不均衡大数据集下的文本特征基因提取方法

孙晶涛[1]，张秋余[2]

(1. 西安邮电大学计算机学院 西安 710121；2. 兰州理工大学计算机与通信学院 兰州 730050)

【摘要】在不均衡大数据集情况下，传统特征处理方法偏重大类而忽略小类，影响分类性能。该文提出了一种文本特征基因提取方法。首先，基于样本类别分布不均衡对特征选择的影响，给出了一种结合信息熵的CHI统计矩阵特征选择方法，以强化小类的特征；然后，在探究多维统计数据高阶相关性的基础上，采取独立成分分析手段，设计了文本特征基因提取方法，用以增强特征项的泛化能力；最后，将这两种方法相融合，实现了在不均衡大数据集下的文本特征基因提取新方法。实验结果表明，所提方法具有较好的早熟性及特征降维能力，在小类的分类效果上优于常见特征选择算法。

关 键 词 CHI统计选择方法； 不均衡大数据集； 独立成分分析； 信息熵； 文本特征基因提取

中图分类号 TN393.098 **文献标志码** A doi:10.3969/j.issn.1001-0548.2018.01.019

Text Feature Gene Extraction on Imbalanced Big Dataset

SUN Jing-tao[1] and ZHANG Qiu-yu[2]

(1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications Xi'an 710121;

2. School of Computer and Communication, Lanzhou University of Technology Lanzhou 730050)

Abstract In the cases of imbalance big datasets, the traditional feature processing method is biased to the large class and ignores the small class, which affects the classification performance. So a text feature gene extraction method is proposed in this paper. First of all, considering the feature selection impact of imbalance distribution of sample categorization, a feature selection method based on the CHI statistical matrix combined with information entropy is used to strengthen the characteristics of the small class. Secondly, based on the high order correlation of multidimensional statistical data, the method of text feature extraction is designed to enhance the generalization ability of feature item. Finally, the two methods are combined to construct a new method of text feature extraction under unbalanced large datasets. The experimental results show that the proposed method has a better performance in early maturity and feature dimension reduction, and is far superior to the common feature selection algorithm in the classification ability of small classes.

Key words CHI statistical selection method; imbalanced big dataset; independent component analysis; information entropy; text feature gene extraction

当前社会正在逐渐步入大数据时代，文本内容分析业已成为实现大数据理解与价值发现的有效手段，文本分类作为大数据内容挖掘的关键技术，广泛应用于互联网舆情监测与预警、网络有害信息过滤以及情感分析等多个领域。而特征选择作为文本分类中至关重要的一环，也直接影响到模型构建及分类效率和准确性。

目前在文本分类中，特征选择方法多采用基于向量空间模型(vector space model)[1-2]的统计方法，但这类方法在实际应用中会出现两方面的问题：1)集内样本的类别分布不均衡，传统特征选择函数如文档频率(ocument frequency, DF)、信息增益(information gain, IG)、互信息(mutual information, MI)等方法[3-5]，往往假定数据集内样本的类别分布相同或相近，使得所确定的特征项大多来自类别数量占优的大类，导致选出的最具区分度的特征子集无法准确反映整个样本空间的真实分布；2) "大数据"[6-7]使得数据维数呈现爆炸性增长，面对超高维度的数据集，不仅意味着巨大的内存需求，而且意味着高昂的计算成本投入。在这些高维数据的特征空间中，繁多的特征点之间存在着很强的相关性，使得采用传统方法选取的特征项泛化能力急剧恶

收稿日期：2016– 09 – 21；修回日期：2017 – 09 – 15

基金项目：国家自然科学基金(61363078)；陕西省科技统筹创新工程–重点产业创新链–工业领域项目(2016KTZDGY04-01)；陕西省自然科学基础研究计划(2016JM6048).

作者简介：孙晶涛(1981 –)，男，博士，高级工程师，主要从事自然语言理解、数据挖掘、机器学习与人工智能等方面的研究.

化。如何从纷繁复杂的表象信息中提取出事物的本质特征、提高特征项的泛化能力就愈显重要。

如同生物基因是生命体最小的功能单位一样，文本特征基因也是文本最小的结构单位，其储存着文本深层次的语义结构以及潜在的语义联系，是全部文本信息的基本载体。本文正是研究如何在文本特征集中提取出稳定、泛化能力强的特征基因子集，从而降低向量空间的特征维数，提高分类识别效果。将信息熵引入特征点类别权重定义中，构造特征点对文本类别的区分度矩阵，消除传统方法对不均衡样本集进行特征提取时的缺陷，提高分类识别的正确性。采用独立成分分析方法(independent component analysis, ICA)[8]，通过分析多维统计数据间的高阶相关性，找出相互独立的隐含信息成分，以此提取文本特征基因，减少特征采样点数量，实现在不均衡大数据集中，较准确提取全面、真实反映文本内容信息的最优特征基因子集，提升文本分类识别的性能。通过在搜狐新闻数据(SogouCS) 20151022语料库上的实验表明，文本特征基因提取方法(text feature gene extraction, TFGE)能够使采用SVM(support vector machine)分类算法构造的文本分类器识别性能显著提升。

1 相关工作

目前，国内外学者针对不均衡数据集下的数据分析进行了相关的研究。文献[9]提出了基于逐级优化的逆随机欠抽样算法，该算法在去除训练样本中噪声和重复信息的同时，使获得的分类器更倾向于小类样本，而采用Bagging方法进行多分类器集成，能够尽可能保留有用信息，提高有效数据的利用率。文献[10]提出了基于核聚类欠抽样集成不均衡SVM分类算法，该算法首先在核空间中对大类样本集进行聚类，然后随机选择出具有代表意义的聚类信息点，在减少大类样本的同时，将SVM算法的分类界面向小类样本方向偏移，并利用集成手段对基于核聚类的欠抽样SVM算法进行集成，最终实现提高不均衡数据SVM算法泛化性能的目的。文献[11]提出了改进的基于核密度估计的数据分类算法。该方法通过引入空间信息以及平滑参数，改善了原有核密度估计的分类方法在处理不均衡问题时所存在的缺陷。但该方法将空间信息仍定义为检测点到类中心的距离，这必然导致方法的鲁棒性较差。还有一些文献提出了分类算法的改进，如boosting[12]、FCM-KFDA[13]、AdaBoost-SVM[14]、代价敏感学习算法[15]，这些方法所选特征子集通常较为优化，但面对“大数据”普遍存在运行效率低的问题。

然而现有处理手段大多集中在对样本类别分布再平衡及分类算法改进等方面，对特征选择、特征提取的研究尚不多见。文献[16]提出了一种迭代的特征选择模型，利用迭代过程中的聚类结果进行数据特征采样排名，以此选取最优特征子集。但该模型中迭代函数以及迭代次数的选择对问题求解影响因素很大，模型性能受到了一定制约。文献[17]提出了一种应用于不均衡数据集下的无监督特征选择方法，该方法在无监督环境下，对不同特征空间利用概率密度分析各特征数据的分布状况，通过特征之间的数据分布关系来进行特征选择。但该方法没有考虑到数据分布的特点，对分类性能的影响较大。

本文研究拟以CHI统计选择方法(χ^2统计)为基础，通过引入信息熵以及文本特征分布矩阵，克服χ^2统计量在处理不均衡数据集上的不足，并采用ICA方法分析原本隐藏在大量数据集中的内在因子或成分信息，利用这些信息描述数据的本质结构，实现降维去噪，提高文本分类的识别率。

2 相关理论

2.1 CHI统计选择方法

CHI统计选择方法是假定在特征与类别之间，具有一阶自由度χ^2分布基础上提出的一种FS方法。χ^2值大小与特征和类别的相关性成正比[18-19]。特征t与类别c_i的χ^2值定义为：

$$\chi^2(t,c_i)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

式中，N表示样本集中样本的总数；A表示包含特征t并且类别为c_i的样本数；B表示包含特征t并且类别不为c_i的样本数；C表示不包含特征t，但类别为c_i的样本数；D表示既不包含特征t，类别也不为c_i的样本数。

可以看出，χ^2值越大，特征t与类别c_i越相关。当类别数较多时，可以分别计算特征t与不同类别的χ^2值，选取其中最大的χ^2值作为特征t的所属类别，即$\chi^2_{\mathrm{MAX}}(t)=\mathrm{MAX}_{i=1}^{m}\chi^2(t,c_i)$。

文献[20]的实验证明，CHI统计选择方法具有较好的特征选择性能。但从式(1)能够看出，χ^2值仅体现特征在样本集中的文档频率，并没有考虑特征的词频，导致CHI方法在处理低频词时具有较大误差，使一部分噪声词被“优选”出来，降低了分类精度。

2.2 信息熵

1948年美国数学家香农(Claude Elwood Shannon)借鉴热力学中熵的概念，系统性提出了信息的度量方法，并定义了信息熵[21]。

定义 1 假设离散随机变量X，其取值$R=\{x_1,x_2,\cdots,x_n\}$为有限可数，取值随机出现的概率分别为$p(x_1),p(x_2),\cdots,p(x_n)$，且$\sum_{i=1}^{n}p(x_i)=1$，则$X$的信息熵$H(X)$为：

$$H(X)=-\sum_{i=1}^{n}p(x_i)\log_2 p(x_i) \tag{2}$$

2.3 独立成分分析

ICA是从多元(多维)统计数据中寻找内在因子或成分的一种方法，有别于其他统计方法的特点在于：不仅能够去除数据中各分量间的一、二阶相关性，还具有发掘并去除数据间高阶相关性的能力，使输出分量不仅相互独立，而且是非高斯分布[21]。

ICA基本模型[22]：假设存在n个随机变量$x_1,x_2,\cdots,x_n$，而这些变量是由另外n个随机变量$s_1,s_2,\cdots,s_n$的线性组合得到：

$$x_i=a_{i1}s_1+a_{i2}s_2+\cdots+a_{in}s_n \tag{3}$$

式中，$a_{ij}\ (i,j=1,2,\cdots,n)$为实系数，在模型中，假设$s_i$在统计上彼此独立。

为方便表示，通常选用向量-矩阵形式，即用随机向量$\boldsymbol{x}$来表示混合向量，其元素分别为$x_1,x_2,\cdots,x_n$，同样地用$\boldsymbol{s}$来表示元素$s_1,s_2,\cdots,s_n$。用矩阵$\boldsymbol{A}$表示混合系数a_{ij}，由此混合模型可以写为：

$$\boldsymbol{x}=\boldsymbol{A}\boldsymbol{s} \tag{4}$$

或写为：

$$\boldsymbol{x}=\sum_{i=1}^{n}a_{ij}s_i \tag{5}$$

ICA的目的就是计算分离矩阵$\boldsymbol{W}$，并得到一组相互独立的随机变量$y_1,y_2,\cdots,y_n$，使随机向量$\boldsymbol{y}$：

$$\boldsymbol{y}=\boldsymbol{W}\boldsymbol{x} \tag{6}$$

3 不均衡大数据集下的文本特征基因提取方法

3.1 不均衡数据集合上的特征选择方法

3.1.1 基于χ^2统计量的文本特征分布矩阵

CHI方法认为，特征的重要性主要由其χ^2的大小决定，低于某一χ^2值的特征一般不含或较少含有类别区分信息。但这一认知建立在数据集合类别分布处于平衡或准平衡状态的条件下，并没考虑类别分布处于不均衡状态时，文档类别分布对分类的影响，也没有考虑特征词频对分类的影响。因此，在不均衡数据集条件下，传统CHI方法存在明显的缺陷。为了避免χ^2统计量的不足，综合考虑特征在每一个类别中的具体分布，需要对数据类别分布不均衡及特征选择两类问题进行处理。本文在现有χ^2统计量中融入信息熵，以此建立新的加权χ^2统计量矩阵，较好地解决了不均衡数据集条件下的特征选择问题。通过对特征的类别分布进行一定程度的修正，不仅能够清楚地反映特征项的实际分布情况，而且能极大地改善CHI统计选择方法的性能。

为了解决不同特征类别之间存在的差异性，本文同时对特征t与类别c_i进行加权，加权后的χ^2统计量可定义为$W\chi^2(t,c_i)$。传统特征选择方法中W=1，如果小类别被分配较大权重，则小类别中χ^2会增大，这些特征被选择的机会也就会增大，从而提高小类别的分类精度，但如果分配给小类别χ^2统计量的权重值过大，则会影响大类别中特征的选择，因此如何设置权重显得尤为重要，本文将权重定义为特征t与类别c_i的信息熵值，即：

$$\begin{cases} W\chi^2(t,c_i)=\chi^2(t,c_i)H(t\mid c_i)H(c_i) \\ H(t\mid c_i)=-p(t,c_i)\log p(t\mid c_i) \\ H(c_i)=-p(c_i)\log p(c_i) \end{cases} \tag{7}$$

式中，$p(t\mid c_i)$为特征t在类别c_i中的出现概率；$p(c_i)$为类别c_i出现的概率；$p(t,c_i)$为类别c_i中特征t出现的概率。式(7)充分利用信息熵的反增长性，综合$H(t\mid c_i)$、$H(c_i)$，一方面通过对类别权重的调节，赋予小类别较大的权重，使χ^2统计量能够客观地反映类别分布对特征选择的影响，有利于小类别特征的选择，另一方面，通过控制特征的权重，有效防止噪声词被选中，确保整体的分类性能。

利用加权后的χ^2统计量建立统计矩阵$\boldsymbol{K}$：

$$\boldsymbol{K}=\begin{bmatrix} W\chi^2(t_1,c_1) & W\chi^2(t_1,c_2) & \cdots & W\chi^2(t_1,c_m) \\ W\chi^2(t_2,c_1) & W\chi^2(t_2,c_2) & \cdots & W\chi^2(t_2,c_m) \\ \vdots & \vdots & \ddots & \vdots \\ W\chi^2(t_n,c_1) & W\chi^2(t_n,c_2) & \cdots & W\chi^2(t_n,c_m) \end{bmatrix} \tag{8}$$

式中，行与列分别表示特征在不同类别和相同类别中的加权概率分布。在此基础上进行特征选择操作，将避免过多考虑特征或过多考虑分类类别的缺陷。

3.1.2 算法描述

输入：加权后的文本χ^2统计量矩阵$\boldsymbol{K}$。

输出：文本特征子集T

算法步骤：

1) $T=\varnothing$;// T 为特征集，初始化操作

2) 依次选择$\boldsymbol{K}$中每一行 t_i，进行如下处理：

① 查找每行中 $t_i^{\max}=\max\left\{W\chi^2(t_i,c_j)\right\}$ 及 $t_i^{\min}=\min\left\{W\chi^2(t_i,c_j)\right\}$；

② 通过 $\dfrac{W\chi^2(t_i,c_j)-\min\left\{t_i^{\min}\right\}}{\max\left\{t_i^{\max}\right\}-\min\left\{t_i^{\min}\right\}}$，将 t_i 转化成对应的隶属度 μ_{ij}；

③ 构造新类别向量 $\boldsymbol{c}_j^*=\left\{b_{ij}\right\}$ // b_{ij} 为 t_i 中，按照降序排列的隶属度 μ_{ij}；

④ 计算 $b_i^2=\sum_{j=1}^{m}b_{ij}^2$ // b_i^2 表示特征 t_i 对各类所提供的贡献总和；

⑤ 计算 $\varphi=\dfrac{\sum_{i=1}^{p}b_i^2}{\sum_{i=1}^{n}b_i^2}$ // φ 表示累计方差贡献率；

⑥ $\varphi\geqslant 0.85$ 时，得到特征子集 T。

算法的时间复杂度主要由步骤2)决定。算法的时间复杂度为 $O(nm)$ (n 为特征个数， m 为类别数)。此外，由算法的具体步骤可知，算法的空间复杂度为 $O(n)$。

3.2 文本特征基因提取模式

ICA的目的就是计算出分离矩阵，并得到一组相互独立的随机变量。本文采用基于负熵的快速固定点算法(FastICA)[21, 23]，通过分析多维数据间的高阶统计相关性，找出相互独立的隐含信息成分，以此提取独立特征基因并完成分量间高阶冗余的去除。

3.2.1 基于负熵的快速不动点算法

定义 2 若随机变量 y 的密度函数为 $p_y(x)$，那么其微分熵的定义为：

$$H(y)=-\int p_y(x)\log p_y(x)\mathrm{d}x \tag{9}$$

定义 3 负熵 J 的定义为：

$$J(y)=H(y*)-H(y) \tag{10}$$

式中， $y*$ 是与 y 具有相同相关(和协方差)矩阵的高斯随机向量。

负熵很难直接计算，需要用近似的方法。负熵近似的经典方法是使用高阶积累量及密度多项式方法。它相应的近似为：

$$J(y)\approx\frac{1}{12}\mathrm{E}\left\{y^3\right\}^2+\frac{1}{48}\mathrm{kurt}(y)^2 \tag{11}$$

式中， $\mathrm{kurt}(y)$ 为 y 的峭度。但是这种估计方法存在非鲁棒性问题，所以在实际问题中，一般使用非二次函数G的期望形式，相应的近似形式为：

$$J(y)\propto[\mathrm{E}\left\{G(y)\right\}-\mathrm{E}\left\{G(v)\right\}]^2 \tag{12}$$

式中，函数 G 可以选择 $G(y)=\dfrac{1}{a}\log\cosh ay$ 或 $G(y)=-\exp(-0.5y^2)$，$1\leqslant a\leqslant 2$，通常取1。

FastICA算法即找到一个单位(长度)向量 $\boldsymbol{w}$，使得对应的投影 $\boldsymbol{w}^{\mathrm{T}}z$ 的非高斯性达到极大化，非高斯性利用式(11)定义的负熵近似 $J(\boldsymbol{w}^{\mathrm{T}}z)$ 来度量。

算法的基本形式描述为：

1) 对数据进行中心化使其均值为0；

2) 然后对数据进行白化得到 z；

3) 选择要估计的独立成分个数 m，令 $i=1$；

4) 选择一个具有单位范数的初始化(可随机选取)向量 $\boldsymbol{w}_i$；

5) 更新 $\boldsymbol{w}_i\leftarrow\mathrm{E}\left\{zg(\boldsymbol{w}_i^{\mathrm{T}}z)\right\}-\left\{g(\boldsymbol{w}_i^{\mathrm{T}}z)\right\}\boldsymbol{w}_i$，函数 g 为非二次型函数G的导数；

6) 标准化 $\boldsymbol{w}_i$，$\boldsymbol{w}_i\leftarrow\boldsymbol{w}_i/\|\boldsymbol{w}_i\|$；

7) 如果尚未收敛，返回步骤5)；

8) 令 $i\leftarrow i+1$，如果 $i\leqslant m$，返回步骤4)。

3.2.2 文本特征基因提取算法描述

在前面的内容中，分别介绍了加权 χ^2 统计量的文本特征分布矩阵和独立成分分析的文本特征基因提取方法，本文将这两种方法结合来进行文本特征基因提取。具体算法描述：

由前面算法得到训练集的文本特征子集 $X=(x_1,x_2,\cdots,x_p)$。

1) 中心化特征子集

计算文本特征子集 $X=(x_1,x_2,\cdots,x_p)$ 的平均向量：

$$\overline{x}=\mathrm{E}(X)=\frac{1}{p}\sum_{i=1}^{p}x_i$$

中心化后的文本特征子集 $\overline{X}=(\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_p)$，其中 $\overline{x}_i=x_i-\overline{x}$，$\overline{x_i}\in R^n$。

2) 白化

计算文本特征子集 $\overline{X}=(\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_p)$ 的协方差矩阵：

$$\boldsymbol{C}_x=\mathrm{E}\left\{\overline{X}\overline{X}^{\mathrm{T}}\right\}=\boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^{\mathrm{T}}$$

式中，$\boldsymbol{E}$ 为 $\boldsymbol{C}_x$ 的特征向量矩阵，且为正交矩阵；$\boldsymbol{D}$ 为 $\boldsymbol{C}_x$ 的特征值矩阵，且为对角矩阵。

线性白化变换为：

$$\boldsymbol{V}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{E}^{\mathrm{T}}$$

白化后的数据为：

$$\boldsymbol{Z}=\boldsymbol{V}\overline{\boldsymbol{X}}$$

3) 利用3.2.1节中的算法计算得到各独立成分。

4　实验与结果分析

4.1　实验设计

本文实验中所有代码均在Matlab R2015a平台上编写，编译运行的PC机参数为：HP Pavilion 15，Intel i7-6500U、8 GB内存、Win10 64位操作系统。

为验证该文所提方法的合理性，选用搜狐新闻数据(SogouCS)20151022语料库作为实验数据集，语料库包含12个大类，共10 902个文本，其类间分布并不均衡，其中，最大类中包含2 254个文本，最小类中只包含130个文本。为了检验各方法的实际处理效果，对语料库进行了一定程度的补充和优化，如：补充了部分类型的文本、去除了一些数据不完整的文本，并从处理后的语料库中选取7个类别：农林渔畜、医学、娱乐、电子游戏、自然科学、艺术、运动休闲，类别选择为有目的的随机，选取类别中文本分布呈现较大的差异性，其中数据集中训练集包含4 360个文本，测试集包含1 336个文本，训练集的文本分布如图1所示。文本预处理阶段采用中国科学院的中文分词工具ICTCLAS对数据进行处理。分类算法采用性能较优的SVM，具体实现利用台湾大学LIBSVM工具包。

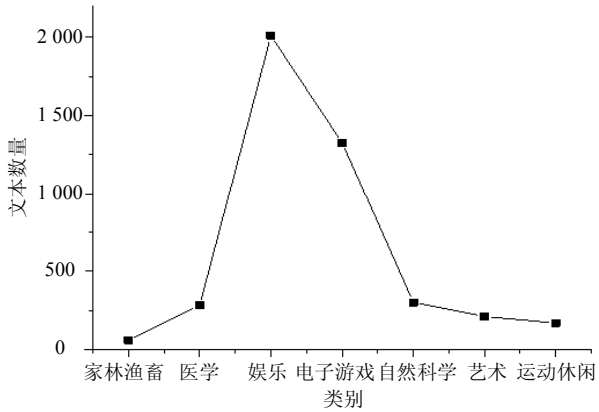

图1　训练集各类别文本数量

目前对分类器性能的评价通常选取查全率、准确率、F_1三项指标。由于在不均衡数据集的分类中，分类结果极易偏向大类，如果仍采用这类传统指标，则无法真实评价分类器的实际性能。本文采用宏平均查全率、宏平均准确率、宏平均F_1以及混合矩阵下的真正类率(true positive rate, TPR)、负正类率(false positive rate, FPR)、精确度(accuracy rate, ACC)及AUC(area under the ROC curve)等指标对分类结果进行评价。

1) 宏平均查全率为：

$$\text{Macro}\_\bar{r}=\frac{1}{|C|}\sum_{i=1}^{|C|}r_i \tag{13}$$

式中，r_i表示第i个类别的查全率；$|C|$表示类别数。

2) 宏平均准确率为：

$$\text{Macro}\_\bar{p}=\frac{1}{|C|}\sum_{i=1}^{|C|}p_i \tag{14}$$

式中，p_i表示第i个类别的准确率。

3) 宏平均F_1为：

$$\text{Macro}\_\overline{F1}=\frac{1}{|C|}\sum_{i=1}^{|C|}F1_i \tag{15}$$

式中，$F1_i$表示第i个类别的$F1$值。

4) TPR、FPR、ACC分别为：

$$\text{TPR}=\text{TP}/(\text{TP}+\text{FN}) \tag{16}$$

$$\text{FPR}=\text{FP}/(\text{FP}+\text{TN}) \tag{17}$$

$$\text{ACC}=(\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{FN}+\text{TN}) \tag{18}$$

式中，TP和TN表示实际正类和负类被正确判定的数目；FP表示将负类判定为正类的数目；FN表示将正类判定为负类的数目。

5) AUC表示ROC曲线下的面积。当ROC曲线不能很好地说明分类器效果时，通过该数值，能够更清晰的说明分类器效果。

4.2　实验测试与分析

为了得到更具统计意义的实验结果，本文采用5折交叉检验方法，对DF、IG、MI以及新提出的文本特征基因提取方法(text feature gene extraction, TFGE)在SVM分类器上进行有效性评估。

表1列出了DF、IG、MI以及加权χ^2统计量矩阵特征选择方法在搜狐新闻数据(SogouCS)20151022语料库上用LIBSVM分类器的实验结果(采用径向基核函数)。可以看出，在选取特征数量较少时，采用加权χ^2统计量矩阵特征选择方法优势比较明显，表明加权方法能够更早地获得较好的分类效果。

表1　不同特征选择方法的性能比较

特征数	DF		IG		MI		加权χ^2均值特征选择方法	
	准确率/%	$F1$/%	准确率/%	$F1$/%	准确率/%	$F1$/%	准确率/%	$F1$/%
50	62.13	52.57	66.11	53.61	55.34	45.43	74.62	65.14
100	63.59	56.10	65.47	56.33	56.41	50.11	73.59	66.01
500	74.14	66.17	75.12	69.42	69.78	63.69	84.34	79.17
1 000	77.13	74.37	80.66	76.13	74.75	71.18	86.44	82.51
3 000	79.59	75.61	82.47	79.05	76.4	72.26	90.45	84.13
5 000	80.05	78.93	83.76	81.87	77.23	73.41	88.13	84.42

表2表示经过FastICA提取后的特征子集在LIBSVM分类器上的正确率变化情况(采用径向基核函数)。可以看出，虽然分类识别准确率有小幅度下降，但特征子集规模减少约41%，实现了在大数据集高维特征空间中高阶冗余的去除，所提取出的特征基因数据泛化能力显著增强。

表2 加权 χ^2 统计量矩阵特征选择与特征基因提取方法的性能比较

加权 χ^2 矩阵+SVM		加权 χ^2 矩阵+FastICA+SVM	
特征数	宏平均准确率/%	特征数	宏平均准确率/%
50	74.62	35	71.66
100	73.59	69	72.86
500	84.34	350	83.83
1 000	86.44	697	84.46
3 000	90.45	2 041	88.6
5 000	88.13	3 431	87.11

表3列出了DF、IG、MI以及新提出的TFGE在搜狐新闻数据(SogouCS)20151022语料库上用LIBSVM分类器的实验结果(采用径向基核函数，特征空间向量维数设置成4 000)。可以看出，TFGE的效果优于DF、IG和MI，宏平均F1值提高显著，且该方法相较DF、IG和MI方法，SVM分类器的运算时间缩短了$\frac{1}{5}\sim\frac{1}{2}$，计算复杂度大大降低。

表3 4种方法的总体性能比较

方法	性能		
	宏平均查全率/%	宏平均$F1$值/%	运行时间/s
DF	75.32	77.13	347
IG	77.57	81.23	201
MI	72.36	75.57	129
TFGE	83.95	87.71	59

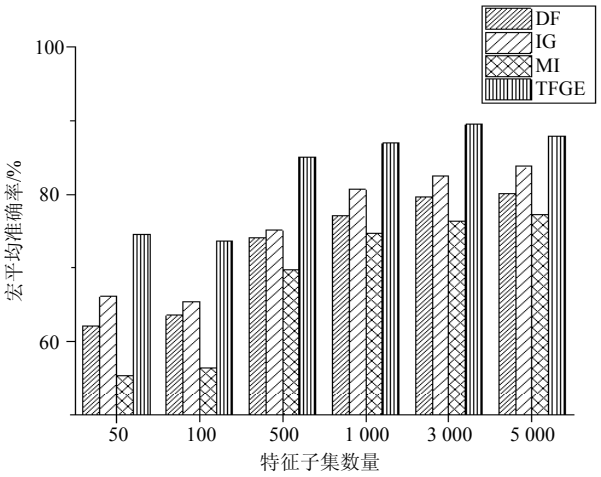

图2 4种方法的分类效果比较

从图2中可以明显看出，TFGE在特征数量为500时，就可以使SVM分类器的宏平均准确率达到85.04%，在特征数量为3 000时，SVM分类器的宏平均准确率达到最高89.47%，之后趋于稳定。由此可以看出，文本特征基因提取方法TFGE所选特征子集泛化性能优于DF、IG和MI。

在搜狐新闻数据(SogouCS) 20151022语料库上，采用将运动休闲类的230个样本作为正例，其余为负例的方式，构造不均衡二分问题。表4列出了DF、IG、MI以及TFGE采用LIBSVM分类器的实验结果。可以看出，DF与IG方法在解决不均衡问题时，两者差异不大，MI方法则相对性能较差，TFGE各项指标均优于其他3种方法，表现出较优的分类性能。

表4 4种方法在二分问题上的性能比较

方法	性能			
	TPR/ %	FPR/ %	ACC/ %	AUC/ %
DF	73.17	24.57	75.3	82.54
IG	75.91	23.78	76.2	83.61
MI	70.46	28.52	71.4	78.69
TFGE	81.37	16.11	83.76	90.71

5 结 束 语

本文对传统特征降维方法及不均衡数据集下的数据分析进行了深入研究，结合 χ^2 统计分布矩阵及独立成分分析思想，提出了一种新颖的文本特征基因子集提取方法，实现了多维统计数据中隐含信息成分的提取，克服了传统分类器因类别分布不均衡导致的性能不佳问题。文中重点介绍了CHI统计选择方法，利用信息熵构建了加权 χ^2 统计量的文本特征分布矩阵，避免采用过抽样或欠抽样方法对原始不均衡数据集的类别分布所做出的改变，通过特征类别分布的修正，极大改善CHI统计选择方法的性能。在不均衡数据集特征选择过程中优先去除冗余和噪声样本，使得在减少数据的同时保留更多的有用信息。最后采用基于负熵的FastICA进行多维数据间的独立隐含信息成分提取，获取文本特征基因用于分类器集成，进一步提高对训练样本中有效信息的利用率。实验结果表明，本文提出的不均衡大数据集下的TFGE使文本分类求解效率和准确率得到一定的提高，新方法有效且实用。另外，特征基因提取选择仅仅是数据预处理的一个步骤，将文本特征基因提取思想用于实际应用，更好地适应其他领域的分类需求，是今后要进行的研究工作。

参 考 文 献

[1] 马力, 刘惠福. 一种改进的文本特征提取算法[J]. 西安邮电大学学报, 2015, 20(6): 79-81.
MA Li, LIU Hui-fu. Study on the extraction of characteristics of chinese text based on the LDA model[J]. Journal of Xi'an University of Posts and Telecommunications, 2015, 20(6): 79-81.

[2] 曾琦, 周刚, 兰明敬, 等. 一种多义词词向量计算方法[J]. 小型微型计算机系统, 2016(7): 1417-1421.
ZENG Qi, ZHOU Gang, LAN Ming-jing, et al. Polysemous word multi-embedding calculation[J]. Journal of Chinese Computer Systems, 2016(7): 1417-1421.

[3] LI A, ZANG Q, SUN D, et al. A text feature-based approach for literature mining of lncRNA-protein interactions[J]. Neurocomputing, 2016, 206: 73-80.

[4] YUAN M. Feature extension for short text categorization using frequent term sets[J]. Procedia Computer Science, 2014, 31: 663-670.

[5] PEREZ-TELLEZ F, CARDIFF J, ROSSO P, et al. Weblog and short text feature extraction and impact on categorisation[J]. Journal of Intelligent & Fuzzy Systems, 2014, 27(5): 2529-2544.

[6] MORENO A, REDONDO T. Text analytics: the convergence of big data and artificial intelligence[J]. International Journal of Interactive Multimedia and Artificial Intelligence, 2016, 3: 57-64.

[7] YUAN S, XIANG Y, SHI J E. Text big data content understanding and development trend based on feature learning[J]. Big Data Research, 2015(3): 1-10

[8] BOUZALMAT A, KHARROUBI J, ZARGHILI A. Comparative study of PCA, ICA, LDA using SVM classifier[J]. Journal of Emerging Technologies in Web Intelligence, 2014, 6(1): 64-68.

[9] 陈睿, 张亮, 杨静, 等. 基于BSMOTE和逆转欠抽样的不均衡数据分类算法[J]. 计算机应用研究, 2014(11): 3299-3303.
CHEN Rui, ZHANG Liang, YANG Jing, et al. Classification algorithm for imbalanced data sets based on combination of BSMOTE and inverse under sampling[J]. Application Research of Computers, 2014(11): 3299-3303.

[10] 陶新民. 不均衡数据SVM分类算法及其应用[M]. 哈尔滨:黑龙江科学技术出版社, 2011.
TAO Xing-ming. Imbalance data SVM classification algorithm and its application[M]. Harbin: Heilongjiang Science and Technology Press, 2011.

[11] 李俊林, 符红光. 改进的基于核密度估计的数据分类算法[J]. 控制与决策, 2010, 25(4): 507-514.
LI JUN-lin, FU Hong-guang. Improved KDE-based data classification algorithm[J]. Control and Decision, 2010, 25(4): 507-514.

[12] LI Q J, MAO Y B, WANG Z Q. An imbalanced data classification algorithm based on boosting[C]// Proceedings of the 30th Chinese Control Conference. [S.l.]: IEEE, 2011.

[13] YIN S. A classification method for imbalanced data sets based on FCM-KFDA discriminant[J]. Journal of Huazhong Normal University, 2013, 47(6): 776-780.

[14] LI P, BI T T, YU X Y, et al. Imbalanced data classification based on AdaBoost-SVM[J]. International Journal of Database Theory & Application, 2014, 7(5): 85-94.

[15] QIONG G U, YUAN L, XIONG Q J, et al. A comparative study of cost-sensitive learning algorithm based on imbalanced data sets[J]. Microelectronics & Computer, 2011, 28(8): 145-146.

[16] KHOSHGOFTAAR T M, GAO K, NAPOLITANO A. Exploring an iterative feature selection technique for highly imbalanced data sets[C]//International Conference on Information Reuse and Integration. [S.l.]: IEEE, 2012.

[17] ALIBEIGI M, HASHEMI S, HAMZEH A. Unsupervised feature selection based on the distribution of features attributed to imbalanced data sets[J]. International Journal of Artificial Intelligence & Expert Systems, 2011, 2(1): 2011-2014.

[18] CHEN G, CHEN L. Augmenting service recommender systems by incorporating contextual opinions from user reviews[J]. User Modeling and User-Adapted Interaction, 2015, 25(3): 295-329.

[19] BULATOV A, BULATOVA N, LOGINOVICH Y, et al. Illusion of extent evoked by closed two-dimensional shapes[J]. Biological Cybernetics, 2014, 109(2): 163-178.

[20] QIU Y F, WANG W, LIU D Y. Research on an improved CHI feature selection method[J]. Applied Mechanics & Materials, 2012 (241-244): 2841-2844.

[21] 芬海韦里恩. 独立成分分析[M]. 北京: 电子工业出版社, 2007.
AAPO H. Independent component analysis[M]. Beijing: Publishing House of Electronics Industry, 2007.

[22] ROUIGUEB A, CHITROUB S, BOURIDANE A. Density estimation of high dimensional data using ICA and Bayesian networks[J]. Intelligent Data Analysis, 2014, 18(2): 157-179.

[23] 吴微, 彭华, 张帆. FastICA和RobustICA算法在盲源分离中的性能分析[J]. 计算机应用研究, 2014, 31(1): 95-98.
WU Wei, PENG Hua, ZHANG Fan. Performance analysis of FastICA and RobustICA on blind sources separation[J]. Application Research of Computers, 2014, 31(1): 95-98.

编 辑 叶 芳