

基于主动学习方法的网络流分类研究

顾名宇

(安顺职业技术学院, 贵州 安顺 561000)

摘 要: 提出了一种基于主动学习方法的网络流分类方法, 采用主动学习技术提取少量高质量的训练样本进行建模, 并提出了一种基于轮盘赌选择的样本筛选方法, 能够有效避免已有主动学习方法中的早熟收敛现象. 实验结果表明, 其相对于已有的流识别方法, 能够在仅依赖少量高质量训练样本的前提下, 保证较高的识别正确率, 更适用于现实网络环境.

关键词: 流分类; 主动学习; 轮盘赌选择; 支持向量机

中图分类号: TP309

文献标识码: A

文章编号: 1000-7180(2011)09-0173-03

Network Flow Classification Based on Active Learning

GU Ming-yu

(Anshun Vocational & Technical College, Anshun 561000, China)

Abstract: This paper introduces an active learning method to select the most qualified data for training and propose a roulette wheel selection method. The experimental results demonstrate that the proposed method is able to guarantee high identification accuracy by using a small quantity of high qualified data. Therefore, it is more suitable for the real network applications than the traditional ones.

Key words: traffic classification; active learning; roulette wheel selection; SVM

1 引言

流分类是增强网络可控性的基础技术之一. 该技术将流的识别问题转化为机器学习研究中相应的二分类或多分类问题处理, 如基于支持向量机 (Support Vector Machines, SVM) 算法的流分类方法^[1-2], 基于贝叶斯算法的流分类方法^[3]等. 目前这些方法在一定程度上取得了较好的分类效果, 但在很大程度上都依赖于训练样本集, 造成流分类器模型获取及维护困难. 同时, 由于互联网的动态特性, 为了避免分类器模型的失效, 经常采集流样本、更新分类器是必不可少的工作.

文中提出并实现了一个基于主动学习算法的网络流分类模型.

2 基于主动学习方法的流分类

主动学习方法的研究目标就是设计能够有效筛

选高质量待标注样本的学习算法, 将筛选出的样本交由督导者标注后构建高质量的训练样本集^[4].

在主动学习算法中, 查询函数建立在分类器模型之上, 因此查询函数是分类器相关的, 对于不同类型的分类器, 查询函数的具体设计方法也有所不同, 确定分类器类型是查询函数设计的首要条件. 在分类器的方面, 考虑到支持向量机是目前处理分类问题的最佳工具之一, 它以能够处理非线性、小样本、高维度和局部最小点等实际问题而被广泛使用^[5], 因此文中采用 SVM 分类器作为主动学习算法的分类器. SVM 建立在结构风险最小化原则基础上的. 对于训练样本为 $\{(x_i, y_i)\}_{i=1}^l \subset R^n \times \{1, -1\}$ 的二分类问题, 根据统计学理论, 可建立如下标准的线性 SVM 模型 A . 其中 $C > 0$ 为正则化参数, $\xi_i (i = 1, 2, \dots, l)$ 为松弛变量, $\omega \in R^n$ 为分类超平面的法向量, $b \in R$ 为阈值.

$$\begin{cases} \min \frac{1}{2}(\omega^T \omega) + C \sum_{i=1}^l \xi_i \\ \text{s. t. } y_i((\omega^T x_i) + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, 2, \dots, l \end{cases} \quad (1)$$

为了适度增加样本选择的多样性,文中采用轮盘赌选择方法改善查询函数的设计.根据轮盘赌选择的原理,质量较高的样本应该分配较大的轮盘块,以提供较高的选中几率.按照样本到分类超平的距离 d 作为不确定性的评估值从大到小对样本进行排名,根据样本排名 j 值设计轮盘块 v_j ($j > 1$) 为 $(a^{j-1}, a^j]$, v_j ($j = 1$) 为 $[0, a]$, 其中 a 是一个大于 1 的常数.将 $V(v_1, v_2, \dots, v_m)$ 归一化后构造轮盘,可以计算轮盘块 v_i ($i > 1$) 区间为 $(\frac{a^{j-1}-1}{a^m-1}, \frac{a^j-1}{a^m-1}]$, v_1 的区间为 $[0, \frac{a-1}{a^m-1}]$. 利用随机函数在 $[0, 1]$ 产生轮盘效应,如果随机函数产生的数据位于轮盘块 v_i 内,则查询函数选择样本 i . a 用于调节选择的多样性,当 a 接近 1 时,算法查询函数近似于随机选择,当 a 取一个极大值时,算法查询函数近似于 SIMPLE 算法,文中依据多次实验设定 a 的经验值为 2.

根据上面构造的查询函数,给出了针对该算法的主动学习方法伪代码.假定有一个少量样本已标注好的训练集 l 和一个未标注的样本池 U ,当算法进入循环阶段,在每一轮中,计算样本池 U 中的每一个样本的评估值 d ,并按照评估值 d 对样本进行排名,然后根据样本排名计算其所在的轮盘块,产生范围 $[0, 1]$ 的随机数,选取随机数落在轮盘块 v_i 的样本 i ,请求标注后并将样本 i 加入以标记好的训练集 l 中再进行训练,直到满足循环终止条件.使用均方预测误差率作为算法循环的终止条件指标,针对训练集 D 在测试样本 x 下的预测函数 f^* ,均方预测误差率的计算方法如下:

$$f_{\text{error}} = \frac{\text{MSE}(f^*(X, D))}{f(X, D)} \quad (2)$$

其中, MSE (Mean Squared Prediction Error) 为计算分类错误率的均方预测误差, f 为使用 SVM 分类算法获取的预测分类错误率.设定经验阈值 ϵ ,当均方预测误差率 f_{error} 小于阈值 ϵ 时,样本选择循环,最后得到的训练集 l 即为选择与精简后的训练.

基于轮盘赌选择的主动学习算法如下:

算法参数: l 为少量以标注样本, U 为大量未标注样本池, x 为测试样本集, ϵ 为算法终止阈值.

将训练集初始化为 l

repeat

{

for each sample j in U do

计算 U 中样本 j 的 d 值;

根据 d 值从大到小对样本进行排序

for each sample j in U do

根据排名计算样本轮盘块 v_j

产生范围 $[0, 1]$ 的随机数,根据随机数所属的轮盘块 v_i 选择样本 i ;

将样本 i 加入训练集 l 并从样本池 U 中删除该样本;

根据测试样本集 x 计算 f_{error}

}

until ($f_{\text{error}} < \epsilon$)

3 实验设计和分析

3.1 数据集及预处理

实验中使用了 Moore 等人专为流分类实验而制作的数据集^[6].在这个数据集中每一条作为一个流样本,样本中包括许多特征,如连接的持续时间、包间隔时间、传输的包数量等.每一个流有 248 个特征.除了 248 个特征之外,每个样本还有一个流的应用类型标注,如 WWW、P2P、MAIL、BULK 等.

数据集共有 10 组数据,包含大量样本.为了方便处理,从中抽取 1676 条 P2P 类别样本以及 6006 条其它类别样本构成实验数据.同时,将其它类别样本归作一类,标注为类别 OTHER,转换为二分类问题进行实验.

由于实验数据中的样本特征维度过高,存在大量冗余无关数据,为了提高分类器性能,减少冗余无关特征的影响,实验首先使用 FCBF (Fast Correlation-based Filter) 算法对独立于实验数据的训练数据进行特征选择,降低数据维度在特征选择后,我们通过多次试验优化 SVM 分类器参数,优化后的 SVM 分类器的参数为:类型为 C-SVM,核函数为 RBF (Radial Basic Function), γ 取值 0, C 取值 2048.参数优化后 5 折交叉验证获得的分类准确率达到 97.4%, P2P 流的检测率为 96.6%,误报率为 1.8%.

3.2 采用主动学习方法后的实验结果

为了验证基于所提出主动学习方法的有效性,在选择相同数量样本的情况下,关注了采用主动学习方法与随机采样方法选择样本后分类效果的差异.在实验中,为了确保实验结果的可信性,实验使用了 5 折交叉验证的方法.在每组实验中都从训练

样本中抽出1个P2P样本与1个OTHER样本构成初始训练集1,将剩余的训练样本作为样本池U,然后筛选200个样本进行对比。

图1与图2显示了实验中基于轮盘赌选择的主动学习算法分别与SIMPLE^[7]和SELF-CONF^[8]两种算法在分类准确率的均值与标准差的比较。可以看出:随着选择样本数量的增加,基于轮盘赌选择的主动学习算法分类准确率稳步提升,标准差也逐渐缩小,在选择160个样本后,平均分类准确率稳定在97.4%附近,而SIMPLE与SELF-CONF算法的平均分类正确率提升相对缓慢,一直未能超过94%,标准差也未出现显著缩小的迹象。这表明,文中算法具有更好的稳定性及平均分类准确率。

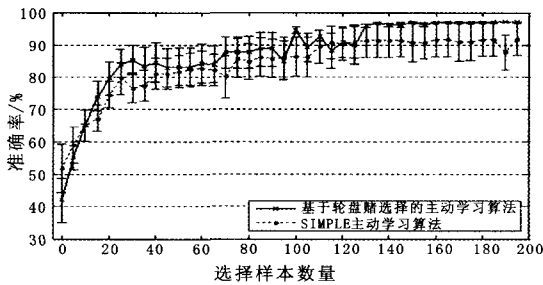


图1 SIMPLE算法与基于轮盘赌选择的主动学习算法实验对比

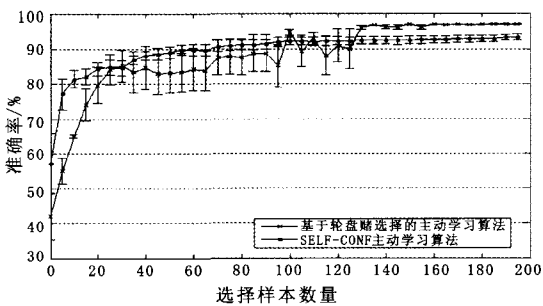


图2 SELF-CONF算法与基于轮盘赌选择的主动学习算法实验对比

图3显示了三种算法在最坏情况下的实验对比。可以看出,SIMPLE算法在最坏的一次评估中出现了明显的早熟收敛现象,分类准确率达到70%附近后停滞不前;SELF-CONF算法的早熟收敛现象出现相对较晚,分类准确率达到90%附近后停滞不前;文中算法由于样本选择具有了多样性,最后获得的分类准确率高达96.87%,与最后97.14%的平均准确率非常接近。

4 结束语

训练数据集是建立流分类模型的基础,但数据集中样本标注工作量大且必须依赖领域专家,这造

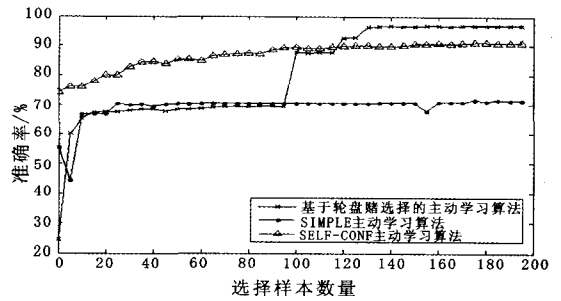


图3 最坏情况下三种算法的实验对比

成了流分类模型建立与维护代价巨大。针对这一问题,引入主动学习算法以减少标注工作量,提高方法可用性。实验表明,文中算法可以有效减少建立流分类模型需要标注的数据量,并且能够避免早熟收敛现象,在稳定性上优于同类方法,更适用于降低流分类器模型建立与维护的代价。

参考文献:

- [1] Yang A M, Jiang S Y, Deng H. A P2P network traffic classification method using SVM[C]// Proceedings of the 9th International Conference for Young Computer Scientists. Hunan: IEEE Computer Society, 2008:398-403.
- [2] 郭亚琴,王正群.一种改进的支持向量机BS-SVM[J].微电子学与计算机,2010(6):52-54.
- [3] Moore A W, Zuev D. Internet traffic classification using bayesian analysis techniques [C]// Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems. Banff, Alberta, Canada: ACM, 2005.
- [4] 宫秀军,孙建平,史忠植.主动贝叶斯网络分类器[J].计算机研究与发展,2002,39(5):574-579
- [5] 邓河,阳爱民,刘永定.一种基于SVM的P2P网络流量分类方法[J].计算机工程与应用,2008,44(14):122-126.
- [6] Moore A W, Zuev D. Discriminators for use in flow-based classification[R]. Intel Research, Technical Report. 2005.
- [7] Tong S, Koller D. Support vector machine active learning with applications to text classification [J]. J Mach Learn Res, 2002(2):45-66.
- [8] Baram Y, El Yaniv R, Luz K. Online choice of active learning algorithms[J]. J Machine Learning Research, 2004(5):255-291.

作者简介:

顾名宇 男,(1960-),学士,副教授.研究方向为计算机应用技术。