软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software

E-mail: jiangjing@buaa.edu.cn
http://www.jos.org.cn
Tel: 13466751061

一种基于支持向量机和主题模型的评论分析方法 RASL*

陈 琪[1]，张 莉[1,2]，蒋 竞[2]，黄新越[2]

[1](北京航空航天大学 软件学院，北京 100191)

[2](北京航空航天大学 计算机学院，北京 100191)

通讯作者：蒋竞，E-mail：jiangjing@buaa.edu.cn

摘 要： 在移动应用软件中，用户评论是一种重要的用户反馈途径。用户可能提到一些移动应用使用中的问题，比如系统兼容性问题、应用崩溃等。随着移动应用软件广泛流行，用户提供大量无结构化的反馈评论。为了从用户抱怨评论中提取有效信息，本文提出了一种基于支持向量机和主题模型的评论分析方法 RASL（Review Analysis method based on SVM and LDA），来帮助开发人员更好更快地了解用户反馈。本文首先对移动应用的中、差评提取特征，然后使用支持向量机对评论进行多标签分类。接下来，使用 LDA 主题模型(Latent Dirichlet allocation)对各问题类型下的评论进行主题提取与代表句提取。本文从两个移动应用中爬取 5141 条用户原始评论，并对这些评论分别用本文方法和 ASUM 方法进行处理得到两个新的文本。与经典方法 ASUM 相比，本文提出方法的困惑度更低、可理解性更佳，包含更完整的原始评论信息，冗余信息也更少。

关键词： 用户评论;分类;主题分析

中图法分类号:

中文引用格式：陈琪，张莉，蒋竞，黄新越. 一种基于支持向量机和主题模型的评论分析方法 RASL. 软件学报.

英文引用格式：Chen Q, Zhang L, Jiang J, Huang XY.a user review topic analysis Approach RASL based on the classification. Ruan Jian Xue Bao/Journal of Software,(in Chinese).

Review Analysis method RASL based on SVM and LDA

CHEN Qi[1], ZHANG Li[1,2], JIANG Jing[2], HUANG Xin-Yue[2]

[1](Software College Beihang University，Beijing 100191)

[2](School of Computer science and Engineering Beihang University，Beijing 100191)

Abstract: In mobile apps (applications), the app reviews by users have become an important feedback resource. Users may raise some issues when they use apps, such as system compatibility issues, application crashes, and so on. With the development of mobile apps, users provide a large number of unstructured feedback comments. In order to extract effective information from user complaint comments, we propose a review analysis method based on SVM and LDA (RASL) which can help developers to understand user feedback better and faster. Firstly, we extract features from the user neutral reviews and negative reviews, and then we use the support vector machine (SVM) to label comments on multiple tags. Next, we use the Latent Dirichlet allocation(LDA) topic model to get topic extraction and representative sentence extraction which are performed on the comments under each question type. We crawl 5141 original reviews from two mobile apps. Then we

* 基金项目: 国家重点研发计划项目(2018YFB1004202)；国家自然科学基金(61732019)

Foundation item: National Key Research and Development Program (2018YFB1004202) ; National Natural Science Foundation of China (61732019)

收稿时间: 0000–00–00; 修改时间: 0000–00–00; 采用时间: 0000–00–00; jos 在线出版时间: 0000–00–00

CNKI 网络优先出版: 0000–00–00

use our methods(RASL) and ASUM to process these comments to get new texts. In comparison with the Classical approach ASUM, RASL has less perplexity, better understandability, more complete original review information and less redundant information.

Keywords：User comments; Classification; Topic analysis

随着移动互联网兴起，产生了大量针对移动应用的在线评论信息。移动应用程序的用户群体广泛，用户的反馈丰富，并且随着版本迭代迅速更新。尤其用户对移动应用的中评和差评（简称中差评），是收集用户问题的重要数据来源。现有的应用分发平台都支持用户对应用进行评论。比如 360 手机助手, 它是国内市场份额较大的移动应用平台，提供应用程序卸载、安装、升级和评论等一系列服务。一些实证性研究表明，用户评论中包含很有价值的信息，例如错误报告、功能需求和用户体验等[1]；对开发者来说，应用市场中的用户评论能够帮助他们更好地理解用户反馈，提高软件质量[2][3]。

随着移动应用的广泛流行，用户评论的数量庞大，并且是无结构的，手动检查耗时且低效。因此，需要信息挖掘对用户的中差评进行处理，使用户抱怨信息的核心内容直观的展现在开发者面前，让开发者更快更有针对性的对软件进行更新。为了分析用户评论，现有研究主要采用分类或者主题提取的方法。首先，是对用户评论进行分类，Panichella 等人[4]通过评估发现文本分析、自然语言处理和情感分析三种技术结合可以得到最好的分类结果。Maalej 等人[5]尝试了多种技术对用户评论进行处理与分类，通过实验发现多个二元分类器优于单一多元分类器。McIlroy 等人[6]对几种机器学习分类器进行了比较，通过评估最终采用支持向量机进行分类。其次，一些研究采用主题提取的方法分析评论。Galvis 等人[7]将意见挖掘领域的 ASUM 模型用于软件应用的用户评论中，来自动地提取评论内包含的主题。姜巍[8]提出了针对意见挖掘问题域的关联 LDA 模型并应用于用户在线评论。

上述研究工作单纯考虑分类或者主题提取的方法，没有结合两种方法来分析评论。我们以两条评论为例说明分类与主题分析的区别：(1)“页面中的按钮没反应。”(2)“应该在页面中添加一个按钮。”这两条评论以分类的方法来处理，将分为两类：(1)软件错误；(2)请求增加功能。以主题分析的方法来处理，将提取出“页面”、“按钮”这样的主题。可以发现，分类方法能够了解用户遇到的问题种类，但很难得到评论中针对的软件特征，而主题分析方法能够得到特征信息，但很难区别用户的意图。如果我们将分类与主题分析结合起来，那么就既能通过分类得到评论指出的问题种类，又能通过主题挖掘得到评论里具体针对的软件特征。对于前面提到的两条评论的例子，结合分类与主题分析得到的结果可能如图 1 所示：

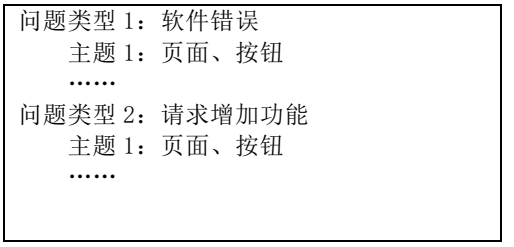

Fig.1 Results based on classification and topic extraction

图 1 基于分类和主题提取的结果

开发者则能由此了解用户遇到了页面中按钮相关的错误，以及用户希望为页面中增添按钮。相比于单独的分类分析或是主题分析，都能更精确地定位需求。

针对该问题本文提出了一种基于支持向量机和主题模型的评论分析方法，该方法从分类和主题分析两个方面对用户评论进行研究，并将这两者结合起来，更好地帮助开发者理解用户中差评中包含的需求，最后将所得结果进行对比试验。首先，该方法对用户评论进行特征提取，使用代价敏感学习减轻不平衡数据带来的影响，并将提取到的文本用线性支持向量机进行多标签分类。然后采用主题模型，对每组分类的用户评论提取主题词和主题句，最终得到了基于分类的主题分析结果。为了证明我们方法的有效性，我们从 360 手机助手随机选取了评分高(今日头条)和评分低(360 云盘)的两个应用，分析了两个应用共 5141 条用户中差评，实验

结果表明，本文提出方法获得的结果优于 ASUM 方法[7]。

本文主要贡献如下：

- 本文提出了一种基于支持向量机和主题模型的评论分析方法，更好地帮助开发人员了解用户反馈。
- 与经典方法 ASUM[7]相比，本文提出方法的困惑度更低、可理解性更佳，包含更完整的原始评论信息，冗余信息也更少。

本文后续的组织方式如下：第 1 节介绍分类方法和主题词提取等相关内容。第 2 节概述本文方法框架并且详细讲解本文提出的 RASL 方法。第 3 节通过定性和定量实验评估验证方法效果。第 4 节进行有效性分析。第 5 节给出结论及未来工作展望。

1 相关工作

近年来，一些研究人员对用户评论进行主题挖掘。提取出评论中的主题并给出主题下的代表性句子能使开发者快速、直观地理解用户反馈。Blei 等人[9]在 2003 年就提出了主题挖掘模型 LDA 模型，这是主题挖掘方面的经典模型。姜巍等人[8]提出了针对意见挖掘问题域的关联 LDA 模型并应用于用户在线评论。Galvis 等人[7]将主题挖掘领域的 ASUM 模型用于软件应用的用户评论中，来自动地提取评论内包含的主题。在结果的呈现上，为每个主题给出了代表性的句子，需求工程师可以查看这些具有代表性的用户评论，来决定主题是否是需求更改的候选项，比单纯的主题词集合更容易理解，是一种较好的主题表现形式.

除了主题挖掘，对用户评论进行分类也是用户反馈获取的主流方法。Panichella 等人[4]认为主题分析技术对于发现评论文本中的主题是有用的，但是它们不能揭示包含特定主题评论的用户的意图。文章基于文本分析、自然语言处理和情感分析，设计了三种不同的技术，从评论中提取出特征，然后使用这些特征来训练机器学习的分类器。通过评估发现结合三种技术可以得到最好的结果。Maalej 等人[5]尝试了多种技术对用户评论进行处理与分类，如：字符串（关键词）匹配；情感分析；二元分类器与多元分类器。最终通过实验发现多个二元分类器优于单一多元分类器。McIlroy 等人[6]则关注评论的多标签问题，认为一条评论可能包含着多个问题。文章中提出了 14 种类型的问题，并认为这些问题是相对于特定应用来说是独立的，并对朴素贝叶斯、J48 决策树、支持向量机这几种机器学习分类器进行了比较，通过评估，最终采用支持向量机进行分类。Pagano 等人[2]在 2013 年调查了苹果应用商店（AppStore）上用户评论的具体内容，并将这些内容按照主题进行了分类，该方法是否适合中文应用市场，还需要进一步验证。在本文的前期工作中，也对用户评论进行了分类，Zhang 等人[10]使用对支持向量机对文本进行分类，评估指标的值优于 McIlroy 等人提出的 Multi-label 方法[6]。

上述研究采用主题挖掘或者分类方法，对用户评论进行分析。在这些工作的基础上，本文先对用户评论进行分类，然后对每个类别的评论进行主题挖掘，产生给出主题的代表词以及代表性句子。这样既能通过分类得到包含用户意图的信息，又能通过主题挖掘得到评论里用户重点关心的问题，使开发者能够快速、方便地理解用户反馈。

2 基于支持向量机和主题模型的评论分析方法 RASL

基于以上分析，本文提出一种基于支持向量机和主题模型的评论分析方法 RASL（Review Analysis method based on SVM and LDA）。方法架构如图 2 所示。

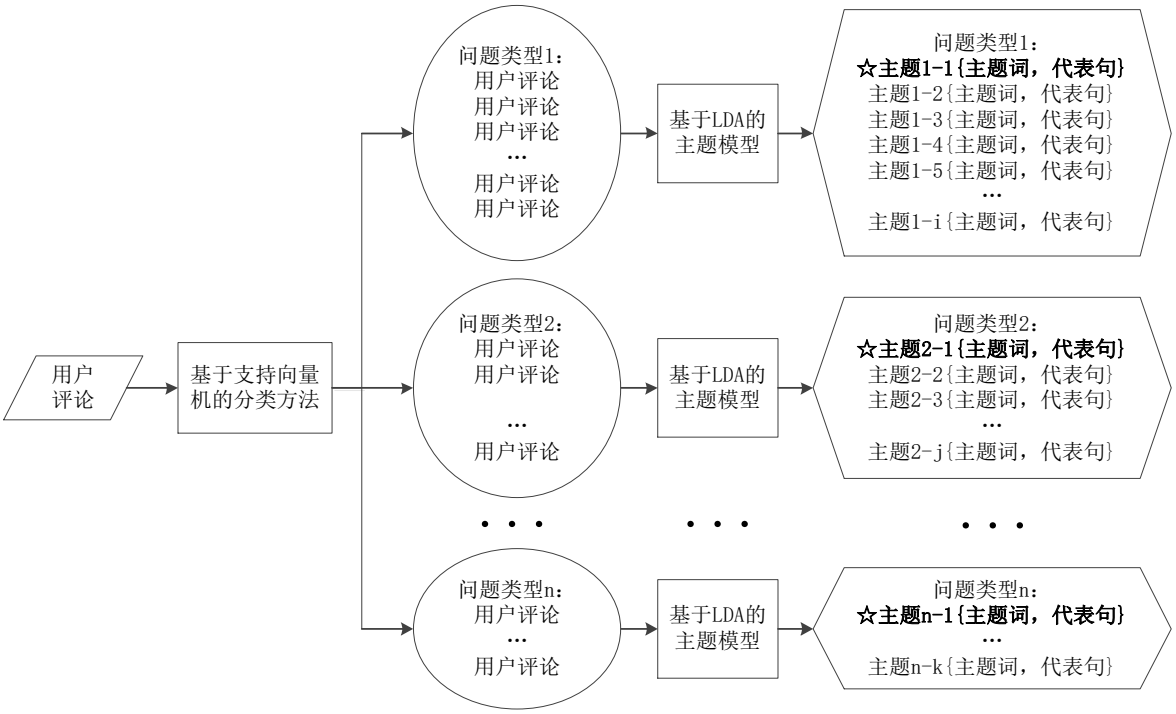

Fig.2 The overall workflow of the RASL method

图 2 RASL 方法的整体工作流程

方法分为两个阶段：分类阶段和主题分析阶段。首先，根据 2.1 节确定的评论类型，本文通过支持向量机的方法[10]对评论进行分类，可以得到包含用户意图的信息。然后，本文将分类好的评论数据分别进行 LDA 主题分析[9]，并给出代表句，从而得到评论里用户重点关心的问题。结合现有方法的优势，使得使开发者能够快速、方便地理解用户反馈。可以注意到每个问题类型的圆圈大小不一，这代表着每个问题类型下的主题个数由该问题类型的评论比例确定。

下面首先讨论用户评论的分类类型，然后对分类方法进行描述，最后将分类的结果作为输入，得到主题词和代表句。

2.1 用户评论的分类类型

为了对评论进行分类，需要确定用户评论的类型。依照 Seaman 等人的提出的方法[11][12]，迭代地对抽取的用户评论进行人工标注，分析评论中包含的问题种类。分析过程如下：首先，选择由 McIlroy 等人定义的问题类型集[6]作为起始集。对于每一条评论，手工检查并标注评论指出的问题类型。如果评论中的问题不包括在问题类型集中，则设定一个新的问题类型并将其添加到问题类型集中，然后基于新的问题类型集重新启动标注过程。这个过程是由三人并行完成，在三人均完成此过程后，比较三人标注的结果。

采用组内相关系数（Intra-class Correlation Coefficient，ICC）对标注结果的可靠性进行度量。ICC 是一个推断统计量，它描述了同一组中元素的相似程度[13]，可以用于评估不同观测者进行相同的定量测量时的一致性或可重复性。如果 ICC 小于 0.4，则表示相似性较差；如果 ICC 在 0.40 和 0.59 之间，则表示相似性一般；如果 ICC 在 0.60 和 0.74 之间，则表示相似性较好；如果 ICC 在 0.75 和 1.00 之间，则表示相似性很好。我们对标注结果之间的 ICC 进行了计算，以衡量人工标注的可靠性。对于每一个问题类型，ICC 都是较好或很好，并且发现，独立标注的结果差异没有特别大。然后进行讨论，并消除差异，所有问题类型的 ICC 均为 1，也即没有区别。最终得到了 17 种问题类型，问题的类型与描述如表 2 所示。详细过程见我们前期工作[10]。

Table 1 Classification of comments

表 1 评论类型

问题类型	描述
额外开销	抱怨需要额外的花费来享受完整的体验
兼容问题	应用在某个设备或操作系统上有问题
内容抱怨	内容没有吸引力或缺乏某些内容
崩溃	应用出现崩溃
移除特性	应用的某些特性非常糟糕
增加特性	希望应用增加某些特性
功能问题	应用的功能出现异常或失败
内容问题	在安装应用时失败
网络连接	应用出现网络连接方面的问题，如延迟
隐私与道德	应用侵犯用户隐私或者不道德
财产安全	应用威胁到用户的财产安全
资源占用	应用占用了太多电量、内存等
响应时间	应用响应缓慢
流量浪费	应用使用了超出用户预期的流量
更新问题	用户抱怨更新带来了新的问题
界面交互	对界面设计、交互、视觉方面的抱怨
其他	无用的，没有指出问题的评论

2.2 用户评论分类方法

为了能自动标记出新的用户评论属于的问题种类，本文采用机器学习的方法来进行用户评论分类。此分类方法包含特征提取部分和模型构建部分。特征提取部分的目标是提取评论文本的特征，使其转换为分类模型可用的形式。然后采用支持向量机，构建用户评论的多标签分类模型。同时，为了减轻不平衡数据的影响，使用了代价敏感学习的方法。

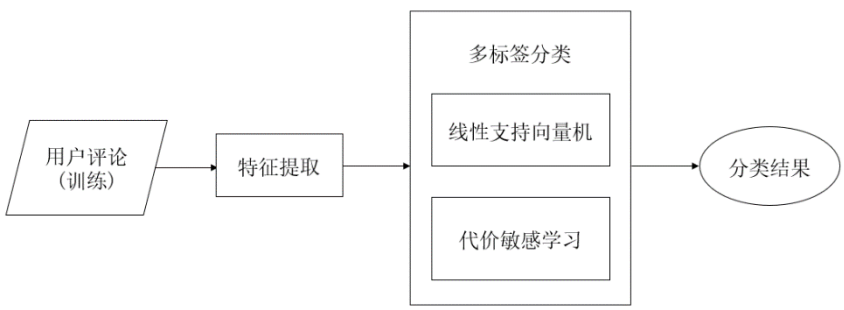

Fig.3 User review classification framework

图 3 用户评论分类框架

在特征提取阶段，目标是提取评论文本的特征。由于文本是非结构化的数据，因此必须首先将其转换为计算机可解析的形式。向量空间模型（VSM）是一种适用于大规模文献的文本表示模型[14]。在该模型中，文本空间被认为是由一组正交特征向量组成的向量空间。矢量的每个维对应于文本中的一个特征，每个维度本身表示文本中对应的特征的权重。使用向量空间模型来描述文本数据需要确定文本的特征与权重。对于英文文本来说，一个词就是一个特征。而中文文本首先需要进行分词。Jieba 分词是一个 Python 分词工具，本文使用它来进行分词，并删除数字和非汉字，但是停止词需要被保留下来，因为其中一些可以帮助确定问题类型，例如“不要”。参考现有工作[6]，过滤掉出现不到三次的词语，去除拼写错误或不重要的词语，降低分类的复杂性。其次，对于特征的权重，tf-idf 算法[15]是计算权重的常用方法。它的主要思想是，如果一个单词或短语多次出现在某一文档中，并且在其他文档中很少见，则该词或短语被认为具有很好的分类能力，比如：“安装”一词会在内容问题这个分类下出现，但它很少在资源占用等其它分类下出现，因此我们可以把“安装”一词作为分类依据之一。为了构建特征向量，本文使用 String To Word Vector filter，这是 WEKA 对 tf-idf 算法的一个实现[16]。

模型构建阶段的目标是为用户评论构建一个多标签的分类模型，对于输入的评论，模型可以输出评论所属的问题类型。由于一条用户评论可能包含多个问题类型，因此需要解决的问题实际上是一个多标签分类问题。Binary Relevance（BR）是解决多标签分类问题的代表性算法之一，它将多标签分类问题转化为多个二分类问题。模型选择使用 BR，因为它是线性复杂度的，较为简单[17]。这意味着需要构建多个二分类器，并且要对分类效果进行整体评估。本文选择支持向量机作为二分类器，支持向量机将数据视为 p 维向量，如果支持向量机用（p-1）维平面分离这些点，则它被称为是线性的。为了处理有的原始问题在有限维中不能线性分离的情况，支持向量机使用核函数将原始有限维空间映射成高维空间，例如径向基函数核。在样本数量少且特征数量非常大的情况下，非线性分类通常不准确，可能错误地划分特征空间，导致比线性模型更差的结果。因此，本文选择使用线性支持向量机，在具体的算法中采用 WEKA 的支持向量机实现，即 SMO 分类器[18][19]。将 PolyKernel 参数设置为 1，使其成为线性支持向量机。

但对于一些问题类型，负样本的数量远大于正样本。这些不平衡的数据可能导致分类器更倾向于将新样本预测为负样本，为了减轻不平衡数据的影响采用代价敏感学习的方法[20]来处理这个问题。代价敏感学习方法的核心是代价矩阵。代价矩阵定义如表 2 所示。

Table 2 Cost matrix

表 2 代价矩阵

真实 / 预测	0	1
0	C_{00}	C_{01}
1	C_{10}	C_{11}

其中，C_{ij} 是把 j 类分类到 i 类的成本。显然，$C_{00} = C_{11} = 0$；而 C_{01}、C_{10} 是两种不同的错分代价。由于数据是不平衡的，可以根据不同的错分代价来给数据重新赋予权重。当将正样本预测为负样本的代价较高时，就增加正样本的权重。具体到算法实现中，本文使用一个元分类器来使基类分类器成为代价敏感的，这个元分类器支持通过抽样来增加样本的权重，而基类分类器也即前面提到的支持向量机分类器。元分类器在 WEKA 中的实现即 CostSensitiveClassifier[18]。另外，还需要为每个问题类型指定代价矩阵。为了确定代价矩阵的具体值，对于每个问题类型，通过遍历来得到一个使分类结果最佳的代价矩阵值，然后采用这个值。详细过程见我们前期工作[10]。

2.3 主题词与代表句的生成

本节针对分类结果，进行统计抽取主题，通过所得到的主题，进一步生成主题词和代表句。

在使用基于支持向量机的分类模型进行分类，得到分类结果之后，需要确定将提取的主题总数（例如期望

每 X 条评论提取 Y 个主题，则根据用户评论数量计算提取的主题总数)。由于“其他”类型包含的是无用的评论，因此不对该类型进行主题提取。其余问题类型则根据分类的结果中各自所占比例（除去“其他”类型）计算出每个问题类型对应的主题数目，后续再对各问题类型下的评论进行主题分析。在这里举例说明多分类下主题数的确定，比如主题总数为 M，那么每个分类的主题则是分类下的评论数量占总数量的比重，具体来说，如果“内容问题”比重为 Ratio，那么分类为内容问题的主题数为 M*Ratio。

在机器学习和自然语言处理中，主题模型是用于发现文档集合中抽象“主题”的一种统计模型，是一种经常用于在文本主体中发现隐藏的语义结构的文本挖掘工具。每条评论都是与某些主题相关的，因而特定的词语也会出现在不同主题的评论中。本文采用了 LDA (Latent Dirichlet Allocation)的模型[9]来进行主题词和代表词的生成。这个模型是一种典型的词袋模型，即一个评论是由一组词构成，不去考虑词语的顺序，因而简化了语义关联问题的复杂性。LDA 模型包含主题的生成、根据阈值挑选关键词、代表句的生成三个部分。图 4 为 LDA 生成过程[9]。LDA 方法将评论的主题以概率分布的形式给出，通过分析评论抽取出它们的主题分布，然后再以一定概率迭代的选取主题下的某个单词作为主题词，最后根据概率选择代表句。

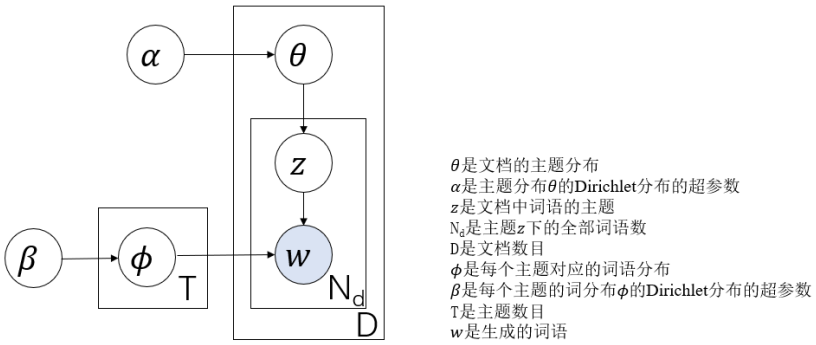

Fig.4 LDA generation process[9]

图 4 LDA 生成过程[9]

在主题生成阶段，本文将通过基于支持向量机的模型得到的 16 个(除去“其他”分类)分类评论作为输入，目标输出是指定数目的主题。由于得到了 16 个分类的结果，下面我们用“内容问题”这个分类举例说明，基本思想是 LDA 中存在主题词库，通过分析“内容问题”这个分类的所有评论，LDA 通过词库自动分析得到“内容问题”对应的主题，由于开始的时候，我们设定主题数为 T，因此 LDA 方法将选择最相关的前 T 个主题作为“内容问题”全部评论的主题。对应到图 4，“内容问题”所有评论与 T 个主题的一个多项分布相对应，将该多项分布记为 θ；α 是主题分布 θ 的先验分布 Dirichlet 分布的超参数，在这里本文选最相关的 T 个主题作为主题的生成部分的结果。

在根据阈值挑选关键词阶段，输入是主题生成阶段得到的“内容问题”分类下的 T 个主题，目标输出是每个主题一下指定数目的主题词。基本思想是由于一个词汇在“内容问题”的其中一个主题中都存在一个概率值，因此在主题词生成时去选择对应主题下概率值最高的 M 个词汇即可，M 取值可由需要确定。对应到图 4，每个主题与评论中的 N_d 单词的一个多项分布相对应，将这个多项分布记为 ϕ。β 是每个主题的词分布 ϕ 的先验分布 Dirichlet 分布的超参数。依据“内容问题”文档所对应的主题分布 θ 抽取一个主题 z，主题 z 所对应的多项分布 ϕ 中抽取一个单词 w，将这个过程重复 N_d（主题 z 下的全部词数）次，就生成了其中主题 z 下的主题词。LDA 通过变分 EM 算法、Gibbs 抽样法等方法，迭代地学习这两个参数，使其最终收敛于某一结果。

在代表句生成阶段，输入是根据阈值挑选关键词阶段得到的主题词，目标输出是每个主题一下的代表句。基本思想是假设 d 向量为 $(d_1,d_2,...,d_n)$，每个 d_i 代表一条用户评论被分配到每个主题的概率，假设主题数为 T，则第一条评论的向量 d_1 为 $(d_{1\_1},d_{1\_2},...,d_{1\_20})$，第二条评论的向量 d_2 为 $(d_{2_1},d_{2_2},...,d_{2_20})$等。对于目标主题 1，如果某一评论在 d_{i_1} 位置的概率值为所有评论中的最大值，则选择该评论作为目标主题的代表句。

最终输出结果，本文用分类后得到类别包含评论的数量从多到少进行排序，然后进行主题分类，得到每个主题下评论的数量，将主题按照评论数量从多到少排序，将每个分类下评论数量最多的主题进行标注（用“☆”并且加粗进行标注），从而方便阅读者对提出较多的问题进行重点关注。

3 实验验证

本节对本文提出的基于支持向量机和主题模型的评论分析方法 RASL 进行评估验证，本文将从定性分析和定量分析两个方面进行实验，以检验 RASL 方法的有效性。

3.1 实验对象

本文的 RASL 方法基于支持向量机分类算法采用 LDA 主题模型提取主题词和代表句。本文使用 Jieba 分词工具，把中文分词后输入经典方法 ASUM[7]。与 RASL 方法不同，ASUM 方法[7]是一种结合情感分析的主题模型。此方法将句子看作文档，句子中每个词都是隐含主题的分布，然后进行主题挖掘。在此基础上，融合主题特征和情感信息来分析用户对这些主题的偏好，并以<主题词，代表句>序对作为输出。本节我们将本文所提出的方法 RASL 与 ASUM 方法进行对比分析。

360 手机助手是国内市场份额较大的应用平台，提供移动应用程序卸载、安装、升级和评价等一系列服务。本文分别从 360 手机助手中随机抽取一个评分高的应用(评分 9 以上)和一个评分低的应用(评分 6 以下)，这两个应用分别是今日头条和 360 云盘，将它们的全部中差评收集起来。360 云盘的中差评共计 3950 条，今日头条的中差评共计 1191 条。将这 5141 条数据作为原始用户评论信息数据，将通过文本预处理得到的结果分别通过 ASUM 方法和 RASL 方法处理，得到实验所需数据。根据统计计算在我们所爬取的 5141 条数据中只存在 0.027%的评论由连续的几个段落组成，而 99.973%的评论是由单独一个段落组成，因此本文没有考虑评论分段问题，将分成多段的评论作为一个单独的评论进行处理。

3.2 研究问题

本文对两个应用的 5141 条评论进行了分析，除了本文的方法之外，也采用 ASUM 方法进行主题提取与代表句提取，和原始评论一起作为比较对象。我们希望通过调查能回答以下研究问题：

RQ1：ASUM 方法和本文方法 RASL 的困惑度如何？

RQ2：和原始评论相比，ASUM 方法和本文方法 RASL 是否包含完整的信息？

RQ3：和原始评论相比，ASUM 方法和本文方法 RASL 是否包含冗余的信息？

RQ4：ASUM 方法和本文方法 RASL 的阅读理解性如何？

RQ5：ASUM 方法和本文方法 RASL 的评论阅读时间如何？

3.3 实验方法

3.3.1 困惑度分析

对于主题模型的评估，Blei 等人[9]在论文中提出了用困惑度（Perplexity 值）作为评判标准。困惑度度量概率分布或概率模型的预测结果与样本的契合程度，在这里是指：对于一个文档 d，所训练出来的模型对于文档 d 属于哪个主题的确定程度。困惑度越小，说明模型效果越好。困惑度的计算公式为：

$$\text{Perplexity}(\text{D}) = \exp\left\{-\frac{\sum_{d=1}^{M} log\ p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{1}$$

M 为文档集合中的文档数目，N_d 为第 d 篇文档中单词的个数，$p(w_d)$为第 d 篇文档的概率（probability），也即这篇文档中每个单词概率的乘积。而对任意一个单词 w，单词概率 p(w)=∑ p(z|d)* p(w|z)，z 代表主题，p(z|d)为各主题下该词所在文档的概率，p(w|z)为该词在各主题下的概率。

对于 ASUM 方法，根据每个应用的评论数量计算出提取的主题数量，然后用 ASUM 方法提取出主题词与代表句，作为这个实验的输入；对于本文所提出的 RASL 方法，首先将每个应用的评论使用基于支持向量机

的分类方法进行分类(问题类型与描述如表 1 所示)，根据每个应用的评论数量计算出提取的主题总数，再根据分类的结果中各问题类型（除去“其他”类型）所占比例计算出每个问题类型对应的主题数目，然后对各问题类型下的评论按照主题数目进行主题提取与代表句提取，将问题类型和主题词、代表句共同作为这个实验的输入。计算困惑度并进行比较。

3.3.2 问卷调查

为研究 RQ2，RQ3，RQ4，RQ5，实验选择了今日头条和 360 云盘作为实验数据。实验邀请了 6 位北京航空航天大学软件工程专业研究生,都具有至少四年的编程经验，并且经常使用手机移动应用，有过对手机移动应用软件打分和作评论的经历。由他们完成调查问卷，以便回答 RQ2-RQ5。未来我们尝试联系手机应用的开发人员，请他们评价不同方法的结果。

对于每个应用都提供给受试者三份文件：1) 原始的中差评集合；2) ASUM 方法提供的分析结果；3) RASL 方法提供的分析结果，结果样例见图 5，图 6 和图 7。链接[1]展示一个实际的例子，包括原始评论、ASUM 生成的结果和 RASL 生成的结果。为了对比出结果的“完整性”（完整性指对比原始评论，是否有内容上的缺失）并且不使原始评论影响方法结果的可理解性，本文要求受试者先对 ASUM 方法和 RASL 方法进行阅读，最后再阅读原始评论。两种方法可能存在相互影响阅读结果的问题。为了减少这种问题带来的不确定性，本文将受试者随机分为两组，一组先阅读 ASUM 方法再阅读 RASL 方法，另一组则先阅读 RASL 方法再阅读 ASUM 方法。

1 原始评论
2 原始评论
3 原始评论
4 原始评论
5 原始评论
…
n 原始评论

Fig.5 Original review example

图 5 原始评论示例

主题1：
主题词
代表句
主题2：
主题词
代表句
…
主题m：
主题词
代表句

Fig.6 ASUM Method results example

图 6 ASUM方法结果示例

分类类型1

☆主题1-1：
主题词
代表句
主题1-2：
主题词
代表句
…
主题1-i：
主题词
代表句
⋮

分类类型16

☆主题16-1：
主题词
代表句
…

Fig.7 RASL Method results example

图 7 RASL方法结果示例

[1] https://github.com/ChenQifromBeihang/Essay.git

表 3 总结了调查问卷中设计的问题。为了不提供给受试者更多信息，问卷中以方法 A 指代 ASUM 方法生成的分析结果，以方法 B 指代 RASL 方法生成的分析结果。首先，问卷对两种主题分析方法的结果的表现力进行了比较：为了回答主题分析方法产生的结果是否包含完整的信息，即：对比原始评论，是否有内容上的缺失，设计了 Q1-1、Q1-2 两项问题；为了回答主题分析方法产生的结果是否包含冗余的信息，即：出现重复性内容，设计了 Q2-1、Q2-2 两项问题；为了回答主题分析方法产生的结果是否具有可阅理解性，设计了 Q3-1、Q3-2 两项问题。然后，问卷对原始评论和主题分析方法之间进行了比较：为了回答主题分析方法如何影响分析用户评论花费的时间的问题，设计了 Q4-1、Q4-2 两项问题。

Table 3 questionnaire

表 3 调查问卷问题

Q1-1	和原始评论相比，方法 A 包含的信息是否完整？请打分(10 分制：1 分为缺失极多信息，10 分为没有缺失任何信息)
Q1-2	和原始评论相比，方法 B 包含的信息是否完整？请打分(10 分制：1 分为缺失极多信息，10 分为没有缺失任何信息)
Q2-1	和原始评论相比，方法 A 是否包含冗余的信息？请打分(10 分制：1 分为有许多冗余的信息，10 分为没有冗余的信息)
Q2-2	和原始评论相比，方法 B 是否包含冗余的信息？请打分(10 分制：1 分为有许多冗余的信息，10 分为没有冗余的信息)
Q3-1	方法 A 是否易于阅读理解？请打分(10 分制：1 分为难以阅读理解，10 分为非常易于阅读理解)
Q3-2	方法 B 是否易于阅读理解？请打分(10 分制：1 分为难以阅读理解，10 分为非常易于阅读理解)
Q4-1	请填写采用方法 A，阅读评论所花费的时间(分钟)
Q4-2	请填写采用方法 B，阅读评论所花费的时间(分钟)

3.4 实验结果

RQ1：ASUM 方法和本文方法 RASL 的困惑度如何？

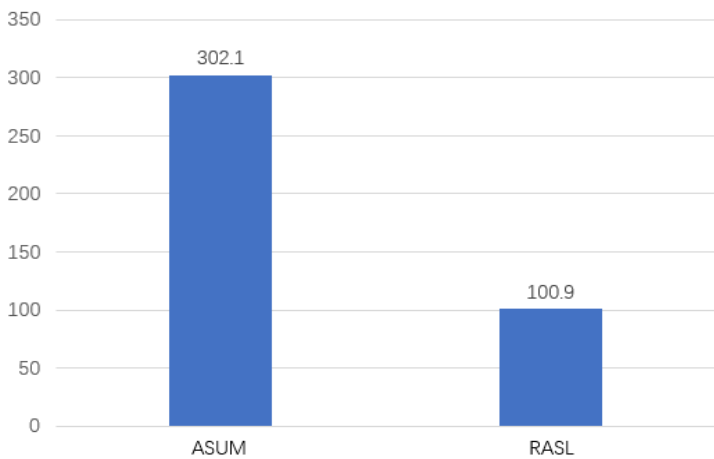

Fig.8 Comparison of ASUM and RASL perplexity

图 8 ASUM、RASL 的困惑度对比

用本文所收集的用户评论作为实验数据，运用 ASUM 方法和 RASL 模型得到结果，对结果计算困惑度并进行比较，如图 8 所示，ASUM 的困惑度是 302.1，RASL 的困惑度是 100.9。本文所提出的方法得到的困惑度小于 ASUM 方法所得到的困惑度，本文提出的方法优于 ASUM 方法。

RQ1：本文方法得到的困惑度小于 ASUM 方法所得到的困惑度

RQ2：和原始评论相比，ASUM 方法和本文方法 RASL 是否包含完整的信息？

为了回答主题分析方法产生的结果是否具有充分性，将 Q3-1、Q3-2 的结果汇总于表 4，1 分为缺失极多信息，10 分为没有缺失任何信息。360 云盘和今日头条的结果都显示，对于六位受试者来说，完整性均为 RASL 优于 ASUM。

Table 4 Integrity of ASUM and RASL

表 4 ASUM、RASL 的完整性

受试者编号	360 云盘		今日头条	
	ASUM 方法	RASL 方法	ASUM 方法	RASL 方法
1	5	8	2	8
2	1	8	2	8
3	7	8	8	9
4	5	8	4	7
5	5	7	3	6
6	6	8	5	7

而后本文进行 Mann-Whitney U 检验检测这种差异的显著性。Mann-Whitney U 检验是由 H.B.Mann 和 D.R.Whitney 于 1947 年提出的[21]，是一种非参数秩和假设检验，这个检验是对独立样本进行的一种不要求正态分布的 t-test 检验方法。主要是对来自除了总体均值以外完全相同的两个总体，检验其是否具有显著差异。最终得到的结果显示，RASL 的完整性在 0.05 的显著性水平下明显优于 ASUM。

RQ2：在 0.05 的显著性水平下，RASL 方法在完整性显著优于 ASUM 方法。

RQ3：和原始评论相比，ASUM 方法和本文方法 RASL 是否包含冗余的信息？

为了回答主题分析方法产生的结果是否具有简明性，将 Q4-1、Q4-2 的结果汇总于表 5，1 分为有许多冗余的信息，10 分为没有冗余的信息。360 云盘的结果显示，对于受试者 1、受试者 2、受试者 4 和受试者 5 来说， RASL 包含的冗余信息相比 ASUM 较少。受试者 3 认为 ASUM 和 RASL 包含多于信息的数量差不多，而受试者 6 评价 ASUM 包含的冗余信息相比 RASL 较少。采访受试者 6，其认为 RASL 存在个别主题与问题类型不匹配的问题，因此评价略低于 ASUM。今日头条的结果显示，对于其中五位受试者来说，RASL 包含的冗余信息相比 ASUM 较少，对于受试者 3 来说，ASUM 和 RASL 得到结果包含冗余信息量相同。

Table 5 Redundancy of ASUM and RASL

表 5 ASUM、RASL 的冗余性

受试者编号	360 云盘		今日头条	
	ASUM 方法	RASL 方法	ASUM 方法	RASL 方法
1	4	9	4	7
2	2	8	1	9
3	9	9	9	9
4	5	9	4	8
5	3	5	4	5
6	7	6	5	6

为了检测这种差异是否具有显著性，本文对两个方法的简明性进行了 Mann-Whitney U 检验，在 0.05 的

显著性水平下，RASL 包含的冗余信息明显少于 ASUM。

RQ3：在 0.05 的显著性水平下，RASL 包含的冗余信息显著少于 ASUM。

RQ4：ASUM 方法和本文方法 RASL 的的阅读理解性如何？

为了回答主题分析方法产生的结果是否具有可阅读性，将 Q5-1、Q5-2 的结果汇总于表 6，1 分为难以阅读理解，10 分为非常易于阅读理解。360 云盘和今日头条的结果都显示，对于六位受试者来说，可阅读性均为 RASL 优于 ASUM。

Table 6 Understandability of ASUM and RASL

表 6 ASUM、RASL 的可理解性

受试者编号	360 云盘		今日头条	
	ASUM 方法	RASL 方法	ASUM 方法	RASL 方法
1	6	9	5	9
2	3	8	2	9
3	3	8	2	8
4	6	9	6	9
5	6	8	5	7
6	7	8	6	7

而后本文进行 Mann-Whitney U 检验检测这种差异的显著性，最终得到的结果显示，RASL 的可阅读性在 0.05 的显著性水平下明显优于 ASUM。

RQ4：在 0.05 的显著性水平下，RASL 方法在可理解性方面显著优于 ASUM 方法。

RQ5：ASUM 方法和本文方法 RASL 的评论阅读时间如何？

为了回答主题分析方法如何影响分析用户评论所花费时间的问题，将 Q2-1、Q2-2 的结果汇总于表 7。360 云盘的结果显示，受试者 3 阅读 ASUM 和 RASL 所花费的时间相等，受试者 2、受试者 5、受试者 6 阅读 RASL 所花费的时间略高于 ASUM，受试者 1 和受试者 3 阅读 RASL 所花费的时间略低于 ASUM。今日头条的结果显示，受试者 1、受试者 4、受试者 5 阅读 ASUM 和 RASL 所花费的时间相等，受试者 2 阅读 ASUM 所花费的时间略高于 RASL，受试者 3 和受试者 6 阅读 RASL 所花费的时间略低于 ASUM。

Table 7 Analysis time of ASUM and RASL (min)

表 7 ASUM、RASL 的分析时间(分钟)

受试者编号	360 云盘		今日头条	
	ASUM 方法	RASL 方法	ASUM 方法	RASL 方法
1	8	7	6	6
2	8	16	6	9
3	4	2	3.5	2
4	5	5	4	4
5	5	6	5	5
6	16	18	18	16

由于所得 ASUM 方法和 RASL 方法所得阅读时间无法直接看出是否具有差异性，本文进行 Mann-Whitney U 检验， 在 0.05 的显著性水平下，ASUM 方法和 RASL 方法之间的时间差异不具有统计的显著性，总的来说，RASL 和 ASUM 所用阅读时间无差异。

RQ5：RASL 方法和 ASUM 方法所用阅读时间无显著差异。

从实验结果分析中，我们可以得出结论：在困惑度方面，RASL 主题分析方法明显优于 ASUM 方法。RASL 主题分析方法和 ASUM 方法相比，可理解性更佳，包含更完整的原始评论信息，冗余信息也更少。

3.5 定性分析

回收问卷后，采访了受试者对于两种主题分析方法的主观感受。受试者们表示，ASUM 的主题阅读起来比较费力，存在较多的无意义主题。本文方法 RASL 由于有问题类型作为基础，相当于具有两层结构，有组织性，阅读起来较为清晰明了，一些用户评论可以帮助开发者发现问题。例如图 9 所示，本文方法 RASL 可以发现“安装”分类下，最常见的问题是更新的版本在部分机器上无法进行安装。

```
--------------------
安装问题
--------------------
☆0.069*"安装" + 0.024*"的" + 0.018*"无法" + 0.018*"了" + 0.018
*"不了"
不能安装这个版本
```

Fig.9 Example

图 9 示例图

4 有效性威胁

4.1 内部有效性威胁

在对主题分析方法 RASL 进行问卷调查时，比较了两种可能的方法，一是分两组，单独评价 ASUM 方法和 RASL 方法，但是在人员不足够多的情况下，难以消除不同人不同判断标准的问题；因此，本文将受试者随机分为两组，每位受试者对两种方法进行阅读，虽有可以在同一标准下给两种方法打分，但是会带来阅读顺序的问题，因此本文采用改变顺序减轻了两种方法间的相互影响。另外，ASUM 和 RASL 产生 200 个主题的结果。主题数量较多，难以请受试者对每个主题进行详细打分。未来我们尝试联系更多的受试者参加问卷调查，减轻阅读顺序、主题差异造成的偏差。最后，本文的研究没有加入好评，这可能会导致一些问题的遗漏，因为在好评中也可能有用户对于应用的一些抱怨意见。不过，即使好评中存在一些抱怨意见，也不会影响本文方法的可用性。今后，我们将进行更多的实验来完善对所有评论的研究。

4.2 外部有效性威胁

评估实验针对两个应用进行，尽管本文选择的两个应用软件分别随机抽取了评分最高的应用软件和评分最低的应用软件，但这两个应用仍然可能不能代表所有应用软件，在未来的工作中，我们计划爬取更多应用软件的评论，并与现有的结果进行比较。而且评估实验中的两个应用都来自于 360 手机助手，得到的结果是否可以推广到其他应用商店是未知的。在今后的工作中，我们将选取不同平台的应用进行实验，并将结果与 360 手机助手的结果进行比较，以巩固我们的发现。

5 结论与展望

移动应用的用户在线评论数量巨大、信息量丰富，是重要的反馈数据来源，通过收集用户使用软件后产生的反馈信息，挖掘其中的各类需求，对软件开发者而言有重要的价值，使开发者能快速、直观地理解用户反馈，本文提出了一种基于支持向量机和主题模型的评论分析方法 RASL。RASL 方法首先进行用户的评论分类，然后对每个类别下的评论进行主题挖掘。

本文方法基于支持向量机的分类模型的分类结果为基础，依照分类结果中每个问题类型所占比例确定每

个问题类型的主题数目。然后，选择了 LDA 模型进行主题分析，使用 LDA 模型对各问题类型下的评论进行主题提取与代表句提取。而后，设计实验对比 ASUM 方法对 RASL 方法进行评估。首先对两种方法不同主题数目下的困惑度进行计算，结果得到 RASL 方法困惑度明显减少。然后用调查问卷进行评估，实验数据是两个应用的全部中差评，邀请软件工程专业的受试者对原始评论、ASUM 主题分析方法和 RASL 主题分析方法的生成结果进行评估。实验结果证明，和 ASUM 相比，RASL 方法可理解性、完整性更佳，包含的冗余信息也更少。

在未来的工作中，我们将邀请足够多的受试者进行实验，并且邀请一些软件开发人员，将本文的方法应用于更多应用软件，从而判断本文提出方法所得结果是否能真正符合开发商需求。

Reference:

[1] 张林. 基于移动在线评论的用户情感及需求挖掘研究[D]. 北京:北京航空航天大学, 2015.

[2] Pagano D, Maalej W. User feedback in the appstore: An empirical study[C]. In: Requirements Engineering Conference (RE), 2013 21st IEEE International. IEEE, 2013: 125-134.

[3] Pagano D, Brügge B. User involvement in software evolution practice: a case study[C]. In: Proceedings of the 2013 internation al conference on Software engineering. IEEE Press, 2013: 953-962.

[4] Panichella S, Di Sorbo A, Guzman E, et al. How can i improve my App? classifying user reviews for software maintenance and evolution[C]. In: Software Maintenance and Evolution (ICSME), 2015 IEEE International Conference on. IEEE, 2015: 281-290.

[5] Maalej W, Nabil H. Bug report, feature request, or simply praise? on automatically classifying App reviews[C]. In: 2015 IEEE 23rd international requirements engineering conference (RE). IEEE, 2015: 116-125.

[6] Mcilroy S, Ali N, Khalid H, Hassan AE. Analyzing and automatically labelling the types of user issues that are raised in mobile App reviews[J]. Empirical Software Engineering, 2016, 21(3): 1067-1106.

[7] Galvis Carreño LV, Winbladh K. Analysis of user comments: an Approach for software requirements evolution[C]. In: Proceedings of the 2013 International Conference on Software Engineering. IEEE Press, 2013: 582-591.

[8] 姜巍, 张莉, 戴翼, 蒋竞, 王刚. 面向用户需求获取的在线评论有用性分析[J]. 计算机学报, 2013, 36(1):119-131.

[9] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[10]Zhang L, Huang XY, Jiang J, Hu YK. Cslabel: an approach for labelling mobile app reviews. 计算机科学技术学报(英文版) , 2017,32(6), 1076-1089.

[11]Seaman CB, Shull F, Regardie M, et al. Defect categorization: making use of a decade of widely varying historical data[C]. In Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement. ACM, 2008. 149–157.

[12]Seaman CB. Qualitative methods in empirical studies of software engineering[J]. IEEE Transactions on software engineering, 1999 (4): 557-572.

[13]Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability.[J]. Psychological bulletin, 1979, 86(2):420 -428.

[14]Salton G. The SMART retrieval system-experiments in automatic document processing[M]. USA: Prentice-Hall, 1971.

[15]Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

[16]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.

[17]Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data[M]. In Data mining and knowledge discovery handbook, Boston: Springer, 2009: 667–685.

[18]Witten IH, Frank E, Hall MA, et al. Data Mining: Practical machine learning tools and techniques[M]. Morgan Kaufmann, 2016.

[19]Platt J. Fast training of support vector machines using sequential minimal optimization - Advances in Kernel Methods[M]. Cambridge, MA, USA: MIT Press, 1998: 185-208.

[20]Elkan C. The foundations of cost-sensitive learning[C]. In International joint conference on artificial intelligence. Lawrence Erlbaum Associates Ltd, 2001, 17(1): 973-978.

[21]Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other[J]. The annals of mathematical statistics, 1947: 50-60.