高技术通讯 2013 年 第23 卷 第6 期:592 ~ 597

doi:10.3772/j.issn.1002-0470.2013.06.007

基于文本挖掘的交互式专利分类①

张晓宇②

(中国科学技术信息研究所 北京 100038)

摘 要 将文本挖掘理论应用于专利信息分析,提出了一种基于多分类器融合与主动学习的交互式专利分类算法,旨在实现高效的专利分类。该算法基于训练集,利用支持向量机,针对不同的专利类别分别训练相应的子分类器,然后通过多分类器融合对各子分类器进行有机结合,以获得性能更优的分类器和形成分类决策。在此基础上,利用主动学习选取最有信息的样本进行标引,从而通过人机交互实现分类模型的更新。针对传统批量选择性采样的缺点,还提出了动态批量选择性采样模式,通过确定度传播策略有效降低标引样本冗余度,以进一步提高主动学习的效率。实验结果表明,这种基于多分类器融合与主动学习的交互式专利分类算法的分类性能显著高于其他算法。

关键词 文本挖掘,专利分类,多分类器融合,主动学习,选择性采样

0 引 言

专利信息记载了人类社会发明创造的成就和轨迹,囊括了大部分领域内的技术成果,是当今时代最重要的技术文献和知识宝库。开展专利信息分析,以指导技术开发,对于促进科技进步有重要实际意义。专利分类作为专利信息分析的基础,是有效管理专利资源、揭示相关信息和发掘专利价值的重要手段,是专利信息分析中不可或缺的重要环节[1]。目前专利分类主要是借助专家智慧通过人工标引完成的,这种方法存在诸多问题:一是分类效率低下,对海量的、激增的专利数据难以胜任;二是专业性要求极高的分类的代价高昂;三是由于分类具有高度的主观性,不同标引人员对相同专利数据给出的分类结果不尽相同,甚至同一标引人员对同一种数据在不同时间也会给出不同的分类结果,因而无普适的、唯一的分类标准。

随着计算机技术的发展,将文本挖掘(text mining)[2,3]技术运用于专利信息分析,从专利文本数据中抽取有价值的信息并通过机器学习获得分类模型以辅助专家进行高效的专利分类,已成为研究的热点和趋势。文本挖掘的引入一方面可以大幅提高专利分类效率,降低标引人员的负担;另一方面也可以针对不同标引人员、不同分析需求形成更具个性化、更有针对性的分类决策。传统的机器学习可分为监督学习和非监督学习两类,前者完全基于已标引样本训练分类模型,而后者则只利用未标引样本的信息。近年来出现了一种半监督学习(semi-supervised learning,SSL)[4]方法,它作为一种介于监督学习和非监督学习之间的方法,旨在综合利用标引样本和未标引样本两者的信息以提高分类模型的性能,受到越来越多的关注。在实际分类问题中,由于已标引样本往往需要通过人工标引获得,因而代价较高,数量也非常有限;而另一方面,未标引样本则大量存在于样本库中且易于获取。因此,如何有效利用有限的已标引样本和大量易于获得的未标引样本,是研究半监督学习的关键。主动学习[5-8]是解决上述问题的一种行之有效的方法。其主要思想是:仅仅选取对模型改进最有价值的样本进行标引,使得根据这些样本的标引信息所训练出来的新模型的性能得到尽可能大的提升。分类模型的选择是决定分类效果的又一关键因素。支持向量机(support vector machine,SVM)[9,10]作为一种性能优越的分类器,在

① 中央级公益性科研院所基本科研业务费专项资金(XK2012-2、ZD2012-7-2)和中国科学技术信息研究所科研项目预研基金(YY201208)资助项目。

② 男,1983 年生,博士,助理研究员;研究方向:模式识别与智能系统;联系人,E-mail:zhangxy@istic.ac.cn

(收稿日期:2012-12-26)

许多分类问题方面得到广泛应用。因此,本文重点研究了SVM分类模型下的主动学习算法在专利分类中的应用,提出了一种基于多分类器融合与主动学习的交互式专利分类算法。

1 分类决策模型构建

为了实现专利分类,本文算法基于训练集,针对各专利类别利用SVM分别训练得到相应的子分类器,然后通过多分类器融合,将各子分类器有机结合起来,从而构建总的分类模型并形成分类决策。

1.1 子分类器构建

用 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}=U\cup L$ 表示整个样本集合,其中 U 和 L 分别表示未标引样本集和已标引样本集。

对于二分问题,样本 $\boldsymbol{x}_n\in X(1\leqslant n\leqslant N)$,$y_n\in\{1,-1\}$ 对应于其类标。如果 $\boldsymbol{x}_n\in L$,则 y_n 已知;如果 $\boldsymbol{x}_n\in U$,则 y_n 未知,需要通过 $f(\boldsymbol{x}_n)$ 去预测 y_n,其中 f 是 X 上的分类器。

专利分类是多分类问题,是对二分问题的扩展,对于样本 $\boldsymbol{x}_n\in X$,其类标用 I 维向量表示为 $\boldsymbol{y}_n=(y_{n1},y_{n2},\cdots,y_{nI})$,其中 $y_{ni}\in\{1,-1\}(1\leqslant i\leqslant I)$ 对应于第 i 个类别的类标,I 表示类别数。相应地,需要针对每个类别 i 分别训练子分类器 f_i,从而获得 I 个子分类器。

在分类器构建方面,本文采用SVM分类模型,其基本思想是学习一个最优的超平面,以最大的分类间隔(margin)将训练数据分开。给定训练集 L,通过一个适当的Mercer核函数 K 所得到的隐式映射 $\Phi:X\to F$[9],SVM分类器可以表示为

$$f(\boldsymbol{x})=\boldsymbol{w}\cdot\Phi(\boldsymbol{x})+b \tag{1}$$

对于样本 $\boldsymbol{x}_n\in X$,$f(\boldsymbol{x}_n)$ 可以看成是一种"带符号距离",其符号表示样本所属的类别,其绝对值反映了样本到分类面的距离。

1.2 多分类器融合

在获得针对不同类别的 I 个子分类器的基础上,为了形成最终的分类决策,可以采用不同的多分类器融合算法。根据专利分类的实际,本文仅讨论单类标的情况,即每条专利能且仅能划分至一个类别中。

1.2.1 非融合

顾名思义,非融合算法直接将各子分类器作为融合分类器参与后续分类决策的形成,其公式表示为

$$F_j(\boldsymbol{x})=f_j(\boldsymbol{x}) \tag{2}$$

其中,$F_j(1\leqslant j\leqslant I)$ 表示第 j 个类别上的融合分类器。

非融合算法固然简单直接,但其缺点也非常突出:由于忽略了不同子分类器之间的相互关系,因而分类精度难以保证。

1.2.2 线性加权融合

为了反映不同分类器分类能力的强弱,可以采用线性加权对子分类器进行融合,其形式如下:

$$F_j(\boldsymbol{x})=\sum_{i=1}^{I}\delta(i,j)\mu_i f_i(\boldsymbol{x}) \tag{3}$$

其中 μ_i 是子分类器 f_i 的权重,是 f_i 分类性能强弱的度量,其取值由 f_i 在验证集上的分类结果确定;$\delta(i,j)$ 为符号函数:

$$\delta(i,j)=\begin{cases}1, & i=j\\ -1, & i\neq j\end{cases} \tag{4}$$

线性加权融合算法的前提假设是融合分类器与各子分类器之间是线性关系,因而对于线性模型较为有效,但是对于较为复杂的非线性模型则不可避免地会出现较大的偏差。

1.2.3 超核融合

为了更灵活有效地描述各子分类器之间的相互关系,本文采用超核融合(super-kernel fusion)[11]的方法实现不同子分类器的融合,其主要思想是:将各子分类器在验证集上的分类结果组成新的特征向量,并以这些新样本作为训练集,将训练得到的超分类器(super-classifier)作为融合结果。

基于 I 个不同类别得到相应的 I 个子分类器之后,算法分别对验证集样本 $\boldsymbol{x}_n$ 进行标引预测,并将分类决策组合成为一个新的 I 维的特征向量 $\boldsymbol{z}_n$:

$$\boldsymbol{z}_n=[f_1(\boldsymbol{x}_n),f_2(\boldsymbol{x}_n),\cdots,f_I(\boldsymbol{x}_n)]^{\mathrm{T}} \tag{5}$$

这样,便获得一个新的训练集 $\boldsymbol{Z}$,进而基于 $\boldsymbol{Z}$ 训练出一个超分类器,这一过程同样是一个多分类问题,本文仍使用SVM来求取超分类器。最终,融合分类器可以表示为

$$F_j(\boldsymbol{x})=S_j(f_1(\boldsymbol{x}),f_2(\boldsymbol{x}),\cdots,f_I(\boldsymbol{x})) \tag{6}$$

其中,S_j 即为训练所得的超核融合函数。

超核融合并不基于线性假设对子分类器之间的关系进行预先限定,而是完全通过训练学习获得融合结果,因而能在更大的自由度内获得更加符合数据自身分布的更优分类模型。

1.3 分类决策

由于分类器函数值本身包含了分类判定及分类确定度(certainty)信息,因此基于融合分类器,算法

将具有最大函数值的融合分类器所对应的类别作为分类判定，将最大函数值作为分类确定度，最终形成如下分类决策：

$$\begin{cases} D.class(\boldsymbol{x}) = \underset{1\leqslant j\leqslant J}{\operatorname{argmax}} F_j(\boldsymbol{x}) \\ D.certainty(\boldsymbol{x}) = \underset{1\leqslant j\leqslant J}{\max} F_j(\boldsymbol{x}) = F_{D.class}(\boldsymbol{x}) \end{cases} \tag{7}$$

分类决策模型构建流程如图 1 所示。

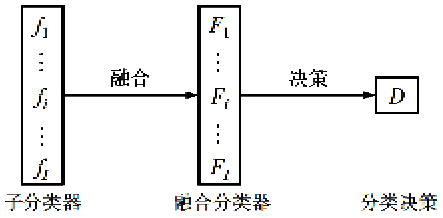

图 1 分类决策模型构建流程

2 主动学习模型更新

为了在现有分类模型基础上不断改进，使得分类决策与分类人员判定和实际分类需求尽可能契合，算法利用主动学习选取最有信息的样本进行标引，从而通过人机交互实现模型更新；同时，针对传统批量选择性采样的缺点，算法提出了动态批量选择性采样模式，通过确定度传播策略有效降低标引样本冗余度，以进一步提高主动学习的效率。

2.1 选择性采样

区别于传统的人工选择标引样本的方式，主动学习由计算机主动地选择最有信息的样本交由分类人员进行标引，从而利用尽可能小的标引量获取尽可能大的分类性能改进。

在主动学习算法中，选择性采样策略是关键。传统的主动学习本质上采用的都是“批量选择性采样”模式[12,13]，即批量地选择确定度最小的 k 个样本进行标引，最后根据这 k 个样本的标引信息对分类模型进行训练更新，这一过程可以概括为“批量标引，批量训练”。事实上，为保证算法效率，选择性采样应遵循以下两个重要准则[14]：(1) 低确定度准则：选取的样本应具有较低的确定度；(2) 低冗余度准则：选取的样本不应与先前已选取的样本之间存在过多的信息冗余。批量选择性采样模式虽然遵循了低确定度准则，却忽略了低冗余度准则。由于样本是以批量的方式同时选取并标引的，因此同一批被选取的样本之间的相互关系被忽略了，从而影响了样本总体信息量。

本文提出“动态批量选择性采样”模式，其主要思想是：每次仅选取一个最有信息的样本进行标引，并根据该样本的标引信息指导下一个样本的选取，如此循环；当一定数目的样本标引完成后，通过一次训练对分类模型进行更新。如图 2 所示，动态批量选择性采样所使用的是“逐一标引，批量训练”的过程。在动态批量选择性采样中，样本的选取不再仅仅取决于当前分类面，同时还在新标引样本信息的指导下动态地进行调整，从而在有效利用样本之间相互关系的同时也有效地弥补了当前分类模型的不足。在执行效率方面，由于动态批量选择性采样的训练过程仍然是批量的，也即在一定批量的样本全部完成标引之后才进行一次训练，因而其计算复杂度与批量选择性采样相当。可见，动态批量选择性采样有效地兼顾了性能和效率。

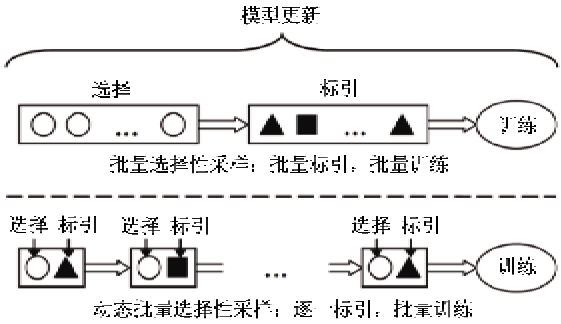

图 2 选择性采样模式比较

2.2 动态确定度传播

本文在动态批量选择性采样框架内提出了动态确定度传播（dynamic certainty propagation，DCP）算法，具体模型更新过程如下：每次选取确定度最低的一个样本进行标引，每当一个新的样本被标引之后，不仅其自身确定度将发生改变，而且这一改变将根据样本之间的相互关系动态传播，进而改变其它未标引样本的确定度，重复这一过程，直到一定批量的样本被标引；最后，利用标引样本进行分类器更新训练。通过考虑未标引样本之间的相互关系，算法在保持低确定度的同时有效地降低了所选取样本的冗余度，从而使得最终获取的样本具有更高的信息量。

2.2.1 算法流程

确定度反映了分类模型对于分类结果的确定程度，确定度越低样本就越有信息。算法用 C 表示样本确定度，初始情况下，每个未标引样本 $\boldsymbol{x}_n \in U$ 的确定度

$$C^{(0)}(\boldsymbol{x}_n) = D.certainty(\boldsymbol{x}_n) \tag{8}$$

由当前分类器决定。

算法采用逐一标引的方法,每次仅仅选取一个具有最低确定度的未标引样本 $x_l \in U$ 进行标引:

$$l = \mathrm{argmin}_n C^{(r-1)}(x_n) \tag{9}$$

其中,r 表示模型更新中第 r 个样本的选取。

对于未标引样本,其确定度并不是一成不变的,而是随着样本的标引过程动态变化的,它不仅取决于当前分类器,同时也受先前标引样本的影响。

在样本 x_l 被选取并标引之后,首先改变其自身确定度:

$$C^{(r)}(x_l) = M \tag{10}$$

其中,M 是一个预先设定的参数用以表示已标引样本的确定度。

然后,x_l 确定度的改变将传播到其它的未标引样本,产生相应的影响。影响的强度取决于各样本与 x_l 之间的在特征空间的相似度。为了描述样本之间的相互关系,算法构建了一个图模型 $G = (V, E)$,其中节点集合 V 即为样本集 X,而边集合 E 表示样本之间的关系,其权重由一个 $N \times N$ 的矩阵 $\boldsymbol{W}$ 决定,$\boldsymbol{W}$ 的元素 w_{mn} 对应于样本 x_m 和 x_n($1 \leqslant m, n \leqslant N$)之间的关系,用热核(heat kernel)[15]表示为

$$w_{mn} = \exp\left(-\frac{\| x_m - x_n \|^2}{t}\right) \tag{11}$$

其中,t 是反映样本之间影响强度的参数。显然,在特征空间内距离较近的样本之间的边权重较高,而随着样本之间距离的增大,其对应的权重将呈指数级衰减。基于 $\boldsymbol{W}$,算法将 x_l 确定度的改变传播到其它的未标引样本:

$$\Delta C^{(r)}(x_n) = w_{nl}\Delta C^{(r)}(x_l) \tag{12}$$

其中,

$$\Delta C^{(r)}(x_n) = C^{(r)}(x_n) - C^{(r-1)}(x_n) \tag{13}$$

可见,x_l 仅仅会对近邻样本产生显著影响,而对较远样本的影响较小。

综合式(12)和(13),对于每一个未标引样本,其新的确定度可以表示为

$$C^{(r)}(x_n) = C^{(r-1)}(x_n) + w_{nl}[C^{(r)}(x_l) - C^{(r-1)}(x_l)] \tag{14}$$

使用式(14),算法可以在每次样本标引之后动态地对其它样本的确定度进行更新。

图 3 总结了动态确定度传播算法流程。

输入:	当前分类器 f,已标引样本集 L,未标引样本集 U,样本相互关系矩阵 W,模型更新中需要标引的样本总数 k.
步骤 1:	用 f 对 U 进行分类,初始化未标引样本的确定度: $C^{(0)}(x_n) = Docertainty(x_n), x_n \in U$
步骤 2:	对 $r = 1, 2, \ldots, k$,重复执行(1)—(4): (1) 选择具有最小确定度的一个样本 x_l 进行标引: $l = \mathrm{argmin}_n C^{(r-1)}(x_n), x_n \in U$ (2) 将 x_l 从 U 中删除,并加入 L: $U = U - \{x_l\}, L = L \cup \{x_l\}$ (3) 改变 x_l 的确定度: $C^{(r)}(x_l) = M$ (4) 将 x_l 确定度的改变传播到其它的未标引样本: $C^{(r)}(x_n) = C^{(r-1)}(x_n) + w_{nl}\left[C^{(r)}(x_l) - C^{(r-1)}(x_l)\right], x_n \in U$
步骤 3:	基于 L 对 f 进行更新.

图 3　动态确定度传播算法流程

2.2.2　参数赋值

动态确定度传播算法包含两个参数:t 和 M。为了保证算法的通用性,参数赋值并不经验性地对其赋以定值,而是根据数据自身的分布自适应地取值。

参数 t 控制着新标引的样本对其周围样本的影响强度,确保只有近邻样本才会被其显著地影响。算法采用最近邻距离来自适应地刻画一个样本集空间分布的紧密程度。对于样本 $x_n \in X$,首先计算其到最近邻之间的距离,记为 $d(x_n)$,然后对所有最近邻距离构成的整个集合 $\{d(x_n)\}$ 计算算术平均:

$$\bar{d} = \frac{1}{N}\sum_{n=1}^{N} d(x_n) \tag{15}$$

最终参数 t 赋值为

$$t = 2\bar{d}^2 \tag{16}$$

参数 M 描述的是已标引样本的确定度,由于样本已标引,因而其确定度 M 相对较高。首先,算法计算所有未标引样本确定度 $\{C^{(0)}(x_n)\}$ $(x_n \in U)$ 的均值和方差:

$$\mu_C = \frac{1}{|U|}\sum_{x_n \in U} C^{(0)}(x_n) \tag{17}$$

$$\sigma_C^2 = \frac{1}{|U|}\sum_{x_n \in U}[|C^{(0)}(x_n)| - \mu_C]^2 \tag{18}$$

其中,$|U|$ 表示未标引样本集合 U 的大小。然后,对 M 进行如下赋值:

$$M = \mu_C + 3\sigma_C \tag{19}$$

这样赋值,一方面保证了 M 足够大,从而可以显示出对于已标引样本足够高的确定度;另一方面也确保其不至于过分大,仍然与大多数未标引样本的确定度具有可比性。

3 实 验

本文将不同算法分别应用于交互式专利分类，通过对比实验，对算法有效性进行验证。

3.1 实验设计

实验数据来源于 Innography 数据库，选取与电动汽车领域相关的约 5000 件专利作为实验集，这些专利经专家标引共分为 5 个类别：电池、电池管理、电机、电机控制、整车控制。

实验基于向量空间模型，从专利数据中提取出 5484 条术语用以表征专利的文本特征，术语在专利中的权重通过计算 TF-IDF 获得；在此基础上，采用主成分分析（principal component analysis，PCA）降维，最终得到 150 维的专利文本特征向量。

实验集中，10% 的数据作为初始训练集用于子分类器的训练，10% 作为验证集用于多分类器融合，在交互式分类中每次标引 10% 的数据用于模型更新，最终通过在 5 个类别上计算平均的 $F1$ 值来对分类结果进行评估，计算公式如下：

$$F1 = \frac{2pr}{p + r} \tag{20}$$

其中，p 和 r 分别代表查准率（precision）和查全率（recall）。

实验比较了在模型构建和模型更新阶段所采用的各种不同算法，如表 1 所示。

表 1 实验算法

算法	说明
算法 1	非融合·非主动学习
算法 2	线性加权融合·非主动学习
算法 3	超核融合·非主动学习
算法 4	超核融合·主动学习（批量选择性采样）
算法 5	超核融合·主动学习（动态批量选择性采样：DCP 算法）

3.2 实验结果

图 4 给出了不同算法在专利分类过程中平均 $F1$ 值随着模型更新次数变化的情况比较。

表 2 列出了不同算法在各轮模型更新后的平均 $F1$ 值。

实验结果分析如下（其中“>”表示“优于”）：

（1）多分类器融合算法：超核融合（算法 3）> 线性加权融合（算法 2）> 非融合（算法 1）。说明不同子分类器有机结合、协同作用可以有效提升最终分类模型的性能；超核融合由于放松了线性相关的假设，因此相对于线性加权融合能够更好地适应复杂的、非线性的数据分布。

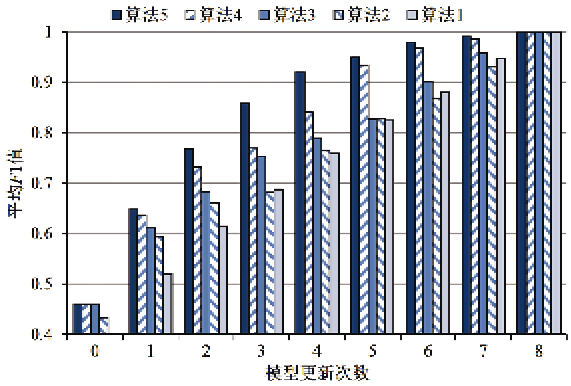

图 4 专利分类算法结果比较

表 2 各轮模型更新后平均 $F1$ 值

	算法 1	算法 2	算法 3	算法 4	算法 5
平均 $F1$ 值	0.737	0.751	0.776	0.814	0.842

（2）模型更新学习算法：主动学习（算法 4、5）> 非主动学习（算法 1、2、3）。说明在主动学习中分类模型可以有针对性地对未标引样本进行选取，相对于只能被动接受标引信息的非主动学习方法，其效果更好。

（3）主动学习算法：动态批量选择性采样（算法 5）> 批量选择性采样（算法 4）。说明动态批量选择性采样过程中考虑了先后标引的样本之间的关系，有效降低了冗余度。

4 结 论

本文提出了一种基于多分类器融合与主动学习的交互式专利分类算法，其通过深入挖掘专利文本信息实现专利的高效分类。本文的贡献在于：（1）利用多分类器融合实现各子分类器的有机结合、协同作用，从而构建总的分类模型并形成分类决策；（2）将主动学习引入专利标引过程中，通过计算机主动选取和分类人员手工标引相结合的人机交互方式实现模型更新；（3）提出动态批量选择性采样模式，综合利用当前分类模型和先前标引样本这两方面的信息对后续样本的选取进行指导，通过确定度传播有效降低标引样本冗余度和提高主动学习的效率。实验结果表明，本文提出的算法能够有效地提升专利分类性能。

参考文献

[1] Fall C J, Benzineb K. Literature survey: Issues to be considered in the automatic classification of patents, V1.0/29.10.02. Geneva: World Intellectual Property Organization, 2002

[2] Feldman R, Sanger J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York: Cambridge University Press, 2006

[3] Tseng Y H, Lin C J, Lin Y I. Text mining techniques for patent analysis. *Information Processing & Management*, 2007, 43(5), 1216-1247

[4] Zhu X. Semi-supervised learning literature survey, Computer Sciences TR 1530. Wisconsin: University of Wisconsin-Madison, 2008

[5] Cohn D A, GhahramaniZ, Jordan M I. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 1996, 4(1):129-145

[6] McCallum A, Nigam K. Employing EM in pool-based active learning for text classification. In: Proceedings of the 15th International Conference on Machine Learning, Bari, Italy, 1998. 350-358

[7] Zhang X Y. Dynamic batch selective sampling based on version space analysis. *High Technology Letters*, 2012, 18(2): 208-213

[8] 张晓宇. 基于多视角二维主动学习的多标签分类. 高技术通讯, 2011, 21(12): 1312-1317

[9] Burges J C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998, 2(2): 121-167

[10] 王朋宇,郭崎,沈海华等. 使用支持向量机的微处理器验证向量优化方法. 高技术通讯, 2010, 20(1): 68-74

[11] Wu Y, Chang E Y, Chang K C C, et al. Optimal multimodal fusion for multimedia data analysis. In: Proceedings of ACM International Conference on Multimedia, New York, USA, 2004. 572-579

[12] Brinker K. Incorporating diversity in active learning with support vector machines. In: Proceedings of the 20th International Conference on Machine Learning, Washington, DC, USA, 2003. 59-66

[13] Tong S, Chang E. Support vector machine active learning for image retrieval. In: Proceedings of the 9th ACM International Conference on Multimedia, Ottawa, Canada, 2001. 107-118

[14] Zhang T, OlesF. A probability analysis on the value of unlabeled data for classification problems. In: Proceedings of the 17th International Conference on Machine Learning, Stanford, USA, 2000: 1191-1198

[15] Belkin M, Niyogi P. Laplacianeigenmaps and spectral techniques for embedding and clustering. In:Proceedings of Advances in Neural Information Processing Systems, Vancouver, Canada, 2001. 585-591

Interactive patent classification based on text mining

Zhang Xiaoyu

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract

This paper introduces the text mining technique into patent analysis and proposes an interactive patent classification algorithm based on multi-classifier fusion and active learning to achieve high classification performance. The algorithm first trains a sub-classifier for each class of the patents by means of support vector machine. Then, via multi-classifier fusion, the sub-classifiers are effectively combined to acquire enhanced classifiers, based on which the classification decision can be made. For refinement of the classification model, active learning is used to select the most informative patents for labeling. Finally, the dynamic batch sampling is presented to address the problem of traditional batch sampling. With dynamic certainty propagation, the selected patents become less redundant and thus more informative for active learning. The experimental results demonstrate the effectiveness of the proposed interactive patent classification algorithm based on multi-classifier fusion and active learning.

Key words: text mining, patent classification, multi-classifier fusion, active learning, selective sampling