的 x_i 和 p_i 限制为 0 或者 1，而速度 v_i 不做这种限制。用速度的 Sigmoid 函数表示位置状态改变的可能性：

$$S(v)=\frac{1}{1+e^{-v}} \quad (3)$$

BPSO 的速度更新公式没有改变，但是位置更新公式(2)改为：

$$\text{if}(rand()<S(v_{id}))\text{ then } x_{id}=1$$

$$\text{else } x_{id}=0 \quad (4)$$

式中的 $rand()$ 为[0，1]之间的随机数。在离散二进制模型中，仍保留 v_{max}，它起着限制 x_{id} 取 1 或 0 的最终概率的作用。实际上，v_{max} 通常被设在±4.0，以便总是至少有一个机会使 $S(v_{max})\approx 0.0180$，即处于即将改变状态的状况下。

利用粒子群算法解决优化问题的关键在于编码和适应度函数的选择。编码要根据问题的实质以及粒子所处的空间来确定，适应度函数体现了实际问题和优化算法之间的联系。

3.2 编码模式

特征选择的实质就是从 M 个特征中，选出 N 个特征构成子集。因此可以把每一个特征定义为粒子的一维离散二进制变量，M 个特征构成粒子的 M 维离散二进制空间。对于每一个粒子，如果第 i 位为 1，表示第 i 个特征被选中，反之表示该特征没被选中。因此，每一个粒子代表了一个不同的特征子集，也就是一个候选解。比如，粒子 $i=001010$，那么表明特征 3 和特征 5 被选中，特征子集为{3，5}。

3.3 适应度函数

在过滤型的特征选择中，适应度评价函数的选择至关重要。虽然人们提出了距离评测[7]、相关性评测[8]等几种不同的建议，但目前还没有能被一致接纳的度量标准。本文采用相关性评测方法[18]，其核心思想在于选择一个属性子集，属性各自与类属性有较大的关联但几乎没有内部关联，以此达到消除无关属性也消除重复属性的目的。两个属性 A 和 B 之间的关系可用对称不定性（symmetric uncertainty）来度量[9]：

$$U(A,B)=2*\frac{H(A)+H(B)-H(A,B)}{H(A)+H(B)} \quad (5)$$

$$H(X)=-\sum_{i=1}^{n}p(a_i)\log_2 p(a_i) \quad (6)$$

$$H(A,B)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(a_i,b_j)\log_2 p(a_i,b_j) \quad (7)$$

式中 $H(X)$ 是熵函数，以每个属性值的概率为基础，$H(A,B)$ 是 A 和 B 的联合熵，由 A 和 B 的所有组合值的联合概率计算出来。基于相关性的属性选择决定一个属性集的优良则用下式度量[9]：

$$\frac{\sum_j U(A_j,C)}{\sqrt{\sum_i\sum_j U(A_i,A_j)}} \quad (8)$$

其中，C 是类属性，i 和 j 包括属性集里的所有属性。公式(8)也就是粒子群算法的适应度函数，显然值越大，粒子的适应度越高。

3.4 算法描述

根据以上的分析，本文算法的主要步骤如下：

Step 1：装载训练数据，设置初始参数；

Step2：随机产生初始群体，为每个粒子生成随机的初始速度，设置粒子的个体极值 pbest 和群体的全局极值 gbest；

Step3：根据公式(8)评价每个粒子的适应值；

Step4：对每个粒子，将其适应值与其经历过的最好位置 pbest 进行比较，如果优于 pbest，则将其作为当前的最好位置 pbest；

Step5：对每个粒子，将其适应值与群体所经历过的最好位置 gbest 进行比较，如果优于 gbest，则将其作为群体最优位置，并重新设置 gbest 的索引号；

Step6：根据公式(1)，(3)，(4)更新粒子的速度和位置；

Step7：如果迭代次数达到最大，转 Step8，否则转 Step3；

Step8：把群体最优位置转化为对应的特征子集返回。

4 仿真实验

4.1 实验设置

为了验证本文特征选择算法法的有效性，实验数据采用 1999 KDD Cup Data 的 IDS 数据库[10]，该库与真实环境的相似性很大，因此被广泛用于测试入侵检测系统的效率。本实验从 10%的训练数据库里选取一部分的数据用于测试算法。为了保持原数据的分类信息，本文对于样本较少的类别全部选中，对于样本多的类别（主要是 neptune、smurf、normal 三大类）随机选取其中 10%的样本，最终产生 57280 个样本构成实验的样本全集。实验分为两组：第一组实验用 PSOSearch（离散粒子群搜索）以及 Weka 内置的 CfsSubsetEval 评估器，也就是公式(8)的评估方法来选取属性，然后分别把样本全集以及经过属性选择的样本子集用 Weka 提供的三种经典分类器 J48（C4.5 决策树分类器）、ID3 分类器、NaiveBayes（标准概率朴素贝叶斯分类器）进行分类，该组实验用来验证属性选择对后续分类算法的影响。第二组实验分别用 PSOSearch 以及 Weka 内置的 BestFirst（贪心式前向和反向搜索）和 GeneticSearch（遗传算法的搜索）在属性空间里进行搜索，然后把各自最终选择出来的属性子集对应的样本子集用 J48（C4.5 决策树分类器）进行分类，该组实验用来比较本文提出的属性选择算法和其他属性选择算法的运行效果。其中 BPSO 算法的参数设置为：惯性权值 w 采用 Shi[11] 提出的线性递减的惯性权值（$w_1=0.9$，$w_2=0.2$），$c_1=c_2=2.0$，$MaxIte=40$，$M=20$，$v_{max}=4.0$。遗传算法的迭代次数也设为 40，其他的参数设置均采用 Weka 的默认值。

4.2 实验结果及分析

表 1 是用 BPSO+CfsSubsetEval 选择出来的最终的属性子集，表 2 是使用分类器对样本全集进行分类的结果，表 3 是使用分类器对属性子集对应的样本子集进行分类的结果。从

表中可看出，使用属性选择后，分类器的构建时间明显减少，而且在不同的分类算法中运行效果都略微有所提高。由此表明，本文提出的特征选择算法可以改善分类算法的性能，而且能提高算法运行的速度，并且适用于多种分类算法。另外需要注意各个分类器得到的分类成功率都较高，是因为本实验中分类器的构建以及评测都是在同一个数据集上。

表1 BPSO+CfsSubsetEval选择的属性子集

属性个数(个)	属性子集
14	{duration, protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, count, srv_count, dst_host_count, dst_host_diff_srv_rate, dst_host_same_src_port rate, dst host srv diff host rate}

表2 样本全集的分类结果

分类算法	样本条数(条)	属性个数(个)	正确分类比率	分类器构建时间(秒)
J48	57280	41	99.8918%	85.69
ID3	57280	41	99.9983%	25.80
NaiveBayes	57280	41	91.6027%	12.34

表3 属性子集对应的样本子集的分类结果

分类算法	样本条数(条)	属性个数(个)	正确分类比率	分类器构建时间(秒)
J48	57280	14	99.9197%	24.72
ID3	57280	14	99.9983%	17.33
NaiveBayes	57280	14	91.9501%	2.94

表4是用各种不同的属性搜索方法利用CfsSubsetEval评估器选择出来的属性子集，表5是用J48分类器对表4里的各个属性子集对应的样本子集进行分类的效果。从表中可以看出，虽然BestFirst后向搜索选出的属性个数最少，但其正确分类比率却是最低；BPSO选出的属性个数较少，但正确分类比率是所有方法中最高的。由此可见，本文提出的BPSO搜索方法能够以较小的属性子集获取较高的分类成功率，是一种比较有效的特征选择算法。

表4 不同的属性搜索方法选择的属性子集

搜索方法	属性个数(个)	属性子集
BPSO	14	{duration, protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, srv_count, dst_host_count, dst_host_diff_srv_rate, dst_host_same_src_port_rate, count, dst_host_srv_diff_host_rate}
GA	19	{protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, num_shells, srv_count, same_srv_rate, diff_srv_rate, srv_diff_host_rate, dst_host_srv_count, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_rerror_rate, count, hot}
BestFirst 前向搜索	15	{protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, root_shell, count, srv_count, diff_srv_rate, dst_host_srv_count, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_rerror_rate}
BestFirst 后向搜索	13	{protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, root_shell, count, srv_count, dst_host_srv_count, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate}

表5 用J48对各个属性子集对应的样本子集的分类结果

搜索方法	样本条数(条)	属性个数(个)	正确分类比率	分类器构建时间(秒)
BPSO	57280	14	99.9197%	24.72
GA	57280	19	99.9162%	47.47
BestFirst 前向搜索	57280	15	99.9057%	25.31
BestFirst 后向搜索	57280	13	99.8045%	21.81

结束语 本文提出一个基于离散粒子群算法和相关性分析的特征选择方法，它能自动地从网络数据描述特征中选择与检测工作相关性较高的特征，为后续的检测工作提供一个精简、准确的特征子集。仿真实验表明了该方法能够提高分类算法分类的成功率以及算法运行的速度，并且适用于多种分类算法；同其它过滤型的特征选择方法相比，该方法能够以较少的特征个数获取较高的分类成功率，是一种比较有效的特征选择方法。

参考文献

1 Lee W. A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems [D]. New York, Columbia University, 1999

2 杨向荣. 人侵检测系统中数据挖掘和人工免疫原理的研究:[博士学位论文]. 西安：西安交通大学，2003

3 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. 电子学报，2006,34(3)：496～498

4 陈彬，洪家荣，王亚东. 最优特征子集选择问题. 计算机学报，1997,20(2)：133～138

5 Kennedy J, Eberhart R C. Particle swarm optimization. In: Proceedings of the IEEE Int. Conf. Neural Networks, 1995. 1942～1948

6 Eberhart R C, Kennedy J. A discrete binary version of the particle swarm algorithm. In: IEEE Conference on Systems, Man, and Cybernetics. Orlando, FL, IEEE Press, 1997. 4104～4109

7 Liu H, Setiono R. Feature Selection with Selective Sampling. In: Proc. 19th Int'l Conf. Machine Learning, 2002. 395～402

8 Ding C, Peng H C. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In: Proc. IEEE Computer Soc. Bioinformatics Conf. (CSB 03), IEEE CS Press, 2003. 523～528

9 董琳，邱泉，于晓峰，等(译). 数据挖掘实用机器学习技术(第2版). 北京：机械工业出版社，2006. 190～195

10 1999 KDD Cup competition. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 2006. 3

11 Shi Y H, Eberhart R C. A Modified Particle Swarm Optimizer. In: IEEE International Conference of Evolutionary Computation, Piscataway, NJ: IEEE, 1998. 69～73