文章编号:1001-9081(2005)03-0664-02

基于语义空间的支持向量机的文本过滤

沈丽虹¹,周昌乐²

(1. 浙江大学人工智能研究所,浙江 杭州 310027; 2. 厦门大学人工智能研究所,福建 厦门 361005)
(li-hong919@hotmail.com)

摘要:传统的基于支持向量机的文本过滤,用向量空间模型来表示文本和用户模板,向量空间模型假设特征项之间是线性无关的,该假设引入了许多因具体用词变化不定而带来的词汇噪音信息,影响了基于支持向量机的文本过滤的过滤性能。提出基于语义空间的支持向量机的文本过滤,用语义来表示文本和用户模板。该方法主要通过奇异值分解提取文本的潜在语义空间,在语义空间上训练支持向量机得到用户模板和过滤阈值,文本流上的文本映射到语义空间上,在语义空间上计算用户模板和新文本的相似度。实验表明:该方法的过滤性能可以达到 98.67%。

关键词:文本过滤;奇异值分解;支持向量机;语义空间

中图分类号:TP391.43 **文献标识码:**A

Text filtering based on support vector machine of semantic space

SHEN Li-hong¹, ZHOU Chang-le²

(1. Institute of Artificial Intelligence, Zhejiang University, Hangzhou Zhejiang 310027, China;
2. Institute of Artificial Intelligence, Xiamen University, Xiamen Fujian 361005, China)

Abstract: Traditionally, text filtering based on support vector machine uses the vector space model to represent the text and user profile. Vector space model draws the noise into the system because it assumes that the word in the text is independent and it influences the performance of the filtering. The proposed method was based on vector support machine of semantic space in which text and user profile were represented by the semantic space. The proposed approach used the singular-value decomposition to derive a latent semantic space. User profile and filtering threshold could be got by training the support vector machine in the semantic space. And the similarity between the user profile and new text was computed by cosine measure, after the new text was mapped into the semantic space. Experimental results show that the filtering rate of our approach can get 98.67%.

Key words: text filtering; singular value decomposition; support vector machine; semantic space

随着在线文本的日益增多,文本过滤技术已经成为信息处理领域中一个相当活跃的分支。文本过滤技术研究的主要问题是^[1]:如何准确地表达用户需求(profile),进而在大规模的文本流中选出与用户需求相关的文本,并把选出的文本传递给用户。文本过滤和文本检索有很大的相似性,其区别在于文本检索有着相对固定的文本库和千变万化的检索需求,而文本过滤则有着相对固定的用户需求和动态变化的文本流^[1]。在文本过滤中,从动态的文本流中选出与用户需求相关的文本可以看作是一个二类的分类问题。文本过滤中使用的分类方法主要有:贝叶斯方法、k-近邻方法、神经网络方法、决策树方法以及支持向量机方法^[2,3],支持向量机的分类效果优于其他的分类方法^[2]。基于支持向量机的文本过滤^[2,3]采用向量空间模型来表示文本,向量空间模型假设特征项之间是线性无关的,这影响了过滤性能,因为在自然语言中词与词之间存在着错综复杂的关系,很难满足假设条件。

本文提出一种基于语义空间的支持向量机的文本过滤方法,这种方法采用文本的潜在语义结构来表示文本,能有效地解决用向量空间模型表示文本的缺点。这种方法降低了支持向量机的误检率,提高了过滤性能。实验表明,用文本的潜在语义结构来表示文本,支持向量机的误检率可降至 1.67% 及过滤性能(用 F 度量)可达到 98.67%。

1 支持向量机

支持向量机^[4]是 Vapnik 等人根据统计学习理论提出的一种新的通用学习机器,它的核心思想是保持经验风险值固定而最小化置信区间。支持向量机的学习问题就是寻找这样的最优超平面,它使不同类别数据之间的分类间隔最大,并且使位于这个平面上的数据 x 到平面 $w \cdot x - b$ (w 是平面的法向量, b 是平面到原点的距离) 的距离为零。

假定训练数据:

$$S = \{x_i, y_i\} (1 \leq i \leq l), L_i \in R^n, y_i \in \{+1, -1\}$$

对于线性分类平面,样本集 S 满足约束条件:

$$y_i(x_i \cdot w - b) - 1 \geq 0, i = 1, 2, \dots, l \quad (1)$$

容易验证,最优超平面就是满足约束条件(1)式并且使得 $\Phi(w) = \|w\|^2$ 最小化的超平面。求最优超平面是一个二次规划问题,可以用拉格朗日方法求解得到:

$$L = \frac{1}{2} \|w\|^2 - \sum_{i=1}^l \alpha_i y_i (x_i \cdot w - b) + \sum_{i=1}^l \alpha_i$$

其中 α_i 是拉格朗日乘子,每个 $\alpha_i > 0$ 所对应的样本叫作“支持向量”。支持向量是训练样本中的关键元素,如果只保留支持向量样本,而删除所有其他训练样本重新进行训练,所

收稿日期:2004-08-11;修订日期:2004-10-26 基金项目:福建省科技计划重点资助项目(001J005)

作者简介:沈丽虹(1980-),女,浙江宁波人,硕士研究生,主要研究方向:人工智能、信息过滤、机器学习;周昌乐(1959-),男,山东人,教授,博士生导师,主要研究方向:人工智能、计算语言学、认知逻辑、多媒体技术。

得到的最优超平面保持不变。但是在实际情况中可能存在一些样本不能被最优超平面正确分类,因此在式(1)中引入松弛因子 ξ_i :

$$y_i(x_i \cdot w - b) - 1 + \xi_i \geq 0, i = 1, 2, \dots, l \quad (2)$$

对于线性不可分数据,支持向量机通过核函数把数据从低维空间映射到高维空间,然后在高维空间为低维数据构造线性可分超平面。支持向量机中最常用的3种核函数是多项式函数、径向基函数和多层神经网络函数。

2 语义提取和文本映射

通过对项/文本矩阵进行奇异值分解,生成一个只包含若干个正交因子的降秩空间,该空间就是文本的语义空间。

2.1 奇异值分解

对于任意的 $t \times d$ 的矩阵 X 都可以分解为:

$$X = USV^T \quad (3)$$

其中 U 由 X 的左奇异向量组成 (t 行 m 列), V 由 X 的右奇异向量组成 (d 行 m 列), S 是一个对角矩阵 (m 行 m 列), S 中的对角元素按序排列,并且单调递减。 U, V, S 是满秩矩阵, m 是矩阵 X 的秩。奇异值分解的优点是可以利用较小的矩阵做到最优的近似,奇异值降维处理后的 X 用 X_k 来表示:

$$X \approx X_k = U_k S_k V_k^T \quad (4)$$

其中 S_k 是由 S 中前 k 个对角线元素组成的对角阵, U_k 和 V_k 分别由 U 和 V 的前 k 列构成。

通过奇异值降维后, U_k 和 V_k 的行向量分别为 k 维语义空间中特征项和文本的表示。其中 k 的选择是至关重要的,一方面希望 k 足够大能够覆盖所有的潜在语义结构,但如果太大,则可能引入噪音,对于过滤将产生一定的影响;如果 k 太小,则保留下来的重要结构太少,无法把握运算的结果^[5]。

2.2 新文本映射

在文本过滤中,如何将原始分析中未出现的新文本映射到语义空间上呢?实际上,对于任意一个新文本,均可以通过下面的转换公式,将它映射到语义空间上^[5]。

$$d = x_n^T U_k S_k^{-1} \quad (5)$$

其中 x_n 表示新文本的在原空间的向量表示, d 表示文本在语义空间中的表示。

3 基于语义空间的支持向量机的文本过滤

传统的基于支持向量机的文本过滤采用向量空间模型来表示文本。向量空间模型把文本表示成 N 维欧式空间的向量(向量的维数就是特征项的个数,向量的每个分量就是特征项在文档中出现的次数)。特征项可以是字、词或短语。数学表示如下:

$$X = [x_{i,j}] \in R^{t \times d}, i = 1, 2, \dots, t; j = 1, 2, \dots, d$$

(t 表示特征项的个数, d 表示文本个数)

矩阵的行向量表示 T 文本集中 t 个不同的特征项,列向量表示文本集中 d 个不同的文档。 $x_{i,j}$ 表示特征项 i 在文本 j 中出现的次数。向量空间模型假设特征项之间是线性无关的,可是自然语言中存在同义词和多义词等现象,所以用向量空间模型表示文本,会引入许多因具体用词变化而带来的词汇噪音信息,影响了支持向量机在文本过滤中的分类性能。本文提出基于语义空间的支持向量机的文本过滤。对矩阵 X 进行奇异值分解,生成一个只包含若干个正交因子的降秩空间 X_k , X_k 与 X 所体现的特征信息一致,同时还能体现了 X 的语义结构,从而剔除了其中因具体用词变化而带来的词汇噪音信息,通过式(5)将新文本映射到语义空间上,在语义空间中进行文本相似度计算,如果新文本和用户模板的相似度大于阈值,则判定为相关,否则判定为不相关。该方法能有效提高支持向

量机在文本过滤中的分类效果,从而提高了文本的过滤性能。算法描述如下:

1) 对项/文本矩阵进行奇异值分解降维后,得到三个矩阵 U_k, S_k, V_k (U_k 和 V_k 分别表示索引项和文本在语义空间中的表示。)

2) 利用 V_k 训练支持向量机得到最优超平面 (w_k, b_k), w_k 表示用户模板在语义空间的表示, b_k 表示在过滤过程中的阈值。

3) 利用 U_k 和 S_k , 根据重构公式,得到新文本 x_n 在语义空间的具体表示 d 。

4) 计算用户模板和新文本的相似度:

$$\text{sim}(w_k, d) = w_k \cdot d$$

5) 过滤: 如果 $\text{sim}(w_k, d) \geq b_k$, 则该文本与用户需求相关, 否则不相关。

4 实验过程及结果

4.1 实验的数据集

本实验使用 Reuters-21578 数据集, 该数据集是由 David Lewis 汇编而成的, 数据集可以在 <http://www.research.att.com/tewis> 得到, 其中主要的切分方式为“ModeApte”。利用其中的 9603 篇作为训练集, 3299 作为测试集。本实验有 2000 篇训练文本 (其中 1000 篇是正例, 另外 1000 篇是反例), 有 600 篇测试文本 (其中 300 篇作为正例, 另外 300 篇作为反例)。文本的预处理包括: 剔除停用词, 提取词干, 使用 IFIDF 进行加权, 并进行归一化。

本试验主要采用查准率 (precision)、查全率 (recall) 和 F 度量 ($F_measure$) 这三个评价指标。其中 $F = 2pr/(p+r)$, p 表示查准率, r 表示查全率。

4.2 实验结果及分析

本实验用基于向量空间模型的支持向量机作为对比参考。在实验中, 我们 k 取值为 300, 因为在一般情况下 $k = 300$ 时, 效果最好^[5]。

表1 SVM与SVM_SS的性能比较

文本过滤方法	查准率(%)	查全率(%)	F(%)
SVM	97.29	96.32	96.80
SVM_SS(k=300)	98.34	99.00	98.67

SVM 表示传统的基于支持向量机的文本过滤方法, SVM_SS 表示基于语义空间的支持向量机的文本过滤, k 表示语义空间的维数。

从表1可以看出, 跟基于向量空间的支持向量机的文本过滤方法相比, 基于语义空间的支持向量机方法的过滤效果明显提高。

参考文献:

- [1] BELKIN N, CROFT BW. Information filtering and information retrieval: two sides of the same coin? [J]. Communications of the ACM, 1992, 35(12): 29-38.
- [2] JOACHIMS T. Text Categorization with Support Vector Machines: learning with many relevant features [A]. Proceedings 10th European Conference on Machine Learning [C], 1998. 137-142.
- [3] LEOPOLD E, KINDERMAN J. Text Categorization with Support Vector Machines, How to Represent Texts in Input Space? [J]. Machine Learning, 1998, 46(1-3): 423-444.
- [4] VAPNIK VN. 统计学习理论的本质 [M]. 张学工, 译. 北京: 清华大学出版社, 2000. 91-101.
- [5] DEERWESTER S, DUMAIS ST, FURNAS GW, et al. Indexing by Latent Semantic Analysis [J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.