学校编码：10384　　分类号_____ 密级_____

学号：19020141152625　　UDC_____

厦门大学

硕 士 学 位 论 文

基于半监督的带Elastic Net正则项BP神经网络在文本分类上的应用

Improved BP neural network combined with semi-supervised algorithm and its application on text classification

陈 欣

指导教师姓名：谭　忠　教授

专 业 名 称：应　用　数　学

论文提交日期：2017　年　月

论文答辩日期：2017　年　月

学位授予日期：2017　年　月

答辩委员会主席：_____

评　阅　人：_____

2017 年 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。本学位论文属于：

（　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

中文摘要

随着大数据时代的来临，挖掘数据潜在的价值成为了各领域学者、工作者致力于研究的课题之一。文本数据承载着大量的信息，垃圾文本分类作为文本挖掘中经典的课题之一，虽然已经有了长足的发展，但提升垃圾文本分类器的精度一直都是人们追求的目标。BP神经网络模型是一种非常有效的非线性模型，它通过模拟生物神经网络，可以较好的拟合线性不可分的数据，是进行分类问题的常用模型之一。现今，文本数据呈爆炸式增长，呈现数据量大、高纬度、“有标签”数据少“未标签”数据多等特点，传统的BP神经网络已不能很好地解决这些问题。本篇论文中，我采用改进后的BP神经网络进行改进，并结合基于图的半监督学习方法在垃圾文本分类问题上进行实证分析。本文的主要研究内容及成果如下：

（1）对于传统BP神经网络模型，我加入了具备良好性质的Elastic Net正则项以防止BP神经网络在面对高维数据时出现的“过拟合”现象，并给出了带Elastic Net正则项的BP神经网络具备组效应性质的证明，同时，在Zhang[7]、范钦伟[8]等人的研究基础上，采取了光滑函数逼近的方法对带Elastic Net正则项的BP神经网络的收敛性进行分析。

（2）讨论基于图的半监督学习算法，使用带Elastic Net正则项的BP神经网络结合基于图的半监督学习的组合模型，在垃圾文本数据集上进行实证分析，从结果中发现，带Elastic Net正则项的BP神经网络在处理“过拟合”现象上相较于BP神经网络表现出了其优越性，再结合基于图的半监督学习算法后，分类模型的精度也得到了提升，验证了该组合方法的有效性。

关键词： BP神经网络，Elastic Net，基于图的半监督

Abstract

With the advent of the era of big data, mining the potential value of data has become one of the popular research topics. Text data is loaded with a large amount of information, although spam text classification is well developed, which is one of the most well-known subjects of text mining, improving the accuracy of the spam text classifier has been a goal pursued by people. BP neural network is a effective nonlinear model neural network, can better fit the non-linearly separable data, is one of the commonly used models of classification problems.

Today, we have a large amount of the text data with high latitude, but there are few labeled data. The traditional BP neural network can not solve these problems well. In this paper, I have done the empirical analysis on the spam text classification with the algorithm of the improved BP neural network and the graph based semi-supervised learning method. The main research contents and results are as follows:

(1) For the traditional BP neural network, I added the Elastic Net regularization, with good properties In order to avoid the over-fitting problem of BP neural network faced with high dimensional data. I gave the prove of the group effect of the BP neural network with Elastic Net regularization. On the basis of Zhang's[7] and Fan Qinwei[8] research, I discussed convergence analysis on the BP neural network with Elastic Net regularization.

(2) I discussed the graph based semi-supervised learning algorithm, combined with BP neural network with Elastic Net regularization to make an empirical analysis carried out on the spam text classification. It is found that the BP neural network with Elastic Net regularization term is more effective than the BP neural network in dealing with the over-fitting problem. After combined with the graph based semi-supervised learning algorithm, the accuracy of the classification model has been improved too.

Key words: BP neural network, Elastic Net, graph based semi-supervised

目 录

Contents

第一章 绪论

1.1 研究背景

随着互联网技术的不断发展，人们逐渐从信息匮乏的时代进入了信息过载(Information Overload)的时代。以互联网为载体，数据呈爆炸式的增长，而如今随着科学技术的不断进步，数据采集技术也日渐成熟，获取海量数据已不再是难题。海量的数据、高维的数据特征、多元的数据类型使我们正式进入大数据时代，如何挖掘数据潜在的价值，如何正确地使用数据造福人类社会已逐渐成为各领域科学家、企业家关注的重点。数据分析与数据挖掘，现如今已成为海量数据处理的重要技术之一，通过数据挖掘算法找出海量数据中潜藏的珍贵信息，并根据海量数据的不同特征以及不同的业务场景，将数据挖掘任务进行了归类，如分类、聚类、回归等。数据挖掘具有自适应的选择合适的算法解决海量数据处理问题的能力，已成为金融、互联网、医疗等领域关注的重点。

分类是数据分析与挖掘中一项非常重要的研究课题，在日常生活中，有许多常见的分类问题，如根据用户的性别、年龄、学历、工作、历史消费行为、银行还款行为等特征预测该用户是否会逾期还款。数据挖掘技术也为解决此类问题提供了多种不同的算法，常用的分类算法有朴素贝叶斯(Naive Bayes)、Logistic回归(Logistic Regression)、支持向量机(SVM)、BP神经网络(Neural Network)等。

文本分类是分类任务中最为经典的研究问题之一，文本分类是基于文本数据的特征，根据“有标记”的样本数据，预测“未标记”的文本所属的类型。比如对于一个论坛上的某一篇帖子或是回复，可根据该帖子或回复所携带的文本信息，如词语、符号、图片等信息将该文本划分为垃圾文本或是正常文本，而垃圾文本分类也是文本分类中最常见的、最重要的任务之一。BP神经网络通过构造多层网络结构，凭借

其优秀的非线性映射能力、自学习自适应能力等优秀性质成为了最为流行的文本分类算法之一。

现如今随着大数据时代的来临，大量的文本信息接踵而至，处理大规模的文本数据已成为一个新的课题，文本数据的特征也发生了重要的变化。高纬度、特征相关度高、样本数据量大但“有标记”样本少已成为了现今文本数据的新特点。传统BP神经网络模型在处理高纬度、特征相关度高的数据上容易出现“过拟合”现象。所谓“过拟合”现象是指模型在训练集的表现很好，分类精度高，但在测试集上的泛化能力不足，分类精度低。Elastic Net作为一种优秀的特征选择算法，它不仅能有效地处理样本数据维度远远大于样本数据容量的特征选择问题，防止模型“过拟合”，还能有效地处理组效应(Group Effect)的问题。对于样本数据量大但“有标记”样本数据少的这一特点，传统的有监督学习算法已经无法很好的解决，从而出现了基于图的半监督学习，基于图的半监督学习的旨在于利用大量“未标记”的样本数据内在的信息，通过标签传播的方式帮助提高模型的精度。

近年来，文本分类的在金融、互联网等领域发挥着越来越重要的作用，如何在文本数据的新特点下，提高分类器的精度，已成为一个新的难题。现有的文本分类研究大多是基于传统的有监督分类算法，本文将研究带Elastic Net正则项的BP神经网络，从而改善传统的BP神经网络分类算法在面对高维度、特征相关度高的文本数据时，容易造成“过拟合”的现象，并结合基于图的半监督学习算法，改善传统的有监督学习在面对文本数据量大但“有标记”文本数据少时模型表现较差缺点，进一步提升文本分类器的精度。

1.2 国内外研究现状

1.2.1 文本分类

随着互联网飞速发展，文本数据作为人类知识的载体，挖掘其中潜藏的宝贵信息，以应用于实际的业务中，意义非凡。其中，文本分类是文本挖掘中重要的任务

之一，提高文本分类器的模型精度具有重要的意义。

相比于国内在自动文本分类已进行了深入的研究，文本分类研究在国内的发展起步较晚。目前比较流行的文本分类算法是基于传统有监督算法，如朴素贝叶斯、Logistic回归、支持向量机、BP神经网络等。随着文本数据特点发生变化，“有标记”的文本数据较少，“未标记”的文本数据较多，也有学者开始研究基于半监督的学习方法在文本分类上的应用，并成功地验证了其有效性。

1.2.2 BP神经网络

BP神经网络算法首先由Werbos(1974)提出，此后由Rumelhart et al.(1986)[1]重新发明。BP神经网络算法实质是LMS(Least Mean Square)算法的推广。LMS试图使网络的输出均方误差最小化，可用于神经元激活函数可微的感知机学习；将LMS推广到由非线性可微神经元组成的多层前馈网络，就得到BP神经网络算法，因此BP神经网络算法亦称广义σ规则[2]。

MacKay(1992)[3]在贝叶斯框架下提出了自动确定神经网络正则化参数的方法。Gori和Tesi(1992)[4]对BP网络的局部极小问题进行了详细讨论。Yao(1999)[5]综述了利用以遗传算法为代表的演化计算(Evolutionary Computation)技术来生成神经网络的研究工作。对BP神经网络算法改进有大量研究，例如为了提速，可在训练过程中自适应缩小学习率，即先使用较大的学习率然后逐步缩小，可参阅[6]。Zhang[7]发表的“Boundedness of a batch gradient method with penalty for feedforward neural networks”讨论了带L_2正则项的BP神经网络的收敛性质，范钦伟(2014)[8]研究了在光滑函数逼近的条件下，进行了带$L_{\frac{1}{2}}$正则项前馈神经网络学习算法的收敛性分析，并给出了相应的数值模拟及收敛性证明，这都为本文提供了理论研究的基础。

1.2.3 Elastic Net

变量选择是数据分析重要的研究领域之一，Tibshirani(1996)[9]提出了模型变量选择发展史上具有重要意义的Lasso方法，它通过压缩回归方程的系数至0进行变量选择。在Lasso基础上进一步发展考虑变量分组结构的Group Lasso(2006)[9]、考虑

变量序结构的Fused Lasso[11]等变体。但Lasso方法也有其缺陷，当变量个数远远大于样本个数时，Lasso方法将容易受到样本个数的限制，无法选出更多变量；当变量之间存在交互作用时，Lasso方法不能讲它们同时选出，因此，Zou和Hastie[12]提出Elastic Net方法，并指出该方法不仅能有效地处理数据维度远远大于数据个数的变量选择问题，还能有效地解决组效应问题。Li和Lin[13]在假设适合的先验分布的前提下，建立了贝叶斯方法与Elastic Net方法的联系，验证了贝叶斯Elastic Net方法相较于传统的变量选择方法在预测方面有更好的表现。Yin Shen[14]等人研究了Elastic Net在稀疏向量重构上的能力，并给出了相应理论的稳定性及组效应分析。黄登香(2014)[15] 证明了运用Adaptive Elastic Net方法在Poisson对数线性模型中具有渐进正态性，即具有Oracle性质以及组效应性质，并通过数值模拟验证其有效性。

1.2.4 半监督学习

半监督学习的研究一般认为始于[16]，该领域在二十世纪末、二十一世纪随着现实应用中利用未标签数据的巨大需求涌现而蓬勃发展。国际机器学习大会(ICML)从2008年开始评选“十年最佳论文”，在短短6年中，半监督学习四大范型(Paradigm)中基于分歧的方法、半监督SVM、基于图的半监督学习的代表性工作先后与2008年[17]、2009年[18]、2013年[19]获奖。

最早的半监督学习方法[20]直接基于聚类假设，将学习目标看做图的最小割(Mincut)。对此类方法来说，图的质量记为重要，高斯距离图以及k紧邻图、ε近邻图都较为常用，此外已有一些关构图的研究[21]，基于图核(Graph Kernel)的方法也与此有密切联系[22]。刘钰峰、李仁发(2015)[23]提出了任意结构的异构信息网络上的半监督学习的正则化分类函数，并得到分类函数的闭式解,以此预测“未标记”节点的类别。郑文静(2016)[24]研究了基于图的半监督文本情感分类算法，将Graph-of-words文本表示模型引入半监督情感分类问题中,提出了两种基于Graph-of-words的半监督情感分类算法。

1.3 论文组织结构

全文大体分为五个部分，主要由绪论、相关理论知识、理论研究、实证分析、总结与展望组成，其组织结构如下：

第一章，介绍了本文的研究背景、国内外研究现状、论文组织结构、论文创新点；

第二章，介绍了BP神经网络的发展历程、BP神经网络相关基础知识，重点介绍了BP神经网络算法及其求解过程；

第三章，介绍了Lasso、Ridge、Elastic Net几种特征选择的方法，讨论了关于Elastic Net的组效应性质；

第四章，介绍了半监督学习的基本假设和学习思想，详细介绍了基于图的半监督学习方法；

第五章，本文的主体，研究了带Elastic Net正则项的BP神经网络模型，讨论了带Elastic Net正则项的BP神经网络的组效应性质的证明；在Zhang[7]、范钦伟[8]、等人的研究基础上，在光滑函数逼近的前提下，给出了带Elastic Net正则项的BP神经网络的收敛性的证明；

第六章，通过实验，将带Elastic Net正则项的BP神经网络及基于图的半监督学习的组合方法应用于垃圾文本分类问题中，并对实验结果进行分析；

第七章，对本文讨论的主要内容及主要结论进行总结，并给出相关研究方向的展望。

1.4 论文创新点

（1）Elastic Net已被广泛的应用于广义线性模型中，在神经网络中的应用较为少见，本文讨论了带Elastic Net正则项的BP神经网络算法，并给出了带Elastic Net正则项的BP神经网络的组效应性质的证明；

（2）在Zhang、范钦伟等人的研究基础上，在光滑函数逼近的前提下，给出了带Elastic Net正则项的BP神经网络的收敛性的证明；

（3）在存在大量“未标记”数据的分类问题场景下，通过组合基于图的半监督学习方法，提升带Elastic Net正则项的BP神经网络模型在垃圾文本分类上的精度，并通过实验进行对比分析。

第 二 章　BP神经网络

本节将从起源、发展、基础理论、基本方法等方面对BP神经网络(Back Propagation Neural Network)进行简单的介绍，其中包括人工神经网络的基础模型感知机模型，重点介绍了BP神经网络模型的概念及参数学习过程等内容，为之后续讨论带Elastic Net正则项的BP神经网络模型提供理论基础。

2.1　人工神经网络的简史

人工神经网络研究的相关工作起始于20世纪40年代，由心理学家W·Mcculloch和数理逻辑学家W·Pitts在分析、总结神经元基本特性的基础上，提出了人工神经网络的基础模型“神经元”模型。20世纪50年代末，由F·Rosenblatt提出的一种多层神经网络——“感知机”模型，标志着将人工神经网络模型理论结合实际应用的开端。但人工神经网络的发展并不是一帆风顺的，在20世纪60年代末期，人们更偏向于使用数字计算机解决人工智能、模式识别等方面问题，而冷落了对人工神经网络的学习与研究。Marvin Lee Minsky＆S.Paper也在他们的著作《感知器》一书中用数学证明指出了“感知机”模型的局限性，至此人工神经网络进入了低谷期。在长时间沉寂后，20世纪80年代初期，Kohen教授提出了自组织神经网络和John Hopfield教授提出了模拟人脑的Hopfield网络，让人们重新认识到了人工神经网络的潜力和研究价值，至此，人工神经网络的研究进入高潮时期。随着BP神经网络算法、Boltzmann机[25]、RBF网络（径向基网络）[26]的出现，人工神经网络的研究工作日渐成熟，也为现在的人们研究领域“深度学习”奠定了坚实的理论基础。

人工神经网络具有以下性质：

1.非线性：人工神经网络具备拟合非线性数据的能力，可以用来处理数据中存在线性不可分的问题，相较于普通的线性模型它具备更丰富的处理问题能力；