● 王东波[1]，苏新宁[1]，朱丹浩[1]，年洪东[2]
(1. 南京大学 信息管理系信息技术开发研究所，江苏 南京 210093；2. 方正集团 技术研发部，江苏 南京 210046)

基于支持向量机的医学期刊文章自动分类研究*

摘 要：基于支持向量机学习模型，使用万方期刊数据库中医学、卫生的有关标题和摘要数据，对医学、卫生大类下的R7中的9个小类进行了自动分类研究。在中文信息处理知识和技术的基础上选取分类特征，在分类过程中主要采取了基于低密度多特征的训练方法。在互信息、卡方统计、交叉熵和证据权值4个不同的统计特征量的开放测试中，自动分类的查全率和准确率都取得了相对令人满意的结果。

关键词：支持向量机；期刊；自动分类

Abstract: Based on the SVM (Support Vector Machine) learning model, and by the use of the data of the titles and abstracts concerning traditional Chinese medicine and sanitation in Wanfang Periodical Database, this paper makes an automatic classification study of 9 sub-classes in R7 of medicine and sanitation class. The classification features are selected on the basis of the Chinese information processing knowledge and technology. During the classification process, the paper mainly uses the training method based on low density and multi-feature. The recall and pertinency ratios of automatic classification is relatively satisfactory in the open test of 4 different statistic features, that is, mutual information, chi-square statistics, cross entropy and proof weight.

Keywords: SVM; journal; automatic classification

随着万方、中国知网和维普等数据资源的急速增加，如何使用文本挖掘的方法和技术从这些海量数据中获取或挖掘知识不仅具备了可能性而且变得日益迫切。文本挖掘又称为文本数据挖掘或文本知识发现[1]，指从文本集中获取隐含的、以前未知的、潜在有用的知识，如关联知识、时间序列信息，甚至科学文献的创新推断和假设等构成[2]。文本挖掘的主要研究内容由文本特征提取、文本检索、文本自动分类、文本自动聚类、本体、自动摘要、语义网和情感计算等。本文主要是在万方期刊数据库中医学、卫生有关标题和摘要数据的基础上，基于支持向量机，对《中国图书馆分类法》下的R7（R7为中医药、卫生中的一类，由妇产科学、儿科学、肿瘤学、神经病学与精神病学、皮肤病学与性病学、耳鼻咽喉科学、眼科学、口腔科学、外国民族医学等各小类组成）中的9个小类进行了自动分类研究[3]。本文的探究一方面在一定程度上有助于医学期刊自动分类的实现，另一方面可以验证基于支持向量机的自动分类方法的性能。

自从20世纪90年代AT & Bell实验室的Vapink等人提出支持向量机（Support Vector Machine，SVM）机器学习算法以来，随着自然语言处理技术和方法的日益成熟，支持向量机被广泛应用到自动分类的研究中并取得了一定的成效[4]。基于支持向量机，M. Corney等人以作者的性别对电子邮件进行了分类[5]。张学工详细介绍了统计学习理论和支持向量机方法的基本思想和特点[6]。萧嵘、王继成和张福炎对支持向量机的原理进行了详细的综述[7]。梁坤和古丽拉·阿东别克提出了一个基于SVM的情感分类方法，并对真实的新闻评论进行了实验，实验表明SVM是一种性能比较好的方法，能满足大量评论知识发现的需求[8]。翟林、刘亚军对SVM的特点进行了分析，并且分别以400到2 700为特征项数在复旦大学提供的分类语料上进行了实验，取得了令人满意的结果[9]。梁秀娟基于SVM，用向量空间模型表示文本，用互信息和词频相结合的方式对文本进行特征提取，并将其用特征向量表示出来，从而来训练各并行的两类分类机，并以500篇交叉和边缘学科类的文档作为测试文本进行验证[10]。胡燕系统研究和分析了基于二叉树的多类支持向量机分类算法，并在此基础上对其作出了改进[11]。在上述研究的基础上，基于支持向量机，采取低密度多特征的训练方法，对《中

* 本文为教育部人文社会科学重点研究基地重大项目“基于智能信息处理的知识挖掘技术及应用研究”的研究成果之一，项目批准号：08JJD870225。

国图书馆分类法》R7 中的 9 个类别进行了自动分类实验，并基于互信息、卡方统计、交叉熵和证据权值 4 个不同的统计特征量进行了实验。

1 支持向量机简介

支持向量机是在统计学习理论基础上发展而来的一种新的学习算法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。SVM 在解决小样本、非线性及高维模式识别问题上具有良好的表现，如人脸识别、手写体识别、文本分配、网页分类、搜索引擎、学科导航、邮件过滤等。

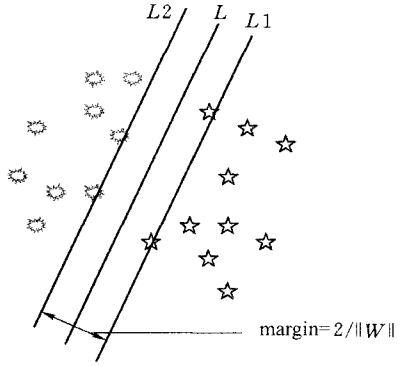

图 1 支持向量机的基本思想体现

支持向量机的基本思想可用图 1 的二维图表示出来，五角形和多边形代表两类样本，L 为分类线，$L1$，$L2$ 分别为各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫分类间隔。最优分类线就是要求分类线不仅能将两类正确分开，而且能使分类间隔达到最大。$x \cdot w + b = 0$ 为分类线方程。分类线经过归一化处理后，使得样本集 (x_i, y_i)，$i = 1, \cdots, n$，$x \in R_d$，$y \in \{+1, -1\}$，满足：

$$y_i[(w \cdot x_i) + b] - 1 \geqslant 0, \quad i = 1, \cdots, n \tag{1}$$

$2/\|W\|$为分类间隔的值，使间隔最大也就是 $\|w\|^2$ 最小。满足公式（1）且使 $1/2\|w\|$最大的分类面就叫做最优分类面，$L1$ 和 $L2$ 上的训练样本点就称作支持向量。

支持向量机是针对两类问题设计的，要实现对多个类进行处理，支持向量机就需要扩展。One-vs-Rest，One-vs-One，ECOC（Error-Correcting Output Coding）和二叉树等是常用的多分类方法。

2 特征选择及分类实验

2.1 自动分类各类语料的选取

在具体的实验中，选取了《中国图书馆分类法》中 R7 下的 9 个类别文章的摘要和标题作为训练和测试语料。9 个类别分别是妇产科学、儿科学、肿瘤学、神经病学与精神病学、皮肤病学与性病学、耳鼻咽喉科学、眼科学、口腔科学、外国民族医学（见表 1）。

表 1 9 个类别基本信息

类别	类别代号	文件数	大小
妇产科学	R71	12960	6200KB
儿科学	R72	14090	7160KB
肿瘤学	R73	17130	9130KB
神经病学与精神病学	R74	15890	7910KB
皮肤病学与性病学	R75	10430	4190KB
耳鼻咽喉科学	R76	14330	6840KB
眼科学	R77	15750	7940KB
口腔科学	R78	18600	8470KB
外国民族医学	R79	1001	643KB

由于标题和摘要基本上包含了某一篇文章的所有主要信息，因此通过摘要和标题来给某一篇文章进行分类是可行的。本文基于支持向量机的自动分类是在词汇的基础上展开的（在具体实验中，使用了李荣陆博士提供的基于 SVM-light 的文本分类工具包，在此表示感谢），这样更能突出文本挖掘的特性，因为每一类别下的词汇含有更多的本类别的特征，而无论是从语义上还是语法上具备的本类别的特征信息相对词汇是少的。在目前的情况下，本文分词是使用中国科学院 ICTCLAS 软件来完成的。

2.2 统计量的确定

在实验中，主要使用了下面 4 个常用的统计量。

1）互信息。互信息是计算语言学模型分析的常用方法，它度量两个对象之间的相互性。在过滤问题中用于度量特征对于主题的区分度。在医学期刊分类中，互信息体现在作者使用的词语上，即作者习惯使用的词语和每一类别整体上具有一定的关联性。

$$MI(c,f) = \log\left(\frac{P(c,f)}{p(c)P(f)}\right) \tag{2}$$

其中 f 是期刊作者具体使用的词语，c 是属于 R7 中 9 个类别中的一个类别，当 f 独立于 c 时，$MI(c, f)$ 为 0。具体一般取的平均值：

$$MI_{\text{avg}}(f) = \sum_{c \in C} P(c) MI(c,f) \tag{3}$$

2）卡方统计（χ^2）。两个变量的相关性主要使用 χ^2 统计来衡量，却比互信息更强，因为同时考虑了特征存与否的情况，即医学期刊作者在自己作品中是否使用某些词语和不使用某些词语 χ^2 统计都会考虑到。

$$\chi^2(c,f) = \frac{P(c,f)P(\bar{c},\bar{f}) - P(c,\bar{f})P(\bar{c},f)}{p(c)p(f)p(\bar{c})p(\bar{f})} \tag{4}$$

当 c 与 f 相互独立时，$\chi^2(c, f)$ 为 0。平均值为：

$$\chi^2_{\text{avg}}(f) = \sum_{c \in C} P(c)\chi^2(c,f) \tag{5}$$

3）交叉熵。语言模型的性能通常用交叉熵（Cross Entropy）和复杂度来衡量。交叉熵仅考虑特征在文本中发生的情况。

$$CE(f) = \sum_{c\in C} p(c,f)\log\left(\frac{p(c,f)}{p(c)p(f)}\right) \quad (6)$$

4）证据权值。类概率通过证据权值（Weight of Evidence）反映出来，文中确定为文章作者的概率。在给定某一特征值下的类概率的差，对于特征 f，其证据权值记为 $WE(f)$，计算公式如下：

$$WE(f) = \sum_{c\in C} P(c)p(f)\left|\log\left(\frac{\text{odds}(c\mid f)}{\text{odds}(c)}\right)\right| \quad (7)$$

当 n 是训练文档实例数目，则有：

$$\text{odds}(X_i)\begin{cases}\dfrac{1/n^2}{1-1/n^2} & P(X_i)=0\\ \dfrac{1-1/n^2}{1/n^2} & P(X_i)=1\\ \dfrac{P(X_i)}{1-P(X_i)} & P(X_i)\neq 0\wedge P(X_i)\neq 1\end{cases} \quad (8)$$

2.3 具体的实验步骤

1）用随机抽取的方法获取训练和测试语料。笔者从120 181个文件中（除 R79 之外）按照 1/10 的比率抽取12 919个文件组成语料，然后再按照一定的比率从中获取训练和测试语料。

2）用低密度多特征方法训练模型。低密度就是训练语料属于某类的特征点覆盖度，并通过对训练语料作进一步的分割来降低特征点的密度，在选取特征维度时要尽量往高取，具体数值在实验中确定，测试语料也采取同样的策略。

3）自动分类的性能评价指标。本文主要使用宏平均查全率（$\bar{r}$）、宏平均查准率（$\bar{p}$）和宏平均调和平均值（$\bar{f}$）这 3 个指标来衡量。具体公式如下：

$$\bar{r} = \left(\sum_{c\in C} r_c\right)/|C|\ ,\ \bar{p} = \left(\sum_{c\in C} p_c\right)/|C|$$

$$\bar{f} = 2\bar{r}\bar{p}/(\bar{r}+\bar{p})$$

2.4 自动分类流程图

与其他机器学习模型一样，本自动分类的流程图主要是由训练和测试两个部分组成。具体见图 2。

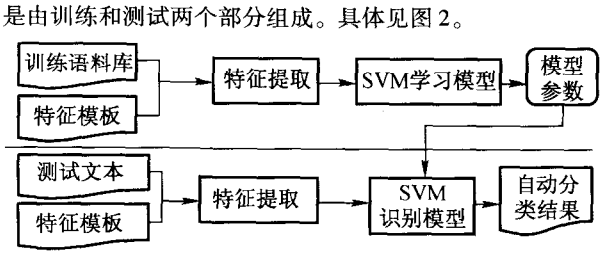

图 2　基于支持向量机的自动分类流程图

3　实验结果及分析

3.1 特征维数的确定

结合已有的实验结果，根据本文的具体实验，特征空间维数与查全率和准确率的关系如图 3 所示。

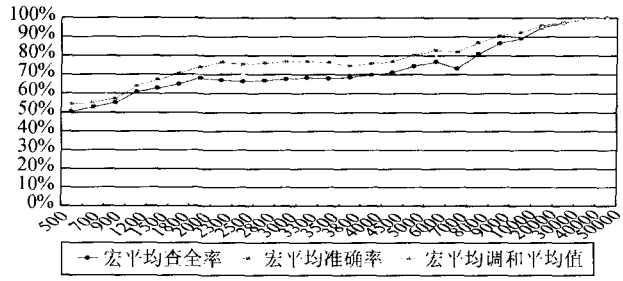

图 3　特征空间维数与查全率和查确率的关系图

把 12 919 个文件，按照 6∶4 的比率分为训练语料和测试语料，特征统计量为互信息，实验获得了不同特征空间维数下的查全率和查准率。从图 3 可以看出，查全率和查准率与特征空间维数是成正比的，当特征空间维数达到 50 000 的时候，查全率和查准率基本达到 100%。因此，在具体的实验过程中，在训练时间和硬件条件允许的前提下，应尽可能地增加特征空间维数。

3.2 4 个统计量下的实验结果

按照 2.3 节的实验步骤，在把 12 919 个文件，按照 6∶4的比率分为训练语料和测试语料，特征空间维数为 4 000的基础上，各统计量下的结果见表 2。

表 2　4 个统计量下的实验结果

统计特征	特征维数	查全率（%）	查准率（%）	调和平均值（%）	训练时间（s）
互信息	4000	80.03	82.14	81.07	4412
卡方统计	4000	92.35	91.32	91.83	6172
交叉熵	4000	93.56	92.45	93.00	5263
证据权值	4000	93.56	93.49	93.52	6265

从表 2 中可以看出，基于医学 R7 中 9 个类别下的自动分类整体上的查全率和查准率效果基本上还是令人满意的，最差的调和平均值为 81.07%，最好的调和平均值为 93.52%。从 4 个统计量的表现来看，由于互信息仅考虑了两个词语之间的相互性，整个效果最差，而证据权值考虑了类概率的信息，无论是在查全率还是在查准率上相比其他 3 个统计量表现良好，同时查全率和查准率之间相差很小。从具体的 9 个类别下的自动分类结果看，前 8 个类下的查全率和查准率相差不大，而最后一个类别的结果比较差，因为前 8 个类自己的特征都比较突出，而最后一个类由于研究的内容涵盖了比较广的领域，造成了自动分类的错误比较严重。

从调和平均值与支持向量机模型在常规新闻语料上的效果来看，实验效果还有待于提高，其中一个主要的原因

是汉语分词不理想造成的。由于 ICTCLAS 主要针对的是通用的语料进行分词，而医学期刊的标题和摘要在使用 ICTCLAS 进行分词时，会造成分词碎片，造成词语本身的语法断层、语义割裂，如“多孔高密度聚乙烯”被分成“多孔/高/密度/聚/乙烯”，从而从根基上造成自动分类的错误。同时，由于一些期刊上的文章并没有严格按照《中国图书馆分类法》进行分类，造成了分类错误，这也会在一定程度上影响自动分类的结果。

4 结束语

本文基于支持向量机，对医学期刊 R7 下的 9 个类别进行了自动分类的研究。在 12 919 个文件组成的训练和测试中，互信息、卡方统计、交叉熵和证据权值等 4 个统计量下的开放测试中，调和平均值取得了相对理想的结果，其中，证据权值下的调合平均值达到了 93.52%。下一步，将会扩大训练语料的规模，增强自动分类模型的健壮性，并同时根据医学期刊的特征，开发有针对性的自动分词软件，确保词汇在语法和语义上的完整性，从而从根本上确保自动分类调和平均值的提高。

感谢北京万方数据股份有限公司提供的数据！□

参考文献

[1] FELDMAN R, SANGER J. The text mining handbook [M]. London: Cambridge University Press, 2009: 5-7, 225.

[2] 马金娜，田大钢．基于 SVM 的中文文本自动分类研究 [J]．计算机与现代化，2006，138 (8)：5.

[3] 中国图书馆分类法 [EB/OL]. [2010-07-20]. http://www.ztflh.com/.

[4] VAPNIK V. The nature of statistical learning theory [M]. New York: Springer—Verlag, 2000: 89.

[5] CORNEY M. Gender-preferential text mining of e-mail discourse [EB/OL]. http://www.acsac.org/2002/papers.

[6] 张学工．关于统计学习理论与支持向量机 [J]．自动化学报，2000，26 (1)：33-42.

[7] 萧嵘，王继成，张福炎．支持向量机理论综述 [J]．计算机科学，2000，27 (3)：1-3.

[8] 梁坤，古丽拉·阿东别克．基于 SVM 的中文新闻评论的情感自动分类研究 [J]．电脑知识与技术，2009，5 (13)：3496-3497.

[9] 翟林，刘亚军．支持向量机的中文文本分类研究 [J]．计算机与数字工程，2005，33 (3)：21-23.

[10] 梁秀娟．基于 SVM 的多类文本分类研究 [D]．武汉：中南财经政法大学，2008：15.

[11] 胡燕，熊浩勇，付香英．线性可分文本的 SVM 算法研究与改进 [J]．计算机与数字工程，2008，36 (3)：18-20.

作者简介：王东波，男，博士生。
苏新宁，男，博士生导师。
朱丹浩，男，硕士生。
年洪东，男，硕士。

收稿日期：2010-12-06

（上接第 101 页）

步骤 4　为了方便起见，这里假设数字图书馆馆藏评价的 7 个指标为等权的情形，即权向量 $w=(1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7)$，利用群评价矩阵 $F_Q(A)$ 中第 i 行代表的第 i 个方案 x_i 在各个评价指标下的指标值，由公式(6) 计算它与理想馆藏方案 x^* 之间的加权平均距离，则有：

$d(x_1,x^*)=8.9224$；$d(x_2,x^*)=14.969$

$d(x_3,x^*)=13.047$；$d(x_4,x^*)=12.469$

步骤 5　因为距离最小对应的方案为利用最优的数字图书馆馆藏方案确定的最优数字图书馆馆藏方案，显然 $d(x_1,x^*)<d(x_4,x^*)<d(x_3,x^*)<d(x_2,x^*)$，所以得到相应的排序结果为：$x_1>x_4>x_3>x_2$，即 x_1 为最优的数字图书馆馆藏方案。

5 结束语

本文引进 C-OWA 算子来处理数字图书馆馆藏评价中出现的区间数信息的情形，定义了评价指标的理想点和理想方案，提出了群评价方案指标值与理想点的距离模型方法，获得了数字图书馆馆藏方案评价结果的新方法，说明评价方法是有效的。本文对区间信息的复杂数字图书馆馆藏群评价的量化研究具有重要的理论和现实意义。□

参考文献

[1] 张健．图书馆评价理论与方法 [M]．成都：西南交通大学出版社，2004.

[2] 王居平．数字图书馆评价的理论和方法 [M]．合肥：安徽大学出版社，2008.

[3] 吴建华．数字图书馆评价方法 [M]．北京：科学出版社，2009.

[4] 盛小平．数字图书馆馆藏评价 [J]．图书情报工作，2005 (5)：40-43.

[5] 赵云华．数字时代评价图书馆馆藏的标准 [J]．图书馆学刊，2007 (1)：113-114.

[6] 王居平．基于纯语言信息的网络信息资源综合评价研究 [J]．情报理论与实践，2008，31 (2)：215-217.

[7] YAGER R R. On ordered weighted averaging aggregation operators in multicriteria decision making [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1988 (18): 183-190.

[8] YAGER R R. OWA aggregation over a continuous interval argument with application to decision making [J]. IEEE Transactions on Systems, Man and Cybernetics-Part B, 2004 (34): 1952-1963.

作者简介：王居平，女，1967 年生，副研究馆员。

收稿日期：2010-11-15