多类文本分类算法 GS-SVDD

吴 德[1] 刘三阳[1] 梁锦锦[2]
(西安电子科技大学计算机学院 西安 710071)[1] (西安石油大学理学院 西安 710065)[2]

摘 要 传统多类文本多分类算法存在计算量大和训练时间长的问题,为此利用黄金分割(Golden Selection,GS)和支持向量域描述(Support Vector Domain Description,SVDD)对多类文本构造一种分类算法。GS-SVDD首先利用词频逆向文件频率(Term Frequency-Inverse Document Frequency,TF-IDF)公式计算词条的相对词频,根据该值将词条降序排列,并对得到的文本向量进行归一化;其次采用黄金分割法对文本向量进行维数约简,使得冗余的样本特征数不超过一个;最后根据支持向量域描述进行多类分类,判断待测文本归属相对类距离之值较小的类。不同数据集的数值实验表明,GS-SVDD比"一对一"和"一对多"支持向量机具有更好的稳定性、更高的分类精度和更短的训练时间,从而更适用于海量文本的多分类。

关键词 文本多分类,黄金分割,支持向量域描述,维数约简,海量文本
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.8.038

Multiclass Text Classification by Golden Selection and Support Vector Domain Description

WU De[1] LIU San-yang[1] LIANG Jin-jin[2]
(School of Computer Science and Technology, Xidian University, Xi'an 710071, China)[1]
(Faculty of Science, Xi'an Shiyou University, Xi'an 710065, China)[2]

Abstract Traditional multiclass text classification methods have disadvantages such as large computation and long training time. An algorithm based on golden selection and support vector domain description (SVDD) was proposed for text classification. The proposed method utilizes TF-IDF formula to compute the relative word frequency for each entry, sorts them in descending order and normalizes the text vector. Then golden selection method is introduced for dimension reduction, where the number of redundant sample features is no more than one. Finally, SVDD is applied for classification, which assigns the test text to the class with the smallest value of the relative class distance. Numerical experiments on various datasets demonstrate that, the proposed method has better robustness, higher classification accuracy and less training time, compared with "one-against-one" and "one-against-all" support vector machine. It is more appropriate for huge text multi-classification problems.

Keywords Multiclass text classification, Golden selection, SVDD, Dimension reduction, Huge text

1 引言

随着社会生活全面向互联网转移,尤其是智能终端、电子商务和社交网络的快速发展,人们可获取的信息呈指数级增长。如何对这些海量数据进行分类和挖掘,已成为人们研究的热点和难点。已有的文本分类技术有贝叶斯、神经网络、聚类和子数匹配算法等[1-4],但在不同程度上存在计算复杂、效率较低和泛化能力差的缺点。

支持向量机凭借其坚固的统计学习理论基础、良好的泛化性能和较高的分类精度受到广泛关注,已有学者[5,6]将其应用于两类文本分类领域。为解决多类文本分类问题,一些学者引入"一对一"[7,8]和"一对多"[9]支持向量机;但前者的子分类器数目较多,而后者需要将所有样本带入训练中,存在计算量大、训练时间长的问题。

文献[10]提出的支持向量域多分类器具有较高的分类精度和较短的训练时间,其构建的子分类器数目较少,且每次训练的样本规模较小。本文借鉴其思想进行文本多分类:运用SVDD求得各类文本的最小包围超球,根据文本到球心的距离与超球半径的比值定义相对类距离,并判断待测文本归属相对类距离之值最小的类。由于文本向量往往是高维的,且并非每个特征都对分类是不可或缺的,因此带入训练之前采用黄金分割法对文本向量进行维数约简,得到的冗余向量特征不超过一个。实验结果表明,该方法具有较好的稳定性、较短的训练时间和较高的分类精度,能够基本满足海量文本分类的需求。

到稿日期:2015-07-06 返修日期:2015-10-03 本文受国家自然科学基金(61373174),陕西省教育厅自然科学基金(2010JK773),西安石油大学博士专项科研基金(Z10027)资助。

吴 德(1979-),男,博士生,高级工程师,主要研究方向为网络安全、数据挖掘和支持向量机,E-mail:wood1979@163.com;刘三阳(1959-),男,教授,博士生导师,主要研究方向为最优化理论方法及其应用、人工智能、传感器理论等;梁锦锦(1983-),女,博士,讲师,主要研究方向为最优化理论方法及其应用。

2 基于 SVDD 的文本多分类算法

文本分类过程首先使用特定的模型来表示文本，再构造分类器，然后使用分类器对测试文本进行分类，然后最后根据评估指标来评价分类器的质量。

2.1 文本特征提取

首先对文本进行预处理，去掉一些标记和停用词等，然后使用向量空间模型(Vector Space Model，VSM)来表示文本。向量空间模型将文本 d 映射为一个特征向量 $V(d)$：

$$V(d)=(\omega_{t_1}(d),\omega_{t_2}(d),\cdots,\omega_{t_n}(d))$$

其中，$t_i(i=1,2,\cdots,n)$为一列互不相同的词条项，$\omega_{t_i}(d)$为 t_i 在 d 中的权值。

权值 $\omega_{t_i}(d)$用相对词频来表示，通过 TF-IDF 公式将词条的绝对词频(即词条在文本中出现的次数)进行归一化处理得到：

$$\omega_t(d)=\frac{tf_t(d)\times\log(\frac{N}{n_t}+0.01)}{\sqrt{\sum_{t\in d}(tf_t(d))^2\times\log^2(\frac{N}{n_t}+0.01)}}$$

其中，$\omega_t(d)$为词 t 在文本中的权值，$tf_t(d)$为词 t 在文本 d 中的词频，N 为训练文本的总数，n_t 为训练文本中出现 t 的文本数，分母为归一化因子。

2.2 维数约简

完成文本的模型表示后，构成文本的词条的数量依然非常多，导致向量空间的维数也相当高。如果能对文本的向量空间进行有效的维数约简，将能有效提高算法的分类性能。张玉芳等综述了传统的降维技术[11]，如文档频率(DF)、互信息(MI)、信息增益(IG)、χ^2 统计量(CHI)等，并指出这些方法存在计算量大和训练时间长的问题，尤其无法处理海量文本信息；夏士雄等提出的局部线性嵌入算法(LLE)[12]操作简单，但是当文本向量的维数较高、数量较多时，涉及到的稀疏矩阵过大，不易于处理；李建林提出主成分分析方法(PCA)[13]，该方法按照累计贡献率不小于 85%的原则选取方差较大的主成分，对相关度较高的文本向量取得了较好的效果；但对一般情况下的文本向量，其降维效果并不明显，且容易损失原始变量含有的部分信息，影响分类性能。

黄金分割法是优化计算中的经典算法，其依据“去坏留好”原则、对称原则和等比收缩原则来逐步缩小搜索范围。由于收敛速度均匀、算法简单、便于实现等优点，黄金分割法被广泛应用到优化、控制、交通等多个领域[15-17]。张丽娜等[18]采用黄金分割法对大样本数据构造一种特征提取方法，取得了较好的结果。

本文将文献[18]的思想推广到文本向量的分类领域，采用黄金分割法对文本向量进行维数约简。具体算法描述如下：

1)利用 TF-IDF 公式计算得到词条的绝对词频，根据绝对词频的值降序排列并进行归一化处理得到向量 $V(d)$。记向量 $V(d)$的维数(特征数)为 N，令 $N_{old}=N$。

2)取前 $N_{new}=I(0.618\times N_{old})$个特征参与训练和测试，其中 $I(x)$为取整函数。

3)若分类器精度优于预定目标值，则令 $N_{old}=N_{new}$，返回步骤 2)；否则执行步骤 4)。

4)若分类器精度低于预定目标值，则令 $N_{new}=N_{new}+I(0.618\times(N_{old}-N_{new}+1))$，用前 N_{new} 个特征参与训练和测试。

5)当$(N_{new}-N_{old})<1$ 时，终止循环。

2.3 多分类算法 GS-SVDD

采用“一对一”和“一对多”进行文本分类，需要把所有样本带入训中练，计算复杂且训练时间长。

采用 GS-SVDD 进行文本分类，子分类器数 k 远小于“一对一”的分类器数 $k(k-1)/2$，且单次训练样本数 l/k 远小于“一对一”和“一对多”的样本数 $2l/k$ 和 l(这里假定每类样本数目均等)。

GS-SVDD 首先用 2.2 节方法对文本向量进行降维；其次，以 SVDD 构造每类样本的最小包围超球，将数据划分为单一超球类内区域 A、B、C、超球重叠区域 D 和超球外区域 E；最后，定义待测样本的相对类距离(Relative Class Distance，RCD)，判断待测样本所处区域，并采取相应分类准则。

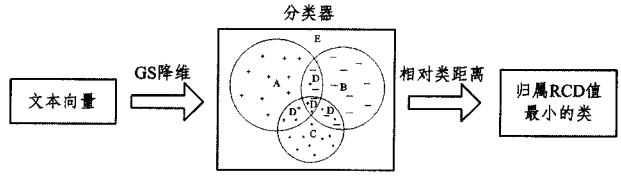

图 1 GS-SVDD 的分类原理示意图

GS-SVDD 算法：

步骤 1 输入 k 类训练文本向量$\{\boldsymbol{x}_i,y_i\}_{i=1}^{l}(\boldsymbol{x}_i\in\mathbf{R}^n,y_i\in\{1,2,\cdots,k\})$，并对其采用黄金分割法降维。

步骤 2 惩罚因子和核参数的初始值为 $C_1,C_2,\cdots,C_k$ 和 $\sigma_1,\sigma_2,\cdots,\sigma_k$。

步骤 3 构造对应 $y_i=1,y_i=2,\cdots,y_i=k$ 这 k 类文本的最小包围超球，记作$(\boldsymbol{a}_1,R_1),(\boldsymbol{a}_2,R_2),\cdots,(\boldsymbol{a}_k,R_k)$。

步骤 4 计算待测文本 z 到球心 $\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_k$ 的距离 $\|\boldsymbol{z}-\boldsymbol{a}_1\|,\|\boldsymbol{z}-\boldsymbol{a}_2\|,\cdots,\|\boldsymbol{z}-\boldsymbol{a}_k\|$。

步骤 5 计算 k 个相对类距离 RCD 的值 d_i，$d_i=\|\boldsymbol{z}-\boldsymbol{a}_i\|/R_i(i=1,2,\cdots,k)$。

步骤 6 根据 3 条分类准则，判断待识别样本的归属。

(1)对位于某单一超球类内的样本，判断样本归属该超球对应的类。

此时，第 i 类样本的相对类距离之值 d_i 满足 $d_i\leqslant 1,d_j>1(i\neq j)$。

(2)对位于超球重叠区域的样本，判断样本归属该值最小的类。

此时，$d_i\leqslant 1(i=1,\cdots,c;c\geqslant 2)$，即有多于两个的相对类距离之值不大于 1。

(3)对位于超球外区域的样本，判断样本归属该值最小的类。

此时，$d_i>1(i=1,2,\cdots,k)$，即所有的相对类距离之值大于 1。

如果出现几个相同的 d_i 而分母 R_i 不同，则比较分子 $\|\boldsymbol{z}-\boldsymbol{a}_i\|$ 的大小，并判断文本归属分子之值最小的类。如果出现几个相同的 d_i 且分母 R_i 亦相同，说明这些类的分布相似，此时判断文本归属任一类都是正确的。

2.4 分类质量评价

文本分类有多种评价指标，常用的评价指标是训练时间、测试时间、查全率和查准率。查全率是分类器找到的正确文本数与应有的正确文本数之比，查准率是分类器找到的正确文本数与分类器找到的全部文本数之比。查全率和查准率反映了分类质量的不同方面，实际应用时应综合考虑两者间的

关系，为此引入如下指标：

$$F_\beta=\frac{(\beta^2+1)PR}{\beta^2P+R}$$

其中，P 为查准率，R 为查全率，参数 β 衡量查全率对于查准率的相对重要性。当 $\beta=1$ 时，表明给查全率和查准率赋予相同的权重，此时评价指标为 $F_1=\frac{2PR}{P+R}$。

对于多分类文本情形，不妨假设共有 n 类。记第 $i(i=1,2,\cdots,n)$ 个类上的评价指标为 $F_1(i)$，本文采用宏平均 F1 的方法，直接计算所有类上评价指标的平均值。宏平均 F1 的数学公式如下：

$$F1=\frac{1}{n}\sum_{i=1}^{n}F_1(i)$$

3 实验结果和分析

仿真实验用 Matlab7.4 编程实现，运行环境为 Celeron T300 Dual-Core/1.8G、1GB 内存的 PC 机。实验采用性能良好的高斯核函数 $K(x,y)=\exp(-\|x-y\|^2/\sigma^2)$，并使用雅虎(www.yahoo.com)网站下载的文本、WebKB 数据集和路透社提供的 RCV1-v2 文档集。

首先，针对从雅虎网站中下载的文本向量，研究不同特征降维方法下多分类器 GS-SVDD 的分类质量表现；其次，对 WebKB 数据集和 RCV1-v2 文档集，对比多分类算法“一对一”(O-A-O)和“一对多”(O-A-A)的数值实验结果。WebKB 数据集选取的文本分为固定的 5 类，分别是 Course、Faculty、Project、Student 和 Staff 等。RCV1-v2 文本集按照主题分类共有 108 种类别，从中选取的文本类别是根据选取样本的数量而动态增加的，选择样本分类最多达到 10 种。

例 1　雅虎网站数据集

从 Business、Science 和 Sports 3 个主题板块下载共 1500 篇文本，每个类别各占 500 篇。随机选取其中 750 篇文本作为训练样本，其余作为测试样本。为避免随机性，取 10 次实验的平均结果。首先使用 DF、MI、IG、CHI 和 GS 不同的降维方法分别对样本进行维数约简，再统一使用 SVDD 对样本进行分类。设定惩罚因子 $C=10$，核参数取 $\sigma=0.5$；比较采用不同特征降维方法得到的分类质量，具体结果如表 1 所列。

表 1　不同文本特征降维方法的分类质量比较

算法	训练时间 (s)	测试时间 (s)	支持向量数	宏平均 F1 (%)
DF-SVDD	156.26	198.84	31	61.23
MI-SVDD	222.02	267.26	46	73.54
IG-SVDD	361.27	429.67	75	75.36
CHI-SVDD	384.56	439.18	81	76.03
PCA-SVDD	294.33	290.08	50	83.76
GS-SVDD	255.28	254.64	52	85.96

从表 1 可以看出：

1)与 DF-SVDD 和 MI-SVDD 相比，GS-SVDD 的宏平均 F1 值分别高 24.73%和 12.42%，训练时间和测试时间稍长些，而支持向量的数目也多些。这是由于采用黄金分割法有效、合理地对文本向量进行了降维。考虑到宏平均 F1 值的显著提高，增加的时间是值得的。

2)与 IG-SVDD 和 CHI-SVDD 相比，GS-SVDD 的宏平均 F1 值分别高 10.6%和 9.93%，但支持向量数明显较少，训练时间和测试时间也显著缩短。

3)与 PCA-SVDD 相比，GS-SVDD 的分类精度高了 2.2%，支持向量机数目相当，而训练时间和测试时间分别缩短了约 13%和 12%。这说明，采用黄金分割法比主成分分析法进行降维更简便快捷。

总体看来，GS-SVDD 在分类时间和分类精度上取得了很好的平衡：当分类精度相当时，GS-SVDD 在训练和测试时间上占有显著优势；当训练和测试时间较长时，宏平均 F1 值又相对较高。

例 2　WebKB 数据集

从 Course、Faculty、Project、Student 和 Staff 5 个类别中随机选取 2000 个文本，其中训练和测试文本各 1000 个。设定惩罚因子 $C=10$，不同分类器对 10 次随机实验的分类质量对比如表 2 所列；其中训练时间、测试时间、支持向量数和宏平均 F1 值均为 10 次分类计算的平均值。

表 2　不同分类器的分类质量比较

算法	训练时间 (s)	测试时间 (s)	支持向量数	宏平均 F1 (%)
GS-SVDD	156.26	198.84	15	80.54
O-A-O	282.02	207.26	46	73.54
O-A-A	365.02	284.64	108	62.03

由表 2 得到以下结论：

1)GS-SVDD 的宏平均 F1 值比“一对一”和“一对多”高 7%和 8.51%。这是由于 GS-SVDD 运用黄金分割法对样本特征进行维数约简，从而提高了分类性能。

2)GS-SVDD 具有最少的支持向量数和最短的训练及测试时间。这是由于 GS-SVDD 每次仅对单一类别的样本进行训练，训练样本规模较小。

例 3　RCV1-v2 文本集

按照主题分类从中分别随机选取 1000、2000 和 3000 个样本，训练样本和测试样本的数量比例是1:1。其中 1000 个样本中包含 3 个类别，2000 个样本包含 5 个类别，3000 个样本中包含 10 个类别。设定惩罚因子 $C=10$，不同分类器的多分类质量结果比较如表 3 所列。

表 3　不同分类器的多分类质量比较

类别数/样本数	GS-SVDD		O-A-O		O-A-A	
	训练速度 (样本数/s)	宏平均 F1(%)	训练速度 (样本数/s)	宏平均 F1(%)	训练速度 (样本数/s)	宏平均 F1(%)
3 / 1000	6.16	85.34	3.54	85.32	2.73	78.29
5 / 2000	6.06	83.05	2.31	79.66	1.15	67.87
10 / 3000	5.92	80.45	1.98	72.43	0.49	58.02

注：训练速度＝训练样本数/训练时间。

由表 3 结果可知，GS-SVDD 的分类质量明显优于“一对多”和“一对一”的分类质量。

1)GS-SVDD 具有最快的训练速度。随着文本向量类别和数量的增多，GS-SVDD 的训练速度基本保持不变，其训练时间基本上呈线性增长；“一对一”和“一对多”的训练速度迅速降低，其训练时间呈指数级增加。

这是由于 GS-SVDD 每次仅将单类样本带入 SVDD 训练，没有重复使用训练样本。

2)GS-SVDD 具有最高的宏平均 F1 值。随着文本向量的类别和数量的增多，GS-SVDD 的宏平均 F1 值可分别比“一对一”和“一对多”高 8.02%和 22.43%。

这是由于 GS-SVDD 定义相对类距离，根据待测文本向量所处的空间位置来判断所属类别。

GS-SVDD 具有快的训练速度和高的宏平均 F1 值这两个特性，保证了其适用于海量文本的分类问题。众所周知，海量文本的特点是样本数量巨大，文本类别也很多；同时类别之间的文本数量不平衡，且文本的长度参差不齐。对海量文本进行分类时，“一对一”支持向量机构造的子分类器数目随文本的类别数呈超线性增长，并且在测试阶段必须计算所有子分类函数，计算量大。“一对多”支持向量机依次选出一类文本向量与其余类构造分类器，两类数据分布不平衡，分类精度较低。GS-SVDD 的训练仅对单一类别样本进行，避免了两类数据分布不平衡的问题；其分类器数目与文本向量的类别数相同，且根据文本向量所处的空间位置判断归属的方法保证了较高的分类精度。

结束语 GS-SVDD 利用黄金分割有效降低了文本向量的维数，没有冗余特征；利用支持向量域描述(SVDD)构造各类文本的最小包围超球，根据最小相对类距离之值判断类重叠区域文本的归属，具有较短的训练时间和较高的宏平均 F1 值。文本分类数据上的数值实验表明所提算法可行有效，分类质量较高；尤其是训练速度随着训练样本的增加而保持稳定。

参 考 文 献

[1] Sebastiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47

[2] Su Jin-shu, Zhang Bo-feng, Xu Xin. Advances in machine learning based text categorization[J]. Journal of Software, 2006, 17(9): 1848-1859(in Chinese)
苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859

[3] Dong Yue-hua, Guo Shi-chuan. Text clustering algorithm with improved weighting factor and feature vector[J]. Computer Engineering and Design, 2015, 35(4): 1051-1057(in Chinese)
董跃华，郭士串. 结合权重因子与特征向量改进的文本聚类算法[J]. 计算机工程与设计，2015，35(4)：1051-1057

[4] Zhang Pei-yun, Chen Chuan-ming, Huang Bo. Texts similarity algorithm based on subtrees matching[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(3): 226-234(in Chinese)
张佩云，陈传明，黄波. 基于子树匹配的文本相似度算法[J]. 模式识别与人工智能，2014，27(3)：226-234

[5] Wan C H, Lee L H, Rajkumar R, et al. A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine[J]. Expert System with Application, 2012, 39(15): 11880-11888

[6] Arun K M, Gopal M. A comparison study on multiple binary-class SVM methods for unilabel text categorization[J]. Pattern Recognition Letters, 2010, 31(11): 1437-1444

[7] Kumar M A, Gopal M. One-against-one fuzzy support vector machine classifier: An approach to text categorization[J]. Expert System with Application, 2009, 36(6): 10030-10034

[8] Lin Xu-dong, Liu Han-xing, Lin Pi-yuan, et al. Chinese question classification using alternating and iterative one-against-one algorithm[J]. Journal of Convergence Information Technology, 2010, 5(3): 61-67

[9] Kumar M A, Gopal M. Reduced one-against-all method for multiclass SVM classification[J]. Expert System with Application, 2011, 38(11): 14238-14248

[10] Wu De, Liu San-yang. Multiple support vector domain classifier [J]. Journal of Xi'an Jiaotong University, 2012, 46(6): 87-91(in Chinese)
吴德，刘三阳. 支持向量域多分类器[J]. 西安交通大学学报，2012，46(6)：87-91

[11] Zhang Yu-fang, Wan Bin-hou, Xiong Zhong-yang. Research on feature dimension reduction in text classification[J]. Application Research of Computer, 2012, 29(7): 2541-2543(in Chinese)
张玉芳，万斌候，熊忠阳. 文本分类中的特征降维方法研究[J]. 计算机应用研究，2012，29(7)：2541-2543

[12] Xia Shi-xiong, Li You-wen, Zhou Yong. Method based on semi-supervised local linear algorithm for text classification[J]. Application Research of Computer, 2010, 27(1): 64-67(in Chinese)
夏士雄，李佑文，周勇. 一种半监督局部线性嵌入算法的文本分类方法[J]. 计算机应用研究，2010，27(1)：64-67

[13] Li Jian-lin. A combination of feature extraction in text classification based on PCA[J]. Application Research of Computer, 2013, 30(8): 2398-2401(in Chinese)
李建林. 一种基于 PCA 的组合特征提取文本分类方法[J]. 计算机应用研究，2013，30(8)：2398-2401

[14] Duan Jie, Hu Qing-hua, Zhang Ling-jun, et al. Feature selection for multi-label classification based on neighborhood rough sets [J]. Journal of Coumputer Research and Development, 2015, 52(1): 56-65(in Chinese)
段洁，胡清华，张灵均，等. 基于邻域粗糙集的多标记分类特征选择算法[J]. 计算机研究与发展，2015，52(1)：56-65

[15] Song Ju-long, Qian Fu-cai. The global optimization method based on golden-section[J]. Computer Engineering and Applications, 2005, 48(4): 95-96(in Chinese)
宋巨龙，钱富才. 基于黄金分割的全局最优化方法[J]. 计算机工程与应用，2005，48(4)：95-96

[16] Yang Wen-chen, Zhang Lun, Rao Qian, et al. Multi-objective optimization for traffic signals with golden Ration based genetic algorithm[J]. Journal of Transportation Systems Engineering and Information Technology, 2013, 13(5): 48-55(in Chinese)
杨文臣，张轮，饶倩，等. 基于黄金分割点遗传算法的交通信号多目标优化[J]. 交通运输系统工程与信息，2013，13(5)：48-55

[17] Zhong Hua, Wang Yong, Shao Chang-xing. Golden-section adaptive control based on disturbances and model error compensations[J]. Application Research of Computer, 2015, 32(8): 2343-2346(in Chinese)
钟华，王永，邵长星. 基于扰动和模型误差补偿的黄金分割自适应控制[J]. 计算机应用研究，2015，32(8)：2343-2346

[18] Zhang Li-na, Zhou Run-jing, Na Ri-su. A method for characteristic extraction from large sample databased on the golden section method's ISODATA Algorithm[J]. Journal of Inner Mongolia University(Natural Science Edition), 2013, 44(1): 93-96(in Chinese)
张丽娜，周润景，那日苏. 基于黄金分割法的 ISODATA 算法的大样本特征数据提取方法[J]. 内蒙古大学学报(自然科学)，2013，44(1)：93-96