摘要 - 软件风险管理是软件开发管理的重要组成部分。 由于潜在损失的严重程度，早期的风险识别和缓解变得至关重要。 包含数百个可能的风险提示的列表在学术文献和实践中都是可用的。鉴于大量的风险记录，扫描列出的风险，并确定相关的风险虽然全面，但是变得不切实际。 在这项工作中，基于软件项目特征和其他因素，开发了机器学习算法来产生风险提示。 这项工作还探讨了风险后分类标记的效用。 索引条款 - 软件风险，软件管理，机器学习，风险提示。

一，引言

本研究旨在建立一个基于机器学习的风险提示系统，该系统根据软件开发项目的规范，情景和分类标记输出最相关的风险提示。与“关键字”搜索相比，提示的数量减少了50％。软件风险管理需要从关键流程路径到宏观经济市场动态等各个方面的多维理解。有几个列表可以记录软件开发项目中可能产生的各种风险。这些清单通常包含几百个风险提示。编译这些列表的例子包括[1]和[2]。文献[3]提供了编制软件开发风险前10名的综合调查。查看风险的一个好的起点是考虑[4]中引入的10个软件项目风险分类法。在文献[5]中，更近期和普遍接受的风险“维度”是：团队，组织环境，需求，计划控制，用户和复杂性。这些风险维度中的每一个都有几个相关的风险提示 - 例如，CMU-SEI 1993列表包含基于13个风险分类法的200个问题[6]。虽然较细粒度的方法需要更少的提示，但对于粒度较小的列表所提供的价值肯定是妥协的。然而，高度细化的列表需要大量的时间进行分析，因此结束时实际使用很少。其他较不通用和更精确的风险分析方法包括基于情景的和基于规范的方法。这些方法被认为更耗时且更不科学。然而，强烈建议将风险分析方法结合到所有3个方面 - 场景，规范和基于分类的编译中[7]。其余的工作安排如下：第二部分介绍研究的背景和数据的整理。第三部分介绍了所采用的机器学习分类算法，接下来是第四部分，在风险分析中提出了后置分类标签的应用。第五部分介绍了算法的结果和这项工作的局限性。最后，第六节总结了这项工作的未来范围。

风险提示列表是从几个清单（包括[3]中的清单提示以及文献[7]中的风险提示）中获得的，然后将这些提示划分为更高的特异性等级。正如前面所指出的，共有433个风险提示被编制成16个分类标准以下的高度细化的清单.40个项目管理高级学生志愿者对433个风险提示进行了排序，确定了400个Oracle成功完成的软件项目的风险。请注意，风险提示是根据16个分类标准独立分类的。这种方法极大地减少了志愿者风险识别的任务，因为某个分类法下的风险提示被认为与另一个分类法下的另一个风险提示无关。由于组织良好的Oracle产品目录[8]提供了网络爬虫程序能够在开发过程中为这400个软件产品提供规范和可能的场景。在这些规范中，有20个最具代表性的，经常发生的规范被选为可能规范的列表。

网络爬虫的输出也产生了更大偏差的产品开发情景。 在减少机器学习的可能场景列表的约束下，选择了25个场景。 有些方案虽然不太经常发生，但为了维护所有400个Oracle产品的代表性，选择了这些方案。 虽然这种方法不够精确，但局限性仅限于这个具体的研究。 增强的情景数据可用性将确保风险提示的质量，精确度和相关性得到提高。

III。 M ACHINE L EARNING A LGORITHM

开发基于场景，分类和规范数据的风险提示的机器学习算法不那么繁琐，但需要一套复杂的分布式机器学习工作。由于假定各分类标准的风险提示是独立的，因此每个分类标准的分析可以相互独立地进行处理。此外，也可以假设，对于所考虑的每个风险提示，包含一个风险提示并不包括包含即使两个提示属于同一个分类，也是另一个风险提示。这些简化显着地减少了这个问题，并且我们有几个多标签分类问题，这些问题通过使用[9]和[10]中描述的多个多标签人工神经网络（ANN）来解决。这个过程的基本细节是：

•每个分类器是一个隐藏层的神经网络，它输出范围从0到1的特定风险提示的相关概率。

•即使这个问题似乎受到维度的诅咒，相关的属性也会使用相关度量来进行筛选。

•平均而言，对于每个风险提示，每个神经网络输出一定程度的相关性，其输入为：5-10个场景和几乎相同的规格数量。

•对应于风险提示的每个神经网络独立于另一个运作。

这是很好的缩放 - 400个实例和20个属性。

IV。 P OST -C分类T更新

分类后标签是风险提示提供的重要实用工具。分类后标签提供了可操作的见解来系统地减轻软件开发项目风险。然而，在分类阶段之前有必要对风险进行标记 - 这不是由于学习设置而产生的优势，而是在适当的风险分类后获得的数据整理优势。这种分类后标签的实用性的一个例子是将风险标记为系统性和特殊性。在这项研究中，所有的433个风险提示都被归类为这两个类别之一。这允许高效的软件定价 - 通过将这些风险的货币风险指标包括在项目成本中来转移系统性风险。分类后标记的另一个特别有利的用途是优先化。风险可以被定义为从高优先级到低优先级风险，这可以为项目经理提供时间表。

五，成果和L模仿

训练的神经网络对40种产品进行测试，相应的手动风险提示由志愿者分类。 表II中提供了随机选择的5种产品的结果。 将结果与“关键字”相关搜索的性能进行比较，根据软件产品的规范和方案中的匹配关键字，这些搜索只返回风险提示。 我们发现，与基于关键词的搜索相比，返回的风险提示的数量至少减少了50％。特别是有以下几个限制：

•计算效率低下 - 需要使用几个人工神经网络。 解决方案可以是使用分层方法。

•假阴性构成真正的问题。 一个解决方案是将决策点设置在0.5以下，例如通知相关度超过0.4的提示。

VI。 进一步的讨论

今后的工作有相当大的空间，并且对这项研究进行详细阐述 - 进一步探讨的可能的要点是：

•机器学习方法 - 找到一个全面的数据集。

编译列表的粒度是否足够？

•时间复杂性 - 更快，更相关的风险提示。