两种基于向量化策略 SVM 分类器的对比分析

薛又岷[1],陈春玲[1],余 瀚[1],王官中[2]
(1. 南京邮电大学 计算机学院、软件学院、网络空间安全学院,江苏 南京 210023;
2. 伦敦玛丽女王大学 商务与金融学院,伦敦 E1 4NF)

摘 要:以股票涨跌趋势预测精度为评价指标,针对传统股票数据特征训练过程中预测精度不高的情况,考虑引入两种不同的向量化策略对股民评论、新闻关键词等文本信息进行非结构化数据特征的捕捉,利用词意的积极、消极程度对客观因素进行处理,进而将向量化后的特征作为新的非线性特征项扩充原有的结构化特征集合。文中分别以词向量化和句向量化为出发点设计两种启发式的 SVM 分类器,其目标是在拟合每支股票的情况下尽可能预测出其未来的走势,挖掘出更具有增长潜力的股票样本。经过 2018 年 6 月至 12 月半年沪市股票数据集的实验结果表明,相比于词向量化策略,采用句向量化策略设计的 SVM 分类器不仅能够更好地预测股票涨跌,并且能够更有效地挑选出潜在增长的股票样本。

关键词:向量化策略;非结构化数据;SVM 分类器;启发式算法
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2020)02-0037-05
doi:10.3969/j.issn.1673-629X.2020.02.008

Comparison Analysis between Two Vectorization Strategy Based SVM Classifiers

XUE You-min[1], CHEN Chun-ling[1], YU Han[1], WANG Guan-zhong[2]
(1. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. School of Economics and Finance, Queen Mary University of London, London E1 4NF, United Kingdom)

Abstract: With the accuracy of stock trend prediction as the evaluation index, two different vectorization strategies are introduced to capture the unstructured data characteristics of shareholders' comments, news keywords and other text information in the light of the low accuracy in the traditional stock data training process. Based on the positive and negative degree of lexical meaning, the objective factors are processed, and the vectorized features are used as new nonlinear features to expand the original structural feature set. We design two kinds of heuristic SVM classifiers from the perspective of word vectorization and sentence vectorization respectively so as to predict the future trend of each stock as far as possible under the condition of fitting each stock and dig out the stock samples with more growth potential. The experimental results of the Shanghai Stock Market data set from June to December 2018 show that compared with the word vectorization strategy, the SVM classifier designed by the sentence vectorization strategy can not only better predict the stock trend, but also pick out the stock samples with potential growth more effectively.

Key words: vectorization strategy; unstructured data; SVM classifier; heuristic algorithm

0 引 言

在机器学习任务中,数据大多可分为结构化数据与非结构化数据[1-3]两类。结构化数据一般又可称为行数据,是指存储在数据库中可以用二维表结构实现逻辑表达的数据,如数字、符号等。而非结构化数据指的是字段的长度不定,且每个字段中又可由其他子字段构成的数据,如文本、图像、多媒体信息等。随着计算机科学领域的多样化和不同学科间的交叉发展,越来越多的机器学习任务需要面对非结构化数据的处理问题,例如计算机视觉中对图像数据的处理,自然语言处理中对词句的处理等。近年随着推荐系统方面研究的不断发展[4-6],针对文本数据的分析常采用的方法

收稿日期:2019-03-06 **修回日期**:2019-07-08 **网络出版时间**:2019-09-25
基金项目:中国博士后基金特别资助(2018T110531)
作者简介:薛又岷(1995-),男,硕士研究生,通讯作者,研究方向为可视化数据分析、机器学习与数据挖掘;陈春玲,硕士,副教授,从事硕士研究生的高级软件工程、算法分析与设计、本科生的数据结构等课程的教学工作,软件技术及其在通信中的应用、网络信息安全等方面的研究。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190925.1525.062.html

为文本表示法,又称文本向量化[7]。其对应于不同场景中所采用的处理方式也有所不同,大致可归为两类:词向量化与句向量化。词向量化的方法主要指的是 word2Vec 技术,而句向量化的方法主要是指 str2Vec 技术[8]。

一般来说,传统的向量化技术用于预测上下文语句,但也可以通过词典形式将每个单词的出现频率记录在其中。词向量化能够在一定程度上保证词意判别的精度,但与此同时也会造成词向量与原文本中单词出现顺序无关的现象。在面对大量文本内容的情况下,词向量化常出现特征空间维度灾难与词序混乱的情况。因此,句向量化在此方面更受众多研究学者的青睐。而在机器学习算法选择方面,针对泛化性能这一重要的评价指标,文中采用十大经典分类算法之一的 SVM 算法[9]。

SVM 算法作为近十年来最有效的分类算法,通过核函数的巧妙思想,将所属不同类别之间的非线性可分数据映射到高维空间,以计算不同支持向量间最大软间隔为目标函数实现其分类目的。

金融界越来越多的学者考虑使用机器学习、量化交易等手段来预测复杂模型。然而传统的股市数据中,大量的数值型数据之间具有较强的线性关联性。若直接使用 SVM 分类器训练数据,很难在有限的特征空间中准确预测股价涨跌趋势。为解决这一问题,文中采用向量化策略扩充特征空间,其目的是为 SVM 分类器提供更多的特征依据,从而提高 SVM 的分类性能。

1 基本知识

1.1 支持向量机

分类算法中最基本的想法就是基于训练集在样本空间中找到一个划分超平面,将不同类别的样本分开。

定义 1:给定训练样本集

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\},\quad y_i \in \{-1, +1\} \tag{1}$$

其中,y_i 为类别标记;x_i 为待分类样本。

SVM 分类器的目的是寻找对训练样本局部扰动容忍性最好的划分超平面。换言之,即挑选分类结果最鲁棒的、对未见示例泛化性能最优的线性方程。

定义 2:对给定的待分类样本空间 x,可构建划分超平面,其线性方程为:

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0 \tag{2}$$

显然由式 2 可见,划分超平面由法向量 $\boldsymbol{w}$ 和位移量 b 决定。其中法向量 $\boldsymbol{w} = (w_1, w_2, \cdots, w_l)$ 决定了超平面的方向,而位移量 b 决定了超平面与坐标原点之间的距离。

定义 3:样本空间中任意一点到超平面的距离用 r 表示。

$$r = \frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b|}{\|\boldsymbol{w}\|} \tag{3}$$

距离超平面最近的训练点被称为“支持向量”。以二分类问题为例,所属不同类别的两个支持向量到超平面的距离之和被称为间隔,用符号 η 表示。SVM 旨在找到具有最大间隔的划分超平面,因此可构建约束问题。

定义 4:在给定间隔 η 的情况下,旨在找到约束参数 $\boldsymbol{w}$ 和 b 使其最大。

$$\max_{w,b} \quad \eta = \frac{2}{\|\boldsymbol{w}\|} \qquad \text{s.t.} \quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1, i = 1, 2, \cdots, m \tag{4}$$

以上便是线性可分情况下 SVM 的基本型。针对式 4,凸二次规划问题采用拉格朗日乘子的方式解决其对偶问题,以更高效地找到最优划分超平面。而对于样本线性不可分的情况,由于原始特征空间维数有限的情况下必然存在一个高维空间使样本可分,SVM 分类器采用核函数[10]的方式,将原始样本空间映射到一个更高维的特征空间进行划分。令 $\varphi(x)$ 表示将 x 映射到高维空间后的特征向量,代替式 4 中的样本输入 x,即为引入核函数概念后的约束目标函数。

用 $<\varphi(x_i), \varphi(x_j)>$ 表示样本 x_i 和 x_j 映射到高维空间后的内积,因为在拉格朗日对偶问题中为避开复杂的内积计算过程,故采用 $k(x_i, x_j)$ 表示两种映射经过核函数计算后的结果。

定义 5:核函数 $k(\cdots)$ 的定义如下:

$$k(x_i, x_j) = <\varphi(x_i), \varphi(x_j)> = \varphi(x_i)^{\mathrm{T}}\varphi(x_j) \tag{5}$$

SVM 分类器常用五大核函数为:线性核、多项式核、RBF 核、Laplacian 核和 Sigmoid 核[11]。对于不同的划分任务,常需实验不同的核函数以获得最优的泛化性能。

1.2 文本向量化

文本表示[12]是自然语言处理和推荐系统中最基本的任务。利用独热编码(one-hot)进行向量化操作可以将文本表示成一系列能够表达语意的向量,从而实现非结构化数据到结构化数据间的转换。目前主流的文本向量化策略分为词向量化(word2Vec)和句向量化(str2Vec)。

word2Vec 计算词语间的相似度有非常好的效果,可用于计算句子或者其他长文本间的相似度。其一般做法是对文本分词后,提取其关键词,用词向量表示这些关键词,接着对关键词向量求平均或者将其拼接,最后利用词向量计算文本间的相似度。

在文本内容庞大的情况下,word2Vec 方法因不具备连贯性,极易丢失文本中包含重要内容的语序信息。在此情况下可以考虑使用 str2Vec 方法。通过在输入层添加句向量(paragraph vector)记忆每个前序词语的判断结果,从而实现更精确的判断和预测。

2　基于向量化策略的 SVM 分类器

由于股票数据特征间关联性较强,且具有较强的不确定性,利用传统的机器学习方法很难准确地预测涨跌趋势。而如今日益发展的网络环境下,诸如用户评论、新闻内容等文本信息都具有潜在的价值。为了获取这些具有潜在价值的文本信息,首先提取每支股票的子论坛或新闻子板块中的文本内容关键词,对其积极、消极程度进行打分;同时,针对变化趋势明显的股票,一定具有更多的用户访问、评论和相关新闻推送量的特点对这些数据进行汇总统计,从而扩充原有的特征空间,提供潜在的特征依据。根据这一思想,文中分别利用 word2Vec 模型和 str2Vec 模型[13-15]设计了基于向量化策略的 SVM 分类器。

基于词向量化策略的 SVM 分类器:

给定一组股票样本集合 X,基于词向量化策略的 SVM 分类器可分为 2 步:首先,利用向量化模型对评论区、新闻板块关键词进行打分统计,并对每只股票的点击量、评论数、新闻数进行汇总,构建 5 种特征项;其次,利用上述得出的特征集中前 5 个月的数据作为训练集,后 1 个月的数据作为测试集对 SVM 分类器进行训练。

执行过程如下所示:

算法:向量化 SVM 分类器

(*以单只股票为例)

输入:待扩充股票半年数据样本集合 X,特征集合 C

输出:C_word、$ACC_{word}(X)$ 和 C_str、$ACC_{str}(X)$

步骤 1:计算每日数据中的评论数、新闻数及点击量,分别利用 word2Vec 模型和 str2Vec 模型对评论和新闻内容进行打分,构建新特征子集$\{C_{com_num}, C_{news_num}, C_{click}, C_{com_rank}, C_{news_rank}\}$并添加到原特征集合 C 中构成 C_word 和 C_str;

步骤 2:利用 SVM 分类器分别对 C_word 和 C_str 前 5 个月内股票数据进行训练;

步骤 3:利用最后一个月的数据作为测试集,计算 $ACC_{word}(X)$ 和 $ACC_{str}(X)$;

步骤 4:输出 C_word、$ACC_{word}(X)$ 和 C_str、$ACC_{str}(X)$。

实验中所选择的词向量化模型是 CBoW(continuous bag-of-words),而选择的句向量化模型是 DBoW(distributed bag-of-words)。通过计算表明,较 DBoW 模型而言,CBoW 模型消耗的时间成本更低一些。

3　实验结果及分析

3.1　实验及分析

为了验证算法的有效性,选取了沪市半年共计 3 486支股票样本进行实验。列举单支股票基本信息如表 1 所示。实验 1 是将提出的词向量 SVM 分类器与传统的 SVM 算法进行对比分析;实验 2 是将提出的句向量 SVM 分类器与传统的 SVM 算法进行对比分析;实验 3 是将提出的词向量 SVM 与句向量 SVM 进行对比分析。实验环境为 PC 机,双核 2.1 GHz CPU,4 GB 内存,Ubuntu16.04 操作系统,python 3.6 实验平台。

用后一天开盘价减去前一天的收盘价,若结果为正,则标记为涨;负则标记为跌。

表 1　数据集的基本信息

数据集(股票代码)	样本数(天数)	原始特征数	类别数(标签)	新增特征数
#000000	181	14	2	5

实验 1:词向量化 SVM 分类器(Word-SVM)与传统 SVM 分类器的比较。

在实验 1 中,将词向量化 SVM 分类器与传统的 SVM 分类器进行了对比分析,采用三种不同核函数对全部股票进行训练。列举 5 支拟合后保持长期增长趋势的股票,结果如表 2 所示。

表 2　词向量化 SVM 分类器与传统 SVM 分类器的比较

股票代码	特征数		原始分类精度/%			Word-SVM 分类精度/%			整体耗时/s	
	原始	Word-SVM	RBF	Sigmoid	多项式	RBF	Sigmoid	多项式	原始	Word-SVM
000001			65.71	62.18	70.11	67.28	63.45	73.09		
000688			58.56	57.22	63.29	59.03	59.80	69.42		
300587	14	19	61.04	58.54	59.37	61.32	66.92	61.51	337.28	381.94
600707			57.55	63.19	64.29	63.48	60.77	68.03		
600823			69.27	65.66	68.51	73.76	65.89	69.64		

从表 2 可以看出,通过使用 CBoW 模型对文本内容进行处理,Word-SVM 方法因添加了特征集合扩充的过程,因此比单纯使用 SVM 方法训练需要消耗更多的时间成本。Word-SVM 方法在三种核函数的试验基础上都可以有效地提升分类精度,其中最有效的方法是使用多项式核函数。这是因为对特征集合进行扩充,为模型训练提供了更多的特征依据的同时,考虑到了将客观因素通过数值形式表示,利用隐藏的客观事件规律实现更好的预测。因此通过引用词向量化策略,可以有效地提升分类器的性能。但同时值得注意的是,股票数据每日之间具有强关联性,原始特征集之间也具有较强的线性相关性。如股票代码 000001 中,后一天的数据由前一天的数据、当日的大盘走势、政策因素等直接影响;五日均线、十日均线等属性之间实际上包含许多隐藏的线性关联性。

实验 2:句向量化 SVM 分类器(Str-SVM)与传统 SVM 分类器的比较。

在实验 2 中,将句向量化 SVM 分类器与传统的 SVM 分类器进行了对比分析,同样是采用三种不同核函数对全部股票数据进行训练。列举 5 支股票结果如表 3 所示。

表 3 句向量化 SVM 分类器与传统 SVM 分类器的比较

股票代码	特征数		原始分类精度/%			Str-SVM 分类精度/%			整体耗时/s	
	原始	Str-SVM	RBF	Sigmoid	多项式	RBF	Sigmoid	多项式	原始	Str-SVM
000001			65.71	62.18	70.11	65.97	63.91	74.22		
000688			58.56	57.22	63.29	63.21	61.99	68.57		
300587	14	19	61.04	58.54	59.37	67.86	72.40	63.76	337.28	393.44
600707			57.55	63.19	64.29	62.43	61.03	63.49		
600823			69.27	65.66	68.51	76.08	68.18	71.32		

从表 3 可以看出,通过使用 DBoW 模型,在消耗一定时间成本的基础上对文本内容进行处理进而扩充特征空间,同样可以有效提升分类器的分类性能。

实验 3:词向量化 SVM 分类器与句向量化 SVM 分类器的比较。

在实验 3 中,将词向量化 SVM 分类器(Word-SVM)与句向量化 SVM 分类器(Str-SVM)进行对比分析,如表 4 所示。

表 4 词向量化 SVM 分类器和句向量化 SVM 分类器的比较

股票代码	特征数	Word-SVM 分类精度/%			Str-SVM 分类精度/%			整体耗时/s	
		RBF	Sigmoid	多项式	RBF	Sigmoid	多项式	Word-SVM	Str-SVM
000001		67.28	63.45	73.09	65.97	63.91	74.22		
000688		59.03	59.80	69.42	63.21	61.99	68.57		
300587	19	61.32	66.92	61.51	67.86	72.40	63.76	337.28	393.44
600707		63.48	60.77	68.03	62.43	61.03	63.49		
600823		73.76	65.89	69.64	76.08	68.18	71.32		

从表 4 可以看出,句向量化 SVM 分类器(Str-SVM)在三种核函数的基础上可以有效提升精度,且利用多项式核的 SVM 分类器进行训练,其分类精度要普遍高于利用正向贪心特征选择出的特征子集进行分类所求出的分类精度。

3.2 实验结论

为了进一步提升传统 SVM 分类器算法在股票预测模型中的精度,利用向量化策略,采用现有的 CBoW 和 DBoW 模型对文本特征进行词向量化与句向量化处理,并结合传统 SVM 分类器设计了两款启发式机器学习算法。由实验结果可知,向量化策略可以有效地将非结构化数据转换为结构化数据,从而扩充特征空间,提供更多的特征依据。且相比于词向量化,句向量化策略更能够有效地将长文本、大文本内容转换为数值数据。因为股票数据具有很多的非确定性,通过向量化非确定因素来增加特征项是目前数据处理阶段的一种重要手段。因此所提出的算法是具有现实意义的。

4 结束语

利用 SVM 分类器对比分析了词向量化和句向量化在股票数据特征处理方面的优劣。实验表明相比词

向量化,句向量化更能够有效生成文本特征实现模型进一步的精确预测。

在文中工作的基础上,笔者将重点考虑数据预处理的方式(如:基本面等因素的介入),同时进一步考虑使用量化交易策略与 SVM 分类器、神经网络等卓越的机器学习算法相结合,并将其应用到优质股票预测与推荐的问题上。

参考文献:

[1] D'AMICO B,MYERS R J,SYKES J,et al. Machine learning for sustainable structures:a call for data[J]. Structures,2019,19:1-4.

[2] 苏金树,张博锋,徐　昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1859.

[3] ZHANG Dongwen, XU Hua, SU Zengcai, et al. Chinese comments sentiment classification based on word2vec and SVM[J]. Expert Systems with Applications,2015,42(4):1857-1863.

[4] 郑英丽,王　新,马　倩,等. 一种结合用户相似度的社会化推荐算法[J]. 云南民族大学学报:自然科学版,2019,28(1):93-99.

[5] 刘建国,周　涛,郭　强,等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学,2009,6(3):1-10.

[6] CHANG Wenbing,XU Zhenzhong,ZHOU Shenghan,et al. Research on detection methods based on Doc2vec abnormal comments[J]. Future Generation Computer Systems,2018,86:656-662.

[7] 赵丽丽,赵茜倩,杨　娟,等. 财经新闻对中国股市影响的定量分析[J]. 山东大学学报:理学版,2012,47(7):70-75.

[8] 于　政. 基于深度学习的文本向量化研究与应用[D]. 上海:华东师范大学,2016.

[9] 杨新斌,黄晓娟. 基于支持向量机的股票价格预测研究[J]. 计算机仿真,2010,27(9):302-305.

[10] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011,40(1):1-10.

[11] 奉国和. SVM 分类核函数及参数选择比较[J]. 计算机工程与应用,2011,47(3):123-124.

[12] MIHALCEA R,TARAU P. Textrank:bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. Barcelona, Spain:[s. n.],2004:404-411.

[13] MIKOLOV T,SUTSKEVER I,CHEN Kai,et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th international conference on neural information processing systems. Lake Tahoe,Nevada:Curran Associates Inc.,2013:3111-3119.

[14] LE Q,MIKILOV T. Distributed representations of sentences and documents[C]//International conference on machine learning. Beijing,China:JMLR. org,2014:1188-1196.

[15] KARVELIS P,GAVRILIS D,GEORGOULAS G,et al. Topic recommendation using Doc2Vec[C]//International joint conference on neural networks. Rio de Janeiro:IEEE,2018:1-6.