基于抽样的两阶段支持向量机训练算法

曹菲菲 朱慕华 朱靖波

（东北大学自然语言处理实验室，沈阳 110004）

摘 要： 本文针对支持向量机计算复杂度过高的问题，提出一种基于抽样的两阶段的快速训练算法，加快支持向量机训练速度。该方法是在序贯最小优化算法的基础上，对训练过程的改进，首先在数据集中随机抽取少量数据，用序贯最小优化算法训练近似的分类函数，按规则选取结果集，再与原始数据集合为新的数据集进行第二阶段的训练。实验结果表明，基于抽样的两阶段的训练方法，可以显著地减少训练的时间和空间，同时保证可比较的分类性能。

关键词： 支持向量机；序贯最小优化；文本分类；两阶段训练

Sampling-based Two-stage Training for Support Vector Machines

Cao Feifei, Zhu Muhua, Zhu Jingbo

(Natural Language Processing Lab, Northeastern University, Shenyang, 110004)

Abstract: Support Vector Machines are limit in practical applications for the sake of high computational complexity of training. In this paper, we proposed a sampling-based two stage faster training for Support Vector Machines. This algorithm is based Sequential Minimum Optimization, by first training a classifier in the first step from sampled data, then filtering the whole training dataset. Experimental results show that the algorithm proposed is much faster than direct application of Sequential Minimum Optimization, meanwhile maintaining the final classification performance.

Key words: Support Vector Machines; Sequential Minimum Optimization; Text Categorization; Two-stage training

1 引言

支持向量机（Support Vector Machine）是当前常用的，基于统计学习理论的机器学习算法，可以用于解决分类，回归以及排序问题[3]。具体到分类任务，支持向量机通过核函数将原始输入空间中的向量隐式映射到高维特征空间，并通过在原始空间中的计算学习特征空间中的线性二类分类函数，保证得到的分类函数以最大间隔正确分类训练数据。当前支持向量机已被广泛应用到图像识别[5]，信号处理[6]，中文分词[8]和自动文本分类[3]等各个领域。

虽然支持向量机因为良好性能而获得广泛关注和应用，在某些实际的工程性应用中，却因为它的训练过程时间和空间复杂度而受到限制。支持向量机的训练过程等价于求解一个二次规划问题，其复杂程度主要由参与训练的实例个数决定。在某些应用，例如基于字的中文分词，其实例个数等于训练数据中的字数（通常字数超过百万），

本文工作部分得到国家自然科学基金（No. 60473140）和 国家教育部新世纪优秀人才计划项目资助

作者简介：曹菲菲，女，山东省德州市，本科，E-mail：caoff@ics.neu.edu.cn

对支持向量机的训练是一个巨大挑战。

针对支持向量机训练过程计算复杂度过高的问题，本文提出了基于抽样的两阶段的训练方法，该方法基于序贯最小优化（Sequential Minimal Optimization）算法[1]。序贯最小优化算法的思想基于优化问题分解，每次迭代只处理数据集合的最小子集。在满足线性约束的条件下，可以处理的最小子集只包含两个数据点。算法的优势在于两个点的优化问题可以采用解析方法进行求解。虽然总体的迭代次数不可避免地有所增加，由于每次迭代需要操作很少，整体上的速度却有数量级的提高。

本文提出的方法利用了序贯最小优化算法的分类函数只由从实例中训练确定支持向量决定，而与其它非支持向量实例无关的性质。在训练数据集中保留支持向量的部分，去除非支持向量的实例，训练数据。为了证明方法有效性，我们以自动分本分类作为应用环境，与直接应用序贯最小优化的训练方法进行了比较实验。实验表明，基于抽样的两阶段的训练方法，可以显著地减少训练的时间和空间，同时保证可比较的分类性能。

在本文剩余内容中，第二部分简单地介绍对支持向量机训练方法的相关研究；第三部分我们分析了目前应用最广泛的序贯最小优化算法的性质；针对前一部分的分析，第四部分详细描述了基于抽样的两阶段训练方法；第五部分的比较实验以及最后第六部分的结论。

2 相关研究

支持向量机的训练过程等价于求解线性约束条件下的凸二次规划问题[3]。研究人员提出多种优化算法用于解决凸二次规划的问题，包括牛顿法（Newton method），梯度变化（Conjugate Gradient），梯度上升（Gradient Ascent）。这些方法要求在内存中存储整个核矩阵，其空间复杂度与训练样例个数的平方成正比，限制了支持向量机在大规模计算中的应用。目前，大部分支持向量机的具体实现都采用序贯最小优化算法（Sequential Minimal Optimization），这也是本文工作的基础。

3 问题分析

支持向量机的训练过程等价于求解如下的二次规划问题：

$$\max W(a) = \sum_{i=1}^{\lambda} \alpha i - \frac{1}{2} \sum_{i,j=1}^{\lambda} yiyj\alpha i\alpha j\left(xi \cdot xj\right),$$

$$\text{s.t.} \quad w = \sum_{i=1}^{\lambda} yi\alpha ixi,$$

$$0 = \sum_{i=1}^{\lambda} yi\alpha i,$$

应用序贯最小优化算法求解上述二次规划问题时，算法必须选择对加快优化速度贡献最大的点最优化。从理论上分析，如果选择数据点所需要的计算时间小于因为迭代次数减少所获得的收益，最终整体的优化速度同样可以获得提高。由此序贯最小优化算法等价于一个二重循环的过程，其中，外层循环寻找不符合 Karush-Kuhn-Tucker 条件的数据点优先优化；在选定第一个点的条件下，第二个点根据使目标函数涨幅最大的原则进行选择。

但是，到目前为止，没有任何理论上的证明可以保证利用上述启发式规则寻找待优化点可以改善整体的优化速度，尤其是大数据量计算的情况下，寻找待优化点所需的计算量将不可以忽略；另一方面，根据支持向量机的性质，最终训练得到的分类超平面只取决于支持向量，而与大量的其它数据无关。根据这一理论分析，本文提出基于抽样的两阶段训练方法，对数据集合中的点进行筛选，减少数据的规模，从而加快支持向量机的训练速度。同时，数据规模减少也可以降低计算所需的空间复杂度。

4 两阶段支持向量机的快速训练方法

针对支持向量机训练速度过慢的问题，基于序贯最小优化算法，我们设计了基于抽样的两阶段训练方法。该算法分两部分：

表 4-1 基于抽样的两阶段训练算法

Table 4-1 Sampling-based Two-stage Training Algorithm

1. 给定训练数据集 D 中，随机的抽取相同数量的正、负训练数据，表示为集合 D_{sample}；
2. 在数据集 D_{sample} 上学习分类函数 f_{step1}；
3. 给定阈值 T，利用 f_{step1} 对 D 进行分类，保留函数输出绝对值小于 T 的实例，表示为集合 D_{train}；
4. 在数据集 D_{train} 上学习分类函数 f_{step2}，作为最后的分类函数。

我们把表 4-1 中的 1-2 称为算法的第一阶段，3-4 为算法的第二阶段。第一阶段中 D_{sample} 包含随机抽取的相同数量的正类样例和负类样例；第二阶段的重点在于利用第一阶段训练得到的分类器 f_{step1} 对整个训练数据进行筛选，其中以实例到分类超平面的距离作为函数输出值，阈值 T 在本文实验中取值为函数输出绝对值的算术平均。

5 实验

为了验证方法的有效性，本文以文本分类作为应用，比较基于抽样的两阶段训练方法与序贯最小优化算法在时间和空间计算复杂度上的优劣。

5.1 实验数据集

本文采用路透社新闻组（newsgroup）数据作为实验的数据集。Newsgroup 数据集包含 UseNet 文本。由 Lang 于 1995 年负责收集。总共有 Usenet 20 个新闻组（20 个类型）的文本，共有 20000 篇，每个类别包含 1000 篇，绝大部分文本（96%）只属于 1 个类型。在本文实验中，首先要进行数据集预处理，只留下 Body 部分，其它部分都去掉（如 Subject、日期等）。

5.2 实验设置

在本文的实验中，我们采用了 SVM^{light} [9]作为支持向量机模型的实现。SVM^{light} 采用序贯最小优化作为优化算法。在本文的实验中，我们简单地使用 SVM^{light} 的默认参数，构造一个线性 SVM。我们将数据集按 4：1 比例随机分割成训练数据集和测试数据集。

支持向量机是一个二类分类模型。本文的实验中的分类任务是多类别分类问题（newsgroup 包括 20 个类别），如何实现多类别支持向量机模型是在实验过程中必须考虑的一个问题。本文采用 one-against-rest 方法[3，11.8]，将多类别分类问题转化为多个二类分类问题。

本文的实验运行环境为处理器主频为 3GHz 的机器上进行。所有时间和空间的性能衡量都基于该配置进行。

5.3 性能指标

在本文中，一个文档只属于一个类别。本文使用传统的召回率、正确率、Macro F1 来评价分类结果。计算公式如下：

$$MacroP = \frac{1}{n}\sum_{j=1}^{n} P_j \qquad MacroR = \frac{1}{n}\sum_{j=1}^{n} R_j$$

$$MacroF1 = \frac{MacroP \times MacroR \times 2}{MacroP + MacroR}$$

其中，n 是类别总数，P_j 为第 j 类的准确率，R_j 为第 j 类的召回率。

5.4 实验结果

直接采用序贯最小优化算法，支持向量机包含的训练数据包括 14997 个文本，最终得到的宏 F1 值为 83.2%，运行时间 14 分 51 秒。采用基于抽样的两阶段训练方法，结果数据如表 5-1 所示。

其中，“抽样个数”表示算法在第一阶段抽样样本个数（例如，10 表示随机抽取正、负训练实例各 10 个）；宏 F1，准确率，召回率表示最终的分类性能；运行时间的单位为分：秒，包括完成整个训练过程（包括 20 个二类分类器的训练过程）；“第二阶段训练数据”表示在第二阶段中，经过筛选以后得到训练数据的规模。

从表中可以看出，与直接使用序贯最小优化算法得到的结果相比，基于抽样两阶段训练方法可以使训练时间大幅减少（当抽样个数为 10 时，只需要一半的训练时间），同时保证最终分类性能仍然具有可比性。

表 5-1 实验结果

Table 5-1 Experimental Results

抽样实例数	10	30	50	70	90	120	150	170	180	190	200
宏 F1（%）	80.1	80.3	80.9	82.0	82.4	82.6	83.5	81.1	83.1	83.2	83.4
第二阶段训练数据	7122	10856	7586	6724	8137	7443	8983	9243	8125	8842	8171
准确率 P（%）	81.3	81.3	80.8	82.1	82.5	82.6	83.4	81.3	83.1	83.2	83.4
召回率 R（%）	79.0	79.5	80.9	82.0	82.3	82.6	83.5	80.8	83.0	83.3	83.5
运行时间（分：秒）	7:00	6:26	6:58	8:10	8.33	7:59	8:37	8:41	10:1	9:28	9:18

与理论分析一致，随着抽样个数的增加，训练时间必然增加，最终的分类性能与直接使用序贯最小优化算法得到的分类性能接近。同时，我们观察到，在抽样个数增加的过程中，训练时间和最终的分类性能并不是随着抽样个数单调增加。其中的原因在于，第一阶段的抽样具有随机性，它对第一阶段训练得到的分类器具有很大的影响。理想情况下，第一阶段训练得到的分类超平面应该与最终得到的分类超平面尽可能的接近；分类平面的近似程度直接影响着第二阶段对训练数据的筛选，从而影响最终的性能。

6 结论

训练过程的计算复杂度过高，使支持向量机的实际应用受到限制。本文基于序贯最小优化算法提出了基于抽样的两阶段训练算法，并在文本分类实际任务中，与直接应用序贯最小优化的过程进行了比较实验。实验结果表明，经过适当的抽样，基于抽样的两阶段训练算法，在明显减少运算的时间与空间复杂度，同时可以获得可比较的分类性能。

参考文献：

[1] J. Platt. Fast training of support vector machines using sequential minimal optimization[A]. In Advances in Kernel Methods -- Support Vector Learning[C]. MIT Press, 1998.

[2] T.Joachims. Text Categorization with support vector machines[A]. In Proceedings of the European Conference on Machine Learning, 1998.

[3] Vladimir N. Vapnik, 张学工译 统计学习理论的本质[M].清华大学出版社. 2000.

[4] Nello Cristianini, John Shawe-Taylor An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M],2001.

[5] Boughorbel, S. and Tarel, J.-P. and Fleuret, F. and Boujemaa, N. GCS Kernel For SVM-Based Image Recognition[A]. Proceedings of International Conference on Artificial Neural Networks (ICANN'05). 595 – 600.2005

[6] S. Poyhonen et al. Numerical magnetic field analysis and signal processing for fault diagnostics of electrical machines[A],IEEE,2003

[7] J. Bredensteiner. Optimization Methods in Data Mining and Machine Learning[D]. PhD thesis, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY,1997.

[8] Xu Yun, Zhang Feng. Using SVM to construct a Chinese dependency parser[A] .Univ SCIENCEA. 2006

[9] SVM^{light}, http://svmlight.joachims.org/