

基于混合 CRFs & SVM 应用于文本信息分类

胡改蝶¹, 马建芬²

(1. 太原工业学院 计算机工程系, 山西 太原 030008; 2. 太原理工大学 信息与计算机学院, 山西 太原 030024)

摘要: 为了对文本信息进行更加有效分类, 研究文本信息的一种新型分类。在了解文本信息分类和讨论条件随机场(CRFs)和支持向量机(SVM)的基础上, 结合 CRFs 对上下文依赖性的优点以及 SVM 对高维问题的特殊解决方法, 分析了二者结合的方法, 并将其中一种结合方法应用到文本信息分类中。通过实验对此方法进行了验证。实验结果表明: CRFs & SVM 方法的查全率、查准率以及 F1-测量的各平均值比 SVM 方法都有所提高, 具有良好分类性能和更加准确的分类率。进一步证明 CRFs & SVM 方法适用于文本信息分类, 具有良好的应用前景。

关键词: 文本信息; 分类; 条件随机场; 支持向量机; 查全率; 查准率

中图分类号: TP311 **文献标志码:** A **文章编号:** 2095-0373(2020)04-0075-05

0 引言

随着网络的日益发展, 网络信息的增长也愈来愈快, 其中文本信息是网络信息中最重要的一部分。如何快速且有效地对文本信息进行合理的组织、管理和使用, 从而进一步促进文本信息分类的发展已成为组织文本数据的关键技术^[1-2]。

分类本质上就是在特征向量 $x = (x_1, x_2, \dots, x_k)$ 给定的条件下, 估计离散类变量 y 。基于机器学习的分类方法有产生式模型(Generative Model, 简称 GM)与判别式模型(Discriminative Model, 简称 DM)。二者的主要区别在于: 前者是将估计联合分布概率作为输入和输出, 它的局限在于维数特别大, 特征也有很大的依赖性, 后者是估计条件概率, 然后进行分类, 以此来解决前面所述的问题^[3]; 前者的观察序列可以是模型的一部分, 后者的观察序列只可以是条件; 前者是针对无穷样本, 后者是针对有限样本^[4]。文本信息分类就是用自动化的技术将文本样例分配到一组预定义的类别中的一个或多个的过程^[5-6]。在文本信息分类中用到了很多技术, 常见的判别式模型有条件随机场(Conditional Random Fields, 简称 CRFs)、支持向量机(Support Vector Machine, 简称 SVM)和最大熵模型(Maximum Entropy Method, 简称 MEM)等。

基于条件随机场和支持向量机 2 种方法, 运用循序渐进的方法, 先通过条件随机场得到特征集, 随后通过支持向量机再进行训练和分类, 最终得出比传统 SVM 分类效果更好的混合方法。将混合 CRFs & SVM 方法运用到文本信息分类中, 从而对文本信息获取更好的分类。最后利用文本集对此方法进行实验, 并与传统 SVM 方法作了相应对比, 验证了 CRFs & SVM 方法的有效性。

1 CRFs & SVM

1.1 条件随机场(CRFs)

条件随机场是基于统计的判别式模型之一, 是在给定一输入序列的前提下, 求出标注序列的概率模型, 本质上是整合了分类和图像模型优点的一种模型。经常使用的是线性链式 CRFs。其定义是^[4]: 设 x 和 y 为随机变量, θ_i 为从训练集中训练所得到的参数变量, $f_t(y_k, y_{k-1}, x_k)$ 为一组真值特征函数, k 表示序列位置, 此时 CRFs 的条件概率

收稿日期: 2019-05-15 责任编辑: 车轩玉 DOI: 10.13319/j.cnki.sjztdxxb.20190092

作者简介: 胡改蝶(1981—), 女, 助教, 研究方向为自然语言处理、机器学习、计算机网络。E-mail: butterfly586@163.com

胡改蝶, 马建芬. 基于混合 CRFs & SVM 应用于文本信息分类[J]. 石家庄铁道大学学报: 自然科学版, 2020, 33(4): 75-79.

$$p(y/x) = \frac{1}{z(x)} \exp \left[\sum_{k=1}^K \sum_{t=1}^T \theta_t f_t(y_k, y_{k-1}, x_k) \right] \quad (1)$$

式中, $z(x)$ 为归一化因子

$$z(x) = \sum_y \exp \left[\sum_{k=1}^K \sum_{t=1}^T \theta_t f_t(y_k, y_{k-1}, x_k) \right] \quad (2)$$

CRFs 对独立性假设和数据归纳偏注问题解决有很大的见解,同时 CRFs 具有产生式模型的特点,对上下文的文本信息描述有特别好的效果,并对所有特征进行全局归一化^[7]。它主要用于分词、词性标注、文本处理、图像分类、实体识别和生物信息学等范畴的研究。

1.2 支持向量机(SVM)

支持向量机是一种结构风险最小化的统计学习方法,是基于分类边界的方法,主要应用于小样本分类。SVM 大致分为线性可分的、线性不可分和非线性 3 种情况。第一种情况是通过最大化边缘的超平面来实现的;第二种情况是通过定义松弛变量,存放到边缘的离差来实现的;第三种情况是将其低维空间中的点映射到新的高维空间,可以用适当的核函数,将其转换成线性可分,然后辨别分类的边界,从而大大避免维数灾难问题^[8]。即支持向量机的主要思想是通过非线性变换将输入空间变换到高维特征空间,再求出最优线性分类面^[9]。

支持向量机是一种神经网络,它对分类做出了巨大贡献,得到国内外诸多研究人员的高度重视,并将其理论在多个领域应用,如在文本分类、语音方面、数据挖掘、图像分类和字符识别范畴都有广泛应用。

1.3 CRFs & SVM

基于 CRFs 和 SVM 的特点,从理论上讲,将二者结合起来是可以应用在多个研究领域的。CRFs 和 SVM 的结合方法有:顺序法,即第二个的分类器的输入是第一个分类器的输出;并列法,即同时用 2 种方法进行分类后,再将其进行整合,以此得到最终的分类结果^[10]。本文采用顺序法进行研究。将文本信息的训练集通过 CRFs 得出特征集,然后将其送入 SVM,再次进行训练,最后得出分类结果^[11]。这种方法的优点是 CRFs 的上下文相关性可以解决文本信息的不确定分类问题。CRFs & SVM 定义

$$p(y/x) = \frac{1}{z(x)} \exp \left[\sum_{k=1}^K \sum_{t=1}^T \theta_t f_t(y_k, y_{k-1}, x_k) + \sum_{k=1}^K H(y_k, x_k) \right] \quad (3)$$

式中, $H(y_k, x_k)$ 为第 k 个序列位置只考虑当前观测值的后验概率。

CRFs 训练是将文本信息数据集转换成适合 CRFs 的文本信息训练集,然后用训练集生成一个模型文件,从而得到特征总数集及每一特征出现次数,最后对参数进行训练^[11]。本文条件随机场的训练是用 CRF++0.53 实现。

SVM 训练主要是解决后验概率问题,此处可由对偶分类耦合来得出多类分类的后验概率。假设观察 x 和类标签 y, k, t 为训练集中类别的位置,用对偶分类的条件概率 $\mu_{kt} = p(y=k/y=kort, x)$ 的耦合类的估计值 r_{kt} 来估计后验概率 $p_k = P(y=k/x), k=1, 2, \dots, K$ (其中 K 为类别的数目)。通过优化后的后验概率^[12-13]

$$\min_r \sum_{k=1}^K \left(\sum_{t, t \neq k} r_{tk} p_k - \sum_{t, t \neq k} r_{kt} p_t \right)^2 \text{ 条件为: } \min_r \sum_{k=1}^K p_k = 1, p_k \geq 0, k=1, 2, \dots, K \quad (4)$$

式中, $r_{kt} = P(y=k/y=kort, x) = 1/(1+e^{Af+B})$, f 为决策函数分类值, A, B 是训练所得的参数。

2 实验与分析

2.1 实验环境、文本集与测量标准

为了对上述 CRFs & SVM 方法进行验证,本实验将同一组文本集分别在传统的 SVM 方法和混合 CRFs & SVM 方法上进行了分类实验。本文的实验环境是 Microsoft Windows7 操作系统, Intel(R) Core(TM) I5-6200U CPU @ 2.30 GHz 处理器, 4.00 G 内存, 64 位操作系统。用到的工具有 Visual Studio 2010、CRF++0.53 和 Visual C++6.0。

用到的实验文本集为复旦大学提供的分类文本集,该文本集包括训练文本集和测试文本集,从中抽

取了 7 个类别,分别是经济、计算机、军事、艺术、医药、教育和政治。训练文本集和测试文本集相互之间不重叠,从中抽取训练文本集 1 052 个,测试文本集有 577 个,共计 1 719 个文本集,其类别及文本集统计数如表 1 所示。

表 1 类别及文本集

文本集	类别							文本集总数
	经济	计算机	军事	艺术	医药	教育	政治	
训练	181	108	124	100	109	131	299	1 052
测试	98	56	73	72	58	63	157	577

分类实验用查全率(Recall,简称 R)、查准率(Precision,简称 P)和 F1-测量(F1-Measure,简称 F1)来衡量的^[14-15],其中 F1-测量是由查全率和查准率 2 个值综合而来的。同时求出了所有类别的查全率、查准率和 F1-测量的平均值(Average,简称 A),用其对分类效果进行衡量。

2.2 实验对比结果与分析

一般情况下,SVM 分类算法优于其他的分类算法,因而实验中用传统 SVM 和混合 CRFs & SVM 2 种方法对文本信息进行分类,并进行比较分析。实验中传统 SVM 方法的特征选择用的是卡方统计法,特征加权算法选择 TF·IDF 算法,特征空间数目选择 800。

为了与传统 SVM 方法进行比较,实验共使用了 3 种方式在相同实验环境下对同一组文本集进行分类实验,实验分类对比效果如表 2 所示。从表 2 中可以看到,采用 CRFs & SVM 方法后,除了在经济和政治类别的查全率略小、经济类别的 F1-测量值略小外,其余类别的查全率、查准率和 F1-测量结果均有明显提升,都得到了较好的分类效果,准确率大大提升,尤其是对军事、艺术和医药类别的分类效果非常好。同时也可看到,CRFs & SVM 方法的查全率、查准率和 F1-测量的平均值分别达到了 0.958、0.964 和 0.961,其平均值均比传统 SVM 方法的分类值高。通过对分类后的查全率、查准率和 F1-测量进行对比,尽管分类后高出的值并不是很突出,但整体来看,CRFs & SVM 方法的分类效果要好。

表 2 分类对比效果

标准	方法	类别							平均值
		经济	计算机	军事	艺术	医药	教育	政治	
R	传统 SVM	0.959	0.982	0.877	0.944	0.966	0.952	0.955	0.948
	CRFs & SVM	0.954	0.982	0.904	0.976	0.971	0.968	0.952	0.958
P	传统 SVM	0.979	0.982	0.928	0.964	0.933	0.967	0.909	0.952
	CRFs & SVM	0.981	0.982	0.970	0.986	0.949	0.968	0.914	0.964
F1	传统 SVM	0.969	0.982	0.901	0.954	0.949	0.960	0.932	0.950
	CRFs & SVM	0.967	0.982	0.935	0.981	0.960	0.968	0.933	0.961

各类别运用 2 种方法分类后的对比效果如图 1、图 2 和图 3 所示。从图 1、图 2 和图 3 的对比效果可知,CRFs & SVM 方法的查全率、查准率和 F1-测量的分类值均高于传统 SVM 方法。

实验结果表明,CRFs & SVM 方法不仅可以消除上下文的不相关性,达到降低特征维数的效果,解决文本信息的不确定分类问题,从而使分类的效果具有一定的优越性,而且 CRFs & SVM 方法可以对文本信息进行分类,进一步验证了 CRFs & SVM 方法适用于文本信息分类,使用此方法进行文本信息分类是有效的、可行的,在文本信息分类中有一定的实用价值。

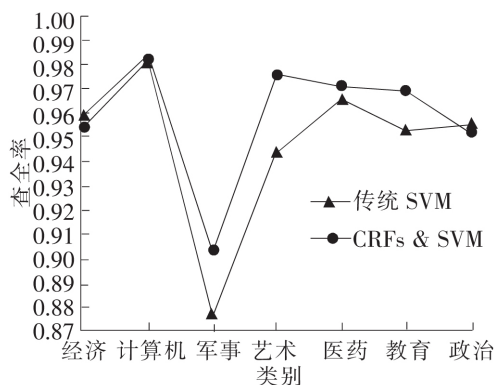


图 1 查全率对比图

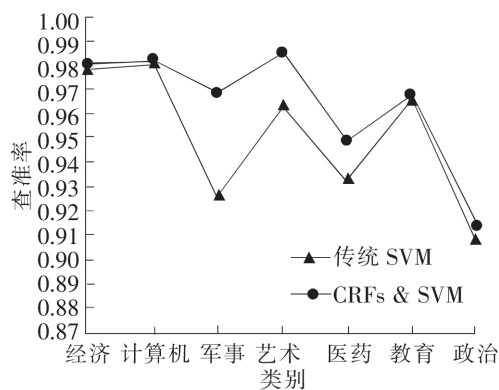


图 2 查准率对比图

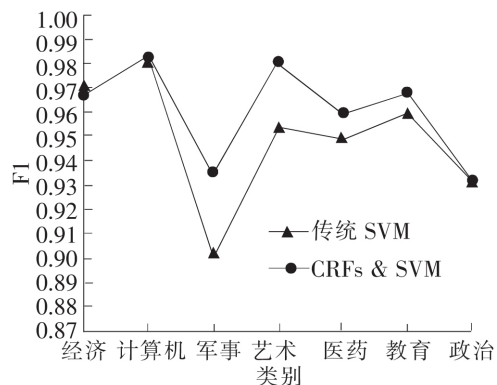


图 3 F1 对比图

3 结语

以文本信息为研究对象,探讨了条件随机场方法和支持向量机方法。采用条件随机场和支持向量机结合起来的方法—CRFs & SVM 方法,并将 CRFs & SVM 方法运用到文本信息分类中。不仅可解决特征集出现次数问题,同时还解决了分类后验概率问题。经过实验将其分类效果与传统 SVM 方法效果进行了对比,实验证明 CRFs & SVM 方法对文本信息分类的优越性,适用于文本信息分类领域,具有分类的实践性和实用性,可以为分类领域提供可靠的研究依据。但是部分文本信息分类效果还有待提高,在今后的研究中,努力探索二者的结合性,达到更好的分类优化,并在更大的文本信息集上实现更好的分类效果。

参 考 文 献

- [1] 刘晓飞, 邱书灵. 基于隐马尔科夫模型的文本分类[J]. 石家庄铁道大学学报: 自然科学版, 2013, 26(1): 101-106.
- [2] Wail Hamood Khaled, Haytham Saleem Al-sarrayih, Lars Knipping. Arabic text categorization using improved k-nearest neighbour Algorithm[J]. Journal of Applied Computer Science & Mathematics, 2014, 8(18): 9-12.
- [3] 江雪莲, 石洪波. 产生式与判别式组合分类器学习算法[J]. 山东大学学报: 理学版, 2010, 45(7): 7-12.
- [4] 汪光亚, 曹国, 尚岩峰. 基于条件随机场的多相遥感影像分类[J]. 计算机应用研究, 2018, 35(9): 2834-2837.
- [5] Vieira A Seaer, Iglesias E L, Borrajo L, et al. A HMM text classification model with learning capacity[J]. Advances in Distributed Computing and Artificial Intelligence Journal, 2014, 3(3): 21-34.
- [6] Guo Xiaoli, Sun Huiyu, Zhou Tiehua, et al. SAW classification algorithm for Chinese text classification[J]. Sustainability, 2015, 7(3): 2338-2352.
- [7] 曾佳妮. 基于条件随机场的中文短文本分类算法研究[D]. 上海: 上海交通大学, 2012.
- [8] 全奇, 叶霞, 李俊山, 等. 基于文本分类和 SVM 的雷达侦察装备故障诊断研究[J]. 电光与控制, 2016, 23(2): 94-98.
- [9] 范玉妹, 赵丽丽. 关于支持向量分类机算法的研究[J]. 石家庄铁道大学学报: 自然科学版, 2007, 20(3): 31-36.
- [10] 强家虎. 基于 SVM-CRF 的多分类器融合的语音情感识别研究[D]. 天津: 天津师范大学, 2012.
- [11] 张利娟. 基于 CRF 和 SVM 的连续语音识别研究[D]. 天津: 天津师范大学, 2013.
- [12] Wu Tingfan, Lin Chihjen, Weng Ruby C. Probability estimates for multi-class classification by pairwise coupling[J]. Journal of Machine Learning Research, 2004(4): 975-1005.
- [13] 李祖传, 马建文, 张睿, 等. 利用 SVM-CRF 进行高光谱遥感数据分类[J]. 武汉大学学报信息科学版, 2011, 36(3): 306-310.
- [14] 齐莹, 刘振丙, 张卫东, 等. 基于多通道特征融合的病理图像有丝分裂检测[J]. 计算机仿真, 2019, 36(4): 383-387.
- [15] 段文影, 饶泓, 段隆振, 等. 基于 IA 参数寻优组合核的 SVM 文本分类研究[J]. 南昌大学学报: 理科版, 2018, 42(3): 289-292.

Categorization of Text Information Based on CRFs & SVM

Hu Gaidie¹, Ma Jianfen²

(1. Department of Computer Engineering, Taiyuan Institute of Technology, Taiyuan 030008, China;

2. College of Information and Computer, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: In order to categorize text information more effectively, this paper adopted a novel method for categorization of text information by combining Conditional Random Fields (CRFs) of context dependence with Support Vector Machines (SVM) of the special solutions for high-dimensional problems. The results show that the average values of the recall, precision and F1-measure of CRFs & SVM method are higher than that of SVM method, presenting much more higher categorization accuracy and better categorization performance. In summary, the CRFs & SVM method is proper for the categorization of text information with wide application.

Key words: text information; categorization; conditional random fields; support vector machine; recall; precision

~~~~~  
(上接第 23 页)

## Design Research on New Prefabricated Appurtenances of Slab for High Speed Railway Bridges

Yang Bin<sup>1</sup>, Yin Jing<sup>2</sup>, Yu Guangzhi<sup>3</sup>, Chen Jinjie<sup>4</sup>, Rong Xueliang<sup>4</sup>

(1. Engineering Management Center, China State Railway Group Co., Ltd, Beijing 100844, China;

2. Institute of Railway Architecture, China Academy of Railway Sciences Co., Ltd, Beijing 100081, China;

3. Xiong'an High-speed Railway Co., Ltd., Baoding 071700, China;

4. School of Civil Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, China)

**Abstract:** In order to realize the full prefabricated construction of bridge superstructure, substructure and accessory structure, this paper put forward a concept design of integral prefabricated bridge deck ancillary facilities suitable for high speed railway bridges. Firstly, by comparing with the present layout scheme of bridge deck in the form of curtain board, this paper presented a full prefabricated plan of installing bridge deck railing and sound barrier in the form of side wall structure, which mainly includes protective wall, side wall, vertical wall and cable trough bottom plate. Then, based on the design functional requirements and safety performance of bridge deck ancillary facilities, this paper put forward the division method of prefabricated components, the structural scheme of standard block and beam end block, and the connection setting scheme of prefabricated bridge deck appurtenances for railing section and sound barrier section respectively. In the end, a structural check calculation was carried out for the full prefabricated bridge deck appurtenances design scheme proposed in this paper, and the calculation results show that: the strength and crack of protective wall, side wall and vertical wall can meet the design requirements, and the tensile strength as well as shear strength of connector bolts meet the requirements.

**Key words:** high speed railway bridges; bridge deck appurtenances; prefabricated components; side wall; horizontal derailment effect