

基于 SVM 文本分类中的关键词学习研究

孙晋文^{1,2} 肖建国²

(航天信息股份有限公司)¹ (北京大学计算机科学技术研究所 北京 100871)²

摘要 文本分类中分类方法与关键词是两个重要的因素,支持向量机作为一种良好的分类方法,在文本分类中表现出很好的性能。本文以支持向量机为基础,重点研究其分类过程中关键词的学习问题,在其增量分类中,同步进行关键词的增量学习调整,实验结果表明,该方法可以更好地提高文本分类的分类性能。

关键词 支持向量机, 增量学习, 分类, 关键词学习

Study on Keyword Learning in SVM-based Text Classification

SUN Jin-Wen^{1,2} XIAO Jian-Guo²

(Aero-Space Information Co., Ltd.)¹ (The Institute of Computer Science and Technology, Peking University, Beijing 100871)²

Abstract For classification, the classification method and keyword are the two important factors. As a good classification method, SVM is of good performance based on SVM. This paper mainly explores the keyword learning during the classification. With the incremental training, the keyword is re-learned and adjusted. The experimental result shows that the new method achieves the more excellent performance.

Keywords SVM(support vector machine), Incremental learning, Classification, Keyword learning

1 引言

支持向量机技术已成为分类的热点,自从 Vapnik^[1,2] 在 1995 年提出以来,其基于统计学习理论的坚实基础以及所具有的良好推广性、防过学习能力,使其在许多应用领域获得了良好的应用^[3,4]。同样,基于支持向量机的分类技术也被证明是当今最好的分类方法之一。1998 年,Joachims^[4] 验证了其在文本分类中的优异性能。此外,一些学者^[5] 也相继对该问题进行了大量的研究与应用,进一步证实了其性能的有效性。为提高 SVM 的分类性能,从提出开始,人们就研究各种不同改进训练算法来改进其训练过程,包括 1995 年 Cortes 和 Vapnik^[2] 提出的“chunking”方法、1997 年 Qsuna^[3] 提出的分解法、1998 年 Joachims^[6] 采用的适用于大规模训练的分解算法、Platt^[7] 的 SMO 算法等,都为提高支持向量机的训练性能提供了良好的手段,同时也为支持向量机在实际中的应用提供了保证。

对于 SVM 的分类,其不仅具有良好的分类性能,同时具有增量训练特性,这是一般模型所不具备的,这为支持向量机的研究、应用提供了巨大的前景。1999 年, Nadeem^[8] 采用增量学习方法对两类支持向量机分类进行了研究。2001 年, Cauwenberghs^[9] 从理论上分析了进行增量训练时全局优化的方法。同年, Ralaivola^[10] 探讨了一种采用径向基函数(RBF)为核函数的支持向量机进行局部增量训练的优化策略。

本文基于 SVM 的分类方法,实现了一个自动分类系统。本文主要介绍在分类实现中所重点考虑的关键词的学习问题,即在 SVM 分类器增量分类训练中,同步进行关键词的自优化、自调整。实验表明,该方法消除了增量训练与一次性训练间在分类关键词上的差异,从而使基于 SVM 的文本分类

具有更好的性能。

2 支持向量机可增量训练性

对于支持向量机,其基本原理是通过选择的非线性映射将输入向量 x 映射到一个高维特征空间 Z ,并在这个空间中构造最优分类超平面,该超平面使类别间的分类间隔最大。通过求解二次规划问题,最优超平面为训练集中向量的线性组合:

$$w = \sum_{i=1}^l y_i \alpha_i x_i$$

最后的分类函数可表述为:

$$f(x) = \text{sgn} \left(\sum_{i=1}^l y_i \alpha_i K(x_i, x) + b \right)$$

其中, $\alpha_i = 0$ 的样本对分类问题不起什么作用。具有 $\alpha_i > 0$ 占整个训练集合的一小部分的支持向量才会影响分类效果,使得支持向量机这一机器学习技术具有了良好的推广能力及内在的可增量训练特性,并且多种用于改善支持向量机训练性能的方法也正是利用其这一特性,如 Vapnik 的“chunking”方法等。

因此,对于训练后的模型来说,训练结束后,在参数确定的情况下,支持向量集可以完全描述整个训练样本集的分类特征。模型中的支持向量集包含着整个训练集中有用信息,对支持向量进行增量学习是可行与有效的,并且,若新增样本中含有原样本集中所不包含的分类信息或可能的新的支持向量,则其必然违背 KKT 条件,通过训练学习将新的分类信息体现在新的支持向量集中。

3 基于 SVM 分类中关键词学习

3.1 问题的描述

对于本文所提出的融合关键词学习的增量分类学习技

术,可以用一四元组表示 $\langle E, N, V, K \rangle$ 。其中, $E = \{E_0, E_1, \dots, E_l\}$ 为训练文档集, E_0 为初始训练文档集, $E_i (1 \leq i \leq l)$ 为增量训练文档集; $N = \{N_0, N_1, \dots, N_l\}$ 为用于训练的训练文档数量, $N_i = \sum_{j=0}^i |E_j|$; $V = \{V_0, V_1, \dots, V_l\}$ 为相应于训练文档集合 $\sum_{i=0}^l E_i$ 而获得的分类模型中的支持向量集; $K = \sum_{k=1}^m \{(t_k, n_k, \bar{f}_k)\}$ 为关键词集,用于在增量训练过程中进行分类关键词的学习与调节以及分类处理时文档向量的生成,其中 $t_k, n_k, \bar{f}_k$ 分别为关键词、训练文档中出现该关键词的文档数量及关键词的倒排文档频率(idf), $\bar{f} = idf_k = \log N/n_k, m$ 为集合 K 中总关键词数, $m = |K|$ 。

3.2 分类关键词的学习与调整

信息处理技术的基础是文档的向量化。文档向量化是建立在文档特征关键词提取之上的,对于一篇文档可形式化地表述为 $d = \{(t_1, w_1), (t_2, w_2), \dots, (t_{m'}, w_{m'})\}$,其中, t_j, w_j 分别为文档的关键词及其权重, m' 为该文档关键词数。同样,对于分类模型来说,一般不论采用何种分类方法,其模型的基础也一般离不开类别关键词及其权重问题, $\Sigma(t_i, w_i)$ 。对于词权重 w 的计算来说,使用最多的是 $tf * idf$ 方法, tf, idf 为关键词的词频与倒排文档频率。

为了使增量训练过程中关键词的处理实现与非增量的一次性训练有一致的效果,在支持向量进行增量优化选择的同时,进行相应的关键词的学习与调整。本文在增量过程进行支持向量的优化调整之前,首先根据新训练文档的特性,实现了关键词及其特性值的动态调整。具体方法描述如下:

初始条件:对初始训练集 E_0 进行训练,产生文档关键词集合 K_0 。 $N_0 = |E_0|$ 为文档集 E_0 的大小。

step1:读入一个新的增量训练学习文档集合 $E_i (i = 1, \dots, l)$,包含的文档数量为 $|E_i|$,则 $N_i = N_{i-1} + |E_i|$ 。对于其中任一文档 $d \in E_i$ 进行切分处理,获得文档词频向量 $d = \{(t'_1, f'_1), (t'_2, f'_2), \dots, (t'_{m'}, f'_{m'})\}$, t'_j, f'_j 分别为文档关键词及其词频;

step2:对于任一切分后的文档 d 进行关键词学习与调整,在原关键词集合 K_{i-1} 中搜索 t'_j 。若 $t'_j \in K_{i-1}$,不妨假定在集合 K_{i-1} 中与关键词 t'_j 对应的关键词为 t_k ,其相应的出现的文档数为 n_k ,则将其进行学习调整, $n_k = n_k + 1$;否则,该关键词为新关键词,将 t'_j 增加到集合 K_{i-1} 中,记为 t_{m+1} ,相应地 $n_{m+1} = 1$;

step3:重复 step2,直到本增量集 E_i 中所有增量文档 d 处理完毕,集合 K_{i-1} 实现了关键词 t_k 及 n_k 更新;

step4:根据变化了的关键词集合 K_{i-1} ,对于任一 $t_k \in K_{i-1}$,重新计算相应的关键词倒排文档频率 $\bar{f}_k = \log N/n_k$,从而完成关键词的学习与调整,关键词集合 K_{i-1} 更新为 K_i 。

3.3 支持向量权重调整

对于训练后获得的支持向量集 V ,是在通过训练而获得的位于类别边界处的文档,其表现形式为文档向量 $d = \{(t_1, w_1), (t_2, w_2), \dots, (t_{m'}, w_{m'})\}$ 以及相应的二次规划因子 α 。对于文档向量中特征词的权重来讲,由于新的增量训练文档的加入、文档关键词集合的变化,导致关键词集 K 中相应 idf 发生变化。因此,在进行支持向量的增量优化训练前,需要对模型中的支持向量集 V 中的原关键词权重进行相应的调整。新的支持向量关键词权重更新方式为:

$$w'_j = \frac{w_j * idf'_j}{idf_j}, w'_j, w_j \text{ 分别为新、旧关键词权重,}$$

idf'_j, idf_j 为新、旧关键词的倒排文档频率。

支持向量权重调整步骤如下。

step1:对任一增量过程,读入未更新前的关键词集合 K_{i-1} ,及新更新后的新关键词集合 K_i ;

step2:从支持向量集合 V_{i-1} 读取任一支持向量 $d^{vec} = \{(t_1, w_1), (t_2, w_2), \dots, (t_{m'}, w_{m'})\}$;

step3:对于任一 $(t_j, w_j) \in d^{vec}$,分别查询新旧关键词集合 K_i, K_{i-1} ,并分别获得其相应的倒排文档频率 $\bar{f}_k, \bar{f}_k^{-1}$,进行新的支持向量关键词权重计算 $w'_j = (w_j * \bar{f}_k) / \bar{f}_k^{-1}$;

step4:重复 step2、step3,直到集合 V_{i-1} 中所有的支持向量更新完毕。

3.4 增量训练

基于以上关键词学习以及支持向量关键词权重的调整,融合关键词学习策略的支持向量机增量分类学习过程可表述如下。为简单起见,此处仅以两类分类模式为例进行描述。

初始条件:系统对初始训练文档集 E_0 进行训练,产生支持向量集 V_0 、关键词集 K_0 。

step1:读入一个新的增量训练学习文档集合 $E_i (i = 1, \dots, l)$,对于任一 $d \in E_i$,进行相应的文档切分,并生成文档词频向量 $d = \{(t'_1, f'_1), (t'_2, f'_2), \dots, (t'_{m'}, f'_{m'})\}$;

step2:读取原关键词集合 $K_{i-1} (i = 1, \dots, l)$,对切分后的增量文档 d 进行相应的新关键词的学习与调整处理,得到新的关键词集 K_i ;

step3:对于任一 $d \in E_i$ 关键词的学习与调整完成后,根据新关键词集 K_i 计算并获得增量训练集合 E_i 中的训练文档 d 的文档向量 $\{(t'_1, w'_1), (t'_2, w'_2), \dots, (t'_{m'}, w'_{m'})\}$, $w'_j = f'_j * \bar{f}_k, \bar{f}_k$ 为文档关键词 t'_j 在关键词集 K_i 中对应的值;

step4:读入原分类模型中的支持向量集 V_{i-1} ,对原支持向量进行权重调整,获得调整后的支持向量集 V'_{i-1} ;

step5:将 V'_{i-1} 与新的增量训练文档向量 E_i 进行合并, $E_i + V'_{i-1}$ 作为新的训练集 $E_i^c, E_i^c \leftarrow E_i + V'_{i-1}$;

step6:将获得的包含新的增量训练集与原支持向量集合的综合训练集 E_i^c 进行支持向量机的二次规划优化训练,以得到新的支持向量集 V_i ,用于新的分类处理。

4 实验分析

为验证融合关键词学习的增量训练的效果,我们进行了相应的实验研究,取得了预期的良好的实验效果,实验数据的来源为人民网2001年、2002年新闻分类语料,共1万多篇,分为体育、信息技术、军事、文娱、科教、环保、经济共7大类别。在此基础上,我们开发了自己的分类系统,并在系统中实现了本文所提出的增量训练算法,实现了系统的可增量训练学习能力。现介绍其中一组实验数据及结果。

分别从中随机抽取2800篇,组成实验数据集。数据在各个类别中均匀分布,通过随机抽取,又分为训练集合与测试集合。同时,为了进行增量训练的需要,训练集合部分又随机分别等分成12部分,用于增量训练。具体划分情况见表1、表2。

表1 实验数据集

	体育	信息技术	军事	文娱	科教	环保	经济	总计
数据集								
训练集	300	300	300	300	300	300	300	2800
测试集	100	100	100	100	100	100	100	

表 2 增量实验训练数据分配

	初始训练集	增量1	增量2	增量3	增量4	增量5	增量6	增量7	增量8	增量9	增量10	增量11
数据集	175	175	175	175	175	175	175	175	175	175	175	175

表 3 分类实验数据结果

分类精度	I-SVM-KL 有关键词学习	I-SVM 无关键词学习	batch-SVM 一次性训练
初始集	79.571	79.571	79.571
增量1	85.429	84.429	85
增量2	85.714	85.714	85.571
增量3	86.571	86.286	86.429
增量4	87.429	87.143	87.857
增量5	87.429	87.429	87.571
增量6	88.143	88.143	88.286
增量7	88	87.571	88
增量8	88	86.714	87.857
增量9	88.143	87	88.429
增量10	88.714	87.571	88.429
增量11	88.714	87.286	88.714

在实验中,为了验证笔者所提出方法对实际分类精度的效果,分别将本文提出的有关键词学习的增量训练学习与无关键词学习的增量训练以及非增量的一次性训练进行实验比较。为表述方便,将本文提出的具有关键词学习的增量学习方法记为 I-SVM-KL(Incremental SVM with Keyword Learning)方法,而无关键词学习的增量训练学习方法,即普通的增量训练方法记作 I-SVM,非增量的一次性训练方法记作 batch-SVM。具体实验数据结果见表 3。图 1 是上述实验结果的实验数据曲线,其中分类精度为各个类别精度的平均值。

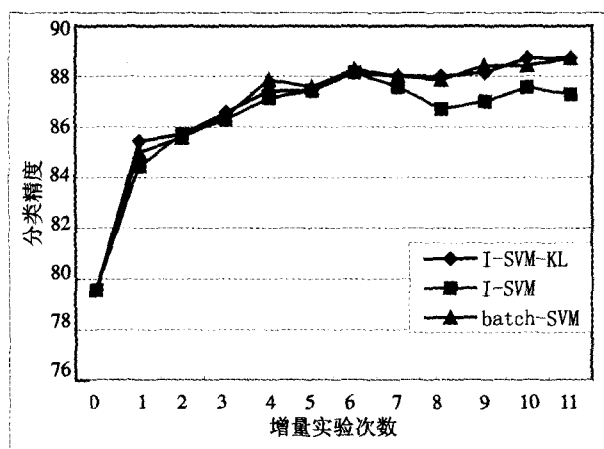


图 1 不同训练方法实验结果比较

根据以上实验结果可以看出,在基于 SVM 的文本分类

中,进一步考虑关键词的学习,可以进一步提高增量训练的性能,大大优于不进行关键词学习而仅仅进行支持向量优化选择的方法。同时,与以前的研究者^[8]的结果一样,不进行关键词学习的增量训练学习方法,则只获得了略低的性能。从实验结果还可以看出,在多数实验数据点,本文提出的方法与一次性训练获得了完全相同或基本相同的结果。有些数据点的差异可能是二者在某次优化中支持向量的选择上略有不同的结果。但从二者的实验数据及性能曲线中可以明显看出,本文提出的改进的增量训练方法在分类精度上达到了与一次性训练相一致的效果,实现了一次性训练的增量式实现,从而为 SVM 方法在文本分类中的具体应用提供了更好的实现途径。

结束语 本文通过 SVM 文本分类的研究,重点探讨了基于 SVM 的增量分类训练中分类关键词的增量训练学习。该方法考虑了基于 SVM 的增量训练过程中,由于新文档的增加而造成的特征关键词及其特征的变化问题。实验结果也表明,本文所提出的增量训练方法在分类精度上获得了更好的效果,从而为 SVM 分类在实际中的应用提供了更好的方法。

参考文献

- Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995
- Cortes C, Vapnik V. Support Vector Networks. Machine Learning, 1995, 20
- Osuna E, Freund R, Girosi T. Training Support Vector Machines: An Application to Face Detection. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. New York, 1997
- Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proceedings of the European Conference on Machine Learning, Berlin, Springer, 1998
- Yang Yiming, Liu Xin. A re-examination of text categorization methods. In: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, 1999
- Joachims T. Making Large-Scale SVM Learning Practical. In: Scholkopf B, Burges C, Smola A, eds. Advances in Kernel Methods - Support Vector Learning. MIT Press, 1999
- Platt J. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. [Technical Report]. MSR-TR-98-14. Microsoft Research, 1998
- Syed N A, Liu Huan, Sung Kah Kay. Incremental Learning with Support Vector Machines. In: Proc. of the Int Joint Conf. on Artificial Intelligence(IJCAI), 1999
- Cauwenberghs G, Poggio T. Incremental and Decremental Support Vector Machine Learning. Adv. Neural Information Processing Systems. Cambridge MA: MIT Press, 2001, 13
- Ralaivola L, d'Alch-Buc F. Incremental Support Vector Machine Learning: a Local Approach. In: Proceedings of ICANN' 01, Springer, 2001