一种用于文本分类问题特征选择的新型多元过滤方法

Mahdieh Labani VDAIL 2018-11-29

摘要：

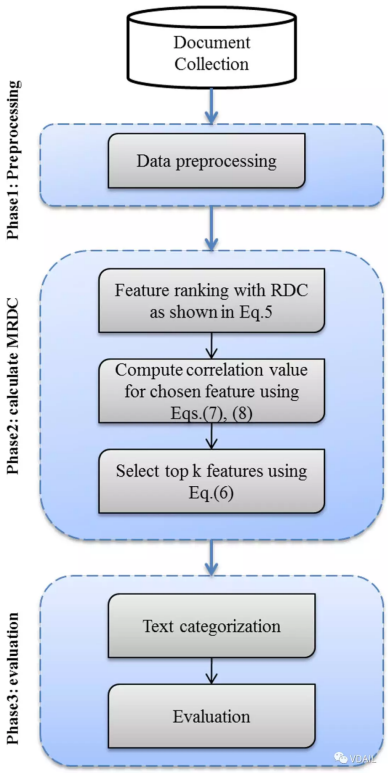
特征项的选择和特征权值的计算是文本分类过程中两个至关重要的环节，对文本分类的结果起关键性作用。随着数字格式文档数量的增加，文本自动分类已成为模式识别问题中的一项重要任务。为了简化分类任务，引入了特征选择方法以减少特征空间的维数，从而提高分类性能。本文提出了一种新的特征选择过滤方法，即多变量相对判别准则（MRDC），该方法使用最小冗余和最大相关性来减少冗余特征，不仅选择具有最大相关性的特征，同时考虑特征词之间的冗余度。实验结果表明，在大多数情况下，MRDC比其他情况产生更好的分类性能。

多变量相对判别准则步骤：

1、预处理。去除停用词、词干化；

2、提出用多变量特征排序评估文本分类问题的特征。第一步，使用RDC来计算每个词的相关性；第二步，Pearson相关性计算特征之间的冗余值。

3、使用有监督学习算法评估所选特征子集。



总结：

本文为文本分类任务引入一种新的多变量特征选择过滤方法MRDC。该方法的目的是对RDC方法进行多元扩展，去除冗余特征和不相关特征，并将该方法应用到三个数据集，结果表明，该方法在精度，召回率和精准度方面都优于传统的特征选择方法。

分享人：陈柳

阅读原文