

一种改进的支持向量机文本分类方法

李 琼, 陈 利

(汉口学院 计算机科学与技术学院, 湖北 武汉 430212)

摘 要:为了解决传统的支持向量机(SVM)多类分类方法中普遍存在的训练和测试时间过长、实际样本输入空间非线性可分等问题,提出了一种改进的支持向量机多类分类方法。首先,利用 Mercer 核函数把实际输入样本向量空间中呈非线性可分分布的样本向量映射到一个高维的特征向量空间,以实现线性可分;然后,采用二叉树来创建高维特征空间中的 SVM 多类分类器,实现分类识别;最后,将该方法应用到网络文本分类中。实验结果表明,该方法较好地解决了多类文本分类中实际输入文本向量呈非线性可分的问题,降低了训练和测试过程中的时间消耗,且在一定程度上提高了多类文本分类的识别准确率。

关键词:文本分类;支持向量机;二叉树;Mercer 核

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2015)05-0078-05

doi:10.3969/j.issn.1673-629X.2015.05.019

An Improved Text Classification Method for Support Vector Machine

LI Qiong, CHEN Li

(School of Computer Science and Technology, Hankou University, Wuhan 430212, China)

Abstract:In order to solve the problem of long time for training and testing, input space nonlinear classification in traditional SVM methods, propose an improved text classification method for SVM. First, the Mercer kernel is used to sample vector with nonlinear distribution in input space, which is mapped into a high dimensional feature vector space to realize the linear separable. Then it constructs the SVM multi-class classifiers in high dimensional space using binary tree, implementing the classification recognition. Finally, use this method into the network text classification. The experimental results show that the method can effectively solve the nonlinear separable problem in input text space, saving the training and testing time, and improving the precision of multi-text classification.

Key words:text classification; SVM; binary tree; Mercer kernel

0 引 言

随着 Internet 和移动技术的迅猛发展,来自互联网上的信息呈海量涌现,而这些网络信息目前绝大多数以文本的形式存在。如何从这些海量网络文本中获取有用信息已成为人们现实生活中的迫切需求。对网络文本进行分类,不仅可以较好地解决网络信息杂乱现象,还可以让用户快速准确地检索到所需信息,实现信息分流。因此,对网络文本进行自动识别分类具有较大的实用价值和学术意义。

文本自动分类的方法有很多,常见的有:贝叶斯方法、KNN 方法以及人工神经网络方法等。贝叶斯方法是基于概率模型的方法,该方法要求所有词在文本中的出现概率应保持相互独立,且文本类别和其长度不相关联。由于 Bayes 方法的假设对大多数文本并不成

立,故在实际应用中的分类效果不是特别稳定。KNN 方法是基于要求的分类识别方法,KNN 存放所有训练样本,直到未知样本需要识别时才为其创建相应的分类器。实际应用中,KNN 方法的分类效果较好。但 KNN 方法的分类时间是非线性的,当训练文本数量增加时,会导致分类时间急剧增加。人工神经网络方法虽然具有较好的容错性和非线性逼近能力,但在实际中,人工神经网络所需的训练时间较长。

支持向量机(Support Vector Machine, SVM)是基于统计学习理论的机器学习方法,由 Vapnik 等人创建。该方法采用结构风险最小化原则替代传统经验风险最小化原则,可以使其在训练样本数量有限的情况下,很好地兼顾分类识别准确率和分类推广能力。因此,SVM 方法被广泛地应用到模式识别和分类问题

收稿日期:2014-07-04

修回日期:2014-10-09

网络出版时间:2015-04-22

基金项目:2012 年湖北省教育科学技术研究计划指导性项目(B20128103)

作者简介:李 琼(1981-),女,讲师,硕士,研究方向为中文信息处理、计算机应用;陈 利,教授,研究方向为中文信息处理、计算机应用。

网络出版地址:<http://www.cnki.net/kcms/detail/61.1450.TP.20150422.1005.018.html>

中^[1]。文中在现有的 SVM 方法上做了改进,提出了一种基于 Mercer 核和二叉树的 SVM 多类分类方法,并将其应用于多类文本的自动分类。实验结果表明,该方法可以避免不可分区域的问题,并在一定程度上提高多类文本分类的准确率。

1 支持向量机

SVM 是从线性可分情况下的最优分类面提出的^[2]。二维空间中两类线性可分情况如图 1 所示。图中实心点、空心点分别代表正反两类训练样本。 H 是将两类样本正确分开的分类线。 H_1 、 H_2 分别代表过两类样本中离 H 最近的点,且平行于 H 的直线。 H_1 、 H_2 之间的距离,定义为两类之间的分类间隔,用 Margin 表示。最优分类线应满足:(1) 能将两类样本准确无误地分开;(2) 保证两类之间的分类间隔(即 Margin 的值)最大。将最优分类线进一步推广至高维向量空间,最优分类线就变为最优分类面^[3]。

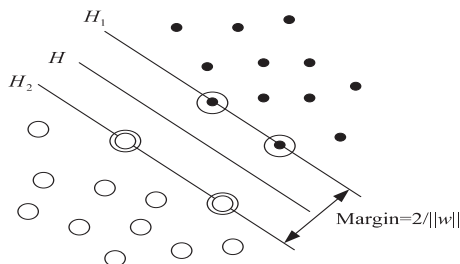


图 1 二维空间两类样本最优分类线

设线性可分的样本集为 $(x_i, y_i), i = 1, 2, \dots, n, x \in R^d, y \in \{+1, -1\}$ 是类别标号。 d 维空间中,线性判别函数为 $g(x) = w \cdot x + b$, 分类面为:

$$w \cdot x + b = 0 \quad (1)$$

将判别函数归一化,让正反两类样本均满足 $|g(x)| \geq 1$, 让离两类分类面最近的样本向量的 $|g(x)| = 1$, 这样分类间隔 Margin 等于 $2/\|w\|$ 。因此,要使两类分类面对所有正反两类样本无错误地分类,必须满足:

$$y_i[(w \cdot x_i) + b] - 1 \geq 0, i = 1, 2, \dots, n \quad (2)$$

故满足式(2),且使 $\|w\|$ 最小的分类面即为所求最优分类面。超平面 H_1 、 H_2 上的训练样本点,称之为支持向量(Support Vectors),即图 1 中圆圈标出的点。

故最优分类面问题可以转换为:在条件(2)的约束下,求函数

$$\varphi(w) = \frac{1}{2} \|w\|^2 = \frac{1}{2} (w \cdot w) \quad (3)$$

的最小值。这是一个约束条件下的二次优化问题,为了解决它,定义如下 Lagrange 函数:

$$L(w, b, \alpha) = \frac{1}{2} (w \cdot w) -$$

$$\sum_{i=1}^n \alpha_i \{y_i[(w \cdot x_i) + b] - 1\} \quad (4)$$

式中, $\alpha_i > 0$ 是 Lagrange 系数。对式(4)中的参数 w, b 求 Lagrange 函数极小值。这样就可以将原求解问题转换成:在以下两个约束条件下

$$\sum_{i=1}^n y_i \alpha_i = 0 \quad (5)$$

$$\alpha_i \geq 0, i = 1, 2, \dots, n \quad (6)$$

对 α_i 求如下函数的最大值:

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (7)$$

若 α_i^* 为最优解,则

$$w^* = \sum_{i=1}^n \alpha_i^* y_i x_i \quad (8)$$

上式表明:所求最优分类面的权系数由训练样本线性组合而成。

由 Kuhn-Tucker 条件,上述优化问题的解必须满足:

$$\alpha_i((y_i(w \cdot x_i) + b) - 1) = 0, i = 1, 2, \dots, n \quad (9)$$

故对多数样本来说, α_i^* 取值应为零,取值不为零的 α_i^* ,则对应于少量的支持向量。

经过求解得到的最优分类函数为:

$$f(x) = \text{sgn} \{ (w^* \cdot x) + b^* \} = \text{sgn} \left\{ \sum_{i=1}^n \alpha_i^* y_i (x_i \cdot x) + b^* \right\} \quad (10)$$

其中, $\text{sgn}()$ 是符号函数; b^* 是分类域值。

2 现有 SVM 多类分类方法比较

用 SVM 实现多类分类,通常有两种做法:一是将多个分类面的参数求解合成一个大的最优化问题以实现多类分类;二是采用某种技术构造多个二值 SVM 分类器,并组合起来以实现多类识别。前者在求解最优化问题时涉及的变量过多,计算消耗量大,故实际应用中一般采用第二种组合二值 SVM 分类器构造多类 SVM 分类模型的方法。

常见的组合构造方法有:1-a-r, 1-a-1 和 DAG-SVM^[4-5]。设总的类别数目为 $k(k \geq 2)$,下面对这几种常用 SVM 多类分类方法作简要比较分析。

1-a-r:是 SVM 用于解决多类问题的最早的方法。该方法每次用一个二值 SVM 分类器将其中一类与剩余类区分识别,总共需要使用 k 个二值 SVM 分类器。测试时,将未知测试样本归类为分类函数值最大的那一个类别。1-a-r 方法需要训练的二值 SVM 分类器个数(k 个)较少,分类速度较快。但该方法可能会将测试样本识别为多个类别或拒绝识别。

1-a-1:是在每两个类别之间训练出一个二值 SVM 分类器,共需训练 $k(k-1)/2$ 个二值 SVM 分类

器。当对未知样本进行分类时,每个二值 SVM 分类器都对其进行类别判别,且为其相应的类别进行投票,将票数最多的类别,作为未知样本的最终类别。其优点是训练速度比 1-a-r 方法快;缺点主要体现在三个方面:一是单个二值 SVM 分类器若不规范化,则会导致构造出来的 k 类 SVM 分类器趋于过学习;二是所需二值 SVM 分类器的数量会随总类别数的增加而急剧增加,进而导致分类识别速度下降;三是泛化误差无界。

DAG-SVM:在训练阶段和 1-a-1 方法一样,也需要构造每两类之间的二值 SVM 分类器,即需创建 $k(k-1)/2$ 个分类器。DAG-SVM 利用一个有向无环图进行分类识别。该有向无环图包含 $k(k-1)/2$ 个内部节点及 k 个叶子节点。其中,一个内部节点代表一个二值 SVM 分类器,一个叶子节点代表一个类别标签。测试时,从其根节点开始,根据二值 SVM 分类器计算出的函数值,决定下一步的行走路线。如此循环,直到叶子节点为止,那么该叶子节点对应的类别值即为未知测试样本的所属类别。DAG-SVM 的优点是分类速度与前面两种方法相比,得到明显提高。缺点是根节点所选的二值 SVM 分类器不同,会导致分类识别结果不同。

上述三种传统的 SVM 多类分类方法对比如表 1 所示。

表 1 传统 SVM 多类分类方法对比

方法	需要训练的 分类器数量	训练 速度	分类时所需 分类器数量	分类 速度
1-a-r	k	较慢	k	较快
1-a-1	$k(k-1)/2$	较快	$k(k-1)/2$	慢
DAG-SVM	$k(k-1)/2$	较快	$k-1$	快

3 改进 SVM 方法

针对上述传统 SVM 多类文本分类方法中存在的问题与缺陷,文中提出了一种基于二叉树和 Mercer 核的 SVM 多类文本分类方法。该方法用二叉树构造 SVM 多类分类器,有效节省了训练和测试时间;用 Mercer 核将非线性可分文本映射到高维特征空间中,使之线性可分,较好地解决了实际多类文本分类中非线性可分问题,并在一定程度上提高了文本的分类精度,经过改进的 SVM 分类过程如图 2 所示。

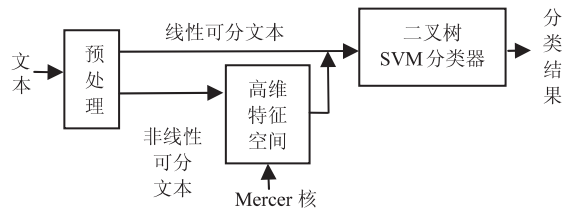


图 2 改进的二叉树 SVM 多类文本分类流程

3.1 输入文本空间线性可分

采用二叉树创建多类 SVM 分类器过程:首先将所有样本向量划分为两个子类别,接着继续把分出的每一个子类别依次细分为两个次级子类别……按此方法逐层划分,直到所有节点仅包含一个类别为止。最后在每两个子类之间训练两类 SVM 分类器实现多类分类。对于 k 类($k \geq 2$)问题,二叉树 SVM 分类器在训练阶段,总共需要训练出 $k-1$ 个二值 SVM 分类器;在测试阶段,不一定需要用到训练出的所有二值 SVM 分类器(最少用 1 个,最多用 $k-1$ 个)。故二叉树 SVM 可以有效节省训练和测试所需时间^[6-7]。二叉树 SVM 分类器属于树形结构的分类器,其分类性能与上层节点中两个子类之间的可分性强度有关^[8-9]。通常,为了使二叉树 SVM 具有最优的泛化能力,应使上层中两个子类之间的可分性尽可能强^[10-11]。因此,在生成二叉树时,应将可分性强度值较大的类别提前分割出来,即将其在二叉树的上层节点处划分。

定义包含每一类的最小超球体的半径 R 为该类别的可分性强度值 SM (Separability Measure), 包含该类的最小超球体半径 R 值越大,该类的可分性强度值 SM 值越大,其可分性越强,越应在上层节点处分割。

定义 1:最小超球体半径 R :设某类 C 有 m 个样本 x_1, x_2, \dots, x_m (x_i :样本的特征向量矩阵, $i = 1, 2, \dots, m$), 此类样本集的重心为 $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$, 则包含类 C 的最小超球体半径 $R = \max_{x_i \in C} \{ |x_i - \bar{x}| \}, i = 1, 2, \dots, m$ 。式中 $|\bullet|$ 表示欧氏距离。

定义 2:类 C 的可分性强度值 SM 定义为:

$$SM = R = \max_{x_i \in C} \{ |x_i - \bar{x}| \}, i = 1, 2, \dots, m$$

对于输入文本空间线性可分的情况,直接采用定义 2 中的 SM 作为类别的可分性强度值,来构造二叉树 SVM 多类分类器进行文本分类即可。

3.2 输入文本空间非线性可分

对于输入文本空间非线性可分的情况,引入 Mercer 核函数。在分类识别前,先将样本映射到高维 Hilbert 特征空间中,让具有同一类别的文本向量更紧凑地聚集在一起,而不同类别的文本之间的间隔更大,从而使原本线性不可分的文本变得线性可分。在此高维 Hilbert 特征空间中,采用包含各类的最小超球体半径 R^H , 作为该类的可分性强度值 SM^H , 来构造高维特征空间中的二叉树 SVM 多类分类器,实现多类文本的自动分类。

输入空间非线性可分文本的分类过程及分类算法:

假设输入空间中非线性可分文本 $x_i \in R^d (i = 1, 2,$

$\cdots, n)$, 用某个非线性的映射函数 φ 映射到高维 Hilbert 空间中后得到 $\varphi(x_1), \varphi(x_2), \cdots, \varphi(x_n)$ 。相应地, 实际输入文本向量空间中的点积在 Hilbert 空间中可以表示为: $K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j), i, j = 1, 2, \cdots, n$ 。

相应高维 Hilbert 空间中的欧氏距离为:

$$d^H(x, y) = \| \varphi(x) - \varphi(y) \| = \{ \| \varphi(x) \cdot \varphi(x) - 2\varphi(x) \cdot \varphi(y) + \varphi(y) \cdot \varphi(y) \| \}^{1/2} = \{ K(x, x) - 2K(x, y) + K(y, y) \}^{1/2}$$

定义 1 中类 C 的重心变为:

$$\varphi(\bar{x}) = \frac{1}{n} \sum_{i=1}^n \varphi(x_i)$$

高维特征空间中包含类 C 的最小超球体半径 R^H 定义为:

$$\begin{aligned} R^H &= \max_{x_i \in C} \{ \| \varphi(x_i) - \varphi(\bar{x}) \| \} = \\ &\max_{x_i \in C} \{ \| \varphi(x_i) - \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \| \} = \\ &\max_{x_i \in C} \{ [K(x_i, x_i) - \frac{2}{n} \sum_{i=1}^n K(x_i, x_i) + \\ &\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n K(x_i, x_j)]^{1/2} \} \end{aligned}$$

高维特征空间中类 C 可分性强度值 SM^H 定义为:

$$\begin{aligned} SM^H &= R^H = \max_{x_i \in C} \{ [K(x_i, x_i) - \frac{2}{n} \sum_{i=1}^n K(x_i, x_i) + \\ &\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n K(x_i, x_j)]^{1/2} \} \end{aligned}$$

其中, 核函数采用 RBF 径向基核函数: $K(x, y) = \exp \{ - \frac{|x - y|^2}{\sigma^2} \}$ 。

改进 SVM 多类文本分类算法步骤(设总的类别数为 $k(k \geq 2)$):

Step1: 用 RBF 径向基核函数将输入文本空间中的训练文本映射到高维 Hilbert 空间。

Step2: 计算各类的可分性强度值 SM^H 。

Step3: 依据 SM^H 值从大到小的顺序, 对所有类别排序。若有两个以上(包括两个)的样本, 其类别的可分性强度值相同, 则把类别标号较小的那个类排在前面。这一步将得到一个类别序列: n_1, n_2, \cdots, n_k 。此处, 类别标号 $n_i \in \{1, 2, \cdots, k\}, i = 1, 2, \cdots, k$ 。

Step4: 利用二值 SVM 分类训练算法创建二叉树中所有内部节点的最优超平面。首先在根节点处, 从训练集中选取第 n_1 类文本作为正样本集, 剩下的作为负样本集。利用 SVM 训练算法创建出相应最优超平面后将属于第 n_1 类的所有文本从训练样本集中删除。按照同样的方法依次处理二叉树中的剩余内部节点, 最后得到如图 3 所示的二叉树 SVM 多类文本分类

模型。

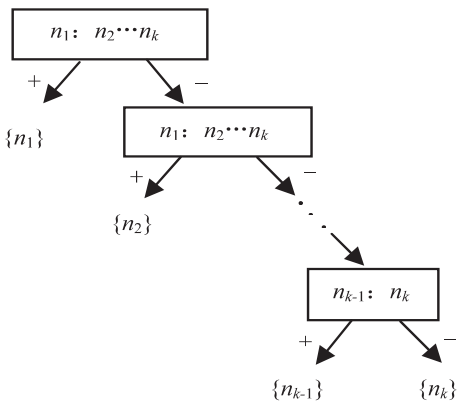


图 3 改进的二叉树 SVM 多类文本分类模型

4 实验结果

实验数据来源于从 Internet 上随机爬下的 583 篇中文网页文本, 类别包含: 教育、环境、经济、军事、民生、旅游 6 类。各类文本的样本数分布见表 2。

表 2 各类文本的样本数分布

文本类别	文本数
教育	100
环境	98
经济	94
军事	95
民生	100
旅游	96

从中任意选取其中的 428 篇作为训练样本, 其余 155 篇作为测试样本。实验文本使用中科院的 ICT-CLAS 中文分词系统进行中文分词, 使用 TF-IDF 算法进行特征向量选择^[12-13], 使用基于 CHI 值的文本特征降维方法等进行预处理后, 利用文中提出的改进二叉树 SVM 多类文本自动分类算法对文本进行分类, 并与 1-a-r、1-a-1 及 DAG-SVM 多类文本分类的结果进行比较。实验中用到的 SVM 分类算法是在 Chih-Jen Lin 和 Chih-Chung Chang 创建的 LibSVM 工具包上修改实现的^[14-15]。SVM 核函数使用 RBF, 其核参数采用网格搜索法及交叉验证法来优化选择, 实验结果如表 3 所示。

表 3 实验结果

方法	核参数(C, γ)	训练 时间/s	测试 时间/s	分类 精度/%
1-a-r	(100, 10^{-4})	67.52	14.23	82.50
1-a-1	(200, 10^{-4})	56.63	34.32	81.30
DAG-SVM	($2^9, 2 * 10^{-5}$)	56.54	12.35	87.60
改进方法	($2^{-1}, 2^4$)	47.71	8.93	93.60

实验结果表明, 改进的二叉树 SVM 多类文本分类

器的整体分类性能和分类精度都有了明显的提高,且节省了文本的训练和测试时间。

5 结束语

文中分析比较了目前常用的几种文本分类方法,指出了 SVM 用于文本分类的优势,并针对现有 SVM 多类文本分类方法中常见的训练或测试过程较长、实际文本空间非线性可分等问题,采用基于二叉树和 Mercer 核的改进 SVM 对多类文本进行识别分类。通过对网络文本进行分类实验后,结果表明,该改进的 SVM 多类分类方法可以较好地解决实际输入文本向量空间呈非线性可分的问题,提高多类文本自动分类精度和泛化能力,缩减文本的训练和测试时间。由于 SVM 是一种基于核的机器学习方法,核函数及相应核参数的选择直接影响到 SVM 的泛化能力,所以在核参数选择方法上还有许多工作要做。

参考文献:


- [1] Vapnik V N. 统计学习理论[M]. 许建华,张学工,译. 北京:电子工业出版社,2004.
 - [2] Li Dawei, Xu Lihong, Goodman E D, et al. Integrating a statistical background-foreground extraction algorithm and SVM classifier for pedestrian detection and tracking[J]. Integrated Computer Aided Engineering, 2013, 20(3): 201-216.
 - [3] 边肇祺,张学工. 模式识别[M]. 北京:清华大学出版社, 2000.
 - [4] Tian Yingjie, Shi Yong, Liu Xiaohui. Recent advances on support vector machines research[J]. Technological & Economic
-
- (上接第 77 页)
- 模及其模型规范化[J]. 计算机科学与探索, 2012, 6(6): 557-576.
- [7] 丁 博,王怀民,史殿习,等. 一种支持软件可信演化的构件模型[J]. 软件学报, 2011, 22(1): 17-27.
 - [8] Wyner A Z. An ontology in OWL for legal case-based reasoning[J]. Artificial Intelligence & Law, 2008, 16(4): 361-387.
 - [9] Labhart J, Rowe M, Matney S, et al. Forward chaining parallel inference[C]//Proceedings of the 2nd IEEE symposium on parallel and distributed processing. Dallas, TX: IEEE, 1990: 455-462.
 - [10] Kerner S M. Eclipse takes 'best' shot at framework[R]. [s. l.]: [s. n.], 2006.
 - [11] Arendt T, Taentzer G. A tool environment for quality assurance based on the eclipse modeling framework[J]. Automated Software Engineering, 2013, 20(2): 141-184.
 - [12] MOSKitt4SPL is a free open-source tool which gives support

- Development of Economy, 2012, 18(1): 5-33.
- [5] Manikandan J, Venkataramani B. Study and evaluation of a multi-class SVM classifier using diminishing learning technique[J]. Neurocomputing, 2010, 73(10-12): 1676-1685.
 - [6] 段 莹. 支持向量机在文本分类中的应用[J]. 计算机与数字工程, 2012, 40(7): 87-88.
 - [7] 崔建明,刘建明,廖周宇. 基于 SVM 算法的文本分类技术研究[J]. 计算机仿真, 2013, 30(2): 299-302.
 - [8] Maksim L, Matthias H, Bernt S. Learning using privileged information; SVM+ and weighted SVM[J]. Neural Networks, 2014, 53(5): 95-108.
 - [9] 黄振龙,郑 骏,胡文心. 基于类间可分性 DAG-SVM 的文本分类[J]. 华东师范大学学报: 自然科学版, 2013(3): 209-218.
 - [10] 许世明,武 波,马 翠,等. 一种基于预分类的高效 SVM 中文网页分类器[J]. 计算机工程与应用, 2010, 46(1): 125-128.
 - [11] Rojo-Álvarez J L, Martine-Ramón M, Muñoz-Mari J, et al. A unified SVM framework for signal estimation[J]. Digital Signal Processing, 2014, 26(3): 1-20.
 - [12] 姜 鹤,陈丽亚. SVM 文本分类中一种新的特征提取方法[J]. 计算机技术与发展, 2010, 20(3): 17-19.
 - [13] 郑 伟,吕建新,张建伟. 文本分类中特征预抽取方法研究[J]. 情报科学, 2011, 29(1): 86-88.
 - [14] Chang C C, Lin C J. LIBSVM: A library for support vector machines[EB/OL]. 2012-04-04. <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>.
 - [15] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification[EB/OL]. 2010-04-15. <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
-
- for modeling Software Product Lines (SPL) [EB/OL]. 2012. <http://www.pros.upv.es/m4spl>.
- [13] Ali N, Nellipaiappan R, Chandran R, et al. Model driven support for the service oriented architecture modeling language [C]//Proc of international conference on software engineering. [s. l.]: [s. n.], 2010: 8-14.
 - [14] Ameller D, Franch X. Service level agreement monitor (SA-LMon) [C]//Proceedings of the seventh international conference on composition-based software systems. Madrid: IEEE, 2008: 224-227.
 - [15] McBride B. Jena: a semantic web toolkit[J]. IEEE Internet Computing, 2002, 6(6): 55-59.
 - [16] Swordfish is a standalone eclipse application supporting deployment of swordfish modules from the eclipse IDE [EB/OL]. 2014. http://archive.eclipse.org/archived_projects/swordfish.

一种改进的支持向量机文本分类方法

作者：[李琼](#)，[陈利](#)，[LI Qiong](#)，[CHEN Li](#)

作者单位：[汉口学院 计算机科学与技术学院, 湖北 武汉, 430212](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2015(5)

引用本文格式：[李琼](#).[陈利](#).[LI Qiong](#).[CHEN Li](#) [一种改进的支持向量机文本分类方法](#)[期刊论文]-[计算机技术与发展](#)
2015(5)