**发明名称**

一种针对公安领域案件文本分类的方法

**摘要**

本发明适用于公安领域案件的数据挖据应用领域，提供了一种针对公安领域案件文本分类的方法，所述方法包括以下几个步骤：A、对案件文本进行预处理，包括分词、筛去停用词、文本向量化等；B、建立层次支持向量机模型对案件文本进行自动分类；C、建立词典规则匹配案件分类器；D、根据决策函数值，对于支持向量机分类所得结果置信度低的案件，利用基于词典的规则匹配案件分类器对其进行二次分类，以其分类结果为最终结果。通过以支持向量机分类为主，以基于词典的规则匹配案件分类为辅的方法，使得案件分类的准确率得到提高。本方法可辅助警务人员处理海量公安案件文本，对案件进行正确分类，为后续的案件研判提供帮助。

**摘要附图**

对案件文本进行预处理，包括分词、筛去停用词、文本向量化等

建立通用型词典匹配案件分类器

建立层次支持向量机模型

对于支持向量机分类所得结果置信度低的案件，利用通用型词典匹配案件分类器对其进行二次分类，得出结果

**权利要求书**

1. 一种针对公安领域案件文本分类的方法，其特征在于，包括以下步骤：
2. 对案件文本进行预处理，包括分词、筛去停用词、文本向量化等；
3. 建立层次支持向量机模型对案件文本进行自动分类；
4. 建立基于词典的规则匹配案件分类器；
5. 根据决策函数值，对于支持向量机分类所得结果置信度低的案件，利用基于词典的规则匹配案件分类器对其进行二次分类，以其分类结果为最终结果。

2、如权利要求1所述的方法，其特征在于：所述步骤A对案件文本进行预处理，主要的文本预处理步骤如下：

1. 先进行案件描述内容提取，利用正则表达式匹配的方法，提取“报警称：”之后的文本内容，然后对该文本内容进行分词；
2. 在对所有案件文本进行分词之后，根据文档频数筛选停用词，满足条件*Ntd >= 70%\*Nd*的词被列为停用词，其中*Ntd*为包含词t的案件文本数量，*Nd*为总的案件文本数量；
3. 在已过滤停用词的条件下，采用TF-IDF公式计算权重，将所有案件文本向量化。

3、 如权利要求1所述的方法，其特征在于：所述步骤B建立层次支持向量机模型对案件文本进行自动分类，前提需要制定案件类别的级别关系，例如:{盗窃：{入室盗窃，公交车盗窃，盗窃电动车、摩托车}，诈骗：{电信诈骗，街头诈骗}}，然后建立层次支持向量机模型对案件文本进行自动分类。

4、如权利要求1所述的方法，其特征在于：所述层次支持向量机模型：所有一级案件类别的分类对应一个一级分类器模型，同属于一类一级类别的所有二级案件类别对应一个二级分类器模型，因此二级分类器模型的数量与一级案件类别的数量一样多；当层次支持向量机模型预测一个案件，先调用一级分类器判断其一级案件类别，根据获得的一级案件类别调用对应的二级分类器，获得其二级案件类别。

5、如权利要求1所述的方法，其特征在于，所述基于词典的规则匹配案件分类器：需要先构造案件分类识别词典：词典由5部分组成：“规则序号”、“一级案件类别”、“二级案件类别”，“关键词”、“排斥词”；“关键词”由一个或多个词组成，词汇间为“与关系”。即案件描述中，同时包含“关键词”中的所有词，才匹配该规则。当“关键词”匹配上后，若该规则的“排斥词”不为空，则还需进一步比对。“排斥词”用于区分两类具备相同“关键词”的案件，该字段同样包含一个或多个词，但词汇间为“或关系”。即案件描述中只要含有“排斥词”中的任意一个词，则表示不符合该规则，继续比对下一条规则；对于匹配上词典的案件，返回“规则序号”、“一级案件类别”、“二级案件类别”。

6. 如权利要求1所述的方法，其特征在于，判断支持向量机分类所得结果置信度低的判断方法为：以各子分类器决策函数值分析置信度的高低，当出现以下情况则判定为置信度低：

1. 各个分类器的决策函数值均为负数；
2. 仅一个分类器的决策函数值为正数，但其值很小，小于1；
3. 出现两个分类器的决策函数值为正数，且数值很接近，相对平均偏差≤5%；
4. 出现三个及以上的分类器的决策函数值为正数。

**说明书**

一种针对公安领域案件文本分类的方法

**技术领域**

本发明涉及将文本挖掘技术应用于公安领域，特别提供了一种公安领域案件文本分类的方法。

**背景技术**

随着社会的发展，公安领域的案件信息以每年百万条的速度递增。目前公安部门虽然有完善的案件信息管理系统，可以方便地对案件进行查询、筛选、统计等处理，为警务人员减少了一定的劳动量，但是依靠这些传统方法，很难挖掘出有价值的线索。因此，许多国内外学者在对公安领域中的信息数据挖掘技术进行深入地研究。

公安领域的案件信息都是以文本的形式记录在案的，这些案件文本中包含着重要的线索，关联相似的案件可以有效地帮助案件的侦破和串并案的发现。因此，对案件文本处理的需求越来越迫切，同时要求的精度和准确度也越来越高，尤其是需要对大批案件进行自动识别案件类别。当前运用较为广泛的文本分类方法有朴素贝叶斯分类、k近邻分类、支持向量机模型分类、基于决策树的分类。其中，支持向量机模型分类具有适用于小样本学习，解决非线性问题，解决高维问题和抗干扰强的能力，总体表现出色，它基于结构风险最小化理论之上在特征空间中建构最优分割超平面。

但是，支持向量机模型分类的缺点有：对缺失数据敏感，对样本的预测依赖于训练所得的分类决策函数，根据决策函数的值判断样本更接近哪个类别，即使在该决策的置信度很低的情况下，依然会给出一个结果，尽管这个结果很有可能是错误的。其次，在公安领域案件分类这一特定场景，案件类别存在层次结构，如果仅采用单一层次的分类器，忽视了案件类别的层次关系，会导致分类的准确率低下。

**发明内容**

本发明提供了一种针对公安领域案件文本分类的方法，用于满足公安领域案件文本处理的需求，并且避免接受置信度低的分类决策结果，提高分类准确。

一种针对公安领域案件文本分类的方法，包括以下步骤：

1、对案件文本进行预处理，包括分词、筛去停用词、文本向量化等。案件文本预处理的内容包括：

1. 先进行案件描述内容提取，利用正则表达式匹配的方法，提取“报警称：”之后的文本内容，并对提取的文本内容进行分词；
2. 在对所有案件文本进行分词之后，根据文档频数筛选停用词，满足条件*Ntd >= 70%\*Nd*的词被列为停用词，其中*Ntd*为包含词t的案件文本数量，*Nd*为总的案件文本数量；
3. 在已过滤停用词的条件下，采用TF-IDF公式计算权重，将所有案件文本向量化；

2、制定案件类别的级别关系，例如:{盗窃：{入室盗窃，公交车盗窃，盗窃电动车、摩托车}，诈骗：{电信诈骗，街头诈骗}}，然后建立层次支持向量机模型对案件文本进行自动分类；

3、建立层次支持向量机模型，所有一级案件类别的分类对应一个一级分类器模型，同属于一类一级类别的所有二级案件类别对应一个二级分类器模型，因此二级分类器模型的数量与一级案件类别的数量一样多；当层次支持向量机模型预测一个案件，先调用一级分类器判断其一级案件类别，根据获得的一级案件类别调用对应的二级分类器，获得其二级案件类别。

4、建立基于规则的匹配案件分类器，主要步骤有：

1. 先构造案件分类识别词典：词典由五个字段组成：“规则序号”、“一级案件类别”、“二级案件类别”，“关键词”、“排斥词”；词典中的每一条记录称之为规则，每条规则具有上述五个字段，所有规则构成案件分类识别词典字典。
2. 制定通用型词典匹配案件分类器分类的过程，具体为：从词典中第一行规则开始比对，检验该行的关键词列中的关键词是否都出现在这个待分类案件文本中，“关键词”由一个或多个词组成，词汇间为“与关系”。即案件描述中，同时包含“关键词”中的所有词，才匹配该规则。当“关键词”匹配上后，若该规则的“排斥词”不为空，则还需进一步比对。“排斥词”用于区分两类具备相同“关键词”的案件，该字段同样包含一个或多个词，但词汇间为“或关系”。即案件描述中只要含有“排斥词”中的任意一个词，则表示不符合该规则，继续比对下一条规则；对于匹配上词典的案件，返回“规则序号”、“一级案件类别”、“二级案件类别”。“规则序号”字段用于人工核验分类识别结果时，能够快速准确找到指定分类规则，进行调整优化。“一级案件类别”即案件分类识别的大类，如“电信诈骗”、“街头诈骗”、“扒窃”等。“二级案件类别”是按照具体作案手段等特征对“一级案件类别”的进一步细分。例如：“扒窃”一级案件类别下，有“车扒”（公交车上扒窃）、“地扒”（街头扒窃）等二级案件类别。

4、案件文本分类的时候，先使用层次支持向量机对其进行分类，再评估支持向量机分类所得结果的置信度：以决策函数值作为衡量置信度大小的标准，当出现以下情况则判定为置信度低：

a）各个分类器的决策函数值均为负数；

1. b）仅一个分类器的决策函数值为正数，但其值很小，小于1；
2. c) 出现两个分类器的决策函数值为正数，且数值很接近，相对平均偏差≤5%；
3. d) 出现三个及以上的分类器的决策函数值为正数。

5、对于支持向量机分类所得结果置信度低的案件，利用通用型词典匹配案件分类器对其进行二次分类，以其分类结果为最终结果。

**附图说明**

图1为本申请实施例中层次支持向量机模型的层次结构示意图；

图2为本申请实施例中通用型词典匹配案件分类的工作过程；

图3为本申请实施例中支持向量机分类与**通用型词典匹配案件分类**相结合的分类方法的实施的过程。

**具体实施方式**

本发明实施例提供了一种公安领域案件文本分类的方法，其包含一种基于层次支持向量机分类和词典规则匹配的案件分类器。

当利用层次支持向量机分类器模型对需要分类的案件文本进行分类时，先执行上级分类器分类，再依据上级分类器分类所得的结果，执行对应的下级分类器分类；比如一个案件，先由上级分类器分类，假设所得结果为“盗窃”，然后根据这个结果由“盗窃”类的下级分类器再进行分类，确定这个案件的最终类别是什么，如“入室盗窃”或“公交车盗窃”。在上级分类层中，案件类别差异较大，而类别数目较少，使得特征选择后的词语集合的类别区分能力强，从原理上说，支持向量机的判决函数只与支持向量有关，如果支持向量差异明显，分类间隔的宽度就比较大，分类准确率较高。在上级分类的准确率得到保障的基础上，从对应同一个上级类别的各个下级类别的案件样本子集中抽取出区分能力强的特征词语，构成新的词语集合。各个下级类别的分类器都有属于自己的特征词语集合，能减少不相关类别的信息干扰，这有利于寻找到最优分类决策面。分类器的层次结构示意图参见图1。

构建层次支持向量机模型的步骤包括：

步骤1：获取原始案件文本数据，以及每个案件进行案件类别标注 ；根据存在案件类别和他们之间的层次关系，构建层次案件类别词典（词典示例如下：{盗窃：{入室盗窃，公交车盗窃，盗窃电动车、摩托车}，诈骗：{电信诈骗，街头诈骗}，……}，字典中的key值如“盗窃”、“诈骗”均为一级案件类别，value值为对应一级案件类别细分后的二级案件类别集合，如{入室盗窃，公交车盗窃，盗窃电动车、摩托车}是一级类别“盗窃”细分得到的下级类别集合；

步骤2：获取案件文本数据和案件类别之后，需要进行案件文本预处理工作：

1. 中文分词是文本预处理的第一步，先进行案件描述内容提取，利用正则表达式匹配的方法，提取“报警称：”之后的文本内容，然后对提取的文本内容进行分词采用"jieba"分词工具进行分词；
2. 预处理的内容还包括：在对所有案件文本进行分词之后，根据文档频数筛选停用词，满足条件*Ntd >= 70%\*Nd*的词被列为停用词，其中*Ntd*为包含词*t*的案件文本数量，*Nd*为总的案件文本数量；
3. 最后是将案件文本转换成计算机能够理解的表示形式；选用适用于文本分类问题的向量空间模型，该模型的主要思想是：将每一个文档都映射为由一组规范化正交词语矢量张成的向量空间中的一个点。对于所有的文档类和未知文档，都可以用此空间中的词语向量来表示（其中，为特征词语；为词的权重）。一般需要构造一个评价函数来表示词语权重，其计算的唯一准则就是要最大限度地区别不同文档。常用的特征项的权重计算方法有TF-IDF方法，布尔模型方法等，本发明选用TF-IDF权重计算方法，某一特定文本内的高词语频率，以及该词语在整个文本集合中的低文本频率，可以产生出高权重的TF-IDF，词*i*在文本*j*中的权重计算公式如下：

为词*k*在文本*j*中出现的次数，为语料库中的文件总数，为包含词语的文本数目。

步骤3：训练支持向量机模型，将经过预处理的案件文本作为支持向量机（SVM）的输入，进行案件分类。SVM基本模型被定义为特征空间上的间隔最大的线性分类器，其学习策略是使间隔最大化，最终转化为一个凸二次规划问题的求解。下面是SVM的目标函数:

求该目标函数,得到:；

根据 ,计算，选择的一个分量，满足 ；

并据此计算出 ，其中，是案件文本和的核函数，为案件文本的案件类别，为惩罚因子，在实验中我们选值为1，多类分类策略选用“一类对余类”策略。

对于核函数，采用的是径向基函数RBF，见下式。通过核函数将特征空间由低维向高维空间映射。



本发明实施例还提供的方法还包括一种通用型词典匹配案件分类器，建立**通用型词典匹配案件分类器**，主要步骤有：

1. 先构造案件分类识别词典：词典由五个字段组成：“规则序号”、“一级案件类别”、“二级案件类别”，“关键词”、“排斥词”；词典中的每一条记录称之为规则，每条规则具有上述五个字段，所有规则构成案件分类识别词典字典；
2. 制定通用型词典匹配案件分类器分类的过程，当给定一个待分类案件文本，通用型词典匹配案件分类器进行分类的过程为（参见图2），从词典中第一行规则开始比对，检验该行的关键词列中的关键词是否都出现在这个待分类案件文本中，“关键词”由一个或多个词组成，词汇间为“与关系”。即案件描述中，同时包含“关键词”中的所有词，才匹配该规则。当“关键词”匹配上后，若该规则的“排斥词”不为空，则还需进一步比对。“排斥词”用于区分两类具备相同“关键词”的案件，该字段同样包含一个或多个词，但词汇间为“或关系”。即案件描述中只要含有“排斥词”中的任意一个词，则表示不符合该规则，继续比对下一条规则；对于匹配上词典的案件，返回“规则序号”、“一级案件类别”、“二级案件类别”。

以上分别介绍了层次支持向量机分类器模型，通用型词典匹配案件分类器的内部结构和功能，下面对以支持向量机分类为主，通用型词典匹配案件分类为辅，两种方法相结合的实现过程进行介绍：

参见图3，本实施例中两种分类结合的方法流程如下：

步骤1：获取待分类样本，提取“报警称：”之后的文本内容；

步骤2：利用“jieba”分词对待分类案件进行分词，筛去停用词和词性为人名的词；

步骤3：假设上级分类器的词语特征集为，采用TF-IDF权重计算方法，将该案件向量化，得到它的特征向量，然后利用上级分类器预测其一级类别，假设得到一级类别为*L1*。

步骤4：根据一级类别*L1*，调用*L1*类别对应的下级分类器，假设该下级分类器的特征集为，采用TF-IDF权重计算方法向量化该案件得到新的特征向量，最后得到由该下级分类器预测出的最终案件类别。

实际情况中，当待分类案件样本实际所属的案件类别是支持向量机模型不曾学习过的，所有分类时，将该案件识别为任意已知类别都是错误的，应拒绝接受支持向量机分类器给出的结果，可以定义该结果的置信度非常低，类似地，当分类器出现给出的结果置信度低的其他情况时，也应当拒绝接受。分类器的置信度是一个值得重视的参量。

本发明中构造的SVM多分类分类器均采用“一类对余类”策略。在决策过程中，对待分类样本分别计算各个子分类器的决策函数值，并选取分类器函数值最大所对应的类别作为测试样本的预测类别。本发明仅对下级分类器的各个子分类器给出的决策函数值进行置信度评估，对置信度低的结果予以拒绝接受。

经过分析和实验，对满足以下情况判断为置信度低：

1. 各个分类器的决策函数值均为负数；
2. 仅一个分类器的决策函数值为正数，但其值很小，小于1；
3. 出现两个分类器的决策函数值为正数，且数值很接近，相对平均偏差≤5%；
4. 出现三个及以上的分类器的决策函数值为正数。

步骤5：通过对分类器给出的结果进行基于决策函数的置信度评估，拒绝置信度水平相对较低的决策结果，接受置信度水平较高的决策结果。对于被拒识的案件，调用通用型词典匹配案件分类器对其进行分类，确定其类别。当出现遍历所有规则仍没有匹配成功，就说明对该案件分类失败。

本发明中的基于规则匹配案件分类器依赖于人工经验积累编写而成的规则集合，适用于识别“纠纷”、“举报”等类别的案件，对于逻辑关系复杂的案件类别容易产生错误，而且由于规则表中规则数量较多，每一步还需迭代关键词和排斥词，所以分类速度远远慢于SVM分类。因此本发明采用以SVM层次分类器为主，规则匹配分类器为辅的方法对案件进行分类。

**说明书附图**

大类分类

盗窃类分类

诈骗类分类

……

……

街头诈骗

电信诈骗

车辆盗窃

扒窃

入室盗窃

图 1

待分类案件

否

对当前类别的关键词是否全都匹配（与关系）

是

否

对当前类别的排斥词是否存在匹配（或关系）

是

否

是否还有下一条规则

是

匹配下一条规则

将当前类别赋予案件

未能成功匹配（无结果）

图2

案件类别

训练模型

字典分类器

训练样本

是

否

案件类别

测试样本

SVM分类器

是否拒识？

文本预处理

图3