面向TRIZ理论使用者的多标签专利分类

袁 力 陈 阳 赵 勇
(华中科技大学自动化学院 武汉 430074)

摘 要 专利是创新的结果,更是再创造的知识源泉,对专利技术知识依据创新需求的分类可有效帮助设计者进行创新设计。依据TRIZ理论对产品专利进行自动分类,以辅助利用专利蕴含的技术冲突进行产品创新设计。TRIZ原始的发明原理过于抽象以及有些原理之间有重叠,文中对40个原始的发明原理进行重组,形成20个新的类别。专利自动分类是一类典型的多标签分类问题,文中从Pro_Techniques和CREAX两个软件中收集了针对发明原理进行具体解释的专利数据,并依据此数据集对问题转换和自适应算法两类多标签分类算法进行对比分析。采用海明损失、测度等评估特性评估了上述算法的性能和质量。结果表明,在使用TRIZ专利数据集时,问题转换方法分类性能要明显优于自适应算法。

关键词 专利分类,发明原理,TRIZ,多标签
中图法分类号 TP181 **文献标识码** A

Multi-label Patent Classification Oriented to TRIZ Users

YUAN Li CHEN Yang ZHAO Yong
(Department of Automation, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract Patent is not only the results but also the resources of products innovation. It can help the designer make innovation effectively, if we classify technique knowledge of patent based on innovative demand. The classification of products patent based on the TRIZ can assistant in using technique contradiction addressed in patent to make innovative design. The original Inventive Principles is so abstract that some principles are overlapped. Paper analyzed the 40 IPs and grouped them into new 20 classes. Patent classification problem is known as multi-label classification problem. Pro-Techniques, CREAX, these two softwares supply patents which explain Inventive Principles in detail. The dataset is used to compare the performance of multi-label classification algorithm: problem transformation and algorithm adaptation. Several measures such as hamming loss, F-measure have been proposed in the literature for the evaluation of multi-lable classifiers. The result shows Problem Transformation performs more excellent than Algorithm Adaptation using TRIZ patent datasets.

Keywords Patent classification, Inventive principle, TRIZ, Multi-label

1 引言

专利文献是集技术、法律和经济信息于一体的、数量巨大且内容广泛的战略性信息资源。它不仅描述了发明所属技术领域中现有技术及其存在的问题,而且说明了发明的基本要点和实施方案。其技术信息对产品改进和创新有重要的参考和启发价值。越来越多的研究开始关注如何挖掘专利中蕴藏的丰富知识来帮助设计者进行快速有效的概念创新设计。但专利的数量是海量的,如何有效地从中找到用户所需要的信息是当前信息科学和技术领域面临的一大挑战。

文档分类作为处理大量文本数据的关键技术,可以在较大程度上解决信息杂乱的问题,方便用户获取所需信息。目前已有一些专利分类系统,比如国际专利分类法(International Patent Classification)IPC[1]、英国的分类法(Europe Classification)ECLA、美国的分类法(United States Patent Classification)USPC、日本的FI和F-term分类法,但是这些专利分类体系都是根据专利描述对象的所属领域来划分的。这样的分类过于宽泛,并不能较好地定位用户所需的技术信息。为了方便设计者更好地利用专利文献中蕴含的技术知识进行产品创新设计,可以借助TRIZ(发明问题解决理论)进行专利分类。TRIZ发明问题解决理论是由前苏联发明家G. S. Altshuler等人在研究世界各国大量高水平专利文献的基础上总结出的各种技术发展进化遵循的规律模式,可以用于指导新产品开发、提升产品质量。这样根据专利所解决的技术冲突和使用的发明原理来分类专利文件,而不是依据专利所属的领域,构建的专利分类系统也有助于TRIZ理论使用者快速有效地获取创新设计方案。特别是,TRIZ囊括总结了40条发明原理,这些发明原理是指导设计方向的重要依据,而现有

本文受高等学校博士学科点专项科研基金项目(20100142120088),国家高技术研究发展计划(2009AA04Z107)资助。
袁 力(1987—),男,硕士,主要研究方向为数据挖掘及TRIZ,E-mail:alvin_green@live.cn;陈 阳(1979—),男,博士,讲师,主要研究方向为决策分析;赵 勇(1967—),男,教授,博士生导师,主要研究方向决策理论、方法及应用。

专利则为这些原理提供了丰富的知识参考和价值，因此，本文主要依据这些发明原理，对专利自动分类技术进行研究和讨论。

传统分类大多采用单标签分类分类方法，如 SVM[2]、KNN[3]等。而对于专利分类，单个专利文件往往可以关联数个发明原理，这是一个典型的多标签分类问题。例如，专利《手扶电动整枝机》[4]用升降杆代替手臂、竹竿、爬树和机械台，从而帮助工作人员轻松获得高度。这个专利就可以同时被标记“多功能原则”和“中介物原则”两条发明原理。多标签分类与单标签分类的不同之处在于：多标签的分类结果是由多个标签组成的一个标签集合，而单标签分类的结果只有一个标签。由于传统单标签分类算法不能直接应用于多标签分类问题中，因此有文献针对多标签分类问题从不同角度提出了不同的解决方法。这些处理方法可分为问题转换方法(Problem Transformation)和自适应算法(Algorithm Adaptation)。第一类型的方法将多标签学习任务转换成一个或多个单标签学习任务。第二种类型的方法扩展了特定的学习算法，使其可以直接处理多标签数据[5-9]。Loh 采用 BR(Binary Relevance)[10]方法处理 TRIZ 领域的多标签分类问题。文中使用目前一些常用的多标签分类方法对面向 TRIZ 理论的数据集作了对比分析。实验中使用了 5 种多标签方法，3 种是问题转换方法 Binary Relevance(BR)、Combination Method (CM)[11]、Pruned Problem Transformation(PPT)[12] 2 种，自适应算法 Multi-Label k Nearest Neighbors(ML-KNN)[13]、Back-Propagation Multi-Label Learning(BPMLL)[14]。另外，再用不同的评估特性对这些方法进行评估，旨在为构建专利文件检索系统提供方案参考。

2 发明原理分组

G. S. Altshuller 坚信解决发明问题的基本原理是客观存在的，这些客观存在可以被整理形成一套理论，掌握该理论的人不仅可以提高发明成功率、缩短发明周期，也可使发明问题的解具有可预见性。TRIZ 理论中提出用 39 个通用工程参数描述问题中出现的技术冲突，然后使用 40 个创新原理来解决这些技术冲突。原始的 40 条发明原理过于抽象以及一些发明原理之间存在一些重叠，也饱受不少诟病。Williams[15]对发明原理的对称性以及不对称性做了分析，并指出有些发明原理之间是对立的。Loh[16]将发明原理做了区分，分为清晰发明原理和模糊发明原理，并在此基础上依据发明原理之间的文本相似性对发明原理做了分组。不管怎样说，以文本特征进行分组没能从本质上给予产品设计者明晰的概念指导。例如，若根据文本特征进行分组，那么可将出现“气体”词汇的这一类专利文件归为一类。发明原理 8(反重量原则)利用气体的物理特性，而发明原理 38(氧化原则)和发明原理 39(惰性环境原则)则利用气体的化学特性，这两类特性拟解决的技术冲突就存在很大的差异性，将其分为一组就有失偏颇。

本文依据 TRIZ 原理对大量中文专利进行研究后，发现某些发明原理解决的技术冲突具有矛盾的普遍性，依据这些共性将这些发明原理分为同一组。而这些矛盾的共性有时也可以通过共享的一些特征词汇表征在专利文件中。产品设计者渴望了解专利文件用何种发明原理解决了何种技术冲突，依据技术冲突的矛盾相似性对原始发明原理进行分组与依据文本特征相似性进行分组相比，能更直观地帮助 TRIZ 理论使用者进行产品发明创新。

比如发明原理 1(分割原理)、发明原理 2(抽取原则)和发明原理 3(局部性质原则)将一个整体拆成多个部分，将部分的功能作用独立突显出来，并且相应的专利文本中也会出现如分割、分开、部分等词汇。发明原理 59(联合原则)、发明原理 6(多功能原则)和发明原理 40(复合材料原则)将多个部分合成一个整体，以发挥整体的功能效用，并且相应的专利文本中会出现诸如联合、组合、复合等词汇。发明原理 18(机械振动)、发明原理 19(周期性作用)和发明原理 20(连续有益作用)突出了动作时间作用的连续性或周期性，相应的专利文本会出现诸如连续、周期、振动等词汇。

根据以上原则，32 个发明原理被重新组合成 20 个新的组合，剩下的 8 个发明原理作为独立的组合，如表 1 所列。后面的专利分类就是基于这 20 个新的类别。

表 1 发明原理分组编号表

类别编号	发明原理	类别编号	发明原理
1	01,02,03	11	18,19,20
2	04	12	22,25
3	05,06,40	13	23
4	07,31	14	24
5	08,29	15	26,28,32
6	09,10,11	16	27,34
7	12,13	17	30
8	14	18	33
9	16,21	19	35,36,37
10	17	20	38,39

3 多标签分类及质量评估

3.1 多标签分类问题描述

多标签分类学习的分类器给新的查询样本指定多个类别，其输出是一个标签集合。标签集合中是样本对于每一种标签是否相关的分类结果。处理多标签分类的方法主要有两种类型，问题转换(Problem Transformation)方法和自适应算法(Algorithm Adaptation)。

为了正式描述这些方法，我们先设定一个样本的特征向量为 $X=\{x_1,x_2,\cdots,x_n\}$，其中 $x_i\in R$；有穷标签集合 $L=\{l_1,l_2,\cdots,l_m\}$，样本所对应的标签集合 $Y=\{l_1,l_2,\cdots,l_p\}$，$p\leqslant m$，即 $Y\subseteq L$；那么包含 N 个样本的多标签训练数据集可以表示为 $D=\{(X_i,Y_i)\mid 1\leqslant i\leqslant N, X_i\in R^n, Y_i\subseteq L\}$，其中 X_i 表示一个样本特征向量，Y_i 表示第 i 个样本的标签集合。

3.2 问题转换方法

问题转换方法将多标签学习任务转换为一个或多个单标签学习任务。它先将多标签数据集转换成单标签数据集，接着再用传统的监督学习分类算法处理转换后的单标签数据集，通过这些方法转换后的数据集与原来的多标签数据有相同的标签集合。

BR(Binary Relevance)[10]方法学习 M 个二分类器，每个分类器只针对标签集合 L 中的一个类别进行分类。分类一个新的样本，BR 输出一个合集，这个合集是由 M 个基分类器输出中包含的正例组成的。该方法的主要缺点在于它假定给样本关联的各个标签之间是独立的，忽略了标签之间的相关关系。

CM(Combination Method)[11]是一种比较简单而有效的

多标签分类算法，它将每个对象所属的标签集合作为一个新的标签。分类一个新的样本，CM方法的单标签分类器输出一个最可能的类别，也就是一个标签集合。CM方法的主要优点在于它考虑了标签之间的相关关系。但如果有大量的类别，而且很多类别都只有非常少的样本，那就使得学习更加困难。

PPT(Pruned Problem Transformation)[12]方法扩展了CM方法，以避免刚才提到的问题。如果有某些标签集合关联的样本数量少于用户定义的样本数量，那么就删除这些标签集合，并选择一些不相交的子集来代替，而这些子集关联的样本数量必须超过用户定义的阈值。

3.3 自适应算法

自适应算法扩展了单表签分类器，使其能够直接处理多标签的分类问题。与问题转换方法相比，自适应算法法虽然是独立算法的实现途径，但由于其需要扩展特定的分类方法，在使用上并没有问题转换方法普遍。下面简要介绍两种自适应算法，与问题转换方法形成对比。

MLKNN(Multi-Label k Nearest Neighbors)[13]使用贝叶斯途径扩展了常见的K近邻(KNN)惰性学习算法。用最大后验概率的原则决定测试样本的标签集合，最大后验概率基于 K 个最近邻居对每个标签的前验和后验概率。

BPMLL(Back-Propagation Multi-Label Learning)[14]是神经网络的多标签算法。该算法修改流行的反转算法来适应多标签数据，主要的修改是引入了新的误差公式来考虑多标签。

3.4 分类质量评估

多标签分类器的评估与处理单标签的分类评估的方法不同。不像单标签问题，一个样本的分类只需要判断它被标注的单个标签是正确或是不正确。在一个多标签问题中，一个样本可能被同时标注多个标签，也就是一个标签集合，而这个标签集合中的标签可能部分是正确的，部分标注是不正确的。也有可能属于样本的所有标签并没有被全部标注出来。

(1)基于样本的评估

基于样本的评价方法可对多标签分类器进行评估，此方法通过计算真实标签集合与预测标签集合的差别来计算。为方便描述，多标签数据集表示为 D，Y_i 是真实的标签集合。给定一个随机的实例 X_i，实例的标签集合可通过多标签分类方法预测表示为 Z_i。

海明损失[13]考虑预测错误和遗失错误。海明损失评估样本标签被错误分类的频率。也就是说，一个样本被关联错误标签或者属于样本的标签没有被预测出来。当海明损失等于0时，分类取得最好效果。海明损失越小，效果越好。其中 Δ 表示两个集合对称差，也就是布尔逻辑中不同兼析取(XOR操作)。

$$\text{Hamming-Loss}=\frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i\Delta Z_i|}{m} \tag{1}$$

精确度：对称地测量 Y_i 与 Z_i 的相近程度。

$$Accuracy=\frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i\cap Z_i|}{|Y_i\cup Z_i|} \tag{2}$$

准确度：所有分类的正确文本与实际分类的文本数之比。

$$Precision=\frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i\cap Z_i|}{|Z_i|} \tag{3}$$

召回率：所有分类正确的文本与应有文本数之比。

$$Recall=\frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i\cap Z_i|}{|Y_i|} \tag{4}$$

测度：准确度和召回率的结合。理想的情况是召回率和准确度这两个特性的值都很高，但是实际情况中两者的表现往往相冲突，于是就用两个特性的调和平均测度来折衷。

$$F\_Measure=\frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i\cap Z_i|}{|Z_i|-|Y_i|} \tag{5}$$

子集正确率[13]：其中 $I(\text{true})=1$ 和 $I(\text{false})=0$，这是一个非常严格的评价方法，因为其需要预测的标签集合与正式的标签集合完全吻合。

$$SubsetAccuracy=\frac{1}{N}\sum_{i=1}^{N}I(|Z_i|=|Y_i|) \tag{6}$$

(2)基于标签的评估

这些对所有标签测量评估方法的计算可以通过两种平均方法实现，一种是宏观平均，一种是微观平均。二分评估算法 $F(t_p,t_n,f_p,f_n)$ 通过计算以下四项的数量来实现，真实的正实例(t_p)、真实的负实例(t_n)、错误的正实例(f_p)、错误的负实例(f_n)。微观平均准确是 L 类被分给 L 类(t_p)，不是 L 类被分给 L 类(f_p)的一个比例表示。微观平均召回率用正确分给类 L 的实例数量与所有实际上属于类 L 的比例表示。微观平均测度表示微观准确度和微观召回率的调和平均。$|L|$ 表示标签的数量。

$$Micro\_F1=\frac{2Mic\_p\times Mic\_R}{Mic\_p+Mic\_R} \tag{7}$$

其中：

$$Mic\_P=\frac{\sum_{l=1}^{|L|}tp_l}{\sum_{l=1}^{|L|}(tp_l+fp_l)},Mic\_k=\frac{\sum_{l=1}^{|L|}tp_l}{\sum_{l=1}^{|L|}(tp_l+fp_l)}$$

宏观平均准确度(Mac_P)首先计算每个单独标签的准确度，然后再对所有标签取平均。计算宏观平均召回率时采用同样的方法。宏观测度表示宏观准确度和宏观召回率的调和平均。

$$Macro\_F1=\frac{2Mic\_P\times Mic\_R}{|L|} \tag{8}$$

其中：

$$Mac\_P=\frac{\sum_{l=1}^{|L|}\frac{tp_l}{tp_l+fp_l}}{|L|},Mac\_R=\frac{\sum_{l=1}^{|L|}\frac{tp_l}{tp_l+fn_l}}{|L|}$$

4 实验

4.1 实验数据集

本文整合了Pro_Techniques[17]和CREAX[18]两个软件的专利资源形成分类训练数据集。这两款软件都是基于TRIZ发明原理进行创意分析和创新研发的计算机辅助创新软件。软件对40条发明原理进行了描述，并对每条发明原理给出一些专利文件来对其进行具体解释。最后形成的数据集有579个文本。用于描述多标签数据集的指标有标签基数(Label Cardinality)、标签密度(Label density)。

标签基数表示每一个样本平均标签数量，定义如下：

$$Label\ Cardinality=\frac{1}{N}\sum_{i=1}^{N}|Y_i| \tag{9}$$

标签密度表示每一个样本标签数量的平均比率，定义如下；

$$Label\ Density=\frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i|}{m} \tag{10}$$

文中实验数据集的标签基数为 2.2，标签密度是 0.11。

4.2 实验设置

由于专利分档都是非结构化的且文档内容是人类使用的自然语言，计算机很难处理其语义，因此有必要对专利文本进行预处理。首先对文本进行中文分词[19]，实验中我们可以用专利文件中的摘要部分来完成对整个文件的表示并采用中科院的 ICTCLAS 分词器对专利摘要进行中文分词；然后去停用词后[20]，参照哈工大自然语言处理实验室归纳的 1028 个停用词，并结合 TRIZ 理论使用者的实际需要，整理出包含 978 个停用词的列表，去除文本语言中一些表意能力很差的辅助性文字；接着用 VSM 向量空间模型[21]对文本的特征信息进行表征，再用信息增益[22]的特征选择方法来降低文本向量模型的维度；最后形成的数据集被表示为一个 579×3975 的矩阵。经过预处理后的文档向量再分别用问题转换方法和自适应算法两种多标签分类方法来处理。

实验中将运用 5 种不同的多标签分类方法，其中 3 种是问题转换方法 Binary Relevance (BR)，Combination Method (CM)，Pruned problem transformation(PPT)，2 种是自适应算法 Multi-Label k-Nearest Neighbor(MLKNN)、Back-propagation for Multi-Label Learning(BPMLL) 。对每个多标签问题转换方法，我们用 5 种监督学习算法：Naive Bayes(NB)、Support Vector Machines(SVM)、Decision Tree(DT)、K Nearest Neighbor(KNN)、Multilayer Perceptron(MLP)[23,24]。这些不同的分类器在它们的分类处理程序中有着各自明显的特征，可通过对这 5 种方法的比较得到一个对数据集更加广泛的对比分析。

本文用 8 种评估方法对实验结果进行评估，其中有 6 种基于样本的评估方法(Hamming Loss，Precision，Accuracy，Recall，F-Measure and Subset Accuracy)，2 种基于标签的评估方法(Micro-F1 and Macro-F1)。实验中所有多标签分类方法和监督学习算法都是基于 Mulan[25] 实现的。Mulan 是 Weka[26] 的一个扩展，可以支持多标签的数据挖掘。算法中的参数值都是软件中默认的设置。

实验采用十层交叉检验法，这样数据集将被随机的分为 10 个部分。在这 10 部分中，选择一部分作为测试集，剩下的部分作为训练集。这样的操作重复做 10 次，每次操作都从数据集中选择一个部分作为测试数据，其余部分作为训练数据进行实验，最后把得到的 10 次结果取平均。

5 实验结果及分析

这部分将列出问题转换方法和自适应算法的实验结果，并对两部分结果作对比分析。

(1)问题转换方法

表 2 显示了用多标签问题转换方法处理面向 TRIZ 理论使用者数据集的结果。这些结果通过 5 种监督学习算法作为分类器类处理多标签分类问题。实验结果使用 8 种不同的方法进行评估，并将不同的学习算法得到的结果作对比，最好的结果以粗体字显示。

从以上表 2 中分析多标签的监督分类算法的效果时，可以观察到没有监督学习分类算法在所有的多标签分类方法下都表现出明显的优势。BR 多标签方法下，没有明显优势的监督学习算法，但在使用 KNN 学习时有 4 个评估特性获得最佳，由此可推荐在 BR 下使用 KNN 进行分类。SVM 在 CM 和 PPT 下显示出最好的分类性能，体现了 SVM 在小样本数据处理方面的优势性能。CM 因为考虑了标签之间的关联关系，总体效果要好于 CM，并且 Subset Accuracy 性能突出。PPT 删除了 CM 方法中样本数量过少的标签，性能因而优于 CM。

表 2 问题转换方法结果

Measure	BR				
	NB	SVM	KNN	DT	MLP
Hamming Loss	0.278	**0.201**	0.207	0.248	0.223
Accuracy	0.433	0.551	**0.573**	0.442	0.463
Precision	0.583	**0.683**	0.679	0.595	0.645
Recall	**0.590**	0.572	0.574	0.543	0.586
F-Measure	0.587	0.620	**0.624**	0.585	0.619
Subset Accuracy	0.238	**0.254**	**0.255**	0.147	0.175
Micro-F1	0.565	0.624	**0.626**	0.575	0.625
Macro-F1	**0.406**	0.342	0.395	0.365	0.395
	CM				
Hamming Loss	0.217	**0.194**	0.241	0.236	0.205
Accuracy	0.655	**0.701**	0.585	0.433	0.699
Precision	0.668	**0.761**	0.667	0.543	0.743
Recall	0.663	0.669	**0.675**	0.556	0.644
F-Measure	0.652	**0.703**	0.670	0.548	0.653
Subset Accuracy	0.254	**0.358**	0.212	0.145	0.286
Micro-F1	0.598	**0.723**	0.639	0.551	0.642
Macro-F1	0.446	**0.714**	0.452	0.387	0.422
	PPT				
Hamming Loss	0.219	**0.195**	0.241	0.276	0.247
Accuracy	0.657	**0.711**	0.594	0.434	0.709
Precision	0.672	**0.809**	0.749	0.551	0.776
Recall	0.641	**0.678**	0.602	0.584	0.617
F-Measure	0.628	**0.712**	0.654	0.592	0.606
Subset Accuracy	0.235	**0.402**	0.343	0.192	0.263
Micro-F1	0.634	**0.676**	0.599	0.581	0.601
Macro-F1	0.515	**0.525**	0.456	0.398	0.457

(2)自适应算法

表 3 用自适应算法的多标签分类方法处理面向 TRIZ 理论使用者的专利数据集。通过表 3 我们可以观察到，ML-KNN 有 8 个评估特性都取得了最好的效果。我们也可以说，对于这 8 个评估特性，ML-KNN 分类性能要优于 BPMLL 算法。故若采用自适应算法处理多标签分类问题，推荐使用 ML-KNN 方法。

表 3 自适应算法结果

	ML-KNN	BPMLL
Hamming Loss	**0.192**	0.285
Accuracy	**0.631**	0.464
Precision	**0.726**	0.459
Recall	**0.612**	0.460
F-Measure	**0.663**	0.457
Subset Accuracy	**0.201**	0.189
Micro-F1	**0.655**	0.216
Macro-F1	**0.505**	0459

(3)问题转换与自适应算法

在表 4 中，我们将自适应算法和问题转换方法的最好分类结果进行比较。因为没有问题转换的诸多缺点，理论上可期望自适应算法优于问题转换方法，但由于实际处理过程中，自适应算法要同时处理所有的类别，实际实验效果要低于理想水平。比如对于海明损失，在 BR 下 SVM 可取得最好结果。综合这 8 个评估特性，问题转换方法取得的结果均优于

(下转第 266 页)

究策略方法。

参 考 文 献

[1] 徐俊,刚裴莹. 数据 ETL 研究综述[J]. 计算机科学,2011,38(4)

[2] Dean J, SGhemawat J. MapReduce: Simplified Data Processing on Large Clusters[C]// Proc. of OSDI. 2004:137-150

[3] Kovoor G, Singer J, Lujan M. Building a Java MapReduce Framework for Multi-core Architectures[C]// Proc. of MULTIPROG. 2010

[4] 王珊,王会举,等. 架构大数据:挑战、现状与展望[J]. 计算机学报,2011,10:1741-1752

[5] 李建江,崔健,等. MapReduce 并行编程模型研究综述[J]. 电子学报,2011,11:2635-2642

[6] Dean J, Ghemawat S. MapReduce: A Flexible Data Processing Tool[J]. CACM,2010,53(1):72-77

(上接第 258 页)

自适应算法。

表 4 问题转换与自适应算法结果比较

	Transformation	Adaptation
Hamming Loss	**0.201**	0.192
Accuracy	**0.711**	0.531
Precision	**0.809**	0.726
Recall	**0.669**	0.612
F-Measure	**0.695**	0.663
Subset Accuracy	**0.358**	0.201
Micro-F1	**0.676**	0.655
Macro-F1	**0.525**	0.505

结束语 为了方便 TRIZ 理论使用者进行发明创新,本文根据 40 条 TRIZ 发明原理所蕴含的冲突矛盾的相似性对原发明进行了分组,并将新生成的 20 个新的组别作为专利文件训练分类的新类别。实验比较多个不同的分类算法的分类性能,并用 8 个评估特性进行评估。从实验结果中也可以得到一些有用的信息,比如在使用 TRIZ 专利数据集时,PPT 多标签问题转换方法和 SVM 监督学习算法关联可以得到最好的效果。还有针对此数据集的问题转换方法分类效果也要优于自适应算法。选定最优的分类方法对中文专利进行面向 TRIZ 理论使用者的自动分类,以便设计者能更好地利用专利知识来辅助产品的创新和改进而服务。

参 考 文 献

[1] 左晶. IPC 和 USC 分类体系下专利检索的对比分析[J]. 现代情报,2007,1:130-132

[2] Meyer D. Support Vector Machines[J]. The Interface to libsvm in package e1071. e1071 Vignette,2012

[3] Chou K C, Shen H B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers[J]. Journal of Proteome Research, 2006, 5(8): 1888-1897

[4] 卢长林. 手扶电动整枝机[D]. 1992

[5] Elisseeff A, Weston J. A kernel method for multi-labelled classification[C] // Advances in neural information processing systems. 2001:681-687

[6] Godbole S, Sarawagi S. Discriminative methods for multi-labeled classification[M] // Advances in Knowledge Discovery and Data Mining. Berlin Heidelberg: Springer, 2004:22-30

[7] Crammer K, Singer Y. A family of additive online algorithms for category ranking[J]. The Journal of Machine Learning Research, 2003, 3:1025-1058

[8] Zhang M L, Zhou Z H. Multilabel neural networks with applications to functional genomics and text categorization[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10):1338-1351

[9] Zhang M L, Zhou Z H. A k-nearest neighbor based algorithm for multi-label classification[C] // Granular Computing, 2005 IEEE International Conference on. IEEE, 2005, 2:718-721

[10] Tsoumakas G, Dimou A, Spyromitros E, et al. Correlation-based pruning of stacked binary relevance models for multi-label learning[C] // Proceeding of ECML/PKDD 2009 Workshop on Learning from Multi-Label Data. Bled, Slovenia, 2009:101-116

[11] Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data [M]. Data mining and knowledge discovery handbook. US: Springer, 2010:667-685

[12] Read J. A pruned problem transformation method for multi-label classification[C] // Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008). 2008:143-150

[13] Zhang M L, Zhou Z H. A k-nearest neighbor based algorithm for multi-label classification[C] // Granular Computing, 2005 IEEE International Conference on. IEEE, 2005, 2:718-721

[14] Gao S, Wu W, Lee C H, et al. A MFoM learning approach to robust multiclass multi-label text categorization[C] // Proceedings of the twenty-first international conference on Machine learning. ACM, 2004:42

[15] Williams T, Domb E. Reversability of the 40 Principles of Problem Solving[J]. The TRIZ Journal, May 1998

[16] Cong H, Tong L H. Grouping of TRIZ Inventive Principles to facilitate automatic patent classification[J]. Expert Systems with Applications, 2008, 34(1):788-795

[17] Pro_Techniques[OL]. http://www.iwint.com.cn/

[18] CREAX http://www.creax.com

[19] 费洪晓,康松林,朱小娟,等. 基于词频统计的中文分词的研究[J]. 计算机工程与应用,2005,41(7):67-68

[20] 王素格,魏英杰. 停用词表对中文文本情感分类的影响[J]. 情报学报,2008,27(2):175-179

[21] Erk K, Padó S. A structured vector space model for word meaning in context[C] // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008:897-906

[22] 任永功,杨荣杰,尹明飞,等. 基于信息增益的文本特征选择方法[J]. 计算机科学,2012,39(11):127-130

[23] Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification[C] // Machine Learning: ECML 2007. Berlin Heidelberg: Springer, 2007:406-417

[24] Yang Y, Liu X. A re-examination of text categorization methods [C] // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999:42-49

[25] Tsoumakas G, Xioufis E S, Vilcek J, et al. MULAN: A Java Library for Multi-Label Learning[J]. Journal of Machine Learning Research, 2011, 12(7):2411-2414

[26] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update[J]. ACM SIGKDD Explorations Newsletter, 2009, 11(1):10-18