文章编号:1001 - 9081(2007)06 - 1382 - 03

使用基于 SVM 的局部潜在语义索引进行文本分类

张秋余[1],刘 洋[1,2]
(1. 兰州理工大学 计算机与通信学院,甘肃 兰州 730050;
2. 渤海大学 信息科学与工程学院,辽宁 锦州 121000)
(jsxxxly@ sohu. com)

摘 要:潜在语义索引(LSI)通过奇异值分解(SVD)获得原始词—文档矩阵的潜在语义结构,在一定程度上解决了一词多义和多词一义问题。但目前文本分类中使用 LSI 方法的效果并不理想,这是因为没有充分考虑分类信息。为解决该问题,提出一种改进的局部潜在语义索引(LLSI)方法,使用支持向量机(SVM)来产生局部区域。实验结果表明,该方法是有效的。

关键词:文本分类;潜在语义索引;支持向量机;局部区域
中图分类号: TP311.13;TP181 **文献标识码:**A

Using SVM-based LLSI for text classification

ZHANG Qiu-yu[1], LIU Yang[1,2]
(1. *College of Computer and Communication, Lanzhou University of Technology, Lanzhou Gansu* 730050, *China*;
2. *College of Information Science and Engineering, Bohai University, Jinzhou Liaoning* 121000, *China*)

Abstract: Latent Semantic Indexing (LSI) uses Singular Value Decomposition (SVD) to obtain latent semantic structure of original term-document matrix, and problems of polysemy and synonymy can be dealt with to some extent. However, the present available methods of applying LSI to text classification are not satisfying, since they do not take full account of classification information. To solve the problem, an improved Local LSI (LLSI) method was proposed, using Support Vector Machine (SVM) to produce the local region. Experimental results suggest that the proposed method is effective.

Key words: text classification; Latent Semantic Indexing (LSI); Support Vector Machine (SVM); local region

0 引言

文本分类是指在给定的分类体系下,根据文本内容自动确定文本类别的过程[1]。向量空间模型(VSM)是文本表示的主要方法,该方法建立在将文本简化为 BOW(Bag Of Words)的基础上[2]。而在实际中一个词可以表达不同的含义,多个词也可以表达同一个含义。本来应属于某一类的文本,因为行文中用词的不同,就可能被错误地归为另一类;相反有些文本只因为其中出现了某类中频率较高的词,就可能被归为含义完全不同的一类。潜在语义索引(Latent Semantic Indexing,LSI)[3]使用统计方法解决了上述问题。通过奇异值分解(Singular Value Decomposition,SVD),LSI 把原始的向量空间转换为低维的潜在语义空间,文本在转换后的语义空间上进行表示和比较。LSI 在文本检索中是有效地方法[3]。

LSI 在文本分类中通常的做法是,对原始词—文档矩阵进行 SVD,选取最大的一些奇异值对应的特征作为潜在语义空间。而对文本分类而言,应保留的特征是那些具有很好分类能力的特征。目前没有理论证明奇异值最大的那些特征具有最好的分类能力。事实也说明了这一点,在该潜在语义空间上进行文本分类,分类效果没有得到改善。究其原因,一些对分类贡献大的语义特征,奇异值较小而被过滤掉了。更进一步,LSI 是一种完全的无监督方法,只注重文本的表示,没有考虑训练集的分类信息。

为提高 LSI 在文本分类中的作用,有两种改进的做法。一种是 LLSI[4],它在每个类别的局部区域上进行 SVD。与在整体词—文档矩阵上进行 SVD 相比,该方法考虑了分类信息,因此提高了分类效果。如何选取合适的局部区域是 LLSI 的关键问题,目前常用的选取局部区域的方法对分类效果的改善非常有限,还有待于进一步改进。该方法可以看成是先考虑分类信息,再进行 SVD。另一种做法可以看成先进行 SVD,再考虑分类信息[5,6]。即在整体 SVD 基础上,选取对分类贡献大的语义特征。除上述两种做法外,还有其他的做法,如文献[7]中提到的引入测试集中的信息,但是该做法在实际应用中的可行性很差。

本文提出一种改进的 LLSI 方法,用支持向量机(SVM)[8]进行局部区域的选取,作为 SVD 的基础。该区域能够更好地反映分类特征,从而能够更好地表示潜在语义空间。

1 潜在语义索引

1.1 LSI

LSI 的基本思想是认为文档中的词与词之间存在某种联系,即某种潜在语义结构,因此用统计方法来寻找该语义结构,并用语义结构来表示词和文档。原始词—文档矩阵 $\boldsymbol{D}$ 通过 SVD 被分解为三个矩阵的积:

$$\boldsymbol{D}_{s\times n} = T_0 \times \boldsymbol{U}_0 \times V_0^T \tag{1}$$

其中,$\boldsymbol{T}_0$ 为 $s\times k$ 的单位正交矩阵,$\boldsymbol{V}_0$ 为 $n\times k$ 的单位正交矩阵,$\boldsymbol{U}_0$ 为 $k\times k$ 的降序正定对角矩阵。选择 $k' \ll k$,保留矩阵 $\boldsymbol{U}_0$ 中的前 k' 最大值,保留 $\boldsymbol{T}_0$ 中的前 k' 行,前 k' 列,保留 $\boldsymbol{V}_0$ 中

收稿日期:2006 - 12 - 13;**修订日期:**2007 - 02 - 12 **基金项目:**甘肃省科技攻关计划资助项目(2GS047-A52-002-03)

作者简介:张秋余(1966 -),男,河北辛集人,副研究员,主要研究方向:信息安全、软件工程、多媒体通信; 刘洋(1979 -),男,辽宁丹东人,讲师,硕士,主要研究方向:文本分类.

的前 k' 列,分别得到矩阵 $\boldsymbol{U}$、$\boldsymbol{T}$、和 $\boldsymbol{V}$。它们的积 D' 是在 k' 阶平方误差内原矩阵的最相似矩阵,在 k' 维正交特征向量空间上描述了原矩阵的潜在语义结构。LSI 通过取 k' 阶近似矩阵,消减了原始词 - 文本矩阵中包含的"噪音" 因素,使词、文档空间向量大大缩减。

1.2　Local LSI

为解决在文本分类中 LSI 不能充分考虑分类信息的问题,Hull 最早提出了 LLSI 的思想。与 LSI 在整体词—文档矩阵上进行 SVD 不同,LLSI 在每类的局部区域上分别进行 LSI。局部区域的产生是 LLSI 的关键问题,对任意类别 c,局部区域中的任意文档向量 $\vec{d}_i$,它的产生过程如下:

$$\vec{d}_i = \vec{d}_i * f(\vec{d}_i) \tag{2}$$

$$\text{where } f(\vec{d}_i) = \begin{cases} 1, \vec{d}_i \text{ 满足类 } c \text{ 的 Local 条件} \\ 0, \text{otherwise} \end{cases} \tag{3}$$

若训练集有 n 个类别,则需要产生 n 个局部区域。与 LSI 只对整体词 — 文档矩阵进行一次 SVD 相比,LLSI 要进行 n 次 SVD,增加了计算开销。但每个局部区域的词 — 文档矩阵同整体词 — 文档矩阵相比要小的多,因此提高了计算速度,并且降低了对系统内存的要求,减少了内存不够用的出现,更适合在实际中应用。

2　支持向量机

SVM 是一种基于统计学习理论的二值分类学习方法。其基本思想是寻找一个能够对样本进行正确分类的最优分类面,同时使分类面两侧的空白区域间隔最大。对于线性不可分的问题,通过非线性变换将输入空间映射到一个高维特征空间,然后在这个新空间中求取最优分类面,而该非线性变换是通过定义适当的内积函数(核函数) 来实现的。

假设有一个两类问题的训练集 (x_i, y_i), $i = 1, \cdots, n$,其中 $x_i \in R^d$ 是第 i 样本的特征向量,$y_i \in \{+1, -1\}$ 是第 i 样本的类别标记。d 维空间中分类面方程为:

$$< w, x > + b = 0 \tag{4}$$

使间隔最大等价于使 $\| w \|$ 最小:

$$\text{Minimize } \frac{1}{2} \| w \| \tag{5}$$

而要求分类面对所有样本正确分类,则要求它满足:

$$y_i [< w, x > + b] - 1 \geqslant 0, i = 1, \cdots, n \tag{6}$$

因此,满足条件(6) 且使(5) 成立的分类面是最优分类面。利用拉格朗日乘子法求解上述问题得到:

$$w^* = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{7}$$

$$b^* = -0.5 < w^*, x_r + x_s > \tag{8}$$

其中,α_i 为拉格朗日系数。x_r、x_s 分别是来自两类的任意支持向量,且满足 $\alpha_r > 0, y_r = -1; \alpha_s > 0, y_s = 1$。

3　基于 SVM 的 LLSI(SVM-LLSI)

对于 LLSI,主要的问题是如何选取合适的局部区域作为 SVD 的基础。最常用的方法可以用某类的所有训练样本构成局部区域。但对该局部区域进行 SVD,相当于考虑了多词一义问题,而没有考虑一词多义问题。另外,考虑到数据集偏斜问题,在一个很小的局部区域上进行 SVD,作用并不明显。实验表明,用某类的所有训练集直接构成局部区域并不合适,该方法对分类效果的提高非常有限。该局部区域本质上是只有正例文档,而没有反例文档,因此,需要增加反例文档来平衡局部区域。

局部区域的产生可以看作是一个分类过程。用事先训练好的分类器,选择与该类别最相近的 n 篇文档作为局部区域。这 n 篇文档应该由两部分构成,一部分是正确分类的文档,它们能够很好地反映该类的特征,这部分占多数;另一部分是错误分类的文档,它们是与该类最相似但不属于该类的文档,这部分对于 LSI 解决一词多义问题是非常重要的,这部分占少数。

文本分类中,SVM 是目前分类效果最好的分类器,使用它来产生 LLSI 的局部区域。首先,使用原始训练集,采用"一对剩余"方法,对每类训练 SVM 分类器。然后,对每个二元分类器执行以下步骤:

1) 用公式(9) 选择 $f(x)$ 值最大的 n 篇文档作为局部区域;

$$f(x) = < w, x > + b \tag{9}$$

2) 在局部区域上进行 SVD,产生潜在语义空间;

3) 把除 n 篇文档以外的其他文档用公式(10) 转换到潜在语义空间,作为训练分类器的反例;

$$\bar{q} = q^T \times T \times U^{-1} \tag{10}$$

4) 用 2) 得到的语义空间作为正例,第 3) 得到的语义空间作为反例,重新训练分类器。

4　实验及结果分析

对于英文文本分类研究,国外有相对标准的语料库,这样就可以在共同的语料库上比较不同分类方法和系统的性能。而就中文文本分类而言,目前国内还没有标准、开放的语料库可供使用。因此,我们从 CNKI 上收集了 5 000 篇文献标题作为分类语料,这些文献标题可以分为法律、教育、农业、数学、医学共 5 个类别,每类有 1 000 篇。训练语料和测试语料按照 1:1 的比例来划分。

文本表示使用向量空间模型,TF-IDF 型权重作为权重表示方式。预处理阶段,中文分词部分,采用中科院计算所开源项目"汉语词法分析系统 ICTCLAS"系统,并按词性只保留了名词、动词和动名词。特征选择采用 χ^2 统计算法[9],在实验中,选取特征数为 1 000。

文本分类常用的评价指标是准确率、招回率和 F_1 测度[1]。准确率是判定属于某类的文本中,正确的判定所占的比例。招回率是对于某类文本,最终给出的判定结果中,正确的判定所占的比例。综合考虑准确率和招回率,可以得到 F_1 测度,计算公式如下:

$$F_1 = \frac{2pr}{p + r} \tag{11}$$

F_1 值越大,则分类器的性能越好。为评价多类分类器的整体性能,使用宏平均和微平均。宏平均是先对每个分类器求上述量度,再对所有分类器求平均,是关于类别的均值。微平均是先合并所有分类器的分类结果,得到一个总的结果,再计算上述量度,是关于文档的均值。

为验证 SVM-LLSI 方法的有效性,在不同特征维数下,对 SVM-LLSI 与 LSI 的微平均 F_1 值进行比较,结果如图 1 所示。其中,虚线表示 LSI,实线表示 SVM-LLSI。从图 1 可以看出,在任何特征维数下,SVM-LLSI 的整体性能较 LSI 都有明显的优势(提高大约 10%),这说明该方法能够充分考虑类别信息,弥补了传统 LSI 方法只注重表示的不足。

对于教育类和法律类样本,图 2 和图 3 分别给出了 LLSI

(类别训练样本直接构成局部区域)和 SVM-LLSI 在特征数目变化下的 F_1 值。从图 2 中不难看出,在开始阶段,随着特征数目的增加,分类效果越来越好,但到了 300 以后,不再有明显变化了;除了在特征数为 700 时,SVM-LLSI 比 LLSI 效果差以外,其余情况下的效果都要好;SVM-LLSI 方法在特征数为 500 时效果最好,而 LLSI 在特征数为 700 时效果最好,说明 SVM-LLSI 方法使用较少的特征就能获得最好的分类效果。为证实这一点,对所有类在两种方法下,取得最佳分类效果时的特征数进行了比较,结果如表 1 所示。

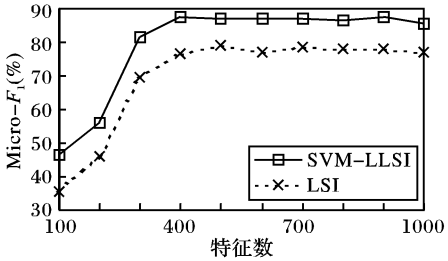

图 1 LSI 与 SVM-LLSI 在不同特征值上的微平均 F_1 值比较

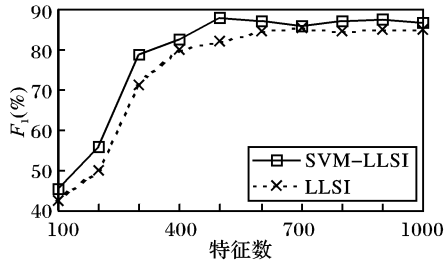

图 2 LLSI 与 SVM-LLSI 在教育类上的 F_1 值比较

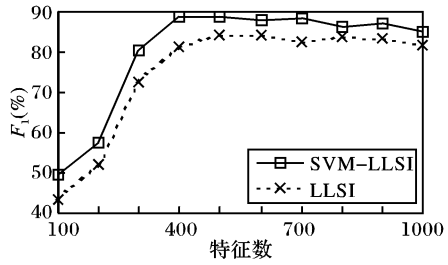

图 3 LLSI 与 SVM-LLSI 在法律类上的 F_1 值比较

表 1 LLSI 与 SVM-LLSI 取得最大 F_1 值时使用的特征数比较

类别	Max F_1 (%)		特征数	
	LLSI	SVM-LLSI	LLSI	SVM-LLSI
法律	84.17	88.73	500	400
教育	86.53	87.88	700	500
农业	85.04	89.68	600	400
数学	84.56	88.74	500	300
医学	86.63	86.23	400	400

从表 1 可以看出,除了医学类以外,其余所有类中,SVM-LLSI 获得最好分类效果时的特征数都要少于另一种方法。这对 LSI 很重要,因为 LSI 的另一个问题是计算复杂性问题,SVD 的时间复杂度和空间复杂度都很高,目前还没有更好的求解算法可以解决该问题,只能从降低问题的求解规模入手。显然,计算出 300 个奇异值的 SVD 比计算出 200 个奇异值的 SVD 要耗费更多的时间和空间,尤其在大规模文本集上体现得更明显。对每个类别,表 2 对两种方法取得最大 F_1 值时,在训练集上进行 SVD 需要的时间进行了比较。从表 2 可以看出,取得最大 F_1 值时,SVM-LLSI 需要的 SVD 时间要明显少于 LLSI,原因在于 SVM-LLSI 取得最大 F_1 值时需要较少的特征数。对于医学类,在特征数相同的情况下 SVM-LLSI 需要更多的时间,是因为 SVM-LLSI 的局部区域中的样本数比 LLSI 的局部区域中的样本数多。

表 2 取得最大 F_1 值时 SVD 需要的时间比较(单位:s)

类别	LLSI	SVM - LLSI
法律	110	90
教育	250	130
农业	160	90
数学	110	55
医学	80	90

5 结语

本文中,针对 LSI 方法在文本分类中的不足,提出一种 LLSI 的改进方法。该方法与以往简单使用某类的样本直接构成局部区域不同,利用 SVM 的分类优势来选取局部区域。这样选择的区域,能够更好地表示某类文档的潜在语义空间。实验结果表明,这是一种有效地文本分类方法,且使用了较少的特征和计算时间。但实验是在文献标题这样的小规模语料上进行的,实验的规模有待于进一步扩大。

参考文献:

[1] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1):1 - 47.

[2] 苏金树,张博锋,徐昕. 基于机器学习的文本分类研究进展 [J]. 软件学报, 2006, 17(9):1848 - 1859.

[3] DEERWESTER S, DUMAIS ST, LANDAUER TK, *et al.* Indexing by Latent Semantic Analysis[J]. Journal of the Society for Information Science, 1990, 41(6): 391 - 407.

[4] LIU T, CHEN Z, ZHANG BY, *et al.* Improving Text Classification using Local Latent Semantic Indexing[A]. Proceedingsof the 4th IEEE International Conference on Data Mining [C]. 2004. 162 - 169.

[5] SHIMA K, TODORIKI M, SUZUKI A. SVM-based feature selection of latent semantic features[J]. Pattern Recognition Letters, 2004, 25(2):1051 - 1057.

[6] 曾雪强,王明文,陈素芬. 一种基于潜在语义结构的文本分类模型 [J]. 华南理工大学学报, 2004, 32:99 - 102.

[7] ZELIKOVITZ S. TransductiveLSI for Short Text Classification Problems[A]. Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference[C]. 2004. 556 - 561.

[8] BURGES CJ. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121 - 167.

[9] 陈涛,谢阳群. 文本分类中的特征降维方法综述 [J]. 情报学报, 2005, 24(6):690 - 695.

[10] YANG YM, LIU X. A re-examination of text categorization methods [A]. Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. 1999. 42 - 49.

[11] MOSCHITTI A, BASILI R. Complex Linguistic Features for Text Classification: A comprehensive study[A]. Proceedingsof the 26th European Conferenceon Information Retrieval Research [C]. 2004. 181 - 196.

[12] CHEN L, TOKUDA N, NAGAI A. A new differential LSI space-based probabilistic document classifier[J]. Information Processing Letters, 2003, 88(5):203 - 212.

[13] KIM H , HOWLAND P , PARK H . Dimension Reduction in Text Classification with Support Vector Machine[J]. Journal of Machine Learning Research, 2005, 6(1):37 - 53.