文本分类中特征提取方法的比较与分析

屈军[1]，林旭[2]

(1、广东省台山电视大学，台山 529200；2、广东省工业贸易学校，佛山 528000)

摘 要：研究了在文本分类中，各种特征提取方法对分类效果的影响，比较了特征提取方法交叉熵(CE)、信息增益(IG)、互信息(MI)、及 χ^2 对文本分类器性能的影响，分析了这几种特征提取方法对 SVM 和 KNN 分类器性能的影响。

关键词：文本自动分类；KNN；SVM；特征提取

0 引 言

文本自动分类是根据一组事先知道类别的文档作为训练样本，建立一个分类模型，来求得未知类别的文档的类别。特征项的选择在文本分类系统中有比较充分的研究，基本方法是根据词汇在文档集中出现的频度来选取，一般的原则是，将文档中出现的词汇按频度排序，选取频度在一定范围内的词汇作为特征词。特征提取方法是文本自动分类中的一项关键技术和瓶颈技术，如何从原始文本特征集合中选择最能表示文本主题内容的特征子集，是文本特征提取算法的研究目标。目前，有多种特征抽取算法被用于文本自动分类的研究中，但这些算法都有其优点和缺点，没有公认的最优方法，需要针对具体系统进行对比来确定最优方法。特征选择可以从两个方面提高文本分类系统性能：一是分类速度，通过特征选择，可以大大减少特征集合中的特征数，降低文本向量的特征数，简化计算，防止过度拟合，提高系统运行速度；二是准确率，通过选择适当的特征，不但不会降低系统准确性，反而会使系统精度提高[1]。

目前，文本分类领域较常用的特征选择算法[2,3]有：文档频率(Document Frequency)、信息增益(Information Gain)、互信息(Mutual Information)、χ^2 统计(Chi-square Statistic)、交叉熵法 (Cross Entropy)、优势率(Odds Ratio)等。

1 常用特征选择算法

1.1 信息增益

信息增益是一种在机器学习领域应用较为广泛的特征选择方法。它是从信息论角度出发，根据特征取值情况在划分学习样本空间时，以所获信息增益的多寡，来选择相应的特征。对于特征 t 和文档类别 c，IG 考察 c 中出现和不出现 t 的文档频数来衡量 t 对于 c 的信息增益。特征 t 对于文档类别 c 的信息增益 IG (t,c)计算公式如下：

$$IG(t,c)=P(t_k,c_i)\log\frac{P(t_k,c_i)}{P(c_i)P(t_k)}+P(\bar{t}_k,c_i)\log\frac{P(\bar{t}_k,c_i)}{P(c_i)P(\bar{t}_k)}$$

其中：C 为某一类文档集合；$\bar{t}$表示特征 t 不出现。信息增益的不足之处在于它考虑了单词未发生的情况，即在式中的 $P(\bar{t}_k,c_i)\log\frac{P(\bar{t}_k,c_i)}{P(c_i)P(\bar{t}_k)}$ 部分。虽然某个单词不出现也可能对判断文本类别有贡献，但实验证明，这种贡献往往远小于考虑单词不出现情况所带来的干扰。特别是在类分布和特征分布高度不平衡的情况下，绝大多数类都是负类，绝大多数特征值都是‘不出现’的，此时信息增益大的特征主要是信息增益公式中后一部分(代表单词不出现情况)大，而非前一部分(代表单词出现情况)大，信息增益的效果就会大大降低了。

1.2 期望交叉熵

$$CrossEntropyTxt(t)=P(t)\sum_j P(C_j|t)\log\frac{P(C_j|t)}{P(C_j)}$$

期望交叉熵是一种基于概率的方法。信息增益要求计算所有特征属性的值，而期望交叉熵则只计算出现在文档中的单词。其中 $P(C_j|t)$表示文本中出现词 t 时，文本属于 C_j 的概率，$P(C_j)$是类别出现的概率。如果词和类别相关，也就是 $P(C_j|t)$大，且相应的类别出现概率又小，则说明词对分类的影响大，相应的函数值就大，就很可能被选中作为特征项。交叉熵反映了文本类别的概率分布和在出现了某个特定词的条件下文本类别的概率分布之间的距离，特征词 t 的交叉熵

越大，对文本类别分布的影响也越大。

1.3 互信息

在统计学中，互信息用于表征两个变量的相关性，常被用来作为文本特征相关的统计模型及其相关应用的标准。特征 t 与 c 类文档之间的相互信息 MI(t, c)的定义如下：

$$MI(t,c)=\log\frac{P(t|s)}{P(t)}=\log\frac{P(t,c)}{P(t)\times P(c)}$$

其近似计算公式为：

$$MI(t,c)\approx\log\frac{A\times N}{(A+C)\times(A+B)}$$

其中 A 为特征 t 与文档 c 类同时出现的次数；B 为特征 t 出现而 c 类文档不出现的次数；C 为 c 类文档出现而特征 t 不出现的次数；N 为文档总数。如果 t 与 c 相互之间独立，那么 MI (t, c)为零。互信息的缺点是受临界特征的概率影响较大，从公式中可以看出，当特征的 P(t|c)值相等时，稀有词比普通词的分值要高，因此，概率相差太大的文本特征互信息值不具有可比性。它与期望交叉熵的本质不同在于它没有考虑单词发生的频度，这是互信息一个很大的缺点，因为它造成了互信息评估函数经常倾向于选择稀有单词。在一些特征词选择算法的研究中发现，如果用互信息进行特征选择，它的精度极低(只有约 30%)，原因是它删掉了很多高频的有用单词。

1.4 χ^2 统计

χ^2 统计也用于表征两个变量的相关性。对特征进行打分时，认为特征 t 与 c 类文档之间非独立关系，类似于具有一维自由度的 χ^2 分布。它计算的是特征 t 与 c 类之间的依赖关系。特征 t 与 c 类文档之间的 χ^2 统计值 χ^2 (t, c)的计算如下：

$$\chi^2(t,c)\approx\frac{N\times(A\times D-C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

其中：A、B、C 和 N 的含义与前面互信息 ((MI)所介绍的相应参量完全相同，而 D 为特征 t 与 c 类文档均不出现的次数。如果 t 与 c 之间相互独立，那么文本特征 t 的 χ^2 估计值为 0。χ^2 估计与互信息的主要区别是 χ^2 为标准值，因此同类中的特征的 χ^2 是可比的。

2 文本分类方法

从文本分类的方法来看，现有的文本分类技术主要采用三种类型的方法：基于统计的方法、基于连接的方法和基于规则的方法。 基于统计的方法有：Naive Bayes[3,4]、KNN[5,6]、类中心向量[7,8]、回归模型[9]、支持向量机[10]、最大熵模型[11]等；基于连接的方法有人工神经网络；基于规则的方法有决策树、关联规则等。现在主要使用基于统计的方法，其中经过大量实验证明，SVM 和 KNN 是性能比较优秀的分类器。

2.1 最近邻(KNN)

最近邻法是基于类比学习的一种方法。每个训练文档代表|F|维空间的一个点，这样所有的训练文档都存放在|F|维空间中。给定一个待分类文档 di，k-最近邻法搜索模式空间，找出最接近待分类文档 d 的 k 个训练文档。这 k 个训练文档称为文档 di 的“近邻”。“临近性”用欧几里德距离定义，其中两个文档 $d_i=(f_{i1},f_{i2},...f_{i|F|})$，$d_k=(f_{k1},f_{k2},...f_{k|F|})$，的欧几里德距离是 $Dist(d_i,d_k)=\sqrt{\sum_{r=1}^{|F|}(f_{i1}-f_{k1})^2}$ 档 di 被分配倒 k 个最邻近训练文档中占比例最大的类别中。当 k=1 时，文档 d_k 被指定到模式空间中与之最邻近的训练文档所属的类别中。

2.2 支持向量机(SVM)

支持向量机方法是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的，它对于一个给定的具有有限数量训练样本的学习任务，如何在正确性(对于给定训练集)和机器容量(机器可无错误地学习任意训练集的能力)进行折衷，根据 V.Vapnil 的结构风险最小化原则，尽量提高学习机的泛化能力，以得到最佳的推广性能。

设给定的训练集为：

$(x_1,y_1),(x_2,y_2)...(x_l,y_l),x\in R^d,y\in\{+1,-1\}$

且可被一个超平面(Hyper Plane)线性分割，该超平面记为：$(W\cdot X)+b=0$。

如果一个训练集中的矢量能被一个超平面无错误地线性分割，且距该超平面最近的矢量之间的距离最大，则该超平面为最佳超平面。其中距离超平面最近的，对于决策面设计起作用的点称为支持向量(SV, Support Vector)，在文本分类中，超平面是对两类分类的划分，对于有大于两类的多类文本分类，就对每个类构造一个超平面，将这一个类与其余的类分开。有多少个类就构造多少个超平面。测试的时候看哪个超平面最适合测试样本，来确定测试样本的类别。

3 实验与分析

数据集分为 10 类，包括环境、计算机、交通、教育、经济、军事、体育、医药、艺术、政治，训练集共有 1882 个文件，测试集共有 934 个文件。

3.1 特征提取方法在 KNN 和 SVM 分类器中的实验结果

对全部文本（1882 个）分词处理后进行特征提取，特征提取的数目由多至少分为 100000、30000、10000、5000、3000、500、200、100，实验的目的是比较几种不同的特征提取方法在训练集相同的情况下随着提取数目的减少在 KNN 和 SVM 分类器中的分类准确率。

表 1 特征提取方法在 KNN 中准确率比较

特征数		100000	30000	10000	5000	3000	1000	500	200	100
提取方法	IG	71.37	71.15	79.17	85.28	87.76	88.89	87.19	82.98	81.20
	MI	71.52	74.44	71.57	62.31	59.31	41.09	41.58	33.67	27.73
	CE	71.51	71.63	79.16	85.81	87.25	88.94	86.55	82.36	80.39
	χ^2	71.49	72.95	78.75	85.28	88.63	89.91	87.14	83.31	81.50
	WE	71.45	73.02	76.12	83.08	85.01	89.07	84.04	81.37	75.75

在表 1 中，KNN 分类器中各种特征提取方法的分类准确率表现出随特征数的减少先增加后降低的变化曲线。IG、CE、χ^2 特征提取方法在特征空间维度为 1000 时，分类正确率达到最大，特征数为 1000 时，CE 特征提取方法在 KNN 分类器中性能是最好的。互信息(MI)特征提取方法随着特征数的下降分类性能下降较快，分类性能也最差。说明特征向量空间过大或过小时，分类准确度都不高。选用的特征词过少时，不能反映各个类别的特征，不能准确地区分各个类别的文档，而选用的特征词过多时，一些区分度很低的冗余词汇也被加了进来，这样那些区分度较高的词在其中被‘稀释’了，不能有效地为区分文档做贡献。

表 2 特征提取方法在 SVM 中准确率比较

特征数		100000	30000	10000	5000	3000	1000	500	200	100
提取方法	IG	95.28	95.18	95.61	95.71	96.14	95.28	94.64	90.25	82.72
	MI	95.28	91.32	86.29	82.44	76.01	55.24	49.25	34.15	25.05
	CE	95.28	95.28	95.71	95.71	96.14	95.61	94.43	89.82	82.86
	$\chi 2$	95.28	95.28	95.07	95.93	95.93	95.61	94.21	89.82	86.18
	WE	95.28	95.71	95.71	95.71	95.51	95.18	93.46	87.47	79.55

在 SVM 分类器中，对于各种特征提取方法，当特征数减小到 1000 以前，分类正确率没有太大变化，都保持在 95%以上，IG、CE、χ^2、WE 这几种特征提取的分类性能比较接近，但是从数据中可以得出在特征数为 3000 时分类的确率是最高的，特征数太多和太少都会影响到分类的效果。当特征数小于 1000 后，分类的准确率变化较为明显。可以看出特征数在 30000、10000、5000、3000 时分类的准确率变化不明显，互信息(MI)特征提取方法随着特征数的下降分类性能下降较其他方法快。在保证分类准确率的前提下可以选择相对比较少的特征数，这样可以提高 SVM 的分类速度。从 KNN 和 SVM 的实验中可以得出，在得到相同的分类正确率时，SVM 分类器的文本特征数可以比 KNN 分类器少。

3.2 样本数量不平衡时在 KNN 和 SVM 的实验结果

在 KNN 分类器中几种提取方法在特征提取为 1000 时都达到一个比较好分类效果，因此在这个实验中选择特征数为 1000，也就是说几种特征提取方法在选择特征数是一样的，然后在训练集中只选择体育和经济两个类作为训练集，体育类中训练文档数为 301 个，经济类中训练文档数为 217 个。在实验中按一定的数量减少体育类别中训练样本的数量，总体样本(体育和经济之和)也会随之减少，每次体育类的训练数减少后，用 KNN 分类器对训练集(总体样本减少后的)训练，然后在测试对体育类分类准确率，研究训练样本数量变化对特征提取和分类方法的影响。实验的结果见表 3。

表 3 体育类样本减少时各种提取方法在 KNN 中的比较

体育类文本数		301	271	241	211	181	151	121	81	51	31
经济类文本数		217	217	217	217	217	217	217	217	217	217
文本训练总数		518	488	458	428	398	368	338	298	268	248
提取方法	IG	99.68	98.98	97.98	95.97	92.62	88.59	79.19	73.15	35.57	0.11
	MI	83.58	76.26	76.26	22.15	40.27	29.53	31.54	30.22	0.08	-
	CE	99.34	97.88	97.88	95.98	93.29	88.56	77.18	75.16	22.82	0.09
	x 2	99.68	99.20	98.13	97.88	95.52	90.02	82.05	68.56	14.09	0.05

IG、CE、χ^2 特提取方法在 KNN 中的性能基本保持一致，MI 的分类性能比较低而且也不稳定。当训练集中体育类样本减少的初期，IG、CE、χ^2 这三种特征提取方法还能对体育类别有较好的分类。用 KNN 分类器，当训练集中体育的样本减少到 150 个左右时，曲线开始明显下降，这时分类器对体育的分类准确率明显下降，如果再减少体育类文档在训练集中的数量，将不能对体育进行分类，这表明训练样本中的各个类别要有一定的数量，当某个类别中的样本数量太少造成整个训练集中各类的样本分布不平衡时，对包含较少数量样本的类别的分类将影响很大。

表 4 体育类样本减少时各种提取方法在 SVM 中的比较

体育类文本数		301	200	100	51	34	21	15	10	5
经济类文本数		217	217	217	217	217	217	217	217	217
文本训练总数		518	417	317	268	251	238	232	227	225
提取方法	IG	100.00	100.00	99.98	97.31	94.62	92.61	86.57	78.52	0.50
	MI	92.94	92.46	91.92	92.88	91.94	81.87	76.51	55.03	0.07
	CE	100.00	100.00	100.00	97.54	96.64	91.22	86.57	79.19	0.50
	x 2	100.00	100.00	100.00	98.11	95.22	87.11	84.33	79.20	0.50

在 SVM 中的实验环境与 KNN 相似，体育类中训练文档数为 301 个，经济类中训练文档数为 217 个。在实验中按一定的数量减少体育类别中训练样本的数量，总体样本(体育和经济之和)也会随之减少，每次体育类的训练数减少后，在要 SVM 分类器对训练集(总体样本减少后的)训练，然后在测试集中测试对体育类的准确率，研究训练样本数量变化对特征提取和

分类方法的影响。在数据分布不平衡时，除 MI 略低一些外，各特征提取方法在 SVM 分类器中对体育类的分类性能比较接近。从实验数据可以看出：SVM 分类器在训练样本不平衡时分类的准确率是比较好的，当体育类样本在训练集减少到只有 10 个时还能对体育类进行分类，而且 IG、CE、χ^2 这三种提取方法在 SVM 中分类性能都接近 80%。MI 提取方法在 SVM 中的分类性能也要比 KNN 中的好，所以当数据分布不平衡时选择 SVM 比 KNN 有更好的分类效果。

4 展望

文本自动分类中的特征提取在训练和测试时的计算量和存储量都较大，如何保证较高精度的同时，又控制算法复杂度是进一步研究高性能算法的目标。

分类器获得最佳分类性能的最佳特征数的确定。在不同的数据集上，采用不同的分类算法及不同的特征提取方法取得最佳分类性能的特征数不同，只能靠经验来确定。本文在对样本不平衡的研究中还有很多不足之处，如何有效地确定最佳特征数及定量地分析数据分布不平衡对分类性能的影响是将来研究的一个目标。

参考文献

[1]Y Yang and 10. Pedersen. A comparative study on feature selection in text categorization. In Proceedings of ICML- 97, 14th International Conference on Machine Learning, pages 412- 20, Nashville, US, 1997

[2]朱明、王军、王俊普. Web 网页识别中的特征选择问题研究. 计算机工程, 2000, 26(8)

[3]Dunja Mladenic, Marko Grobelink. Feature selection on hierarchy of web documents. Decision Support Systems, 2003, 35: 45287. 51

[4]D. D. Lewis. Naive (Bayes) at forty: The Independence Assumption in Informationc Retrieval. In Proceedings of the 10th European Conference on Machine Learning, New York, 1998, 4~15

[5]S. Eyheramendy, D. D. Lewis and and D. Madigan. On the Naive bayes model fortext categorization. Artificial Intelligence&Statistics 2003

[6]Y Yang. An evaluation of statistical approaches to text categorization. Information Retrieval, 1999, 1(1): 76~88

[7]李荣陆、胡运发. 基于密度的 KNN 文本分类器训练样本裁剪方法. 计算机研究与发展. 2004, 41(4): 539~545

[8]W. Cohen and Y Singer. Context- sensitive learning methods for text categorization. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996: 307~315

[9]Y Yang and C.G. Chute. A linear least squares fit mapping method for information retrieval from natural language texts. In Proceedings of the 14th Conference on Computational Linguistics (COLING92), 1992

[10]C. Hsu, C. Lin. A comparison on methods for multi- class support vector machines, IEEE Transactions on Neural Networks. 2002, 13: 415425

[11]K. Nigam, A. K. McCallum, S. Thrun and T. Mitchell. Text classification from labeled and unlabeled documents using EM. Machine Learning, 2000, 39(2/3): 102~134

（收稿日期：2007- 03- 12）

Comparison and Analysis of Feature Extraction Methods for Text Categorization

QU Jun[1] , XU Lin[2]

(1.Taishan Television University Guangdong, 5292002 China; 2.Technical Trade School Guangdong, 528000 China)

Abstract: Studies feature extraction in text categorization, compares cross entropy (CE), information gain(IG), mutual information(MI), χ^2- test(CHI) and class selection these four method, analyzes the influence of performance of these feature extraction methods on SVM and KNN these two classifiers.

Key words: Text Categorization; KNN; SVM; Feature Extraction