基于不平衡数据的中文情感分类*

王中卿，李寿山，朱巧明，李培峰，周国栋
苏州大学 计算机科学与技术学院，江苏 苏州 215006
E-mail: {wangzq870305, shoushan.li}@gmail.com; {qmzhu, pfli, gdzhou}@suda.edu.cn

摘 要：近些年来，情感分类在自然语言处理研究领域获得了显著的发展。然而，大部分已有的研究都假设参与分类的正类样本和负类样本一样多，而实际情况中正负类数据的分布往往是不平衡的。本文收集四个产品领域的中文评论文本，发现正类样本的数目远远多于负类样本。针对不平衡数据的中文情感分类，我们提出了一种基于欠采样和多分类算法的集成学习框架。在四个不同领域的实验结果表明，我们的方法能够显著提高分类性能，并明显优于目前主流的多种不平衡分类方法。

关键词：中文信息处理；情感分类；不平衡分类；集成学习

Chinese Sentiment Classification on Imbalanced Data Distribution

Wang Zhongqing, Li Shoushan, Zhu Qiaoming, Li Peifeng, Zhou Guodong
School of Computer Science and Technology, Soochow University, Suzhou 215006
E-mail: {wangzq870305, shoushan.li}@gmail.com; {qmzhu, pfli, gdzhou}@suda.edu.cn

Abstract: Sentiment classification has undergone significant development in recent years. However, most existing studies assume the balance between the numbers of negative and positive samples, which may not be true in reality. In this paper, we collect product reviews from four domains and find that the positive samples are much more than negative ones. To handle the imbalanced classification in Chinese sentiment classification, we propose a novel approach to combine both sampling and classification algorithms under an ensemble learning framework. Evaluation across different domains shows the proposed approach performs better than several existing imbalanced classification methods.

Keywords: sentiment classification; imbalanced classification; ensemble learning

1 前言

目前，人们越来越习惯于在网络上表达自己的观点和情感，从而使得网络上渐渐出现了大量带有情感的文本。由于传统的基于主题的文本分类方法已经无法很好分析这些情感文本，人们开始关注针对情感文本分类（简称：情感分类）的方法研究[1,2]。虽然情感分类的研究已经开展多年，但是目前大部分情感分类的研究假设正类样本和负类样本是平衡的[1,2,3,4]。该假设和实际情况并不相符，在实际收集的产品评论语料中，我们发现正类样本和负类样本的数目差距很大。样本分布的不平衡往往会使传统的机器学习分类方法在分类过程中严重偏向多样本类别，从而使分类的性能急剧下降。因此，不平衡数据的情感分类问题是一个迫切需要解决的实际问题。

不平衡分类本身在机器学习领域是一个很有挑战性的研究问题[5,6]。为了表述清楚，在下文中我们将样本集合中样本数较多的一类称为多类（majority class），样本数较少的一类称为少类（minority class）。目前已经存在一些用来解决不平衡分类问题的方法，例如，重采样技术（re-sampling）[7]，单类别分类（one-class classification）[8]和代价敏感学习（cost sensitive learning）[9]。但是由于还没有针对情感分类特别是中文情感分类的不平衡问题的研究，很多基本问题仍待

* 本文承国家自然科学基金（90920004，61070123，61003153，60970056，61003155），模式识别国家重点实验室开放课题基金，江苏省自然科学基金（BK2008160）和江苏省高校自然科学重大基础研究项目（08KJA520002）资助。

研究，例如，哪种方法更适合中文情感分类任务。

本文将以欠采样技术（under-sampling）为基础，通过集成学习（ensemble learning）解决情感分类中的不平衡问题。欠采样技术是指从初始的多类标注样本中随机取出和少类标注样本一样规模的样本，与少类样本一同构建分类器。欠采样方法存在一个明显缺点：由于欠采样只是从多类中选择部分样本，使得大量未选中的多类样本在后面的分类过程中未能发挥作用，从而丢失了很多可能对分类有帮助的样本。因此，为了充分利用所有标注数据，可以首先在多类样本中进行多次欠采样，构建多个欠采样基分类器，最终融合这些基分类器进行集成学习[21]。我们称该方法为基于欠采样融合的集成方法。

影响集成学习性能的一个重要因素是参与集成的分类器之间的差异性。一般来说，分类器之间的差别越大，集成学习的性能提高会越明显[22]。然而，在欠采样融合的集成方法中，所有基分类器中参与训练的少类样本是完全一样的。为了进一步增加基分类之间的差异性，我们提出基于欠采样和多分类算法的集成方法。具体来讲，每个欠采样基分类器是由随机分配的分类算法训练得到。由于不同分类算法的分类机理是不同的，这样参与的基分类器之间的差异性进一步扩大，有利于进一步提高分类性能。实验结果表明，基于欠采样和多分类算法的集成方法能够进一步的提高分类的效果。

本文其他部分安排如下：第 2 节详细介绍情感分类以及不平衡分类的相关工作；第 3 节提出基于集成学习的不平衡分类方法；第 4 节给出实验结果及分析；第 5 节给出相关结论。

2 相关工作

2.1 情感分类

早期情感分类研究主要集中在无监督学习方法上面。无监督学习一般是通过两个词之间的关系以及一些资源比如 WordNet/HowNet 或者未标注数据来判断文本的情感倾向 [10]。由于无监督学习方法的分类效果比较差，并不能很好满足实际应用的需求。

基于监督学习的情感分类方法是当前的主流方法，与无监督学习方法相比，基于词袋模型（bag-of-words model）的全监督情感分类方法总是能够获得更好的分类效果[1]。后续的大量研究在基本的词袋模型上面给出了多种方式的改进，进一步提高了分类的性能。例如，采用上下文特征[11]，使用文档子成分（document subcomponent）信息[12]，考虑极性转移[15]等。然而，已有的方法基本都是基于样本分布平衡的假设，不平衡数据的情感分类方法研究还很缺乏。

2.2 不平衡分类

目前，主流的不平衡分类方法主要分为三类：重采样技术、单类别分类和代价敏感学习。

其中，重采样技术应用最为广泛。重采样技术主要分为两类方法：欠采样（Under-sampling）和过采样（Over-sampling）。具体来讲，过采样技术通过重复少类样本使得少类样本数和多类样本数平衡；欠采样技术通过减少多类样本使得两类样本数平衡。除了简单的随机重采样，其他多重采样方法通过启发式的策略来扩展/选择样本。例如，Yen 和 Lee 提出基于聚类方法的欠采样方法，该方法通过聚类的方式在采样的过程中选择更具代表的样本[16]。

代价敏感学习方法的主要思想是在构建分类器过程中修改训练过程中的分错代价函数，让少类分错的代价远远大于多类分错的代价[9]。单类别分类是指在构建分类器过程中，只使用一个类别里面的标注样本，在应用到不平衡分类中，仅仅多类样本作为单类别分类的训练样本[8]。该方法适合样本非常不平衡情况的分类问题。

3 基于集成学习的不平衡分类方法

3.1 情感分类中的不平衡分布情况

假设N个样本的训练数据中包含有N_+个正类样本和N_-个负类样本。目前大多数研究总是假设正类样本数和负类样本数是平衡的，即$N_+ = N_-$，但实际情况并非如此。通常来说，更一般的情况是训练数据中一类样本要远远多于另一类样本，即$N_+ >> N_-$或者$N_+ << N_-$。

为了更好的理解情感分类中的不平衡现象，我们从卓越网[1]上收集来自四个领域的中文评论语料并统计它们在两个类别里面分布情况。这四个领域分别是箱包、化妆品、相机和软件。

表1 各领域正类样本和负类样本分布情况

领域	N_+	N_-	N_+ / N_-
箱包	4864	1185	4.10
化妆品	3568	1102	3.24
相机	2133	749	2.85
软件	971	467	2.08

表1给出了四个领域的类别分布情况。从表中可以看出，各个领域不平衡比（N_+ / N_-）介于2和4。显而易见，在每个领域中，负类样本数目都要明显少于正类样本数目。

3.2 基于欠采样融合的集成方法

集成学习是组合多个基分类器的一种学习机制。为了产生多个不同的基分类器，一种常用的方式是通过训练不同的数据集合产生不同的基分类器，称之为基于样本融合的集成学习。针对不平衡分类问题，可以在多类样本中进行多次欠采样并将每一次采样的样本同少类样本训练获得一个基分类器。

获得基分类器后，组合分类器方法需要特别的融合方法去融合这些结果。融合方法可以分为两种，固定的融合方法（fixed rules）和可训练的融合方法 （trained rules）。本文选择基于固定融合方法的贝叶斯规则融合基分类器的结果。贝叶斯规则可以描述为[18]：

$$assign\ \mathrm{y} \rightarrow \mathrm{c}_j$$

$$j = \arg\max_i p(c_i) \prod_{k=1}^{R} p(c_i \mid d_k)$$

总体来说，基于欠采样的集成学习的实现步骤如下：（1）通过在多类样本中进行多次欠采样的方式和少类样本组成多个训练样本集合，（2）对于每个训练样本集合训练一个基分类器，（3）通过贝叶斯规则融合各个基分类器的结果。

3.3 基于欠采样和多分类算法的集成方法

在构建集成学习系统中除了可以通过训练不同的样本集合产生不同的分类器之外，还可以通过不同的分类方法产生不同的分类器。由于很多分类方法是基于不同的原理的，如 *k*-近邻(*k*-NN)方法是基于记忆的方法，支持向量机方法(SVM)是基于结构风险最小理论的方法等。因此，不同的分类方法实现的分类器实现分类的效果往往是不一样的[2]。所以通过为不同的训练样本集合随机分配不同的分类算法，可以减少由于存在相同的少类样本造成的样本冗余现象，从而进一步提高集成学习的效果。因此，我们提出一种基于欠采样和多分类算法的集成方法。具体实现步骤如下：

[1] http://www.amazon.cn/

（1）通过在多类样本中进行多次欠采样的方式和少类样本组成多个训练样本集合，（2）对于每个训练样本集合随机分配一个分类算法组成基分类器，（3）通过贝叶斯规则融合各个基分类器的结果。

基于欠采样和多分类算法的集成学习系统需要使用多种分类算法用来构建基分类器。本文采用三种不同的分类方法，分别为朴素贝叶斯、最大熵和支持向量机。

4 实验

4.1 实验设置

我们在卓越网上收集了来自四个领域的中文评论语料。这四个领域分别是箱包、化妆品、相机和软件。3.1 节已经分析了每个领域的不平衡情况，具体分布可参考表 1。实验过程中，我们选择 80%的样本作为训练样本，剩余的 20%样本作为测试样本。分类算法包括最大熵、SVM 和朴素贝叶斯。其中，SVM 是使用标准工具 *light*-SVM[1]，朴素贝叶斯和最大熵是使用 MALLET 机器学习工具包[2]。在使用过程中，这些工具的所有参数都设置为它们的默认值。

在进行分类之前首先采用中国科学院计算技术研究所的分词软件 ICTCLAS[3]对中文文本进行分词操作。给定分好词的文本后，我们选取词的 unigram 作为特征，用以获得文本向量的表示。

在平衡数据的情感分类中，通常使用准确率（accuracy，*acc.*）作为分类效果的衡量标准。 而在不平衡分类中，由于分类结果很容易偏向多类，所以使用准确率作为分类效果的衡量标准对于少类变得非常不公平。因此，一般使用几何平均数（*G-mean*）作为衡量分类效果的标准。几何平均数的计算方法为：

$$G-mean=\sqrt{TP_{rate}\times TN_{rate}}$$

其中：TP_{rate}和 TN_{rate}分别代表了正类样本的召回率和负类样本的召回率[6]。

4.2 不同分类算法的比较

表 2 和表 3 分别是各个分类算法基于欠采样以及基于欠采样集成学习的分类的分类结果(*G-mean* 值)。从结果中可以看出：（1）最大熵，SVM 和朴素贝叶斯在分类结果上的差别不明显，这一点同文献[1]中的结果一致。（2）虽然不同的分类算法在实现上有很大的不同，但是基于欠采样集成学习的分类效果都比基于欠采样的分类有很大的提升，充分显示了基于欠采样集成学习的分类在不平衡中文情感分类中的优势。

表 2　基于欠采样的分类的分类结果

领域	ME	SVM	NB
箱包	0.850	**0.851**	0.847
化妆品	0.834	**0.849**	0.818
相机	0.802	0.809	**0.811**
软件	0.745	0.734	**0.755**
平均	0.808	**0.811**	0.808

表 3　基于欠采样集成学习的分类结果

领域	ME	SVM	NB
箱包	**0.864**	0.862	0.857
化妆品	0.849	**0.859**	0.820
相机	0.818	**0.822**	0.816
软件	0.763	0.742	**0.775**
平均	**0.823**	0.821	0.817

4.3 不平衡分类方法比较

为了进行充分的比较，我们实现了了多种主流的不平衡分类方法：

[1] http://svmlight.joachims.org/

[2] http://mallet.cs.umass.edu/

[3] http://ictclas.org/

1）完全训练（Full Training，FullT），直接将所有训练样本进行训练。

2）随机过采样（Random Over-sampling，OverS），在少类样本中使用过采样技术随机选择样本。

3）随机欠采样（Random Under-sampling ，UnderS），在多类样本中使用采样技术随机选择样本。

4）基于聚类的欠采样（Clustering-based Under-sampling，ClusterU），我们根据文献[16]的方法实现了基于聚类的欠采样。

5）基于最邻近的欠采样（Neighbor-based Under-sampling，Neighbor），在多类样本中进行随机欠采样，只是在欠采样时，每次选择一个样本需要在样本集合中去除和它最邻近的 k 个样本（*k*NN）[20]。在我们的实验中，该方法通过在多类样本中去除“多余的”，“边界的”样本从而提高欠采样的效果。

6）单类别分类（One-class Classification，OneClass），我们根据文献[8]的描述，利用 *lib*SVM[1]实现单类别分类。

7）代价敏感分类（Cost-sensitive Classification，CostSensitive），我们根据文献[9]的描述，利用 *lib*SVM 实现代价敏感分类。在这里代价的权重根据每个领域中训练样本集合中多类样本和少类样本的比例进行调整。

8）基于欠采样和多分类算法的集成学习（OurApproach），在基于欠采样的集成学习的基础上为每组训练样本随机分类不同的分类器，也就是本文提出的方法。

从上一节结果可以看出，三个分类算法的分类性能差异不大，我们关于不平衡学习方法的比较研究中仅以最大熵分类算法作为基准系统实现前四种不平衡分类方法。表 4 是各种不平衡分类方法在基于不平衡数据的中文情感分类中的分类效果。从表中可以看出，单类别分类方法表现最差，可能的原因是单类别分类适合不平衡程度非常大的不平衡分类问题（例如，正负或者负正样本比例超过 1000），而中文情感分类的正负比例仅仅介于 2 到 4 之间。完全训练（FullT）方法表现也不理想，主要原因就是分类算法严重趋向多类，使得少类的召回率非常低。几种采样方法的比较可以发现，欠采样方法优于过采样方法，但是几种欠采样的方法的性能基本类似。代价敏感分类方法（CostSensitive）相对于其他方法有明显优势，能够比随机欠采样方法有明显提升。我们的基于欠采样和多分类算法的集成方法明显优于其他各种不平衡分类方法，平均比代价敏感分类方法提高超过 2 个百分点。

表 4　不同不平衡分类的结果

领域	FullT	OverS	UnderS	ClusterU	Neighbor	OneClass	Cost Sensitive	**Our Approach**
箱包	0.790	0.819	0.850	0.846	0.850	0.574	0.862	**0.875**
化妆品	0.807	0.834	0.834	0.830	0.839	0.512	0.854	**0.861**
相机	0.756	0.781	0.802	0.804	0.799	0.514	0.792	**0.830**
软件	0.708	0.713	0.745	0.739	0.736	0.507	0.741	**0.773**
平均	0.766	0.787	0.808	0.805	0.806	0.527	0.812	**0.835**

5　结语

本文研究中文情感文本分类任务中的不平衡数据分类问题，提出一种基于欠采样和多分类算法的集成学习方法。实验结果表明，该方法能够很好的提高中文情感分类任务中的不平衡分类问题。比较研究发现，我们的方法明显优于传统的采样方法、单类别分类和代价敏感分类方法。

情感文本分类任务中的不平衡数据分类问题才刚刚起步，有许多问题还有待进一步深入研究。例如，不平衡情感分类问题中的特征选择是明显区别于传统特征提取方法[23]，如果更有效地进行

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

不平衡数据分类问题中的特征提取是一个值得探讨的问题。另外，情感分类领域适应（Domain Adaptation）[3]中的不平衡数据分类问题也是一个急需解决的问题。在实际应用中，存在目标领域里面的样本分布不平衡的情况。这些问题将作为我们下一步的研究方向。

参 考 文 献

[1] Pang B., Lee L., and Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP*. 2002.

[2] 李寿山, 黄居仁. 基于 Stacking 组合分类方法的中文情感分类研究. *中文信息学报*, 2010, 24(5), 56-61.

[3] Blitzer J., Dredze M., and Pereira F. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of ACL*. 2007.

[4] Li S., Huang C., Zhou G., and Lee S. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In *Proceedings of ACL*. 2010.

[5] Barandela R., Sánchez J. S., García V., and Rangel E. Strategies for Learning in Class Imbalance Problems. *Pattern Recognition*, 2003, 36, 849–851.

[6] Kubat M. and Matwin S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *Proceedings of ICML*. 1997.

[7] Chawla N., Bowyer K., Hall L., and Kegelmeyer W. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, 2002, 16, 321-357.

[8] Juszczak P. and Duin R. Uncertainty Sampling Methods for One-Class Classifiers. In *Proceedings of ICML, Workshop on Learning with Imbalanced Data Sets II*. 2003.

[9] Zhou Z. and Liu X. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. *IEEE Transaction on Knowledge and Data Engineering*, 2006, 18, 63-77.

[10] Turney P. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of reviews. In *Proceedings of ACL*. 2002.

[11] Riloff E., Patwardhan S., and Wiebe J. Feature Subsumption for Opinion Analysis. In *Proceedings of EMNLP*. 2006.

[12] McDonald R., Hannan K., Neylon T., Wells M., and Reynar J. Structured Models for Fine-to-coarse Sentiment Analysis. In *Proceedings of ACL*. 2007.

[13] Somasundaran S., G. Namata, and J. Wiebe. Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification. In Proceedings of EMNLP-09, pp.170-179.

[14] Nakagawa T., Inui K., and Kurohashi S. Dependency Tree-based Sentiment Classification using CRFs with Hidden. In *Proceedings of NAACL*. 2010.

[15] Li S., Lee S., Chen Y., Huang C., and Zhou G. Sentiment Classification and Polarity Shifting. In *Proceedings of COLING*. 2010.

[16] Yen S. and Lee Y. Cluster-Based Under-Sampling Approaches for Imbalanced Data Distributions. *Expert Systems with Applications*, 2009, 36, 5718-5727.

[17] 翟云, 杨炳儒, 曲武. 不平衡类数据挖掘研究综述. *计算机科学*, 2010, 37. 27-32.

[18] Kittler J., Hatef M., Duin R. P.W., and Matas J. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20, 226-239.

[19] Vilalta R. and Drissi Y. A perspective view and survey of meta-learning. *Artificial Intelligence Review*, 2002, 18(2), 77-95.

[20] Laurikkala J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. In *Proceeding of 8th Conference on artificial intelligence in medicine in Europe*, 2001.

[21] Liu X., Wu J., and Zhou Z. Exploratory under-sampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2009, 39(2), 539-550.

[22] Kittler J., Hatef M., Duin R., and Matas J. On Combining Classifiers. *IEEE Trans. PAMI*, 1998, 20, 226-239.

[23] Li S., Xia R., Zong C., and Huang C. A Framework of Feature Selection Methods for Text Categorization. In *Proceedings ACL-IJCNLP*. 2009.