文章编号:1001－9383(2018)01－0001－10

基于ESVM的科技政策文本标签分类研究

吴 峰[1],李银生[1],聂永川[1],范通让[2],赵文彬[2],张 博[2]

(1.河北省科学技术情报研究院,河北省科技信息处理实验室,河北 石家庄 050021;
2.石家庄铁道大学 信息科学与技术学院,河北 石家庄 050043)

摘 要:文本标签作为一种文本关键词,能够简化科技政策中有效信息的挖掘。本文从科技政策类别角度,将标签类别分为科技投入、知识产权、农村科技和税收四类,针对传统SVM算法的缺点和标签数据不平衡的缺点,结合欧式距离思想,提出一种带有惩罚因子的ESVM科技政策文本标签分类方法。最后,对比SVM和ESVM两种分类方法,验证了本文方法在处理科技政策文本标签数据上的有效性。

关键词:文本标签分类;科技政策;SVM;不平衡数据

中图分类号:TP391.1 **文献标识码**:A

DOI:10.16191/j.cnki.hbkx.2018.01.001

Research on text label classification of science and technology policy based on ESVM

WU Feng[1], LI Yin-sheng[1], NIE Yong-chuan[1], FAN Tong-rang[2], ZHAO Wen-bin[2], ZHANG Bo[2]

(1. *Institute of Scientific and Technical Information of Hebei Province, Shijiazhuang Hebei* 050021, *China*;
2. *School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang Hebei* 050043, *China*)

Abstract: Text label is a kind of text keywords, can simplify extraction of effective information from science and technology policy. For science and technology policy, this paper divides text label into four kinds, such as science and technology investment, intellectual property rights, rural science and technology, tax. Aimed at the shortcoming of the traditional SVM algorithm's label data unbalance, this paper provides a text label classification method of science and technology policy, which combines the Euclidean distance algorithm and ESVM algorithm with penalty factor. Finally, with comparing SVM and ESVM, the validity of the propose method on science and technology policy text label is verified.

Keywords: Text label classification; Science and technology policy; SVM; Unbalanced data

---

收稿日期:2017－12－17

基金项目:国家自然科学基金(＃61373160),河北省科技厅科技支撑计划项目(17210113D),(179676334D)

作者简介:吴 峰(1971－),男,正高级工程师,硕士研究生导师,研究方向为大数据与信息处理.

通信作者:范通让(1965－),女,教授,研究方向为网络技术及其信息处理.

1 概述

随着科学技术的迅猛发展，科学日益社会化，社会日益科学化，从而使科技政策的研究和制定显得日益重要。科技政策是国家为实现一定历史时期的科技任务而规定的基本行动准则，是确定科技事业发展方向，指导整个科技事业的战略和策略原则。对于科技政策研究人员来说，快速检索到所需和关注的文件是进行政策分析研究的基础。而每个人所关注和研究的方向并不一致，因此在科技数据信息膨胀的情势下，如何从复杂多样的科技政策信息中快速检索用户所关注的信息越来越受到人们的关注。

为了便于主题信息检索与发现，科技政策加入了文本标签功能。科技政策文本标签是对科技政策消息主题观点的标记，本质上是一种文本关键词，可从语义上揭示文本的主要内容[1]。标签功能为信息的主题标注、分类检索等提供了更为便捷的方式，避免了在大量科技政策数据中搜索主题观点的繁琐，成为了内容大爆炸信息时代中的救命稻草。然而，科技政策文本标签极其简短，并且数据分布不平衡的问题给传统的数据挖掘方法带来了新的挑战。

本文以科技政策文本标签为研究对象，改进传统 SVM 算法，提出了一种科技政策文本标签分类方法。该方法根据科技政策文本标签的特点，在分类特征选取上加入用户基本信息；结合欧式距离公式，对不同分布样本采取不同分类方法，克服 SVM 算法缺点；引入惩罚因子 C 加权补偿特征权重，减小数据分布不平衡带来负面影响。

2 相关工作

SVM 算法是数据挖掘领域目前比较常用的机器学习方法，一种求全局最优解的算法，在解决非线性、高维数据问题上具有明显优势。但不可避免的是，SVM 算法也存在一定的缺点。

SVM 二分类的理想状态是：两类样本被最优分类超平面 H 正确分开，两类的支持向量都无争议地落在 H^+ 和 H^- 上，那么，对于待分类样本 x，只需在 H^+ 和 H^- 上各取一点（x^+ 和 x^-），分别计算 x 与 x^+、x 与 x^- 的距离，取距离最近的那个点的类别作为 x 的类别[2]。但问题是类别代表点的选取是随机的，不能保证每次选取的代表点都能准确的代表该类别。当待分类样本 x 距离分类超平面 H 非常近时，可能存在 x 与随机选取的类别支持向量点之间距离差别不明显的情况，此时，极有可能造成样本类别错分。

针对传统 SVM 算法在最优分类超平面附近易错分的问题，众多学者展开了研究。翟永杰[3]结合模糊理论，利用模糊隶属度方法修正最优分类超平面；岑涌[4]采用遗传算法进行 SVM 参数寻优，够造广义的最优分类超平面来获得较好的整体预测性能；李蓉[5]将 KNN 算法融入 SVM 算法，提出一种 K-SVM 的分类方法，结合 KNN、SVM 两者算法的优点，无需对最优分类超平面进行修改，计算量小、参数少，是一种较稳定的算法。陈丽[6]等指出虽然 K-SVM增加了 KNN 算法的计算量，但确实能够提高最优分类面附近样本点的分类准确率。

但是，根据 KNN 算法，单纯比较 k 个邻居中正、负样本数目，仅以绝对数量作为类别判断依据，在不平衡数据下是不成立的。因此，使用 K-SVM 算法对分布不平衡数据分类，分类精确度反而会下降。

对于训练数据不平衡问题，常用的解决方法有两种：一是，在数据层面上，对训练样本的重构，即采样。可以通过添加少数或去除多数样本来完成[7-8]。但这两种方法都有缺点，前者可

能加入一些冗余,而后者可能去除一些对分类有用的信息。所以,在算法层面上,对算法进行改进[9-10]。通常可以加入一些成发机制或利用集成学习来对数据的非平衡做些补偿,这种方法不会破坏现有的数据组成。

因此,本文进行算法层面的研究,提出了带有惩罚系数的 ESVM 分类方法,对分类超平附件样本点采取计算平均欧式距离值的方法,对剩余样本点采取带有惩罚系数的 SVM 算法,并通过仿真实验进行验证。

3　语料预处理

科技政策语料数据存在信息冗余和格式不规范等问题,在文本标签类型分类之前,需要对数据集进行预处理,过滤无关数据,转换数据格式。语料预处理过程共分为三个步骤:删除无关数据信息,文本分词和去除停用词。

删除无关数据信息包括:删除不包含文本标签的科技政策语料;删除重复文本标签的科技政策语料;转换数据时间格式,删除时间错误的文本标签语料。

对数据集完成基本无关信息删除后,提取数据中的标签文本,使用中科院分词工具 ICTCLAS 完成文本标签的分词及词性标注。新版本的 ICTCLAS 中除了中文分词、词性标注、命名实体识别等基本功能外,还增加了科技政策分词及新词添加功能,是目前较为完善的分词工具。

最后,过滤分词之后数据中的停用词。考虑到科技政策文本标签的特性,仅过滤文本中频繁出现但无实际意义的介词、感叹词和连词。

4　科技政策文本标签分类方法 ESVM

4.1　特征选取

分类模型的分类准确性关键在于分类特征的选取。传统文本分类的特征主要基于统计学与概率论,选取出现频繁,高概率的词语作为特征项。科技政策文本标签本质上就是文本关键短语,内容简短精炼。因此,传统的文本特征项选择方法不适合科技政策文本标签分类研究。

根据科技政策标签文本简短的特点,综合考虑主题关键词和用户基本信息,选取以下几个属性为分类特征项:

(1) 主题词:文本的主题词是文本主旨的体现,主题词的话题类别也就是文本的话题类别。而科技政策文本标签一般在 14 个字以内的短文本,大概在 3 到 4 个词以内,本身就是一个或多个主题词的整合。因此,不必对标签文本进行主题词的筛选,将去除停用词之后的分词结果,全部作为主题词特征项。

(2) 词性:词语的歧义性造成了文本标签语义不明的特性,导致标签类别错分,影响分类器性能。在这些多义词中,一些词语,意义不同,词性也不同。将词性作为分类特征项,能够在一定程度上减小词语歧义性带来的误差。

(3) 地理位置:地理位置是标签发布用户的所在地。不同地理位置,社会氛围、风俗习惯不同。因此,受当地主流文化价值的影响 ,不同所在地的用户其关注的科技政策主题倾向也会存在差异。

(4) 认证:科技政策中包括普通用户和认证用户两种。认证是科技政策赋予的一种身份

肯定，往往在一定领域有一定权威的人士或机构才能通过认证申请。认证用户具有较强的领域性，并且拥有比普通用户更高的信任度。本文选取类型特征比较明显的个人认证、官方认证和自媒体认证作为认证特征项。认证用户分类表如表 1 所示。

表 1 科技政策认证类别分类表

认证类别	认证描述	认证范围
个人认证	科技政策用户个人身份确认	行业名人，权威人士
官方认证	政府官方认证	政府机构等社会组织
	企业官方认证	公司账号、分支机构账号等
	媒体官方认证	报纸、杂志、电视等公共媒体
自媒体认证	帮助优质作者变现，提升影响力	影响力达到媒体效果人士

科技政策文本标签特征项分类表如表 2 所示。

表 2 特征项分类表

特征类别	特征项内容
主题词特征项	文本标签预处理后词语
词性特征项	主题词词性
地理位置特征项	文本标签发布用户地理位置
性别特征项	文本标签发布用户性别
认证特征项	用户是否为认证用户，认证用户类别

4.2 权重计算

特征数字化是采用特征权重计算方法将标签分类特征项数字化，最终将标签表示成数字向量形式。TF－IDF 算法计算简单，运行速度快，是常用的权重计算方法之一。

根据 TF－IDF 核心思想可知，若包含某一特征项的数据数量越少，则逆向文件频率值越大，说明该特征项具有很好的区分能力。反之，则说明该特征项区分能力较弱。与事实矛盾的是，若某一特征项高频率出现于某类数据中时，说明该特征项能够很好地表示该类，应赋予其高权重。文献[11]对传统 TF－IDF 方法进行了改进，使用特征项在某一类别中词条个数占总类别中包含该特征项词条数的比值代替某一类别中词条频率，改进后公式如下：

$$w_{ij}=tf_{ij}\times log\left(\frac{n_i}{n_i+k}\times N\right) \tag{1}$$

其中，n_i 表示训练集该类别中包含特征项 t_i 词条数，k 表示其它类别中包含特征项 t_i 的词条数，N 表示训练集中总词条数。

改进后的 TF－IDF 算法能够很好克服原有算法的缺点，提高分类准确度。本文采用改进后 TF－IDF 公式计算科技政策文本标签特征项的权重。

4.3 基于 ESVM 的科技政策文本标签分类方法

科技政策文本标签分类问题是根据标签类型的不同对科技政策中的文本标签的分类。本

文将文本标签类型分为科技投入、知识产权、农村科技、税收四大类,因此,本文中的文本标签分类实质上是个多分类问题。

ESVM(Euclidean-SVM)分类算法,其基本思想为:采用"一对一"方法构造多分类器,即在任意两类样本之间设计一个 SVM 二分类器,k 个类别的样本就需要设计 $k(k-1)/2$ 个 SVM 二分类器。根据"一对一"思想,训练六个 SVM 分类器模型,保存其最优分类超平面 H 和所有支持向量;设置分类阈值 ε,计算待分类样本点 x 到分类超平面 H 的距离,若大于阈值 ε,直接使用 SVM 方法分类;否则,使用欧式距离方法分类。

4.3.1　SVM 模型训练

SVM 算法实现工具包有很多种,包括 Mysvm,SvmLight,Libsvm 等。其中,台湾大学林智仁教授等开发设计的 Libsvm[12] 是一款采用 one-versus-one(一对一)方法解决多分类问题的开源软件包。该软件使用简单易学,提供了多个默认参数,并提供交互检验(Cross Validation)功能进行最优参数选择。此外,Libsvm 具有修正不平衡样本的功能,提供样本惩罚系数 C 加权设置。本文采用 JAVA 版本开源工具包 Libsvm 进行 SVM 建模。

科技政策文本标签 SVM 分类器训练过程包括以下几个步骤:

(1) 采用 FormatDataLibsvm. xls,将分类好的训练数据转换成 libsvm 所用的数据格式;

(2) 调用函数 svm_scale. java 对数据进行归一化;

(3) 设置参数 s(svm 类型),t(核函数类型),c(损失函数),g(核函数 γ 值),w_i(惩罚系数 C 权重),v(交叉验证);

(4) 采用交叉验证选择最优参数;

(5) 调用 svm_train 函数训练数据,共生成 6 个二分类模型,分别是科技-知识分类器 S-T,科技-农村分类器 S-C,科技-税收分类器 S-T,知识-农村分类器 T-C,知识-税收分类器 T-T。

针对不平衡数据问题,Libsvm 中采用惩罚系数 C 的加权设置来解决,其中权重 wi 的计算决定补偿性能的好坏,本文采用公式(2)计算 w_i:

$$w_i = \frac{N/n_i}{\sum_{i=1}^{n} N/n_i} \quad i = 1,2,\cdots,n \quad \text{其中,} \begin{cases} w_i \in (0,1) \\ \sum_{i=1}^{n} C_i = 1 \end{cases} \tag{2}$$

其中,N 表示训练集中总样本数,n_i 表示训练集第 i 个类别中样本数,n 表示训练集样本类别数。某一类别样本数 n_i 越小,C_i 就越大,反之亦然。上述公式可以很好的展示训练集中数据类别的稀疏分布,对少数类数据权重进行补偿,优化模型训练。

4.3.2　ESVM 算法

ESVM 算法基本步骤描述如下:

(1)使用已标注好的训练数据集 DS1 训练 SVM 二分类器 Classfieri,得到每个分类器最优决策函数 $f(x)i=\sum_{j=1}^{n}(\alpha_j^o K(x_j,x)+b^*$,$i=1,2,\cdots,6$,表示分类器编号;

(2)将测试数据集 DS2 中待分类样本点 x 带入决策函数 $f(x)_i$ 中,计算样本点到每个分类超平面距离 d_i,即分类距离 $d_i=f(x)$;

(3)比较分类距离 d_i 与分类阈值 ε 的大小,若 $|d|>\varepsilon$,执行步骤(4),否则执行步骤(5);

(4)采用训练好的 SVM 分类器 Classfieri,根据其判别函数 $g(x)=sgn(f(x))$ 结果作为样本 x 的类别 a_i;

(5) 计算样本点 x 到 Classfieri 中两个类别 SVs 间的平均欧式距离 $D=\sum_{l=1}^{N}(\sqrt{\sum_{l=1}^{n}(x_i-y_i)})_l/N$,其中,$l$ 表示支持向量的个数,n 表示数据维度,x 表示待分类样本点,y 表示支持向量,比较两个类别 D 的大小,选择较小值的类别为 x 类别 a_i;

(6) 统计 6 次分类结果 a_i,采取投票形式,选取得票数最高的类别作为样本点最终类别。

结果判定投票机制:假设共有 A,B,C,D 四个类别,a_1 表示(A,B)-classifier 分类结果,a_2 表示(A,C)-classifier 分类结果,…,a_6 表示(A,C)-classifier 分类结果。当(A,B)二分类是,若Awin,则 a_i=A,A=A+1;否则,a_i=B,B=B+1,以此类推,最终投结果为:Max(A,B,C,D)。

算法流程如图 1 所示。

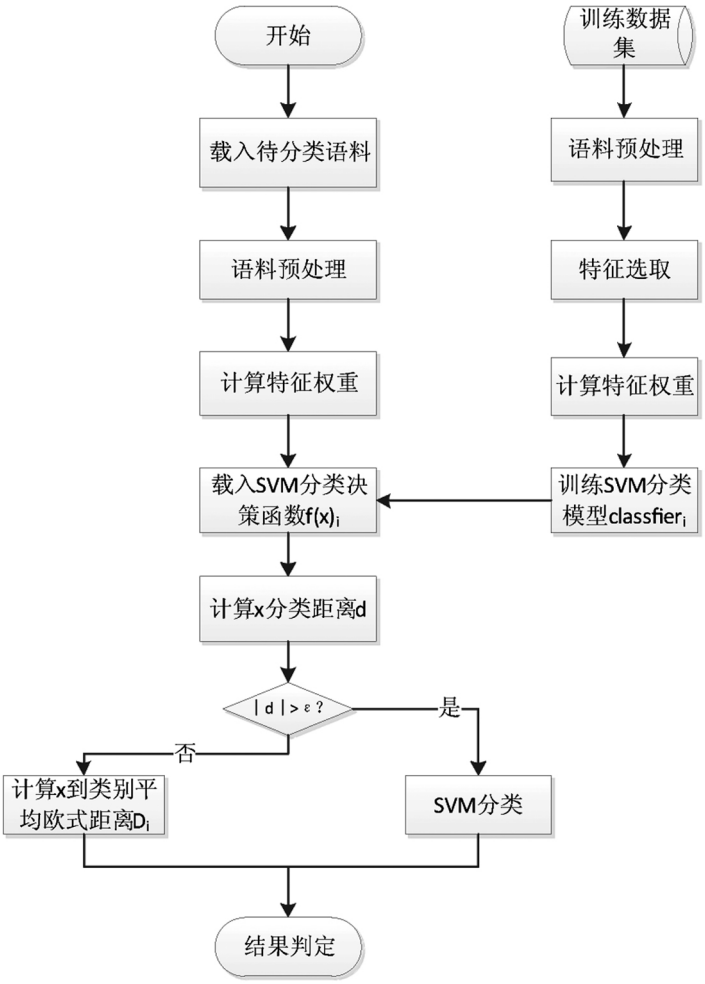

图 1 ESVM 科技政策文本标签分类方法流程图

5 实验

5.1 评价指标

分类评价指标是用于评价分类器性能好坏的指标。常用的分类器性能评价指标主要包括精确度(Precision),召回率(Recall)和 F 值(F-measure)三个。

精确度是指某一类别中分类器正确分类的数据数占分类器判定属于该类别数的比例,用于衡量类别的查准率,其公式如 3 所示。

召回率是指某一类别中分类器正确分类的数据数占真正属于该类别的数据数的比例,用

于衡量类别的查全率,其公式如 4 所示。

$$\text{Precision}=\frac{A}{A+B} \tag{3}$$

$$\text{Recall}=\frac{A}{A+C} \tag{4}$$

上述公式中,A 表示该类样本中分类器正确分类数,B 表示分类器错分为该类的数据样本数,C 表示分类器将该类数据错分为它类样本数。

精准度和召回率从两个不同角度衡量了分类器的分类质量,不能综合评估分类器性能好坏。F 值是指精确度和召回率的调和平均数,综合考虑了分类器的查准率、查准率及其偏向程度。F 值计算公式(5)所示。

$$F=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{5}$$

5.2　实验结果与分析

5.2.1　实验数据准备

本实验数据共分为科技投入类,知识产权类,农村科技类,税收等四类。其中,科技投入类主要包括财政科技拨款、科技基金、科技信贷等类型;知识产权类主要包括专利权、商标权、著作权等类型;农村科技类主要包括先进的育种技术、先进灌溉技术、先进的农业机械技术等类型;税收类主要包括增值税、消费税、营业税等类型。

网络爬取数据集中的科技政策语料,经过数据预处理后,共 16823 个科技政策文本标签。人工对科技政策文本标签进行类别标注,分别使用 1,2,3,4 代表科技投入、知识产权、农村科技和税收四个类别。其中,科技投入类标签 5638 个,知识产权类标签 6745 个,农村科技类标签 3583 个,税收类标签 857 个。采取 70%的语料作为训练数据,30%作为测试数据。

5.2.2　ESVM 分类实验

本实验的硬件实验环境为 Windows 7 操作系统,Intel Xeon E5-2609 V2 处理器,主频 2.5GHz,内存 8GB,采用 Eclipse kepler 为编程开发环境。

SVM 核函数采用 RBF,类型为 c-svm,采取 10 折交叉验证,进行参数寻优。调用 svm_train 进行交叉验证时,返回值是交叉检验下的平均分类准确度,根据交叉效率值和平均准确率找到最优核参数。

交叉验证结果如图 2 所示。

```
nSV = 2232, nBSV = 981
...*
optimization finished, #iter = 3849
nu = 0.4276952061010316
obj = -902.8659209857783, rho = -0.6411811194665614
nSV = 2339, nBSV = 665
.*
optimization finished, #iter = 1598
nu = 0.5682957307679432
obj = -572.4359774002763, rho = 0.48037191138508967
nSV = 1143, nBSV = 520
Total nSV = 3299
Cross Validation Accuracy = 68.5%
Accuracy = 5.625606207565471% (58/1031) (classification)
```

图 2　交叉验证结果图

图 2 所示的值为交叉验证返回的交叉效率值和平均准确率。

选取最优分类阈值 ε，不同 ε 值下，算法分类精确度如图 3 所示。

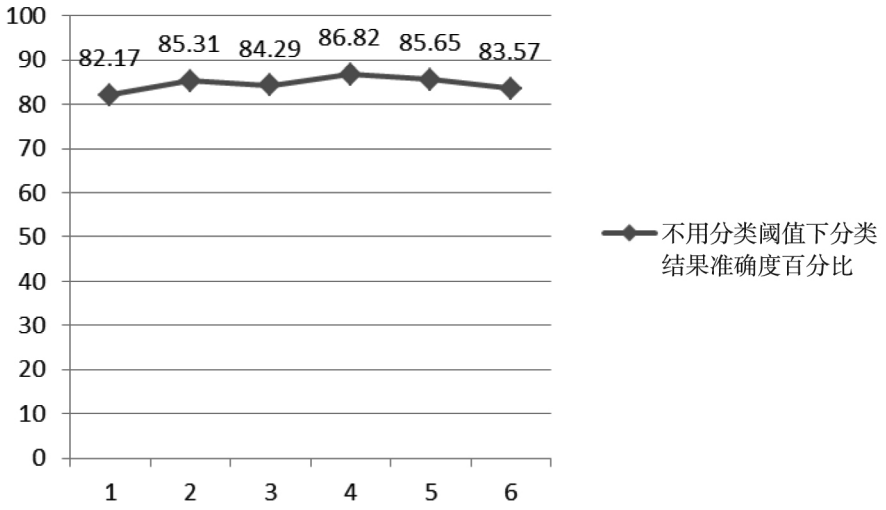

图 3 不同分类阈值下分类准确度

最终，本实验参数设置距离分类超平面分类阈值 ε＝0.6，C＝5000，采用本文提出的 ESVM方法完成科技政策文本标签四分类任务，分类结果如表 3 所示。

表 3 ESVM 方法的分类结果(%)

	Precision	Recall	F
科技投入	89.26	80.02	83.62
知识产权	95.28	89.33	91.21
农村科技	83.51	79.24	81.26
税收	80.35	72.16	76.01

从分类结果上可以看出，本文提出方法对知识产权类标签数据最为敏感，分类效果最好。农村科技类和税收类稍差，虽然进行了参数补偿，但是样本缺失带来的影响还是存在，分类精确度还是不能达到多数类的效果。

5.2.3 ESVM 和 SVM 对比实验

在相同数据集，相同实验参数下，分别使用本文提出 ESVM 分类方法和 SVM 分类方法进行分类实验，对比实验结果的 Precision ，Recall 和 F 值三个评价指标。

表 4 不同惩罚系数下两种方法 Precision 值对比 (%)

c / 类别	C＝1		C＝10		C＝100		C＝500		C＝5000	
	SVM	ESVM	SVM	ESVM	SVM	ESVM	SVM	ESVM	SVM	ESVM
科技投入	84.21	87.34	84.53	87.81	84.98	88.29	85.43	88.79	85.75	89.26
知识产权	88.87	93.27	89.01	93.43	89.30	94.01	89.62	94.66	91.03	95.28
农村科技	78.04	81.92	78.26	82.04	78.88	82.82	79.13	83.04	79.25	83.51
税收	72.18	78.21	73.46	78.96	73.82	79.21	74.56	79.97	74.98	80.35

为解决非平衡数据问题,Libsvm 采取惩罚系数 C 加权对分布不均匀的数据进行修正。由表 4 可以看出,随着惩罚系数 C 的增大,整体上两类算法的分类 Presicion 值在不断提高。其中,税收类别分类 Presicion 上升幅度较大。惩罚系数 C 的加入,确实弥补了数据分布不均的缺点,尤其对少数类别补偿明显,提高了分类器性能。比较两类算法的四个类别分类精确度,ESVM 的分类 Precision 要高于 SVM。可见,ESVM 算法有效提高了分类器分类精度,减小类别特征不明显的样本($\varepsilon<0.6$)错分率。

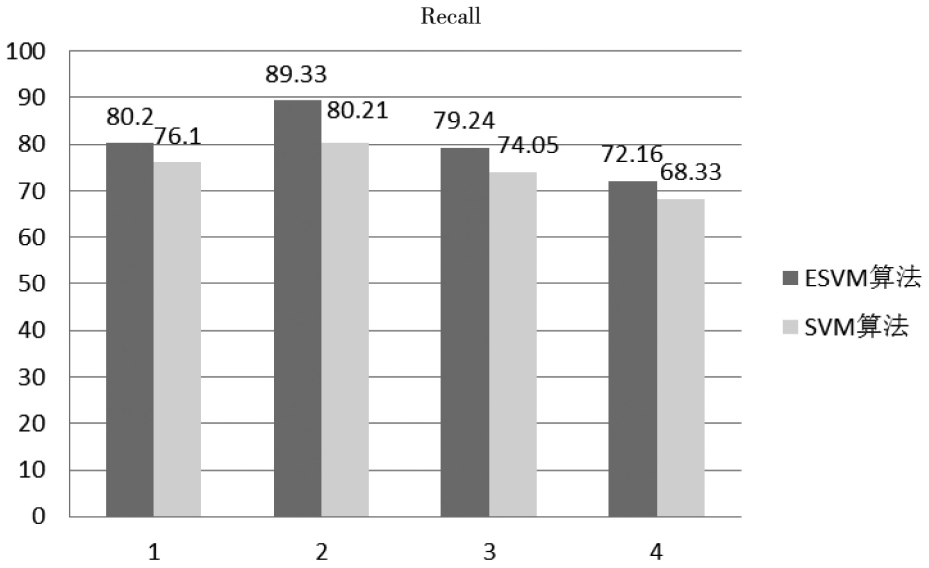

图 4　两种分类方法 Recall 对比

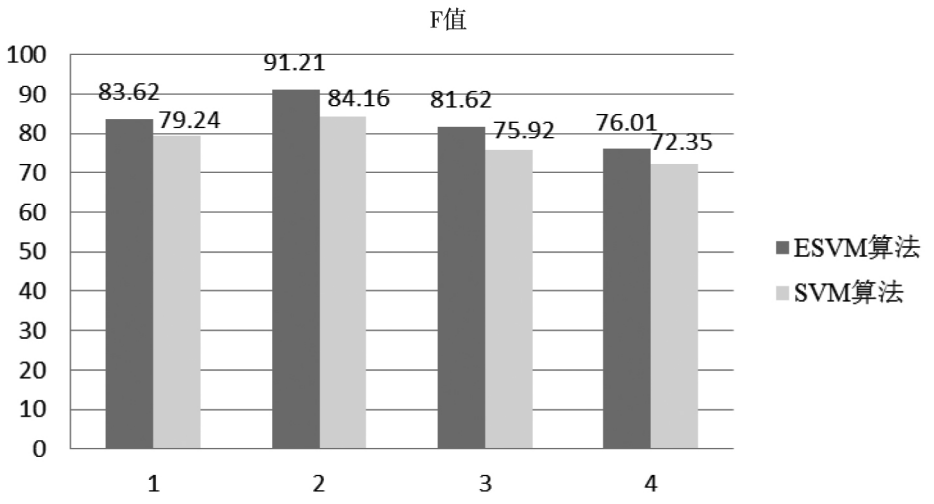

图 5　两种分类方法 F 值对比

图 4、图 5 中显示,本文提出的 ESVM 的查全率及综合分类效果均要高于传统 SVM 方法,可以得出 ESVM 方法针对不同分类决策距离的样本采取不同分类方法的策略能够很好的克服传统 SVM 分类缺点。

6 结束语

针对科技政策文本标签信息简短特点，提取关键词和用户信息作为分类特征。结合欧式距离思想，提出一种 ESVM 方法，对科技政策文本标签进行分类。ESVM 方法对于距离分类超平面近的样本点，通过计算平均距离的方法，消除了多数类在数据样本在数量上的优势；对于距离分类超平面远的样本点，采取带有惩罚系数的 SVM 算法，减小了数据不平衡带来的负面影响。该算法在解决数据分布不平的问题之外，对分类距离不同的样本采取不同分类方法，也弥补了传统 SVM 分了超平面附近样本易错分的缺点。

本文首先分析了目前本体以及信息检索的研究现状，阐述了对科技政策领域的研究意义。将本体技术引入信息检索，并实现个性化检索服务，实现一个基于本体的语义个性化检索模型。具体分析各模块实现的关键技术与优化方法，最后利用实验验证本文方法的有效性。

参考文献：

[1] Tsur O, Rappoport A. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities [C]//Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012: 643－652.

[2] Zanchettin C, Bezerra B L D, Azevedo W W. A KNN-SVM hybrid model for cursive handwriting recognition[C]//Neural Networks (IJCNN), The 2012 International Joint Conference on. IEEE, 2012: 1－8.

[3] 翟永杰，韩璞，王东风，等. 基于损失函数的 SVM 算法及其在轻微故障诊断中的应用[J]. 中国电机工程学报，2003，23(9)：198－203.

[4] 岑涌，钟萍，罗林开. 基于 GA-SVM 的企业财务困境预测[J]. 2008.

[5] 李蓉，叶世伟，史忠植. SVM-KNN 分类器——一种提高 SVM 分类精度的新方法[J]. 电子学报，2002，30(5)：745－748.

[6] 陈丽，陈静. 基于支持向量机和 k-近邻分类器的多特征融合方法[J]. 计算机应用，2009，29(3)：833－835.

[7] Dawson J C, Endelman J B, Heslot N, et al. The use of unbalanced historical data for genomic selection in an international wheat breeding program[J]. Field Crops Research, 2013, 154: 12－22.

[8] Zhang Y, Fu P, Liu W, et al. Imbalanced data classification based on scaling kernel-based support vector machine[J]. Neural Computing and Applications, 2014, 25(3－4): 927－935.

[9] Saad R, Halgamuge S K, Li J. Polynomial kernel adaptation and extensions to the SVM classifier learning[J]. Neural Computing and Applications, 2008, 17(1): 19－25.

[10] Liu J. RESEARCH ON CLASSIFYING UNBALANCED DATA BASED ON PENALTY-BASED SVM AND ENSEMBLE LEARNING[J]. Computer Applications & Software, 2014.

[11] 孔振. 基于 VSM 的文本分类系统的设计和实现[D]. 哈尔滨工业大学，2014.

[12] 林智仁. Libsvm[J]. (2009－04－01)[2009－06－03]. http://www.csie.ntu.edu.tw/-cjlin/libsvm.