

# 基于聚类核的半监督情感分类算法研究

郑文静,李 雷

(南京邮电大学 理学院,江苏 南京 210023)

**摘 要:**在互联网快速发展的今天,人类已经进入“大数据”时代,其中文本数据作为人类知识的载体,对于人类的进步与发展意义重大。如何运用大量未标记样本来提升文本情感分类的精度,也变得愈发重要。将半监督学习中的聚类核算法应用到情感分类问题中,给出基于聚类核的半监督情感分类算法。在标记样本和未标记样本上,建立加权无向图,求解聚类核,然后将该核函数用于 SVM 的情感分类器的训练上,完成情感分类工作。该方法直接将未标记样本所蕴含的信息融合到核中,不需要建立多个分类器,有效利用了未标记样本。实验结果表明,CKSVM 算法在分类精度上明显优于基于 Self-learning SVM 和 Co-training SVM 的半监督情感分类算法,且在不同数据集上都有较好的适应性。

**关键词:**半监督学习;聚类核;图;情感分类

**中图分类号:**TP301.6

**文献标识码:**A

**文章编号:**1673-629X(2016)12-0087-05

**doi:**10.3969/j.issn.1673-629X.2016.12.019

## Research on Semi-supervised Sentiment Classification Based on Cluster Kernel

ZHENG Wen-jing, LI Lei

(School of Science, Nanjing University of Posts and Telecommunications,  
Nanjing 210023, China)

**Abstract:** In the rapid development of the Internet today, mankind has entered the era of big data. Text data as the carrier of human knowledge, is of great significance for human progress and development. So the usage of a large number of unlabeled samples to improve the accuracy of sentiment classification, has become more and more important. The kernel clustering method in semi supervised learning is applied to the emotion classification problem, and a semi supervised sentiment classification algorithm based on kernel clustering is proposed. A weighted undirected graph is built according to the labeled samples and unlabeled samples, solving the clustering kernel, and then the kernel function is used for the training of classifier SVM. This method directly uses the information contained by unlabeled samples into the kernel, no need to set up multiple classifiers, effective usage of the unlabeled samples. Experimental results show that the CKSVM is better than that based on Self-learning SVM and Co-training SVM in classification accuracy, with better adaptability on different data sets.

**Key words:** semi-supervised learning; clustering kernel; graph; sentiment classification

### 0 引 言

随着互联网的发展,越来越多的消费者在网上发表评论<sup>[1]</sup>,这些评论以主观的文本形式表达了消费者对于消费产品或服务的满意度。这不仅可以帮助其他消费者做出更好的判断,还可以帮助制造商跟踪和管理这些意见<sup>[2]</sup>。在自然语言处理(NLP)中,情感分类作为一种特殊的文本分类问题正受到越来越多的重视<sup>[3-6]</sup>。情感分类的标准是挖掘文本中蕴含的极性情感

态度,如“positive” or “negative”, “thumbs up” or “thumbs down”, “favorable” or “unfavorable”<sup>[7]</sup>,而不是主题。如今,情感分类技术广泛应用于商业智能系统、推荐系统、公众的意见收集和挖掘系统等等。在上述领域,存在着丰富的未标记文本数据,标记的文本数据很少,且需要通过人工标注获得。这使得使用许多传统的算法训练数据的代价过高,因为这些分类器要求足够的标记数据来保证实现高精度。

收稿日期:2016-02-27

修回日期:2016-06-15

网络出版时间:2016-11-22

**基金项目:**国家自然科学基金资助项目(61070234,61071167,61501251);南京邮电大学引进人才科研启动基金资助项目(NY214191)

**作者简介:**郑文静(1990-),女,研究方向为机器学习、情感分类;李 雷,博士,教授,研究方向为智能信号处理、非线性分析与计算智能、机器学习。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20161122.1228.040.html>

为使用未标记样本提高情感分类的精度,提出一种基于聚类核的半监督情感分类算法。基于聚类假设,Chapelle 提出聚类核概念<sup>[8]</sup>,使用核函数,而不是明确的特征向量,重新表示给定的数据,以反映未标记数据透露的结构。首先通过 bag-of-words 模型将数据集中的评价转化为向量形式,并将这些向量作为图的顶点,将各评价间的相似度作为边上的权重,构建加权无向图。然后引入线性转换函数,将该图上的相似矩阵重新表示,使得在同一集群中两点间的距离更小,建立半监督聚类核,并将其用于 SVM 分类器的训练上。

实验结果表明,该算法分类精度较高,且在分类精度上优于基于 Self-learning SVM 和基于 Co-training SVM 的半监督情感分类算法。

## 1 相关研究

### 1.1 情感分类

一般来说,情感分类技术可分为基于机器学习(Machine Learning, ML)的情感分类方法、基于词典的方法和混合方法<sup>[9]</sup>,具体见图 1。

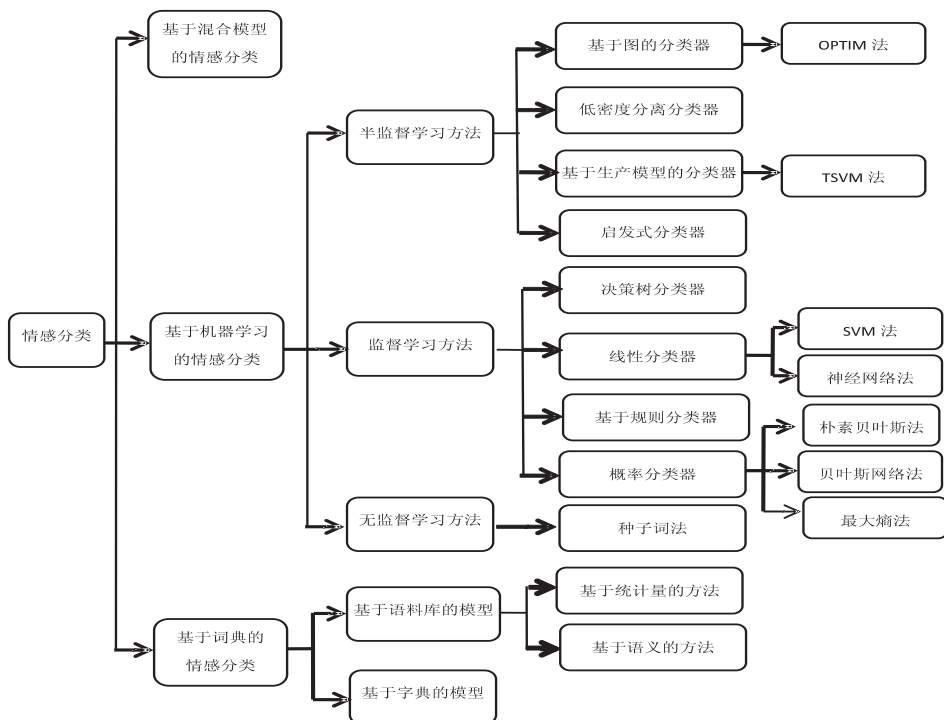


图 1 情感分类算法

基于机器学习的情感分类算法,主要是经典机器学习算法结合语义特征的应用<sup>[10]</sup>,大体上可分为基于监督学习的情感分类方法、基于半监督学习的情感分类方法和基于无监督学习的情感分类方法。基于监督学习的情感分类方法依赖于已存在的标记训练样本,包括概率分类器<sup>[11-13]</sup>、线性分类器<sup>[14-18]</sup>、决策树分类器<sup>[19-20]</sup>和基于规则的分类器<sup>[10,21]</sup>等模型。2002 年, Pang 等<sup>[6]</sup>率先使用监督式机器学习方法中的支持向量机(Support Vector Machine, SVM)对电影评论进行情感分类,并且对比了朴素贝叶斯(Naïve Bayes, NB)、最大熵分类法(Maximum Entropy Classification, MEC)和 SVM 三种监督学习算法在情感分类问题中的性能<sup>[6]</sup>。由于情感分类问题一般是将文档划分到一定数量的预定义类别中,且标记样本较难获得<sup>[9]</sup>,因此基于无监督学习的情感分类方法引起了广泛关注。例如,Turney 使用 PMI-IR(Pointwise Mutual Information and Information Retrieval)方法进行消费者评论的情感分类<sup>[22]</sup>。

半监督学习使用少量标记样本和大量未标记样本,可以有效提高分类精度,大大减少人工标记的工作量,因此越来越多的研究者将其与情感分类问题相结合,取得了较好效果。Goldberg 和 Zhu<sup>[23]</sup>提出了一种基于图的半监督学习算法来处理评分系统的情感分析问题,其中不同数值的评分对应着给定的情感。Sindhwani 和 Melville<sup>[24]</sup>将文本先验信息同未标记样本相结合,给出了一种半监督情感分类算法。Zhou 等运用主动学习来解决半监督情感分类问题,提出了主动深度网络方法(Active Deep Networks, ADN)<sup>[25]</sup>,并进一步提出了模糊深度置信网络法(Fuzzy Deep Belief Networks, FDBN)<sup>[26]</sup>。基于基本假设:具有相似情感倾向的情感词有较高的概率出现在同样情感倾向的消费者评论中,文献[27]提出一种基于特征聚类的半监督式情感分类方法。该方法根据情感特征的共现关系构建共现矩阵,利用 spectral 聚类方法生成分类用的扩展特征,结合原有训练域内的分类特征来训练新的情感

分类器形成两个分类器,共同完成最后的情感分类工作。

## 1.2 聚类核

聚类假设<sup>[28]</sup>是半监督学习的核心,是建立目标函数与未标记样本分布之间关系的枢纽<sup>[5]</sup>,指的是同一聚类中的样本点很可能具有相同的类别标签,即在高密度区域里,如果某两个点可以通过区域内某条路径相连接,那么这两个样本点的标签相同的可能性比较大。而聚类核<sup>[8]</sup>依赖聚类假设思想,使用核函数重新表示给定的数据,从而将未标记数据中的结构加入到分类器中<sup>[29]</sup>。其主要思想在于改变距离度量,使同一群集中两点挨得更紧,距离更小<sup>[30]</sup>。

构造聚类核的整体框在文献[8]中提出。本质上,聚类核来源于核矩阵的能量本征谱,其中两种比较典型的方法是随机游走核<sup>[8]</sup>及谱聚类核<sup>[8]</sup>。在一个标准化且对称化的随机游动过程中,随机游走核是它的  $t$  步转移矩阵。Szummer<sup>[31]</sup>指出,在以  $x_i$  为顶点的图上,随机游动过程的转移矩阵可以是 RBF 核,由此定义的随机游走,就可以通过一步转移矩阵求解出  $t$  步随机游走核。谱聚类核的主要思想依据是谱聚类,即在特征空间中,计算出样本间相似度矩阵的谱分解后,就可以重新表示样本点了。这样一来,位于同一聚类区域中的样本点更加紧凑地分布在新的空间中。文中将聚类核算法应用到情感分类中,在样本集上求解出核函数后,与 SVM 分类器的训练相结合,提出了基于半监督聚类核的情感分类算法(Cluster Kernel based SVM for sentiment classification, CKSVM)。

## 2 基于聚类核的半监督情感分类算法

为了更好地满足聚类假设,减少分类器的使用,文中提出基于聚类核的半监督情感分类算法。在构建基于文本数据集的加权无向图之后,使用线性分段转换函数将图上的相似矩阵重新表示,利用该半监督聚类核训练的基于 SVM 情感分类器有着更好的分类效果。

假设给定一个来自某未知分布的文本数据集  $X = \{x_i\}_{i=1}^n$ , 样本总数为  $n$ ;  $C = \{c_j\}_{j=1}^c$  表示样本的类别标签集合,  $c_j$  表示某一样本的类别,  $c$  表示所有样本的类别总数。设  $X \subset \mathbb{R}^m$ , 即每个评价  $x_i$  都是  $m$  维的,  $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$ 。对于样本集  $X = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ , 其中  $y_i$  ( $y_i \in C$ ) 为文本样本的情感类别,  $Y = \{y_i\}_{i=1}^n$ 。设  $T = \{x_1, x_2, \dots, x_l\}$  为给定的标记样本集合, 即样本点  $x_i$  ( $i = 1, 2, \dots, l$ ) 的类别标签  $y_i$  是已知的,  $U$  为未标记样本集合, 即样本点  $x_i$  ( $i = l+1, l+2, \dots, l+u$ ) 的类别标签是未知的, 其中  $l+u = n$ ,  $X = T \cup U$ 。由于文中只研究二元的情感

分类问题,故  $y_i \in \{-1, 1\}$ 。 $y_i = 1$  说明评价  $x_i$  表达的情感是正向的,  $y_i = -1$  说明评价  $x_i$  表达的情感是负向的。假定所有的标记样本都是训练样本,所有的未标记样本都是测试样本。

为了构建可以给出样本集中  $x_i$  类别的分类器,需要得到决策函数  $f(x): f: X \rightarrow Y$ 。

基于样本相似度,文中在所有标记和未标记样本间建立一个加权无向图  $G = (V, E, W)$ 。其中,  $V$  表示所有标记和未标记的样本点  $x_i$  ( $i = 1, \dots, l, l+1, \dots, l+u$ ),  $E$  表示用以连接两节点的边,  $W$  表示边的权重,通常用相似度或距离来度量。

各顶点间边的权重通过以下相似矩阵来度量。相似矩阵  $W$  为:

$$W_{ij} = \begin{cases} e^{-\frac{d(x_i, x_j)^2}{2\sigma^2}} & (i \neq j) \\ 1 & (i = j) \end{cases} \quad (1)$$

其中,  $x_i$  和  $x_j$  表示两个评价的特征向量;  $d(x_i, x_j)$  取  $\cos(x_i, x_j)$ ;  $\sigma$  为给定的控制参数。

接下来计算对角矩阵  $D, D_{ii} = \sum_j W_{ij}$ , 其元素是  $W$  的行和,可以得到图拉普拉斯矩阵  $L = D - W$ 。 $L$  的谱分解为:

$$L = \sum_{i=1}^{l+u} \lambda_i \varphi_i \varphi_i^T \quad (2)$$

其中,  $\varphi_i$  为  $L$  的特征向量;  $\lambda_i \geq 0$  为  $L$  的特征值。

文中采用标准图拉普拉斯矩阵  $\tilde{C} = D^{-1/2} L D^{-1/2}$ , 并对其进行特征分解得到  $\{\lambda_i, v_i\}_{i=1}^n$ , 使得  $\tilde{C} = V \Lambda V^T$ 。其中,  $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n, \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n), V = [v_1, v_2, \dots, v_n]$  为特征值对应的特征向量矩阵。由聚类核的定义框架可知,引入转换函数  $\varphi$ , 使得  $\tilde{\lambda}_i = \varphi(\lambda_i)$ 。并由  $\tilde{\lambda}_i$  建立  $\tilde{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ , 使得  $\tilde{C} = V \tilde{\Lambda} V$ 。文中取的转换函数为:

$$\varphi(\lambda_i) = \begin{cases} \lambda_i, \lambda_i \geq \lambda_{\text{cut}} \\ 0, \lambda_i < \lambda_{\text{cut}} \end{cases} \quad (3)$$

令  $\tilde{D}$  是对角函数,  $\tilde{D}_{ii} = 1/\tilde{C}_{ii}$ , 则有基于上述图构建的聚类核  $\tilde{K} = \tilde{D}^{1/2} \tilde{L} \tilde{D}^{1/2}$ 。使用上述聚类核训练 SVM 分类器,得到基于半监督聚类核的情感分类器,不用建立多个分类器,也不用依赖监督学习对未标记样本进行分类预测,一次性使用未标记样本中隐含的信息,以指导分类。

从上述描述,可以得到 CKSVM 的步骤:

Step1: 根据 bag-of-words 模型,将文本数据表示为向量,进行初步的特征提取之后,得到标记样本集  $T$  和未标记样本集  $U$  的特征矩阵  $M, M \in \mathbb{R}^{m \times n}$ 。其中,  $m$

为训练数据集的文本个数,  $n$  为特征项的个数。

Step2: 根据式 (2) 计算相似矩阵  $W$ 。

Step3: 计算对角矩阵  $D_{ii} = \sum_j W_{ij}$ 。

Step4: 计算  $\tilde{C} = D^{-1/2} L D^{-1/2}$ , 并对其进行特征分解得到  $\{\lambda_i, v_i\}_{i=1}^n$ , 使得  $\tilde{C} = V \Lambda V^T$ 。其中,  $0 \leq \lambda_1 \leq \lambda_2 \cdots \leq \lambda_n, \Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n), V = [v_1, v_2, \cdots, v_n]$  为特征值对应的特征向量矩阵。

Step5: 引入转换函数  $\varphi$ , 使得  $\tilde{\lambda}_i = \varphi(\lambda_i)$ 。并由  $\tilde{\lambda}_i$  建立  $\tilde{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$ , 使得  $\tilde{C} = V \tilde{\Lambda} V$ 。

Step6: 令  $\tilde{D}$  是对角函数,  $\tilde{D}_{ii} = 1/\tilde{C}_{ii}$ , 计算聚类核  $\tilde{K} = \tilde{D}^{1/2} \tilde{L} \tilde{D}^{1/2}$ 。

Step7: 使用上述聚类核训练 SVM 分类器, 得到基于半监督聚类核的情感分类器, 对数据集上的评价提取出的特征向量进行训练。

3 实 验

为了验证 CKSVM 算法的有效性, 分别在数据集上进行测试。并且与较早实现的基于 Self-learning SVM 和基于 Co-training SVM 的半监督情感分类算法进行比较, 这两种算法的实现过程见文献[32]。算法

均在 32 位 Python 集成环境 Anaconda 中进行, 调用了多个用于科学计算的 Python 库, 如 numpy、sk-learn 等。都选用交叉验证法找出的最优参数。

3.1 数据集选取及预处理

(1) 文中选取 movie-reviews 影评数据集和 20 Newsgroups 数据集。其中, movie-reviews 数据集由康奈尔大学 (Cornell) 提供, 包括 2 000 条电影评价, 其中肯定和否定态度的各 1 000 条。目前影评库被广泛应用于各种粒度的 (如词语、句子和篇章级) 情感分析研究中。20 Newsgroups 数据集包括接近 20 000 种报纸的数据, 每种报纸选出 1 000 篇文章。文章的主题包括计算机、政治、宗教、运动和科学。

(2) 文中对英文文本的预处理, 主要依赖 Python 的 NLTK 库。NLTK 是用来处理和自然语言处理相关事件的工具包, 包括分词 (tokenize)、词性标注 (POS)、文本分类等现成工具。文中将数据集中的文本进行分词, 用 VSM 模型将一个个的文本表示成向量。

3.2 实验结果分析

对 movie-reviews 影评数据集, 在不同训练样本比例情况下, 各半监督情感分类算法的分类精度如表 1 所示。

表 1 movie-reviews 影评数据集三种半监督情感分类算法的分类精度

算法	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Self-learning	0.664	0.712	0.759	0.770	0.821	0.832	0.828	0.839	0.848	0.852
Co-training	0.622	0.687	0.724	0.762	0.787	0.817	0.830	0.829	0.841	0.847
CKSVM	0.705	0.749	0.789	0.818	0.831	0.857	0.870	0.875	0.880	0.885

三种算法在 movie-reviews 数据集上的分类准确度如图 2 所示。

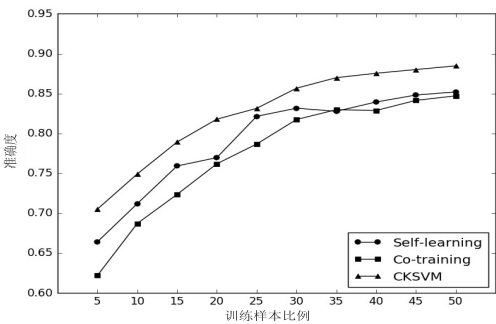


图 2 三种半监督情感分类算法在 movie-reviews 数据集上的分类准确度

表 2 20 Newsgroups 数据集三种半监督情感分类算法的分类精度

算法	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Self-learning	0.464	0.582	0.659	0.697	0.721	0.741	0.747	0.744	0.757	0.772
Co-training	0.402	0.514	0.604	0.648	0.686	0.712	0.732	0.741	0.750	0.758
CKSVM	0.465	0.592	0.679	0.717	0.741	0.776	0.787	0.794	0.797	0.801

从图 2 可以看出, 随着标记样本比例的增加, 各半监督情感分类算法的分类准确度都有提升, 其中 CKSVM 算法提升最快, 且其分类精度几乎一直高于基于 Self-learning SVM 和基于 Co-training SVM 的半监督情感分类算法, 说明 CKSVM 算法更好地运用了未标记样本中的信息。

对 20 Newsgroups 数据集, 在不同训练样本比例情况下, 各半监督情感分类算法的分类精度如表 2 所示。

三种算法在 20 Newsgroups 数据集上的分类准确度如图 3 所示。

从图 3 可以看出, 随着标记样本比例的增加, 各半监督情感分类算法的分类准确度都有提升, 其中 CKSVM 算法提升最快, 说明在该数据集上 CKSVM 算法也



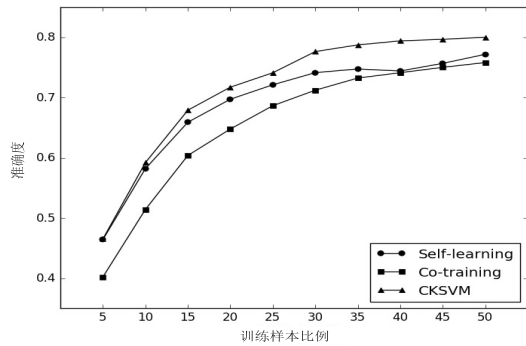


图 3 三种半监督情感分类算法在 20 Newsgroups 数据集上的分类准确度

同样很好地运用了未标记样本中的信息。另外,由于 20 Newsgroups 数据集更加复杂,三种算法的分类精度都有下降,但是 CKSVM 下降最少,说明 CKSVM 有较好的适应性,在不同数据集上依然可以得到较好的结果。

4 结束语

文中提出了基于聚类核的半监督情感分类算法。该方法直接将未标记样本所蕴含的信息融合到核中,可以直接用于 SVM 的情感分类器的训练上,有效利用了未标记样本中蕴含的信息。在两个数据集上的实验表明,该算法在分类精度上明显优于基于 self-learning SVM 和 Co-training 的半监督情感分类算法,且 CKSVM 在两个数据集上表现都最好,有较好的适应性。

参考文献:

[1] Pan S J, Ni X, Sun J T, et al. Cross-domain sentiment classification via spectral feature alignment[C]//International conference on world wide web. [s. l.]: [s. n.], 2010:751-760.

[2] Wei W, Gulla J A. Sentiment learning on product reviews via sentiment ontology tree[C]//Proceedings of meeting of the association for computational linguistics. Uppsala, Sweden: [s. n.], 2010:404-413.

[3] Li S, Huang C R, Zhou G, et al. Employing personal/impersonal views in supervised and semi-supervised sentiment classification[C]//Proceedings of meeting of the association for computational linguistics. Uppsala, Sweden: [s. n.], 2010: 414-423.

[4] Dasgupta S, Ng V. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification[C]//International joint conference on ACL. Singapore: [s. n.], 2009:701-709.

[5] Brosius J. Biographies, bollywood, boomboxes and blenders: domain adaptation for sentiment classification[J]. Association for Computational Linguistics, 2012, 31(2):187-205.

[6] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classi-

fication using machine learning techniques[C]//Proceedings of EMNLP. [s. l.]: [s. n.], 2002:79-86.

[7] Lee S Y M. Sentiment classification and polarity shifting[C]//International conference on computational linguistics. [s. l.]: Association for Computational Linguistics, 2010:635-643.

[8] Chapelle O, Weston J, Scholkopf B. Cluster kernels for semi-supervised learning[C]//Proceedings of the 16th annual conference on neural information processing systems. Massachusetts: MIT Press, 2003:321-328.

[9] Maynard D, Funk A. Automatic detection of political opinions in tweets[C]//International conference on the semantic web. [s. l.]: Springer-Verlag, 2011:88-99.

[10] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey[J]. Ain Shams Engineering Journal, 2014, 5(4):1093-1113.

[11] Kang H, Yoo S J, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews[J]. Expert Systems with Applications, 2012, 39(5):6000-6010.

[12] Ortigosa-Hernández J, Rodríguez J D, Alzate L, et al. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers[J]. Neurocomputing, 2012, 92(3):98-115.

[13] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 2002, 22(1):39-71.

[14] Chen C C, Tseng Y D. Quality evaluation of product reviews using an information quality framework[J]. Decision Support Systems, 2011, 50(4):755-768.

[15] Li Y M, Li T Y. Deriving market intelligence from microblogs[J]. Decision Support Systems, 2013, 55(1):206-217.

[16] Moraes R, Valiati J F, Neto W P G. Document-level sentiment classification: an empirical comparison between SVM and ANN[J]. Expert Systems with Applications, 2013, 40(2):621-633.

[17] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3):273-297.

[18] Ruiz M E, Srinivasan P. Hierarchical text categorization using neural networks[J]. Information Retrieval Journal, 2002, 5(1):87-118.

[19] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1):81-106.

[20] Li Y H, Jain A K. Classification of text documents[J]. The Computer Journal, 1998, 41(8):537-546.

[21] Hu K, Lu Y, Zhou L, et al. Integrating classification and association rule mining: a concept lattice framework[M]//New directions in rough sets, data mining, and granular-soft computing. Berlin: Springer, 2003:443-447.

[22] Turney P. Thumbs up or thumbs down?: semantic orientation

AR 和 Oxford Flowers17 人脸数据库上的实验结果表明,MVKDA 与 MVFS、HOA 以及 KDA 相比,有效地提高了识别率。

参考文献:

[1] 朱长仁,王润生. 基于单视图的多姿态人脸识别算法[J]. 计算机学报,2003,26(1):104-109.

[2] Xiong N,Svensson P,Svensson P. Multi-sensor management for information fusion:issues and approaches[J]. Information Fusion,2002,3(2):163-186.

[3] Lai P L,Fyfe C. Kernel and nonlinear canonical correlation analysis[J]. International Journal of Neural Systems,2012,10(5):365-377.

[4] Shon A,Grochow K,Hertzmann A,et al. Learning shared latent structure for image synthesis and robotic imitation[C]//Advances in neural information processing systems. [s. l.]:[s. n.],2005:1233-1240.

[5] Li S Z,Zhu L,Zhang Z Q,et al. Statistical learning of multi-view face detection[C]//European conference on computer vision-part IV. [s. l.]:[s. n.],2002:67-81.

[6] Tang J,Hu X,Gao H,et al. Unsupervised feature selection for multi-view data in social media[C]//SDM. [s. l.]:[s. n.],2013:270-278.

[7] Jing X,Liu Q,Lan C,et al. Holistic orthogonal analysis of discriminant transforms for color face recognition [C]//17th IEEE international conference on image processing. [s. l.]:

IEEE,2010:3841-3844.

[8] 赵振勇,王保华,王力,等. 人脸图像的特征提取[J]. 计算机技术与发展,2007,17(5):221-224.

[9] Belhumeur P N,Hespanha J P,Kriegman D J. Eigenfaces vs. fisherfaces:recognition using class specific linear projection [J]. Pattern Analysis and Machine Intelligence,1997,19(7):711-720.

[10] Yang M H. Kernel eigenfaces vs. kernel fisherfaces:face recognition using kernel methods[C]//Proceeding of international conference on automatic face and gesture recognition. [s. l.]:[s. n.],2002:215.

[11] Mika S,Ratsch G,Weston J,et al. Fisher discriminant analysis with kernels[C]//Proceeding of IEEE international workshop on neural networks for signal processing IX. [s. l.]:IEEE,1999:41-48.

[12] Martinez A M,Benavente R. The AR face database[EB/OL]. 2009. [http://cobweb.ecn.Purdue.edu/~aleix/aleix\\_face\\_DB.html](http://cobweb.ecn.Purdue.edu/~aleix/aleix_face_DB.html).

[13] Nilsback M E,Zisserman A. A visual vocabulary for flower classification [C]//IEEE computer society conference on computer vision and pattern recognition. [s. l.]:IEEE,2006:1447-1454.

[14] Qian G. Similarity between Euclidean and cosine angle distance for nearest neighbor queries[C]//ACM symposium on applied computing. [s. l.]:ACM,2004:1232-1237.

(上接第 91 页)

applied to unsupervised classification of reviews[C]//Proc of the 40th annual meeting on association for computational linguistics. Stroudsburg,USA:Association for Computational Linguistics,2002:417-424.

[23] Goldberg A B,Zhu X. Seeing stars when there aren't many stars [C]//TextGraphs: the first workshop on graph based methods for natural language processing. [s. l.]:[s. n.],2006:45-52.

[24] Sindhvani V,Melville P. Document-word co-regularization for semi-supervised sentiment analysis[C]//Eighth IEEE international conference on data mining. [s. l.]:IEEE Computer Society,2008:1025-1030.

[25] Zhou S,Chen Q,Wang X. Active deep networks for semi-supervised sentiment classification [C]//International conference on computational linguistics. [s. l.]: Association for Computational Linguistics,2010:1515-1523.

[26] Zhou S,Chen Q,Wang X. Fuzzy deep belief networks for semi

-supervised sentiment classification [J]. Neurocomputing, 2014,131(9):312-322.

[27] Li S,Hao J. Spectral clustering-based semi-supervised sentiment classification[C]//Proc of the 8th advanced data mining and applications. Berlin:Springer,2012:271-283.

[28] Zhou Z H. Co-training paradigm in semi-supervised learning [C]//Proceeding of the Chinese workshop on machine learning and applications. Nanjing,China:[s. n.],2007.

[29] 郑文静,李雷. 基于图的组合半监督 SVM 聚类核算法研究[J]. 计算机技术与发展,2014,24(5):109-112.

[30] Weston J,Leslie C,Ie E,et al. Semi-supervised protein classification using cluster kernels [J]. Bioinformatics,2005,21(15):3241-3247.

[31] Szummer M. Partially labeled classification with Markov random walks [J]. Advances in Neural Information Processing Systems,2002(14):945-952.

[32] 李素科,蒋严冰. 基于情感特征聚类的半监督情感分类 [J]. 计算机研究与发展,2013,50(12):2570-2577.