基于隐含语义索引的文本情感序列回归方法①

刘贤友 [1,2], 孙丙宇 [1,2], 李文波 [2], 汪超永 [1,2]

[1](中国科学技术大学 自动化系, 合肥 230036)
[2](中国科学院 合肥智能机械研究所, 合肥 230031)

摘 要: 传统上, 文本情感分析技术仅限于情感分类, 即仅局限于简单的将评论分为正面或负面两类. 而在实际中,有时更需要将评论进行分级, 比如把商品划分为“好”、“中”、“差”、“极差”等若干个级别, 以便更准确表达评论者的情感; 现有的情感分类方法无法解决评论分级问题. 为此, 提出了基于潜在语义索引的评论文本情感序列回归方法, 首先采用潜在语义索引对评论文本进行特征变换, 并在此基础上采用核判别学习序列回归方法进行序列回归, 实现对评论文本的情感分级. 通过在 Movie Reviews 数据库的实验, 验证了提出方法的有效性.

关键词: 情感分析; 序列回归; 隐含语义索引

Review Text Sentiment Ordinal Regression Based on Latent Semantic Index

LIU Xian-You[1,2], SUN Bing-Yv[1,2], LI Wen-Bo[2], WANG Chao-Yong[1,2]

[1](Department of Automation, University of Science and Technology of China, Hefei 230026, China)
[2](Institute of Intelligent machine, Chinese Academy of Sciences, Hefei 230031, China)

Abstract: Traditionally, text sentiment analysis is only limited to sentiment classification. The review is simply divided into two types: positive and negative comments. In practice, we sometimes need to rank the review, which cannot be solved by traditional sentiment classification methods. To solve this problem, this paper proposes a novel review text sentiment ordinal regression based on Latent Semantic Index. Firstly latent semantic indexing is used to extract features for review texts and then an ordinal regression method is used for review text sentiment analysis. The experimental results on Movie Reviews database proved the effectiveness of the proposed method.

Key words: sentiment analysis; ordinal regression; latent semantic index

1 引言

随着互联网技术, 特别是随着 Web2.0 应用的增多, 网络上对各种产品以及热点事件的评论变得更加方便. 评论文本情感分析是指通过挖掘和分析评论文本中的立场、观点、情绪等主观信息, 对评论文本的情感倾向作出分析判断. 一方面, 针对产品的评论分析对于商家或买家具有较大价值; 一方面, 热点事件的评论分析对于政府了解网民对特定事件的观点也十分重要. 情感分析作为一项新兴技术已经在这些领域得到了大量的研究.

目前情感分析的研究主要集中在情感分类, 即将人们的情感分为正面情感和负面情感. 情感分类采用的主要方法分为两种: 基于机器学习的方法[1-3]和基于语义的方法[4-5]. 基于机器学习的方法将情感分析问题看作是一个分类问题, 标注好的训练集通过机器学习算法训练得到分类模型, 用于以后的情感分类. 基于语义的方法将表示情感的词语分为正面情感词语和负面情感词语, 构造一个情感词典, 然后通过计算一个句子中的正负情感词语的相对数量决定句子的情感倾向. 当前很多研究结果表明[1-2], 基于机器学习的方法性能优于基于语义的方法. 目前应用于文本情感分类的机器学习方法主要有支持向量机、朴素贝叶斯、K-

① 基金项目:国家科技支撑计划(2012BAH89F02)
收稿时间:2013-11-21;收到修改稿时间:2013-12-16

近邻、决策树等, 并且在实际中均取得一定效果.

然而, 尽管基于机器学习的情感分类研究已经取得较大进展, 但这些方法仅局限于简单的将评论分为正面或负面两类. 在实际中, 我们有时更需要将评论进行分级, 尤其是当需要根据推荐结果对产品或事件进行排序时. 比如我们可能需要对电影的评论进行“好”、“中”、“差”、“很差”四个级别, 从而在推荐时让用户有更多选择. 由于评论分级具有明显的序列信息, 传统的分类方法无法建立不同类别之间的序列关系.

为解决上述问题, 本文将序列回归方法引入到评论情感分类问题中: 不是简单的预测评论 “正面”或者“负面”, 而是对评论结果进行分级. 在评论文本情感序列回归方法研究中, 一个简易的方法是采用标准的分类或回归模型, 但一方面由于分类方法只能简单的预测属于哪个类别, 无法考虑不同类别的序列信息; 另一方面由于不同类别的序列之间的真实距离在大多数情况下是未知的, 因此采用传统的回归方法难以建立合适的样本到序列标记之间的映射[6].

本文其余部分安排如下: 第 2 节介绍基于语义的评论文本情感序列回归方法, 第 3 节介绍具体的实验设计, 第 4 节对实验结果进行分析和讨论, 最后一节是本文结论和未来工作的展望.

2 基于隐含语义索引的评论文本情感序列回归方法

2.1 方法流程

图书评论情感序列回归分为训练与测试两部分. 在训练时, 需要对已标注标签样本进行特征提取、特征选择等预处理, 并定义相应的相似度计算函数; 然后用这些特征集建立序列回归模型. 在测试阶段, 测试样本经过特征提取与特征选择后, 输入到序列回归模型, 得到预测结果.

2.2 评论文本表示方法

评论情感序列回归的第一步需要把数据集中的文本表示成特征向量, 这样才能使用序列回归算法进行序列回归. 目前文本表示中一个常见,X- .=,,, - .., -, ..= .的方法是向量空间表示模型(Vector Space Model, VSM), 相关研究主要集中在两点: 1)以什么语意单元作为特征; 2)如何计算特征的权重[7,8]. 基于 VSM 的语句相似度计算是一种统计的方法, 考察的是问题和答案句子中关键词的词频信息, 没有考虑词在上下文中的语义信息, 而且由于进行句子之间的相似度计算, 句子通常比文档含的词少, 采用 VSM 方法存在严重的数据稀疏问题, 因此具有一定的局限性. 为此本文采用基于潜在语义索引(Latent Semantic Index, LSI)的方法进行语句的相似度计算[9].

在 LSI 模型中, 一个文档库可以表示为 $m\times n$ 词-文档矩阵 $A=\alpha_{ij}$, 其中 n 为文档库中的文档数, m 为文档库中包含的所有不同的词的个数, 也就是说, 每一个不同的词对应于矩阵 A 的行, 而每一个文档则对应于矩阵 A 的一列; α_{ij} 表示第 i 个词在第 j 个文档的权重. 客观上, 由于词和文档的数量都很大, 而单个文档中出现的词非常有限, 所以 A 一般为稀疏矩阵. 词-文档矩阵 A 建立后, 可以利用奇异值分解计算 A 的 k 秩近似矩阵 $A_k(k<<min(m,n))$. 经奇异值分解后, 矩阵 A 可表示为三个矩阵的乘积: $A=UEV^T$. 式中, U 和 V 分别是与矩阵 A 的奇异值对应的左、右奇异向量矩阵, E 为矩阵 A 的奇异值按递减顺序排列构成的对角阵, 取 U 和 V 的前 k 行 k 列构成 A 的 k 秩近似矩阵 $A_k=U_kEV_k^T$, 其中 U_k, V_k 均为正交向量. k 只要取一个比较小的值, 得到的语义空间就可以表示原始矩阵 A 的大部分信息.

LSI 将关键词和文档映射到同一个 k 维空间内, 反映出词语在文档中的使用模式, 在保持了大部分信息的同时, 很大程度上克服了传统文本分类系统中多义词、同义词和单词依赖现象, 同时, 在新的语义空间中进行相似度分析是由于它是基于语义层而不是词汇层的分析, 因此, 比传统的向量分析法具有更好的效果.

对于一个新的测试文本 d_j 要在基于潜在语义分分析模型中进行比较, 需要将 d_j 投影到 LSI 的语义空间中去, 具体如下式所示:

$$\overline{d_j}=d_jU_kE_k^{-1} \tag{1}$$

2.3 评论文本序列回归方法

本文采用基于核判别学习的序列回归方法(Kernel Discriminant Learning for Ordinal Regression, KDLOR)[6]进行序列预测. 设目前已有 N 个文档, 每个文档经过LSI变换后可以表示为 $\{x_i,y_i\}(i=1,\cdots,N)$, 其中 x_i 为第 i 个文档, $y_i\in\{1,\cdots,k\}$ 为相应的序列标签, $X_k=\{x_i\,|\,y_i=k\}$. 序列回归算法可以视为通过寻找一个投影方向, 在此投影方向不同样本的序列标记能够保持, 具体如下式所示:

$$\min J(w,\rho)=w^T s_w w-C\cdot\rho \tag{2}$$

$$\text{s.t.}\quad w^T(m_{k+1}-m_k)\geq\rho,k=1,\cdots,k-1$$

其中 C 为惩罚系数，w 为要寻找的投影方向，$S_\omega=\frac{1}{N}\Sigma_{k=1}^{K}\Sigma_{x\in X_k}(x-m_k)(x-m_k)^T$ 为类间散布矩阵，$m_k=\frac{1}{N}\Sigma_{x\in X_k}x$.

上述最优化问题一方面最小化同类样本的类内散射矩阵，同时最大化不同类别样本之间的距离. 如果 $\rho>0$，则不同类别样本投影后可以保持其原有序列信息. 为求解上述最优化问题，我们可以通过构造相应的拉格朗日方程进行求解. 另一方面，为了解决非线性问题，我们可以通过核技巧把上述最优化问题进行转化. 对于样本 x，可以通过一个映射函数把其映射到一个高维特征空间 $\varphi:x\to\varphi(x)$；投影方向可以表示为:

$$w=\Sigma_{i=1}^{N}\beta_i\varphi(x_i),\beta_i\in R \tag{3}$$

将(3)代入最优化问题(2)，可得到:

$$\min J(\beta,\rho)=\beta^T\cdot H\cdot\beta-C\cdot\rho \tag{4}$$

$$\text{s.t.}\quad \beta^T\cdot(M_{k+1}-M_k)\geq\rho,k=1,\cdots,K-1$$

其中 $(M_k)_j=\frac{1}{N}\Sigma_{x\in X_k}\varphi(x_j)\cdot\varphi(x)$，

$H=\Sigma_{k=1}^{K}P_k(I-\mathbf{1}_{N_k})P_k^T$，$P_k$ 是一个 $N\times N_k$ 矩阵，

$(P_k)_{i,j}=\varphi(x_i)\cdot\varphi(x_j),x_j\in X_k$，$I$ 是单位矩阵，$\mathbf{1}_{N_k}$ 是所有元素值为 $1/N_k$ 的矩阵. 通过构造核函数 $S_{i,j}=\varphi(x_i)\cdot\varphi(x_j)$，上述问题可以转换为线性空间求解.

3 实验设计

为了验证序列回归方法在评论文本情感分析领域应用的有效性，本文选择经典的语料库 Movie Reviews[10]进行验证. 语料库中所有的评论文本都通过自动预处理消除文本中含有明显情感级别的语句. Movie Reviews 数据集共有 4 个评论文本库，每个文本库包含的评论文本数分别为 1770, 902, 1307, 1027. 同一个文本库的评论源于同一个作者，因此 4 个评论文本库对应于 4 个作者. 每个文本库的评论文本被划分为 4 个级别{0,1,2,3}. 具体实验流程如下:

1)文本预处理. 主要是对评论文本进行停用词处理，即去停用词. 在对英文文本进行情感分析过程中，去停用词指的是过滤掉一些频繁使用但没有实际意义的词，目的是为了降低特征选择属性维度，以减少系统的计算量，提高分析的效率;

2)维数约减. 利用一定方法，从预处理数据中抽取出若干最有利分类的特征项，并把每个文档表示成特征向量的形式. 本文是采用 LSI 的方法进行维数约减.

3)学习训练. 选择采用的文本库的若干评论文本构成训练样本集进行训练，得到序列回归模型. 本文随机选择 60%样本进行训练, 40%样本进行测试，重复 10 次. 模型中各算法参数采用交叉验证法进行选取;

4)测试评价. 用学习建立的序列回归模型对实验中的测试样本进行测试，并选择合适的评价指标进行评价. 本文实验采用平均绝对误差(Mean Absolute Error, MAE)来进行评价，即 $\frac{1}{N}\Sigma_{i=1}^{N}|y_i'-y_i|$，其中 y_i'，y_i 分别为样本 x_i 的预测类别与真实类别.

4 结果分析

本研究的主要目的在于验证隐含语义索引与序列回归方法在文本情感分析中的有效性，根据上节的实验设计，主要实验结果如表 1 所示. 在实验中，我们分别对比了采用 VSM 与 LSI 进行特征表述情况下, SVM 回归、SVM 分类与序列回归等不同方法的性能[11-12]. 由表一可以看到，在特征表达方面，LSI 方法的性能要优于 VSM 方法. 同时，序列回归算法在两种特征表述方法下的性能均优于 SVM 回归和 SVM 分类. 由于 SVM 分类方法仅考虑样本的不同类别，没有考虑类别之间的序列信息，因此其预测效果比另外两种方法要差. 因此，采用隐含语义索引与序列回归方法可以提高文本情感分析的性能.

表 1　序列回归性能对比

数据集	VSM			LSI		
	SVM 分类	SVM 回归	序列回归	SVM 分类	SVM 回归	序列回归
作者 A	0.54	0.58	0.61	0.59	0.62	0.67
作者 B	0.44	0.49	0.54	0.47	0.52	0.58
作者 C	0.59	0.59	0.60	0.64	0.65	0.68
作者 D	0.46	0.48	0.52	0.49	0.51	0.55

5 结论

随着网络技术的发展，大量用户在互联网上的评

论信息迅速膨胀, 给人工收集和处理海量评论信息带来极大困难. 评论文本情感分析技术可以通过计算机辅助完成相关评价信息的快速整理和分析, 给商家对产品的了解与政府对特定事件的分析带来了极大的便利. 传统上, 文本情感分析技术仅限于情感分类, 即仅局限于简单的将评论分为正面或负面两类. 而在实际中, 我们有时更需要将评论进行分级, 现有的情感分类方法无法解决分级问题. 为此, 本文提出了基于潜在语义索引的评论文本情感序列回归方法, 首先采用潜在语义索引对评论文本进行特征变换, 并在此基础上采用 KDLOR 方法进行序列回归. 在进一步的研究中, 还需要在其它数据集以及实践中对本文结论进行验证.

参考文献

1 Bo P, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning chniques. Proc. of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2002. 79–86.

2 Tan SB, Zhang J. An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications, 2008: 2622–2629.

3 Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources. Proc. of Methods in Natural Language Processing. Barcelona, Spain, 2004. 412–418.

4 Hatzivassiloglou V, McKeown K. Predicting the semantic orientation of adjectives. Proc. of the 35th Annual Meeting of the Association for Computational Linguistics (ACL), 1997. 174–181.

5 Turney PD. Thumbs up or thumbs down?Semantic orientation applied to unsupervised classification of reviews. Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002. 417–424.

6 Sun BY, Li J, Wu DD, Zhang XM, Li WB. Kernel discriminant learning for ordinal regression. IEEE Trans. Knowl. and Data Eng., 2012. 906–910.

7 吕林涛,董迎.基于上下文的概念语义相似度计算模型.计算机工程,2010,36(2l):59–61.

8 苏金树,张博锋,徐昕.基于机器学习的文本分类研究进展.软件学报,2006,17(9),1848–1859.

9 Letsche TA, Berry MW. Largescale information retrieval with latent semantic indexing. Information Sciences, 1997, 100(1): 105–137.

10 Pang B, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, Proc. of ACL. 2005.

11 Vapnik V. 1995. The Nature of Statistical Learning. Theory. Springer.

12 Joachims T. Text categorization with support vector machines: earning with many relevant features. Proc. of 10th European Conference on Machine Learning, 1998. 137–142.