计算机工程与应用 2009 45 (16): 31-35 ISSN: 1002-8331 CN: 11-2127/TP

博士论坛

参数嵌入算法在文本分类可视化中的应用

张 莹, 王耀南, 万 琴

湖南大学 电气与信息工程学院，长沙 410082

收稿日期 2009-2-10 修回日期 2009-3-20 网络版发布日期 2009-5-27 接受日期

摘要 如何对文本分类的结果进行可视化研究一直是模式识别中研究的重点。在假设文本类别在低维嵌入空间服从高斯分布的前提下，通过朴素贝叶斯分类算法得到数据类别属性的后验概率矩阵，然后运用参数嵌入算法在低维空间可视化文本分类结果。参数嵌入算法是使嵌入空间数据的类后验概率与高维空间的条件概率Kullback Leibler散度和最小化的算法，属于同一类的数据在低维空间中分布较为集中，性质相似的数据之间的距离较近，而不同性质的数据之间距离则较大。其优点在于计算复杂度是数据的类别和相应个数的乘积，非常适合于数据量大，类别数较少的数据分类可视化。20新闻组数据集和微型新闻组数据集的实验结果证明了该算法的有效性。

关键词 朴素贝叶斯分类 参数嵌入 文本分类 后验概率 分类可视化

分类号

Application of parametric embedding algorithm to text classifier visualization

ZHANG Ying,WANG Yao-nan,WAN Qin

College of Electrical and Information Engineering，Hunan University，Changsha 410082，China

Abstract

How to visualize the text classifier result is one of the focus field in pattern recognition.On the assumption that each class can be represented by a Gaussian distribution in the embedding space，through Naive Bayes classification algorithms posterior probability for data over classes was got，Parametric Embedding（PE） algorithm was applied into the visualization of classification result in low-dimensional.PE algorithm tries to preserve the structure in an embedding space by minimizing a sum of Kullback-Leibler divergences in high-dimensional space.Data that are located at the center of cluster are typical data for the class，and data that are located between clusters have multiple topics，different data are located in the cluster of different classes.The outstanding advantage is that computing complexity is just the type of data and the corresponding number of the product，is well suited to large volume of data，fewer types of classified data visualization.Experimental result on 20 Newsgroups data sets and MiniNewsgroups data sets proves the effectiveness of the method.

Key words Naive Bayes classifier parametric embedding text classification posterior probability classification visualization

DOI: 10.3778/j.issn.1002-8331.2009.16.008

通讯作者 张 莹 gdutzy@hotmail.com