Journal of Computer Applications
计算机应用,2012,32(2):403-406,410
ISSN 1001-9081
CODEN JYIIDU
2012-02-01
http://www.joca.cn

文章编号:1001-9081(2012)02-0403-04
doi:10.3724/SP.J.1087.2012.00403

基于LDA主题模型的标签传递算法

刘培奇*,孙捷烩
(西安建筑科技大学 信息与控制工程学院,西安 710055)
(*通信作者电子邮箱 peiqiliu@163.com)

摘 要:标签传递算法是一种半监督分类方法,由于该算法存在要求数据分类结果符合流行假设、数据维数较高时计算复杂度高等问题,在文本分类中效果较差。针对这些问题,经过对LDA主题模型和标签传递算法原理及复杂度的分析,将两者结合,提出一种基于LDA主题模型的标签传递算法LPLDA。该算法用LDA主题模型中的主题表示文本数据,一方面使用LDA主题模型表示文本保证分类结果符合流行假设,另一方面有效减少标签传递算法相似度计算时间。经过实验证明,该算法在标记数据少于待测样本时,分类效果优于传统的有监督分类方法。

关键词:LDA主题模型;标签传递算法;半监督学习;数据降维;流行假设
中图分类号:TP181;TP391.4 **文献标志码**:A

Label propagation algorithm based on LDA model

LIU Pei-qi*, SUN Jie-han
(*School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an Shaanxi* 710055, *China*)

Abstract: Label Propagation (LP) algorithm is one kind of semi-supervised learning methods. However, its performance in text classification is not good enough, because LP algorithm demands manifold assumption and it has high computational complexity in calculating the similarity of high dimension data. A new method was proposed to combine Latent Dirichlet Allocation (LDA) model with LP algorithm to solve the above problems after analyzing their principles and complexities. It represented documents with latent topics in LDA. On one hand, it reduces the dimension of matrixes; on the other hand, it can help LDA model lead to the classification results with manifold assumption. The experimental results show that the new method performs better than traditional supervised text classification methods in testing sets when labeled data is less than unlabeled data.

Key words: Latent Dirichlet Allocation (LDA) model; Label Propagation (LP) algorithm; semi-supervised learning; dimensional reduction; manifold assumption

0 引言

半监督分类方法可以利用大量无标签数据指导分类,在减少数据标注的同时提高分类效果[1-4]。标签传递(Label Propagation, LP)算法通过在已标记和未标记数据间根据相似度进行标签的传递进行半监督分类,当标签达到稳定时根据类别概率为未标记数据分配标签。将标签传递算法[5-8]用于文本分类,需要解决两个问题:1)文本数据的维数较高,标签传递算法在计算数据之间相似度时需要花费大量时间;2)标签传递算法要求分类结果符合流行假设。针对这两个问题,提出用LDA(Latent Dirichlet Aollocation)主题模型[9]的主题分布表示文本。LDA主题模型的低维空间表示可以帮助标签传递算法减少图中相似度的计算时间。同时,标签传递算法通过在近邻数据间传递类别标签,可以帮助LDA主题模型根据流行假设划分文本类别。

1 基本理论

标签传递算法构建一个基于标签平滑性假设的图,将有标签和无标签的数据定义为图上的节点,边定义为相似度,假设相似度较高的边连接的顶点具有相同的标签,通过图的扩散将顶点归到不同的类中。

1.1 标签传递算法

标签传递算法是一类基于流行假设并与随机游走结合的直推式学习方法,包括调和高斯场[5]、局部和全局一致性[6]、线性邻居传播[7]等。本文使用Zhu等[8]提出的一种简单的标签传递算法对文本分类,该方法采用径向基函数(Radial Basis Function, RBF)核函数作为顶点间的相似度构建图,对图中的所有顶点都进行连接。图中所有顶点根据概率转移矩阵进行标签传递,过程与随机游走相似,当标签达到稳定状态后,未标记顶点选择标签矩阵中概率最大的类别标签。式(1)定义了概率转移矩阵[8]$\boldsymbol{T}$中的元素T_{ij}:

$$T_{ij} = P(j \rightarrow i) = w_{ij} / \left(\sum_{k=1}^{l+u} w_{kj} \right) \tag{1}$$

其中:w_{ij}为顶点i和顶点j的相似度,$\sum_{k=1}^{l+u} w_{kj}$为所有与顶点j相连的顶点相似度之和,T_{ij}表示从顶点j到顶点i的转移概率。

标签传递算法的步骤如下:

1)设置$t=0$,初始标签矩阵$\boldsymbol{Y}_0$;

2)$\boldsymbol{Y}_{t+1} = \boldsymbol{T}\boldsymbol{Y}_t$,修改新的标签矩阵为$\boldsymbol{Y}_{t+1}$;

3)对每行归一化;

4)固定已标记数据的标签值,重复2)以后的操作直到收敛;

5)根据标签矩阵$\boldsymbol{Y}$判断未标记数据标签。

其中$\boldsymbol{Y}$为标签矩阵,每一列为相应类别,每一行为该顶点属于相应类别的概率。

标签传递算法的结果符合流行假设。流形假设是指处于

收稿日期:2011-08-08;**修回日期**:2011-09-21。

作者简介:刘培奇(1959-),男,陕西西安人,副教授,博士,主要研究方向:机器学习、数据挖掘、自然语言处理; 孙捷烩(1988-),女,山东济南人,硕士研究生,主要研究方向:机器学习、数据挖掘。

一个很小局部邻域内的数据具有相似的性质，因此其标签也应该相似。流形假设主要考虑模型的局部特性，标签传递算法沿着未标记数据定义的稠密区域传递标签，使较小邻域范围内的数据获得相同标签。

标签传递算法较好地解决了 K 最近邻（K-Nearest Neighbor，KNN）算法不能很好处理具有流行特征的数据的分类问题，在手写数字识别、文本情感分类等应用中取得了较好的结果。

1.2 LDA 主题模型

向量空间模型（Vector Space Model，VSM）用词表示文档，假设词之间是独立的，容易忽略一词多义和同义词的情况，导致了分类精度不高。潜在语义索引（Latent Semantic Indexing，LSI）[10]用奇异值的分解将同义词压缩到一个语义空间中，但是不能直接说明语义空间的含义。概率潜在语义索引（Probabilistic LSI，PLSI）[11]假设每个文档由隐含的主题构成，将 LSI 扩展到概率统计的框架下，但是在生成文档时缺少概率模型，当文档数目增加时，容易出现过拟合的情况。针对这个问题，Blei 提出了 LDA 主题模型，通过假设主题分布符合 Dirichlet 分布，解决了过拟合问题。

LDA 主题模型是一种统计语言模型，通过求出每个词对应的隐含主题，利用 Dirichlet 先验得到主题分布，反复抽样产生文档中的每一个词，生成文档。图 1 表示 LDA 主题模型的组成。

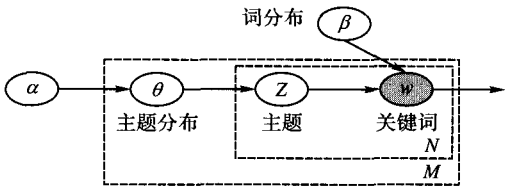

图 1 LDA 主题模型

图 1 中，α 为主题分布的 Dirichlet 超参数，β 为词分布的 Dirichlet 超参数，N 为每个文档中词的个数，M 为文档个数。

文档生成过程如下：

1）根据 α 从主题的多项分布中抽取一个主题；

2）根据 β 从词的多项分布中抽取这个主题对应的一个词；

3）重复第 1）、2），直到抽取出文章中所有的词。

LDA 主题模型可以应用到多种分类器上进行文本、图像分类等。

2 基于 LDA 主题模型的标签传递算法

标签传递算法利用未标记数据结合已标记数据提高分类的效果，是一种直推式的方法。LDA 主题模型是一种无监督的概率生成模型，通过文档集上共享主题的不同分布表示文档。两种算法虽然基于不同的理论，但是不存在相斥的假设，可以结合到一起。本章将介绍基于 LDA 主题模型的标签传递算法（Label Propagation based on LDA model，LPLDA）。

2.1 算法基本思想

将标签传递算法用于文本分类，只需要标注少量的文本，但是要求分类结果符合流行假设。LDA 主题模型使用主题的分布表示文本，符合流行假设。标签传递算法处理高维的文本时使用 LDA 主题模型表示文本可有效降低文本数据的维数，利于减少标签传递算法计算相似度的时间。

2.1.1 流行假设

标签传递算法在迭代过程中按照数据的流行假设进行标签传递。当一个标记顶点向未标记顶点传递自己的标签时，近邻的未标记顶点连接权值大于其他的未标记顶点权值，标记顶点的标签容易转移到这些近邻顶点上，标签在近邻顶点达到局部平滑。

已经证明基于流行假设的空间比欧氏空间更适合作为文档的表示空间来寻找隐藏的主题，LTM（Locally-consistent Topic Modeling）[12]用加入流行假设的 PLSI 表示文本，与传统主题模型相比，可以取得更好的分类效果。LTM 通过构造文档—主题的条件概率分布在数据的几何学上的局部平滑函数并最小化，使邻近的文档—主题条件分布相似。LTM 先对文本表示加上流行假设，然后用支持向量机进行分类，使在近邻分布相似的文本归为同一类。用基于 LDA 主题模型的标签传递算法分类时，同样可以使较小邻域范围内主题分布相似的文本具有相同的类别。不同于 LTM，LDA 主题模型本身并不构造符合流行假设的结构，而是通过标签传递算法使文本的主题分布达到满足流行假设的要求。

标签传递算法可以看作加入流行假设的 1NN，用 LDA 主题模型表示文本，分别用标签传递算法和 1NN 对样本集为 400，类别数为 4 的文本分类。图 2 表示随着未标记数据比例的变化两种算法分类的正确率。

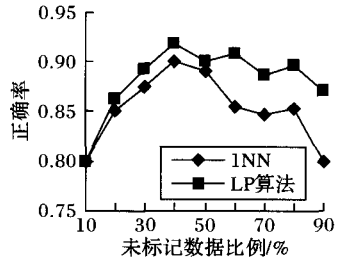

图 2 1NN 与标签传递算法效果对比

从图 2 可以看出，在利用 LDA 主题模型表示文本的基础上，标签传递算法的效果要优于 1NN，并且标签传递算法可以帮助 LDA 主题模型根据文本的流行假设对文本划分类别。

2.1.2 文档的维数

在标签传递算法中，构造图时包含表示文本的顶点和表示文本间相似度的边。计算图中的相似度时，时间复杂度为 $O(m \times n^2)$，m 是数据的维数，n 是顶点的个数，表示文本的顶点维数越少越利于降低计算的时间代价，但同时必须保证文档的信息损失比较小。LDA 主题模型中的每个主题由文档集中所有词的不同分布表示，语义相近的词被划分到同一个隐含主题下。主题的数目和词的数目没有固定的关系，但是一般主题数目远小于词的数目。设文档的数目为 m，文档集的主题数目为 k，共有 v 个词，使用主题分布和词频表示文本数据时，标签传递算法计算相似度的时间复杂度分别为 $O(k \times n^2)$ 和 $O(v \times n^2)$。在有限的文本下，经预处理后得到的 v 随着 m 的增加不断增长，直到新文档的所有词已经包含在现有的词集中。图 3 表示文档的词数与个数之间的关系。

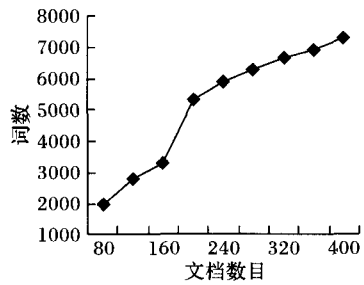

图 3 文档词数

实验表明，有效主题数不超过 50，$k < n \ll v$，可得 $O(k \times$