第10卷 第1期 2010年1月
1671—1815(2010)1-0238-03

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 10 No. 1 Jan. 2010

计算机技术

一种改进的半监督增量SVM学习算法

吕宏伟

(武警工程学院,西安 710086)

摘 要 通过分析现有SVM的两种改进算法:半监督学习算法和增量学习算法,给出了对现有的增量学习算法的改进,提出了一种新的半监督增量SVM学习算法,将其应用于Web文本分类中,并验证了半监督增量SVM学习算法的有效性和可行性。

关键词 支持向量机 半监督学习 增量学习 Web文本分类

中图法分类号 TP181; **文献标志码** A

Vapnik V. 提出的支持向量机(Support Vector Machine,SVM)是在统计学习理论[1]基础上发展起来的一种新的学习算法。它是针对结构风险最小化原则提出的,改变了传统的经验风险最小化原则,具有很好的泛化能力。通过引入核映射方法,有效克服了维数灾难,较好地解决了非线性问题[2]。

支持向量机在模式分类中具有其它方法不可比拟的优越性,但也还存在一些问题:(1)理论上样本数量越多,分类器精度越高,但随着样本数据的增加,时间的增长也是非常迅速的;(2)训练SVM分类器,使用的是标注过的样本(labled examples),但要人工地标注大量样本,既是对人力、时间的耗费,同时也可能带来人为的出错率。解决第一个问题的方法,就是引入增量学习方法,解决第二个问题的办法,就是使用半监督学习方法。

2009年9月25日收到 国家科技部高新司项目(2005EJ000006)资助

第一作者简介:吕宏伟(1965—),男,山西永济市人,副教授,硕士,研究方向:计算机网络,信息安全。

1 增量学习算法(Incremental Learning)

1.1 α-ISVM

萧嵘等提出了SVM的增量学习算法α-ISVM,萧嵘等提出了SVM的增量学习算法α-ISVM,主要思想是:构造了一个再分类-再训练循环,每次总是把误分样本引入和SV样本一起进行训练,直到误分样本比例小于系统设定的阈值。

1.2 Syed等提出的增量学习算法

Syed等提出了的增量学习算法,主要思想是:增量训练由SV样本和新样本组成,再训练只需要进行一次即可完成,所以的非SV样本点都被抛弃。

2 半监督学习算法(Semisupervised Learning)

传统的SVM算法采用的是人工分类好的文档集D′来训练分类器,属于有监督的学习算法。要提高它们的分类精度就要增大训练集D′,使D′中包含的文档数目|D′|增加。但分好类的文档集D′相对没有分类的文档集D″是一种非常昂贵的资源。半监督算法的主要思想是:将未分类文档的类别看成为不完整数据,通过迭代将D″自动转化为D′,从而增大了训练集D的规模,使原来的训练集数据D =

{D′}变为现在的 D = {D′,D″},这样在 D′不变的情况下提高了分类器的精度。

3 半监督增量 SVM 学习算法

增量算法和半监督算法都在经典 SVM 算法基础对分类器性能有较大的提高。但也存在一些可以改进的地方。

3.1 对增量算法的改进

在 1.1 节中提到的 α-ISVM 增量学习算法,每次把误分样本加入训练集中进行训练,这样是不全面的。由 KKT 条件,只有违背 KKT 条件的新增训练样本才能使原 SV 集发生变化。即 Lagrange 乘子 $\alpha_i = 0, y_i f(x_i) \geq 1$ 的样本。违背 KKT 条件的样本可以分为 3 类:

(1)位于分类间隔中,与本类在分类边界同侧,可以被原分类器正确分类的样本,满足 $0 \leq y_i f(x_i) \{ 1$;

(2)位于分类间隔中,与本类在分类边界异侧,被原分类器分类错误的样本,满足 $-1 \leq y_i f(x_i) \leq 0$;

(3)位于分类间隔外,与本类在分类间隔异侧,被原分类器分类错误的样本,满足 $y_i f(x_i) \{ -1$。

由此可见,加入新增样本得到新的 SVM 分类器时,分类错误只是样本违反 KKT 条件的特定情况,并且由于错分样本中给系统引入噪声的可能性较大,所以单纯依靠错分样本很有可能降低训练精度。所以 KKT 条件比分类函数的分类判断更合理,应该选择所有违背 KKT 条件的样本作为下一步训练集。

在 1.2 节中提到 Syed 等提出的增量学习算法,它将训练集中的所有非 SV 样本抛弃,然后加入新增样本进行训练,这样也是不全面的,因为可能会丢失原来样本集中的信息。虽然 SV 集在某些情况下可以代替原来的训练样本集,但随着新样本的加入,最优分类面会发生变化,原样本集中非 SV 可能转化为 SV。

基于对以上两种算法的分析,在设计算法时,我们以是否违背 KKT 条件为判断依据,违背则加入新的训练集,否则就放入测试集中;将训练集中的非 SV 样本不做丢弃处理,而是把它放入测试集中,使它有可能再次符合条件进入训练集。

3.2 半监督增量 SVM 学习算法描述

在保证精度的前提下,为了尽可能地节约人力和时间,在半监督学习和增量学习的基础上的,提出了半监督增量学习算法。它克服了半监督学习训练时间较长、训练量较大的问题,也弥补了增量学习算法全部需要已标记样本的不足。

它的算法可以表示如下:

D′为已标记样本,D″为未标记样本,T_i 为训练集,X_i 为新增样本集,Ω_i 为训练的分类器。

(1)用 D′作为初始训练集 T_0,得到 Ω_0,T_0^{sv} 表示 Ω_0 的 SV 集;

(2)用 Ω_0 判断出 D″中各样本的类别,得到初始新增样本集 X_0;

(3)判断 X_0 中的样本是否有违反 Ω_0 的 KKT 条件的,分 X_0 为 X_0^x 和 X_0^s,X_0^x 表示违反 Ω_0 的 KKT 条件的样本集合,X_0^s 表示满足 Ω_0 的 KKT 条件的样本集合;

(4)用 $T_j = \{T_{j-1}^{sv} \cup X_{j-1}^{s}\}$ 训练得到 Ω_j,其余样本作为新增样本集 $X_j = \{(T_{j-1} - T_{j-1}^{sv}) \cup X_{j-1}^{s}\}(j = 1)$;

(5)用 Ω_j 判断出 X_j 中各样本的 $x_{ji}(x_{ji} \in X_j)$ 的类别,如果 x_{ji} 的类别发生变化,则判断 X_j 中的样本是否有违反 Ω_j 的 KKT 条件的,分 X_j 为 X_j^x 和 X_j^s,X_j^x 表示违反 Ω_0 的 KKT 条件的样本集合,X_j^s 表示满足 Ω_j 的 KKT 条件的样本集合;

(6)$j = j + 1$,返回第(4)、(5)步,直到新增样本集中样本类别不再发生变化或误分率满足某个阈值。

4 实验分析

本文将半监督增量 SVM 算法思想运用到 Web 文本分类中。实验数据为从网上获取的网页,为财经类和非财经类两类分类。共计样本 1400 个,其中已标记样本 400 个(财经类 150 个,非财经类 250 个);未标记样本 1 000 个。用已标记样本中的 200 个(财经类 70 个,非财经类 130 个)做为训练集,未标记样本 1000 个作为新增样本集,其余 200 个已标记样本作为最后分类器性能测试集。对上述三种算法做了比较:Syed 等提出的增量 SVM 学习算法、半

监督 SVM 学习算法、半监督增量 SVM 学习算法。实验结果如表1所示:

表1 三种算法性能比较

算 法	训练时间/s	正确率/%
Syed 等的增量 SVM 学习算法	6.05	85
半监督 SVM 学习算法	21.32	82
半监督增量 SVM 学习算法	10.50	86

由实验数据可知,Syed 等提出的增量 SVM 学习算法由于只需要一次迭代,所以时间较短;而半监督 SVM 学习算法由于需要多次迭代,且样本量较大,所以时间较长。我们提出的半监督增量 SVM 学习算法较增量学习算法来说,节约了很多的用于标注样本的人力成本,时间代价上增长并不是很大,正确率也没有下降;较半监督学习算法来说,时间和正确率都有较大提高。

5 总结

本文提出了一种新的 SVM 算法——半监督增量 SVM 学习算法。它融合了半监督算法和增量算法的长处,又克服了它们各自的不足。实验证明半监督增量 SVM 学习算法的有效性和可行性。

参 考 文 献

1 Vapnik V. The nature of statistical learning theory. New York:Springer-Verlag,1995

2 Burges C J C. A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining,1998;2(2):121—167

3 Platt J C. Fast training of support vector machine using sequential minimaloptimization, In: Advances in Kernel Methods——Support Vector Learning,MIT,2002

4 Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other Kernel-based learning Methods. Publishing House of Electronics Industry,2004

An Refinement Algorithm of Semisupervised-incremental SVM Learning

LÜ Hong-wei

(Engineering College of the Armed Police Forces,Xi'an 710086, P. R. China)

[**Abstract**] SVM has a good performance in pattern classification. According to the analysis of the two present refinement algorithms about SVM: Semisupervised Learning and Incremental Learning, an improved algorithm of present Incremental Learning algorithm is gavn. A new algorithm of semisupervised-Incremental SVM learning is given out. Also this new algorithm is used in Web text classification.

[**Key words**] SVM semisupervised learning incremental learning Web text classification