基于 boosting 算法的新闻文本分类研究

文/熊魏

摘要

人类历史的发展已经进入到网络时代。现在社会信息的发布量和使用量随着网络的发展突飞猛进，这么大的信息量，我们不可能全部的接受。此时，对有用信息快速、精确的掌握就显得尤为重要。方法是随着困难一起产生的，为了解决这个问题，文本自动分类系统就产生了，它的工作原理是对文本的内容在指定的分类体系下进行自动区分类别的过程。目前在所有分类算法中，有一种新兴的机器学习算法，即 Boosting 算法，这种算法经过科学验证后，其效果是非常理想的，且本身有着其它分类算法无可比拟的优点。

【关键词】boosting 算法 新闻 文本分类 研究

所谓文本分类（简称 TC），是一种定性文本内容类别的过程，其具体做法是在确定好的文本类别的前提下，对指定的文本内容进行判别归类。随着网络技术的发展，从上世纪 90 年代开始，传统的文本分类法（知识工程分类法）慢慢的被以计算机学习为基础的自动文本分类法所取代，成为 21 世纪初进行文本分类的主导技术。这种新的文本分类方法包括最近邻分类、回归模型、决策树、推导规则、贝叶斯分类、神经网络、支持向量机以及相关反馈等内容。另外，近几年比较流行的一种分类方法是组合分类器方法。

1 新闻文本预处理

所谓 boosting 算法，就是是通过机器学习方法构建自动文本分类器，根据文本训练集的类集 C 的特征进行学习，使用归纳过程进行分类的一种算法。以计算机学习为基础的自动文本分类法在对文本进行分类时需要一定的形式，称之为特征向量。由于文本内容都是以自然语言来进行表示的，计算机难以对其语义进行理解，为此需要对指定的新闻文本做一下预处理，其具体做法如下：

1.1 对指定新闻文本进行分词

文本包括西文文本和中文文本两种形式，对这两种文本进行分词的方法是不一样的，西文文本分词所采用的方法是用空格作为分隔符放在单词之间；中文文本（包括新闻文本）的分词方法按照依据的基础不同有很多种，例如以字符串匹配为基础的分词方法，以理解为基础的分词方法和以统计词频为基础的分词方法等。其中适合本系统的中文文本分词方法是以统计词频为基础的分词方法。分词完毕后，将会得到一本文本表征词典，此词典是由文档中的词组成的表。

1.2 对指定新闻文本进行粗降维

为了提高文本分类器的训练和分类效率，必须对指定文本在转化特征向量之前进行粗降维。所谓的粗降维，就是删除掉指定文本中的停用词（对分类没有意义且反复出现在文本中

<< 上接 173 页

费目的。

3.2.2 定额监控预警

人次定额是当前使用最广的第三方支付结算方式，通过模拟结算结合人次定额进行对比，可以知道该患者的医保费用是否超过了人次定额。

3.2.3 单病种支付结算方式预测

从以前单一的付费结算方式，演变为按病种、按总额控制、按人次付费等多样化结合的第三方付费结算方式，系统根据各种付费方式的规则算法进行预测，在患者尚未出院之前帮助医生选择最佳的出院支付结算方式，保证医院和医生的切身利益。

3.2.4 其他预警

其他费用预警和与 CIS 交互的诊疗项目预计等，例如药占比预警、材料占比预警、贵重药品预警、平均住院日预警、耗材使用预警、转科患者预警等等。诊疗项目预计如图 2 所示。

3.3 事后统计分析挖掘

由于医保结算系统与医院 CIS 系统是分离的，所以系统通过文件交换、数据交换和 API 等多种数据交换方式，将 CIS 数据与医保结算数据通过 ETL（抽取、转换和家族）进行集成和清洗。系统设置医保类型、科室、时间、收费类别、人员类型等多个维度，可以了解全院、科室医疗费用总额、收治人次、平均住院日、平均住院费用、个人自费额、药费比例等量度值的情况及其趋势变化。

3.3.1 门诊统计分析

门诊医保根据不同的医保、人员类别及待遇类型进行分析，分析监控门诊医保人数、人均限额来监控门诊费用，从科室、医生等维度监控医保人数、参保人发生的总费用、参保人自费费用、乙类个人先自付费用、参保人起付标准费用＋共付段（个人支付＋统筹记账金额）费用，定额费用、医保超定额比例及各费用构成比。

3.3.2 住院统计分析

住院医保根据不同的医保、人员类别及待遇类型进行分析，根据时间维度、医保人员类别维度，统计分析全院 / 指定所有待遇类型的总人数、平均住院天数、总费用、纯自费费用、基本医疗费用、超 4 倍定额费用、人均总费用、人均纯自费费用、人均基本医疗费用等数据。合理掌握科室、医生下医保病人的情况。

4 结语

本文设计并实现了基于 .net 平台，使用了 BI、规则引擎、视图构造引擎、即时消息服务等技术的医保费用调控系统平台，这一平台为医院而设计，按照医保支付规则，结合科学合理使用医保基金的核心思想，其有效控制不合理医疗的同时科学合理使用医保基金，实现医院、医保患者、医保局三方共赢。

参考文献

[1] 莫少雄．深圳市某医院住院医保病人费用管理研究 [J]．武汉：华中科技大学，2008.
[2] 刘涛．医保监控系统的设计与实现 [J]．数字技术与应用，2012(05):132-132.

作者单位

1. 广东医科大学信息工程学院　广东省东莞市　523808
2. 广州坤硕医疗科技有限公司　广东省广州市　510000

表 1：几种分类算法测试数据（%）

算法	封闭测试			开放测试		
	查全率	准确率	F1 值	查全率	准确率	F1 值
Boosting	94.7	96.8	95.7	88.1	87.8	87.9
简单向量距离	87.1	87.1	87.2	80.3	80.3	80.4
贝叶斯	82.4	83.8	83.1	76.2	77.3	76.8
KNN	89.2	91.4	90.3	83.3	85.2	84.3

的词）和低频词（使用频率极低的词）等，并合并数字和人名，从而使表征词典的规模缩小，避免掉分类时给分类器带来噪音。

1.3 文本表示

我们通常把用向量形式表示文本表征词典的方法称之为文本表示。在进行信息处理时，文本表示采用的方法是向量空间模型。

2 boosting算法下新闻文本的分类

在 boosting 算法下，新闻文本的分类设计主要由两大系统架构组成。

2.1 自动分类系统的设计

该系统主要的主要任务是对新闻文本进行自动的分类，即通过对文本进行扫描，实现新闻文本的粗降维；同时，通过自动分类的预处理新闻文本，分类完毕后，进行相应的文本输出。该系统虽属于计算机的前台系统，但此系统还可以根据计算机后台系统传递出的分类器号形成新的分类器。

2.2 训练学习子系统的设计

此系统的设计主要是为了通过训练语料库而形成新的分类器。即对语料库进行更新时，该系统会使语料库的训练重新开始，已达到信号能传递至自动分类系统，从而更新分类器的效果。与自动分类系统相对，此系统隶属于计算机的后台运行系统。

3 基于boosting算法的新闻文本分类设计的构成模块

基于 boosting 算法的新闻文本分类设计的构成模块包括文本预处理、人工分类、文本分词、文本降维和分类器训练五部分。其每个模块有着特定的作用：文本预处理的主要作用是指对文本进行中英文识别，以及转换文本的格式；人工分类的主要作用是指由专家对文本标上类别标签予以分类；文本分词的主要作用是指通过对经过预处理的新闻文本进行高精度的分词，以满足后续算法的需要，并提高后续的分类速度；文本降维的主要作用是通过删除停用词和低频词等对文本分类贡献小的词汇，且避免过匹配问题，来提高程序的效率和运行速度；分类器的主要作用是指对指定的新闻文本的语料进行预处理、分词和降维训练后，得到分类器，并将成功的信号传递到前台系统。

4 基于boosting算法的新闻文本分类试验数据及比较结果

本文算法同常用的分类算法在准确率、查全率以及 F 测试上的表现如表 1 所示。

由表 1 可以看出，在基于 boosting 算法下新闻文本分类系统的设计是否合理，需要通过准确率、查全率以及 F 测试值这三个指标来来进行验证。通过反复的测试与试验，其大致实验过程如下：首先，根据试验所需，从相关计算机数据库中抽取并下载 600 篇新闻文本，以人工分类的方式将这些文本主要分为 3 类。同时应注意，语料库有大小之分，为此我们又将这些新闻文本通过交叉验证的方式，对“熟”语料进行了平均分配，分为 10 份，并将其中的 9 份作为训练集和封闭测试集，1 份作为开放测试集。然后按照此方法，将每一份都作为康芳测试集，进行一次分类操作，共计 10 次。最后，对这 10 次得到的结果记性平均值的计算，与其他的新闻文本的分类方法所得结果进行相应数据的比较。结果显示，即使在训练语料库规模较小的情况下，新闻文本的分类通过 boosting 算法依旧可以达到预期的效果进度。

综上所述，时代在进步，科技在发展，人们每天接触的新闻信息量是越来远大。我们需要对这些新闻信息进行分门别类，去粗取精。为了实现快速、准确掌握必要新闻信息的目的，我们设计了一个基于 boosting 算法的新闻文本分类的实验，经过实验结果数据的对比，证明了基于 boosting 算法的新闻文本分类的方法是可取的，其效果是良好的，可以满足人们的需求。

参考文献

[1] 肖江，张亚非 .Boosting 算法在文本自动分类中的应用 [J]. 解放军理工大学学报自然科学版 ,2003,4(02):25-28.

[2] 董乐红，耿国华，周明全 . 基于 Boosting 算法的文本自动分类器设计 [J]. 计算机应用 ,2007,27(02):384-386.

[3] 张文生，于廷照 .Boosting 算法理论与应用研究 [J]. 中国科学技术大学学报 ,2016(03):222-230.

[4] 赵春兰 . 一种单一编码多分类 boosting 优化算法 [J]. 计算机与现代化 ,2015(08):121-126.

[5] 李诒靖，郭海湘，李亚楠，等 . 一种基于 Boosting 的集成学习算法在不均衡数据中的分类 [J]. 系统工程理论与实践 ,2016(01):189-199.

[6] 罗军，况夯 . 基于 Boosting 算法集成遗传模糊分类器的文本分类 [J]. 计算机应用 ,2016,28(09):2386-2388.

[7] 肖江，张亚非 .Boosting 算法在文本自动分类中的应用 [J]. 解放军理工大学学报自然科学版 ,2016,4(02):25-28.

[8] 刘川，廖士中 . 矩优化 Boosting 算法 [J]. 模式识别与人工智能 ,2015,28(12):1067-1073.

[9]DONG Lehong,GENG Guohua,ZHOU Mingquan, 等 .Design of auto text categorization classifier based on Boosting algorithm 基于 Boosting 算法的文本自动分类器设计 [J]. 计算机应用 ,2017,27(02):384-386.

作者简介

熊魏（1994-），男，长江大学计算机科学学院在读研究生。主要研究方向为数据挖掘。

作者单位

长江大学计算机科学学院　湖北省荆州市 434023