第48卷第6期 中国科学技术大学学报 Vol.48,No.6
2018年6月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jun. 2018

文章编号:0253-2778(2018)06-0447-11

基于模拟退火半监督学习的信用预测研究

张 杰,李 琳,朱 阁

(武汉理工大学计算机科学与技术学院,湖北武汉 430070)

摘要:金融机构结合消费者和商业信息来为企业进行信用打分.我国的企业特别是小微企业信用信息少,造成了只有少量企业拥有信用信息,而大量企业没有信用信息的局面.半监督支持向量机可以利用标记数据和未标记数据进行学习,同时可以克服信用数据类别不均衡和样本信息不足等问题.由于半监督支持向量机的参数对算法效果有较大影响,实际参数选取往往根据经验所得.为此提出了一种利用模拟退火(SA)优化基于确定性退火半监督支持向量机(DAS3VM)参数的SAS3VM算法.该算法在少量有标记信用数据的基础上,利用大量无标记信用数据辅助学习,使用模拟退火寻找最优参数.最后在两组企业信用数据集和三组个人信用数据集上进行对比实验,结果表明,半监督学习方法(DAS3VM和SAS3VM)优于监督学习方法,SAS3VM在准确率上比DAS3VM最大提升了13.108%.

关键词:半监督学习;确定性退火;模拟退火;信用预测

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2018.06.003

引用格式:张杰,李琳,朱阁. 基于模拟退火半监督学习的信用预测研究[J]. 中国科学技术大学学报,2018,48(6):447-457.

ZHANG Jie,LI Lin, ZHU Ge. Simulated annealing based semi-supervised support vector machine for credit prediction[J]. Journal of University of Science and Technology of China, 2018,48(6):447-457.

Simulated annealing based semi-supervised support vector machine for credit prediction

ZHANG Jie,LI Lin, ZHU Ge

(*School of Computer Science & Technology, Wuhan University of Technology, Wuhan* 430070, *China*)

Abstract: In the mid-1990s financial institutions began to combine consumer and business information to create scores for business credits. Enterprises in China, especially small and micro enterprises, have less credit information, resulting in the situation where only a small number of enterprises have credit information, while a large number of enterprises have none. However, semi-supervised support vector machines (S3VM) can learn from labeled data and unlabeled data and solve the problems of imbalanced credit data categories and insufficient sample information. The parameters of S3VM have a great influence on the effect of the algorithm, and the actual parameter selection is often based on experience. An SAS3VM algorithm was proposed to optimize the parameters of deterministic annealing based semi-supervised support vector machine (DAS3VM) with simulated annealing. Based on the small number of

收稿日期:2017-09-09;**修回日期**:2018-04-10

基金项目:国家社会科学基金(15BGL048),武汉理工大学科研基金(2017II39GX),武汉理工大学研究生优秀学位论文培育项目(2016-YS-068),湖北省科技支撑计划(研发与示范)2015BAA072)资助.

作者简介:张杰,男1993年生,硕士生.研究方向:机器学习.E-mail:icecream@whut.edu.cn

通讯作者:李琳,博士/教授.E-mail:cathylilin@whut.edu.cn

labeled credit data, the algorithm takes advantage of the unlabeled credit data to help study and use the simulate annealing to find the optimal parameters. Experiments were conducted on two categories of enterprise credit data and three categories of personal credit data. The results show that semi-supervised learning (DAS3VM and SAS3VM) performs better than supervised learning. The maximum accuracy of SAS3VM has been increased by 13.108% compared with DAS3VM.

Key words: semi-supervised learning; deterministic annealing; simulated annealing; credit prediction

0 引言

多年以来,许多大型银行都已采用结构化的正式系统的贷款,投资组合监测和管理制度,分析贷款损失准备金或资本的充足性以及还款率和贷款率分析[1].目前,信用相关业务产业逐渐变得多样化和复杂化.例如,在线P2P(peer-to-peer)贷款方便了个体消费者之间的直接借贷,网络购物的盛行也使许多小企业得以快速发展.对于许多小微企业而言,他们仅仅需要少量的短期资金,面对银行贷款的复杂流程和各种证明材料,他们望而生畏[2-3].

机器学习从人工智能的模式识别和计算机学习理论出发,从数据中探索学习并构建算法来预测数据,从输入数据(训练数据)中构建的模型来做数据驱动预测或决策.通常,训练数据是有标记数据.在借贷行业,一系列的借贷申请将会根据其申请项目、申请金额、借贷风险和其他特征来判断其是否适合借贷.通常为了达到较高的分类效果,训练数据中的正负样本应当是平衡的,并尽可能丰富,然而以下两种情形却对此有影响:

(Ⅰ)实际生活中,大多数的公司都会努力保持良好的信用记录,只有少数公司因为未能及时还款或其他原因而留下了不好的信用记录.当运用机器学习时,使用传统的监督学习方法会因为数据中正例样本的主导地位产生偏见,导致结果产生偏差.

(Ⅱ)金融机构对公司的一次借贷申请作出回应后,就可以得到一条有标记的数据样本.随着互联网的迅猛发展,小企业也如雨后春笋般不断涌现,并且通过互联网可以很便捷地获取企业的基本信息、在线交易评论信息、产品评论等.数量众多的小企业都没有信用记录,因此与标记过程相关的成本过高,使得人工不可能标记足够多的训练样本.在这种情形下,监督学习方法预测精度较低.

半监督学习可以解决上述问题.半监督学习旨在利用未标记样本辅助学习,扩充训练样本,提高分类效果.本文主要工作是在少量已标记信用数据和大量无标记信用数据的背景下,利用模拟退火优化半监督支持向量机的参数,提高信用预测质量.本文的工作主要包括以下两个方面:①在确定性退火半监督支持向量机(DAS3VM)的基础上,提出了使用模拟退火优化DAS3VM参数的SAS3VM,并给出了求解过程.②在三组个人信用数据集和两组企业信用数据集上进行了实验,与已有算法相比,在准确率上比DAS3VM最大提升了13.108%.

1 相关工作

1.1 信用评估现状

企业信用评估主要用以分析企业的资产状况、负债偿还能力、发展前景、经济交往中的信用状况、经营管理情况及领导层水平等.个人信用评估主要用以分析个人的资产状况、职业类型、教育背景、银行信用、个人消费等.

我国中小企业由于自身原因,银行不轻易借贷.主要原因包括:①中小企业经营规模小,风险大,信用差;②资本实力弱,缺乏抵押担保资产;③中小企业管理制度不规范,财务报表失真等.

目前,信用评估的方法包括线性判别式分析,逻辑回归,遗传算法,神经网络,支持向量机(support vector machine, SVM).这些方法中神经网络和SVM的评估准确率较高.文献[4-5]研究表明,监督学习的SVM可以克服神经网络的结构选择和小样本泛化不足,因此更加适用于信用评估.

1.2 半监督学习现状

生成式(generative methods)半监督学习方法出现最早[6],模型假设每一类数据都服从某个混合分布,如高斯混合分布.一旦模型假设不正确,学习效果将很差,未标记数据的加入会进一步降低学习效率. Nigm 等[7]将生成模型用于文本分类, Baluja[8]将其用于人脸方向定位.

最早的图半监督学习方法(graph-based semi-supervised learning)由 Blum[9]提出,基于聚类假设,将学习目标看作找出图的最小分割(Mincut).基

于图的半监督关键点在于它的一致性假设:相邻的点具有相同的标签,具有同样结构的点具有同样的标签.图半监督学习方法的时间复杂度高,计算时间长.相关工作包括 Zhou 等[10]的局部全局一致性方法,Wang 等[11]对于构图的研究.

协同训练最早由 Blum 等[12]提出,假设数据特征集可以生成两个独立的学习分类器,利用对方的分类结果改进分类.Zhou 等[13]提出三体训练,使用三个学习器,通过"少数服从多数"产生伪标记样本,并将学习器集成.Jones[14]将此方法应用于文本中的信息检索,运用多视图数据,与主动学习相结合.

半监督支持向量机(semi-supervised support vector machines)是支持向量机在半监督学习上的推广,包括 S3VM、TSVM、CS4VM、meanS3VM 等方法,主要思想为:分类超平面同时将标记数据与未标记数据分隔开,且穿过数据稀疏区域.相关工作包括:Joachims[15]提出 TSVM 并应用于文本分类;Chapella 等[16]使用连续统方法,将优化目标从一个简单的凸目标函数开始,逐步变形为非凸的 S3VM 目标函数;Sindhwani 等[17]运用确定性退火(deterministic annealing)将非凸问题转化为一系列凸优化问题,逐步跟踪最优值;Belkin 等[18]通过图的拉普拉斯矩阵来探索数据的流行结构;Li 等[19]的 meanS3VM 通过使两个类别的样本的类别平均值之间的间隔最大化来提高效率.

由于 SVM 在信用评估中表现突出,因此本文主要研究半监督 SVM 方法.基于确定性退火的半监督 SVM 方法中参数取不同的值会造成模型过拟合或欠学习.本文提出基于模拟退火的半监督 SVM 来优化模型的参数选择.

2　基于模拟退火的半监督支持向量机

2.1　DAS3VM

直推式支持向量机(transductive support vector machine,TSVM)是支持向量机在半监督学习方向的扩展.由于 TSVM 的损失函数在未标记样本上是非凸的,这使得它不具有全局最优解,并且求解过程具有较高的时间复杂度.基于退火过程的确定性退火(deterministic annealing,DA)[17,20],通过构造一个自由能函数,将最优化过程转换为一系列温度依赖的物理系统.在某一温度下,系统总会朝自由能减少的方向进行,当自由能达到最小值时,系统达到平衡,因此 DA 可以避开局部最小,达到全局最小.

TSVM 利用未标记样本来驱动分类超平面通过真实的低密度数据区域[17,21].假设分类超平面穿过低密度数据区域,将数据划分为正例和负例.少量的标记样本或分布不完整的标记样本不能准确反映出数据样本的真实分布,但大量未标记样本的加入,丰富了样本,使样本分布趋向于真实分布;同时,分类超平面也要通过未标记样本的低密度区域,因此未标记样本可以辅助学习.

数据集为 l 个标记样本 $\{x_i,y_i\}_{i=1}^{l}$ 和 u 个未标记样本 $\{\acute{x}_j\}_{j=1}^{u}$,$x_i$,$\acute{x}_j\in R^n$,$y_i\in\{+1,-1\}$,$l\ll u$.我们的目标为同时使用标记样本和未标记样本构建分类器 $\omega^{\mathrm{T}}x$.

标准 TSVM 的优化目标如下:

$$\min_{\omega,\{\acute{y}_j\}_{j=1}^{u}}\frac{\lambda}{2}\|\omega\|^2+\frac{1}{2l}\sum_{i=1}^{l}l(y_i\omega^{\mathrm{T}}x_i)+\frac{\acute{\lambda}}{2u}\sum_{j=1}^{u}l(\acute{y}_j\omega^{\mathrm{T}}\acute{x}_j)$$

$$\text{s.t.}\quad\frac{1}{u}\sum_{j=1}^{u}\max[0,\operatorname{sign}(\omega^{\mathrm{T}}\acute{x}_j)]=r\qquad(1)$$

式中,使用 Hinge Loss $l(f(x))=\max(0,1-f(x))$,λ 是惩罚参数,$\acute{\lambda}$ 为用户提供的参数.$\acute{\lambda}$ 控制未标记样本数据的影响.如果将 $\acute{\lambda}$ 设为 0,式(1)就是标准 SVM 的优化目标函数.r 控制了未标记样本中的正、负样本比例,其初值可以从训练数据中获得,也可以通过实验来调整.

文献[22]详细说明了 TSVM 的优化过程.首先,使用标准 SVM 来训练数据,得到初始分类面,标记未标记样本.其次,指定临时因子 $\acute{\lambda}^*$,并迭代切换未标记样本的标签来最小化式(1).然后,均匀地增加 $\acute{\lambda}^*$ 的值,重新用SVM来训练,直到 $\acute{\lambda}^*>\acute{\lambda}$.最终优化问题得以解决,输出最终结果.

TSVM 目标函数重写为

$$\omega^*=\underset{\omega,u_j\neq-1}{\operatorname{argmin}}\frac{\lambda}{2}\|\omega\|^2+\frac{1}{2l}\sum_{i=1}^{l}l_2(y_i\omega^{\mathrm{T}}x_i)+\frac{\acute{\lambda}}{2u}\sum_{j=1}^{u}(u_jl_2(\omega^{\mathrm{T}}\acute{x}_j)+(1-u_j)l_2(-\omega^{\mathrm{T}}\acute{x}_j))\qquad(2)$$

式中,$u_j=\dfrac{1+y_j}{2}$,l_2 为 Hinge Loss 的平方,$l_2(f(x))=\max(0,1-f(x))^2$,如图 1 所示.通过 $\max[0,1-|\omega^{\mathrm{T}}x|^2]=\min[l_2(\omega^{\mathrm{T}}x),$

$l_2(-\omega^{\mathrm{T}}x)$]放松了未标记的惩罚.

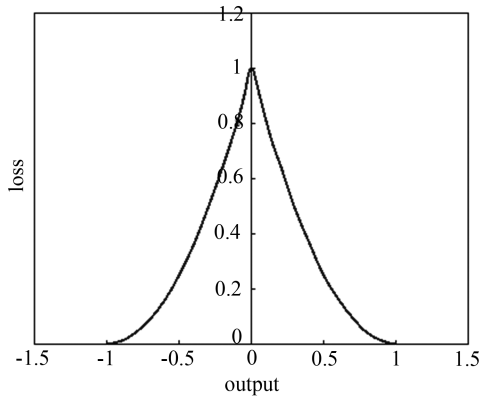

图 1 l_2 损失函数

Fig.1 Loss function of l_2

将式(2)进行重写为

$$\omega^{*T}=\underset{\omega,p_j}{\operatorname{argmin}}\frac{\lambda}{2}\|\omega\|^2+\frac{1}{2l}\sum_{i=1}^{l}l_2(y_i\omega^{\mathrm{T}}x_i)+\frac{\acute{\lambda}}{2u}\sum_{j=1}^{u}(p_jl_2(\omega^{\mathrm{T}}\acute{x}_j)+(1-p_j)l_2(-\omega^{\mathrm{T}}\acute{x}_j))+\frac{T}{2u}\sum_{j=1}^{u}(p_j\log(p_j)+(1-p_j)\log(1-p_j))$$

$$\text{s.t.}\quad \frac{1}{u}\sum_{j=1}^{u}p_j=r \tag{3}$$

式中,我们将二元变量 $u_j\in\{0,1\}$ 放大到概率变量 $p_j\in[0,1]$,且 p_j 表示未标记的样本 $\acute{x}_j$ 属于正例的概率. r 表示未标记样本中正例的比例,这与 TSVM 相同.式(3)加入了 DA 的损失函数,T 表示“温度”.如图 2 所示,图中的每条曲线表示在温度下的损失函数曲线,T 从一个很大的值开始,按规律下降,曲线逐渐变化至图 1 所示的损失函数.当 T 很大时,上式的最优问题很容易求解,当 T 降到 0 时,式(3)就是 TSVM 的优化函数.

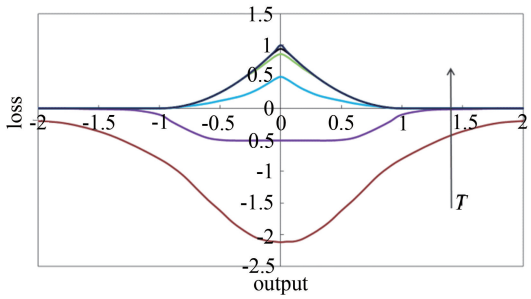

图 2 不同温度下的 DA 损失函数

Fig.2 DA loss function at different temperatures

2.2 SAS3VM

模拟退火(simulated annealing,SA)[23]是根据固体退火过程得到的一种参数优化算法.不同于确定性退火在每一温度 T 下用传统方式来求解最小自由能的方式,模拟退火在每一温度下引入 Metropolis 准则,以一定概率接受拟合精度偏低的参数.随着温度的下降,SA 逐渐趋向全局最优.

Metropolis 准则表示在温度 T 下,SAS3VM 接受新解的概率.新解(s_{new})的能量(E_{new})小于旧解(s_{old})的能量(E_{old}),则接受新解;旧解的能量小于新解的能量,则以概率接受新解,即

$$P(\text{accept } s_{\mathrm{new}})=\exp\left(-\frac{E_{\mathrm{new}}-E_{\mathrm{old}}}{kT}\right).$$

随机生成一个数 $P_{\mathrm{temp}}\in[0,1]$,若 $P>P_{\mathrm{temp}}$,则接受新解;否则不接受.

DAS3VM 公式(3)的超参数有两个,λ 和 $\acute{\lambda}$. 通常根据经验选取 $\lambda=0.001$,$\acute{\lambda}=1$.面对实际问题时,不同的取值会对结果产生较大的影响,造成过拟合或欠学习现象,因此可使用模拟退火优化 DAS3VM 模型的参数选择.本文中,λ 和 $\acute{\lambda}$ 的初始值按经验选取,即 $\lambda_0=0.001$,$\acute{\lambda}_0=1$.第 k 次的超参数的产生具体表达式为

$$\lambda_k=\lambda_{k-1}+\tan\left(\frac{\pi}{3}(r-0.5)\right)T_k.$$

式中,λ_k 表示第 k 次的超参数,T_k 是第 k 次迭代时的温度,$r\in[0,1]$ 是随机数.由此可知,λ_k 的分布受 T_k 影响,T_k 越大,则 λ_k 分布越广;反之,随着 T_k 的降低,λ_k 范围变窄,最优值也逐渐收敛. $\acute{\lambda}$ 的生成同理.温度生成规则为 $T_k=\rho T_{k-1}$,开始时,T_k 快速下降,随着迭代次数的增加,T_k 下降速度减缓.

优化问题通过 T 从一个很大的值递减到 0 来完成.在每个“温度”T 下,由扰动函数确定 λ 和 $\acute{\lambda}$ 的值,先固定式(3)的 p,用 L_2-SVM-MFN[24-26]优化 ω;再固定 ω,构造式(3)的拉格朗日函数.将 p_j 的偏导函数置 0,并用牛顿-拉夫森迭代法和二分法的混合组合优化 p_j [27];最后通过 Metropolis 准则确定是否接受新的参数,判断此时是否达到平衡状态或最大循环次数.

L_2-SVM-MFN 的优化目标如下:

$$\min_{\omega}f(\omega)=\frac{1}{2}\sum_{i\in I(\omega)}c_il_2(y_i\omega^{\mathrm{T}}x_i)+\frac{\lambda}{2}\|\omega\|^2 \tag{4}$$

式中添加了支持向量集合约束 $I(\omega)=\{i:y_i(\omega^{*\mathrm{T}})<1\}$ 和损失成本 c_i.同时,如果 $I(\omega)$ 中的 ω 对于数据集中的每个点都不依赖,此公式将会是 RLS(regularized least squares)的目标函数[17]. $f(\omega)$ 的梯度向量由 X 如下所示:

$$\nabla f(\omega)=\lambda\omega+X_{I(w)}\,C_{I(w)}\,[X_{I(\omega)}-Y_{I(\omega)}] \quad (5)$$

式中，$X_{I(w)}$ 是一个矩阵，它的行是训练样本相对于索引集合 $I(\omega)$ 的特征向量；$Y_{I(\omega)}$ 是一个包含了样本标签的列向量；$C_{I(w)}$ 是一个对角元素为这些样本的损失成本 c_i 的对角矩阵. f_I 的定义如下：

$$f_I(\omega)=\frac{\lambda}{2}\|\omega\|^2+\frac{1}{2}\sum_{i?\ I}c_i l_2(y_i\omega^{\mathrm{T}}x_i) \quad (6)$$

显然，f_I 是一个严格的凸二次函数，具有唯一最小值.

文献[28]运用牛顿法进行迭代更新 ω 如下：

$$\omega_{k+1}=\omega_k+\delta_k n_k \quad (7)$$

式中，$\delta_k\in R$，搜索方向 $n_k\in R^n$，n_k 由 ω_k 给出：

$$n_k=-\nabla f(\omega_k)/\nabla^2 f(\omega_k) \quad (8)$$

式中，$\nabla f(\omega_k)$ 是梯度向量，$\nabla^2 f(\omega_k)$ 是 f 在 ω_k 处的海森矩阵(Hessian matrix). L_2-SVM-MFN 伪码描述如算法 2.1 所示，L_2-SVM-MFN 将在有限次迭代后收敛，这在文献[24]中已被证明.

算法 2.1　L_2-SVM-MFN

输入：训练集 $\{x_i,y_i\}_{i=1}^{l}$

输出：优化结果 ω

初始化一个合适的 ω_0，k=0
while(true)
　if ω_k是(4)的最优解
　then
　　exit with ω_k
　end then
　end if
　$I_k=I(\omega_k)$，$\bar{\omega}=\arg\min_{\omega} f_{Ik}(\omega)$
　$L=\{\omega=\omega_k+\delta(\bar{\omega}-\omega_k):\delta>0\}$
　$\delta^*=\arg\min_{\omega\in L} f(\omega)$
　$\omega_{k+1}=\omega_k+\delta^*(\omega-\omega_k)$
　k==k+1
end while

模拟退火半监督支持向量机的具体过程如下：

(a)确定 λ 和 $\acute{\lambda}$ 的范围，给出 λ_0、$\acute{\lambda}_0$ 和 T_0.

(b)固定 p，用 L_2-SVM-MFN 优化 ω；固定 ω，构造式(3)的拉格朗日函数，将 p_j 的偏导函数置 0，并用牛顿-拉夫森迭代法和二分法的混合组合来优化 p_j. 计算此时的系统能量 E_{k-1}.

(c)通过扰动产生新的超参数，此时状态为 s_k，通过步骤(b)计算系统能量 E_k. 通过 Metropolis 准则概率接受 s_k. 重复此步骤直到系统达到平衡状态，并设置最大内循环次数.

(d)按照降温函数降低温度 T，回到步骤(b). 重复此步骤直到达到终止条件. 终止条件为系统此时自由能已稳定，或已达到最大外循环次数.

3　实验过程与结果分析

我们对 5 组信用数据进行了监督学习方法和半监督学习方法的实验研究，实验数据包括两组爬取的中国企业信用数据(credit-one，credit-two)和三组公开的个人信用数据(German，Australian，Japanese).

3.1　实验数据

个人信用数据来源于 UCI[①]，分别为 German，Australian，Japanese. 表 1 显示了三组数据集的细节. $l+u$ 项为标记数据样本(l)与未标记数据样本(u)的总和. Ratio 项为数据集中正样本所占比例；f 项为数据集属性数. 属性包括个人信息和信用信息.

表 1　个人信用数据集

Fig.1　Personal credit data sets

数据集	$l+u$	Ratio	f
German	1 000	0.700	24
Australian	690	0.445	14
Japanese	690	0.445	15

考虑到 UCI 数据集的时效性，我们同时收集了一些中国企业信用信息. 企业信用数据从国家企业信用信息公示系统[②]和阿里巴巴[③]爬取. 我们已经收集了几百万条的企业基本信息、产品评论信息等，并随机挑选生成两个不同数据集，如表 2 所示.

由于收集的数据不完整，一些企业的基本信息缺失，因此我们在已收集的企业完整数据中，随机挑选具有信用记录的企业作为实验数据. 为了避免一些属性过大而覆盖其他属性，进行归一化处理.

表 2　企业信用数据集

Fig.2　Enterprise credit data sets

数据集	$l+u$	Ratio	f
credit-one	4 255	0.674	17
credit-two	2 589	0.464	17

3.2　实验方法

本次实验采用反 k 折交叉验证的方式，可以有效地避免过拟合现象，使结果更加真实准确. 反 k 折交叉验证是一种类似于 k 折交叉验证的方法，但训练数据和测试数据的数量相反. 具体来说，反 k 折交叉验证将数据集划分为 k 组，每一组轮流当训练集，剩下的 $k-1$ 组作为测试集，实验结果为这 k 次实

① UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/
② http://www.gsxt.gov.cn/index.html
③ https://s.1688.com/company/company search.htm

验的平均值.

本文比较了监督学习算法,即 RLS 和 SVM (L_2-SVM-MFN),半监督学习算法,即 TSVM(使用 L_2-SVM-MFN)、DAS3VM(使用 L_2-SVM-MFN)、NBEM[29](naive Bayes EM algorithm)、HF[30]和本文提出的 SA(使用 L_2-SVM-MFN 的 SAS3VM).对于每组数据集,我们都设定 $k=5,10,20,30,40,50,60,70,80,90,100$.

本文的评测标准包含分类的准确率(Accuracy),正例的精度(Precision),正例的召回率(Recall),F-1 值(F-1 measure)和计算运行时间(computation runtime).

3.3 实验结果

对两个监督学习算法和 5 个半监督学习算法在 5 个不同数据集上进行实验.对于每个实验,我们通过控制训练样本和测试样本的不同比值,即用不同的 k 显示性能变化.由于实验设置采用反 k 折交叉验证,因此测试数据将大于训练数据,这与现实生活中的问题相匹配.

3.3.1 准确率对比试验

图 3 是不同 k 值下的企业数据集的准确率.

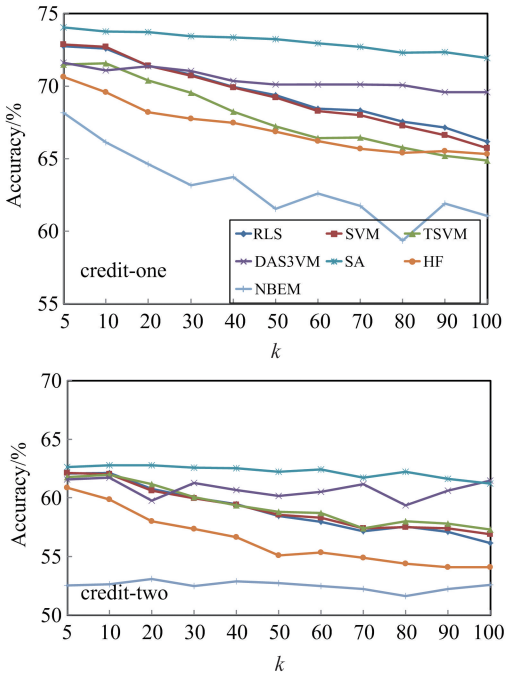

图 3 对 credit-one 和 credit-two 在不同 k 值下的准确率

Fig.3 Accuracy of credit-one and credit-two at different k

随着 k 的增加,训练数据样本(标记样本)数量逐渐减少,测试数据样本(未标记样本)数量逐渐增加.在图 3 中,本文提出的 SA 算法的准确率曲线最为平滑,变化较小,准确率最高.DAS3VM 和 SA 这两个半监督算法表现明显优于其他算法.在 k 比较小时($k=5,10,20$),这两个算法优势并不明显,但随着 k 的增加,两者的性能与其他算法的差距越来越明显.当 $k=5,10,20$ 时,标记样本已足够表现数据的大部分分布特征,因此未标记样本的加入并不会使算法结果有较大的提升;当 k 值较大时,标记样本减少,使得监督学习并不能从少量的标记样本中学习得到正确的模型.以 credit-one 为例,当 $k=50$ 时,训练数据样本只有整个数据集的 2%,即只有 85 个标记样本,而剩下的 4 170 个样本作为未标记的测试样本.在这种情形下,半监督的 DAS3VM 和 SA 显然比其他算法更好.未标记样本的使用,能够辅助学习,弥补了标记样本不足的缺陷.

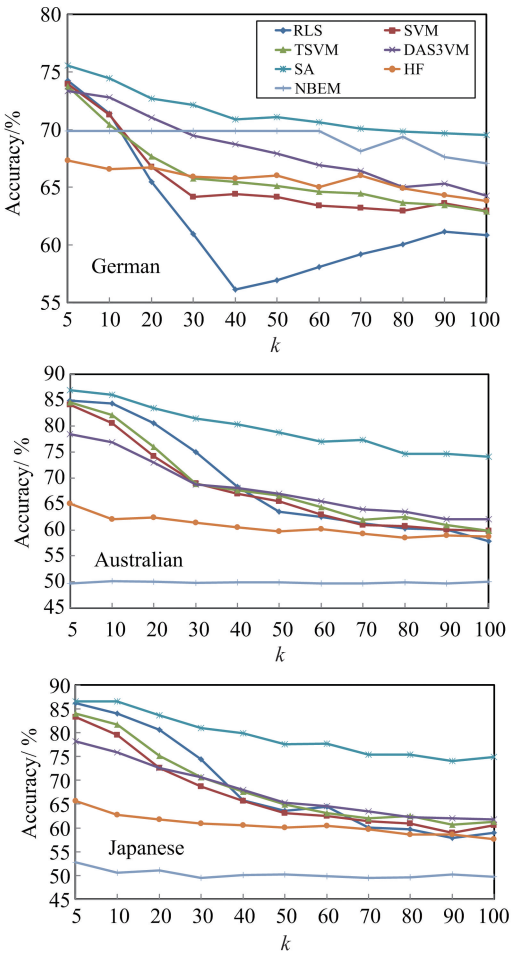

图 4 German, Australian 和 Japanese 在不同 k 值下的准确率

Fig.4 Accuracy of German, Australian and Japanese at different k

图 4 显示了 3 个国家个人信用数据集上的实验结果.与图 3 相似,当训练数据样本数量远小于测试样本数量时,本文提出的 SA 的性能最优.在

German 数据集上，NBEM 要优于 DAS3VM；在 Australian 和 Japanese 数据集上，随着 k 的增大和训练样本的减少，DAS3VM 不如 SA，与其他算法相比也没有明显优势.

3.3.2　召回率对比实验

图 5 显示了 5 种算法在企业信用数据上的正例召回率；图 6 显示了个人信用数据上的召回率.在 German 以外的数据集上，本文提出的 SA 的召回率最高.在 German 数据集上，NBEM 的召回率接近 100%，但是其正确率在 70%左右，这与 German 中正例比(0.7)相同，因此我们认为此方法将全部的未标记样本都分为正例，才会有接近 100%的召回率，说明此方法对未标记数据的处理比较简单，在不同数据集上表现不一致.当 k 较小时，训练样本较多，SA 和 DAS3VM 并不具有优势.随着 k 的增加，标记样本减少，SA 的波动不大且召回率最高.

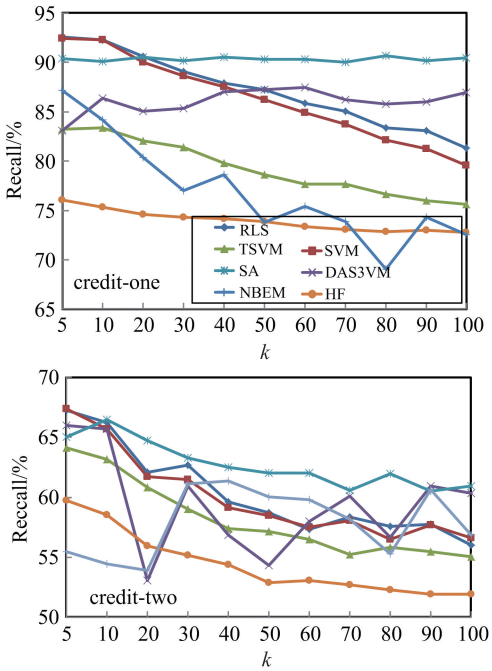

图 5　credit-one 和 credit-two 在不同 k 值下的正例样本召回率

Fig.5　Recall of positive samples of credit-one and credit-two at different k

3.3.3　精度对比试验

图 7 和图 8 显示了正例样本的精度.SA 在 Australian 和 Japanese 数据集上具有绝对优势，HF 方法表现最差，相差 30%左右.虽然 HF 方法在 credit-one，credit-two 和 German 数据集上具有一定的优势，SA 次之，但是两者仅相差 4%左右；图 7 中，当 k 大于 30 后，SA 的精度具有明显优势，与 HF 方法相差不大，曲线平滑，实验效果较好.DAS3VM 的精度虽然与 SA 相差无几，但曲线波动较大.在 Australian 和 Japanese 上，SA 与其他方法优势明显.综合来看，SA 的效果在不同数据集和不同指标下表现最优.

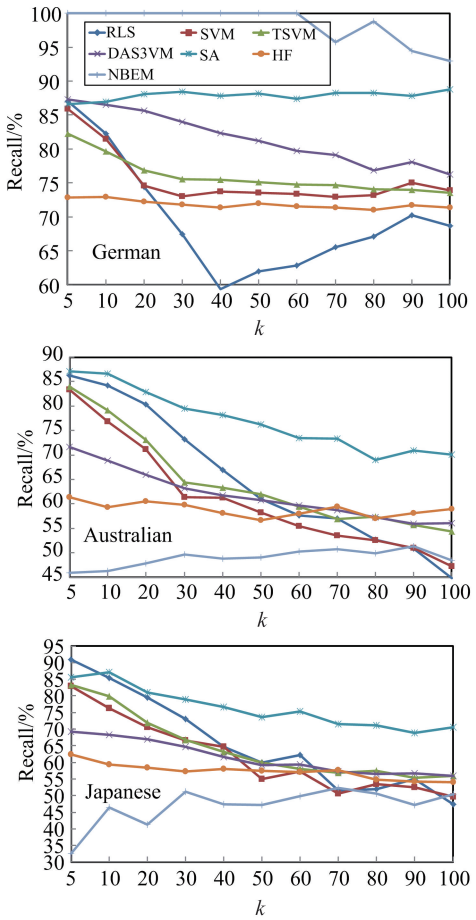

图 6　German，Australian 和 Japanese 在不同 k 下正例样本的召回率

Fig.6　Recall of positive samples of German, Australian and Japanese at different k

3.3.4　F-measure 对比试验

图 9 和图 10 显示了不同 k 值下的 F-1 值.F-1 值融合了精度和召回率，能够更加准确地评价算法的性能.如上所述，随着 k 的增加，本文提出的 SA 的性能明显优于其他算法，DAS3VM 的性能次之，DAS3VM 在 credit-one，credit-two 和 German 数据集上的性能明显优于其他算法，在 Australian 和 Japanese 数据集上的表现也与其他算法近似.总体来看，SA 在信用预测中表现出高性能且稳定.特别是当测试数据远小于训练数据时，传统的监督学习算法不能很好地处理这种情况.

图 11 中，我们对 SA 在 credit-one，credit-two，

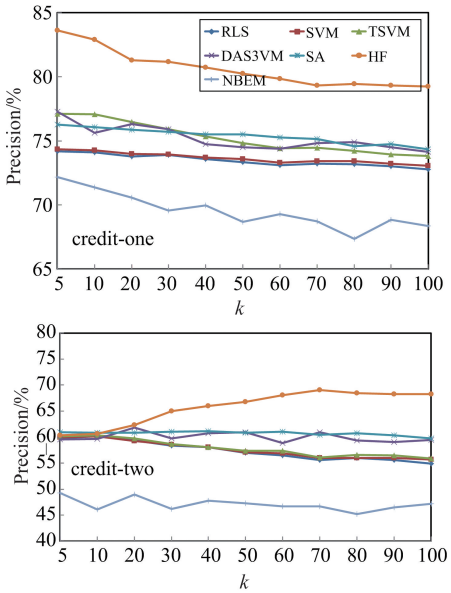

图 7 credit-one 和 credit-two 在不同 k 下的正例样本精度

Fig.7 Precision of positive samples of credit-one and credit-two at different k

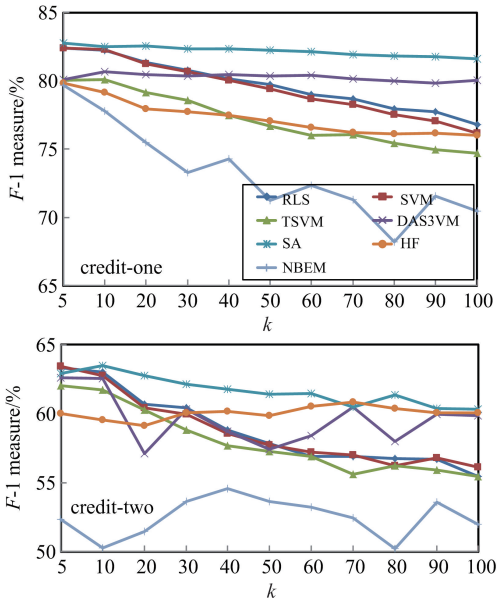

图 9 credit-one 和 credit-two 的正例样本在不同 k 下的 F-1 值

Fig.9 F-1 of positive samples of credit-one and credit-two at different k

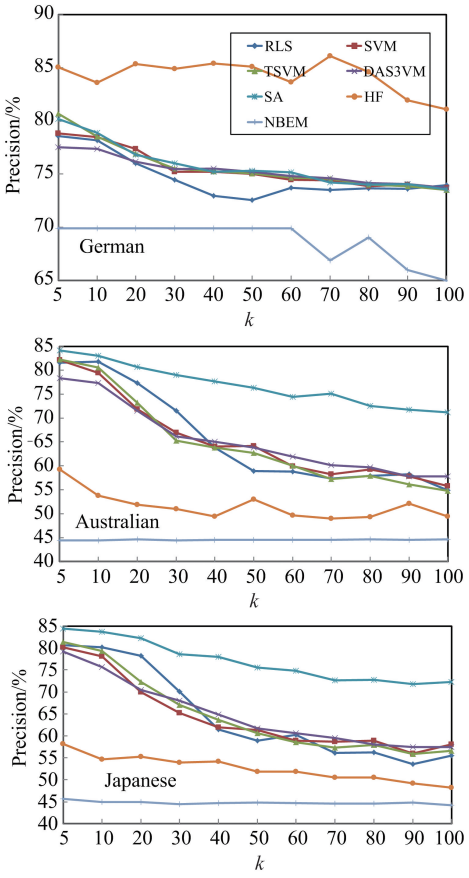

图 8 German，Australian 和 Japanese 的正例样本在不同 k 下的精度

Fig.8 Precision of positive samples of German, Australia and Japanese at different k

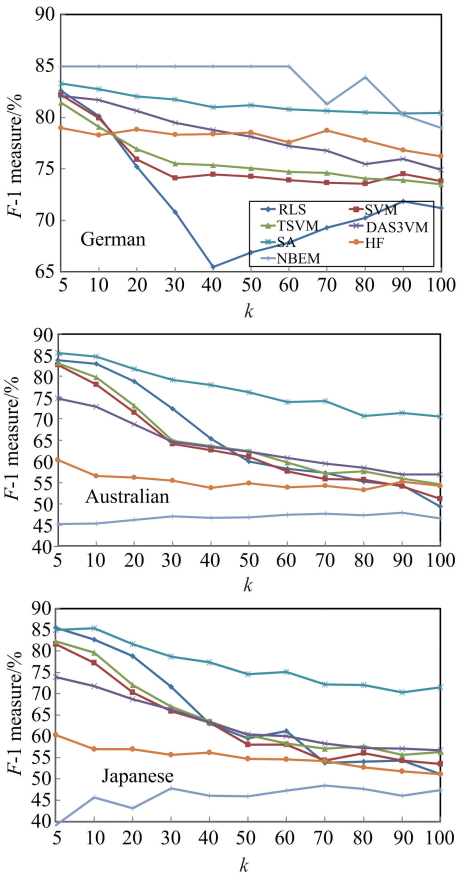

图 10 German，Australian 和 Japanese 的正例样本在不同 k 下的 F-1 值

Fig.10 F-1 of positive samples of German, Australia and Japanese at different k

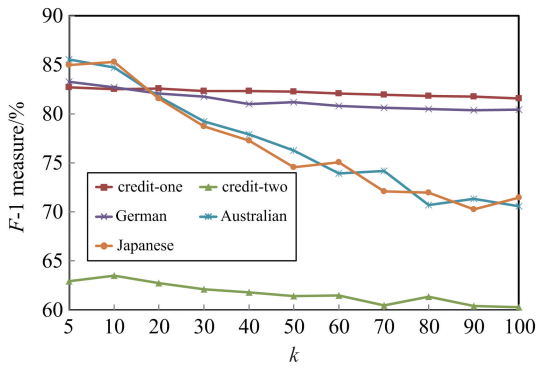

图 11　在不同 k 下，SA 对 credit-one，credit-two，German，Australian 和 Japanese 数据集正例样本的 F-1 值

Fig.11　F-1 of positive samples of German, Australia and Japanese for credit-one and credit-two with SA at different k

German，Australian 和 Japanese 数据集上的 F-1 值进行了比较.credit-one，credit-two 和 German 的曲线整体上一直很平坦、稳定.即使在标记样本数量差距巨大时(当 $k=10$，只有 10%的标记样本，当 $k=100$，只有 1%的标记样本)，变化也非常小；当 $k>50$，Australian 和 Japanese 的走向也趋于平缓.由此可认为 SA 是一个相对稳定和良好的信用预测算法.

3.3.5　计算时间对比试验

表 3～7 显示了本次试验中的 5 个半监督算法的平均运行时间.从表中可以看出，不论在什么数据集上，不论 k 值大小，SA 的计算时间最长，DAS3VM 次之，监督学习算法时间最短.确定性退火和模拟退火都有一个退火的过程，因此计算量会增加，计算时间相应增长.

表 3　credit-one 计算时长(s)

Tab.3　Computing time of credit-one(s)

k	TSVM	DAS3VM	SAS3VM	HF	NBEM
5	1.331 1	19.040 6	277.410 0	0.945 0	0.197 0
10	2.109 6	27.031 1	160.400 0	1.599 0	0.196 0
20	2.675 8	16.188 9	140.671 0	1.702 0	0.166 0
30	1.556 8	18.055 6	154.520 0	1.861 0	0.180 0
40	2.077 7	19.922 3	180.810 0	2.021 0	0.172 2
50	2.366 4	15.569 0	198.355 0	2.042 0	0.198 0
60	1.426 5	14.872 2	172.594 0	2.152 0	0.196 0
70	1.291 8	11.903 7	164.757 0	2.389 0	0.195 0
80	1.136 7	11.017 7	240.143 0	2.293 0	0.207 0
90	1.139 9	11.239 4	150.613 0	2.314 0	0.223 0
100	1.051 0	11.509 1	156.131 0	2.060 0	0.197 0

表 4　credit-two 计算时长(s)

Tab.4　Computing time of credit-two(s)

k	TSVM	DAS3VM	SAS3VM	HF	NBEM
5	0.818 8	6.014 1	104.254 0	0.519 2	0.100 0
10	0.765 8	5.114 9	86.259 0	0.652 0	0.115 9
20	0.800 6	6.330 6	115.385 0	0.699 6	0.119 1
30	0.790 9	7.459 3	136.060 3	0.716 1	0.130 9
40	0.703 7	6.805 4	121.958 6	0.728 4	0.117 5
50	0.639 2	5.152 7	91.548 5	0.737 5	0.130 5
60	0.624 3	6.291 2	95.470 5	0.753 4	0.154 0
70	0.597 2	4.882 4	99.411 3	0.775 6	0.198 5
80	0.578 7	6.372 5	113.565 1	0.773 4	0.131 0
90	0.545 1	5.293 8	134.945 0	0.793 3	0.146 3
100	0.498 4	5.599 9	100.730 0	0.743 9	0.120 3

表 5　German 计算时长(s)

Tab.5　Computing time of German(s)

k	TSVM	DAS3VM	SAS3VM	HF	NBEM
5	0.199 1	0.456 1	20.185	0.197	0.105
10	0.158 2	0.549 5	30.64 6	0.162	0.095
20	0.101 1	0.563 7	38.980	0.136	0.063
30	0.093 8	0.684 4	24.753	0.130	0.059
40	0.101 4	0.714 1	17.605	0.135	0.058
50	0.115 8	0.664 2	20.491	0.146	0.082
60	0.118 7	0.801 7	19.95 0	0.129	0.082
70	0.199 8	0.819 7	18.910	0.127	0.084
80	0.111 7	0.823 1	19.90 2	0.122	0.069
90	0.231 1	0.896 7	19.544 0	0.121	0.107
100	0.190 9	0.835 8	17.259 0	0.159	0.073

表 6　Australian 计算时长(s)

Tab.6　Computing time of Australian(s)

k	TSVM	DAS3VM	SAS3VM	HF	NBEM
5	0.061 2	0.741 8	18.205 0	0.182 2	0.054 6
10	0.055 9	0.588 9	11.788 4	0.112 2	0.058 3
20	0.044 0	0.864 3	24.963 6	0.111 6	0.055 8
30	0.041 3	1.898 6	28.988 5	0.108 5	0.058 2
40	0.041 8	1.213 3	31.235 0	0.107 4	0.063 5
50	0.040 4	1.415 1	26.000 7	0.110 1	0.064 1
60	0.043 9	1.101 3	35.550 2	0.126 2	0.069 1
70	0.135 0	1.927 0	25.104 7	0.221 3	0.069 2
80	0.138 5	2.646 1	28.565 8	0.224 6	0.070 3
90	0.120 6	2.131 7	15.884 6	0.112 2	0.071 8
100	0.124 1	2.066 3	18.438 1	0.110 8	0.054 3

表 7 Japanese 计算时长(s)

Tab.7 Computing time of Japanese(s)

k	TSVM	DAS3VM	SAS3VM	HF	NBEM
5	0.070 0	0.555 8	17.519 2	0.288 9	0.108 4
10	0.052 8	1.412 7	15.322 4	0.147 3	0.072 2
20	0.046 3	1.372 0	27.869 6	0.127 2	0.062 7
30	0.044 4	1.382 1	18.734 5	0.121 3	0.054 8
40	0.044 8	1.281 9	19.142 7	0.116 5	0.076 8
50	0.148 5	1.973 8	19.467 8	0.115 4	0.049 9
60	0.047 5	3.442 8	19.286 1	0.114 4	0.050 0
70	0.140 9	2.699 0	18.420 1	0.113 3	0.052 0
80	0.144 2	2.115 9	11.179 9	0.112 8	0.086 8
90	0.134 5	2.512 6	18.309 2	0.111 7	0.073 8
100	0.040 9	2.279 0	17.369 1	0.111 8	0.071 3

4 结论

本文针对企业信用数据和个人信用数据的信用预测提出了一种半监督算法.在 5 组实验数据上的实验结果表明,本文提出的基于模拟退火的半监督算法在信用数据集上总体表现最优,特别是在标记样本远小于未标记样本的情况下.为了了解信用数据的哪些属性或属性子集可以显示更好的性能,我们将研究特征选择和降维技术,进一步提高信用预测的性能.同时,对于退火阶段的时间过长的缺点,我们将引入耦合模拟退火,加快温度 T 下达到平衡状态的速度,减少总体运行时间.

参考文献(References)

[1] EISENBEIS R A. Recent developments in the application of credit-scoring technique to the evaliation of commercial loan[J]. IMA Journal of Management Mathematics, 1996, 7(4): 271-290.

[2] BERGER A N, FRAME W S. Small business credit scoring and credit availability[J]. Journal of Small Business Management, 2007, 45(1): 5-22.

[3] BERGER A N, UDELL G F. Small business credit availability and relationship lending: The importance of bank organizational structure[J]. The Economic Journal, 2002, 112(477): F32-F53.

[4] HUANG C L, CHEN M C, WANG C J. Credit scoring with a data mining approach based on support vector machines[J]. Expert Systems With Applications, 2007, 33(4): 847-856.

[5] BELLOTTI T, CROOK J. Support vector machines for credit scoring and discovery of significant features[J]. Expert Systems with Applications, 2009, 36(2): 3302-3308.

[6] SHAHSHAHANI B M, LANDGREBE D A. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. IEEE Transactions on Geoscience and Remote Sensing, 1994, 32(5): 1087-1095.

[7] NIGAM K, GHANI R. Analyzing the effectiveness and applicability of co-training[J]. Proceedings of the Ninth International Conference on Information and Knowledge Management. New York, NY, USA: ACM, 2000.

[8] BALUJA S. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data[C]// Proceedings of the 11th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 1998: 854-860.

[9] BLUM A, CHAWLA S. Learning from labeled and unlabeled data using graph mincuts[C]// Proceedings of the Eighteenth international Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001: 19-26.

[10] ZHOU D Y, BOUSQUET O, LAL T N, et al. Learning with local and global consistency[C]// Proceedings of the 16th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2003: 321-328.

[11] WANG F, ZHANG C S. Label propagation through linear neighborhoods[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(1): 55-67.

[12] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the Eleventh Annual Conference on Computational Learning Theory. New York, NY, USA: ACM, 1998: 92-100.

[13] ZHOU Z H, LI M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[14] JONES R. Learning to extract entities from labeled and unlabeled text[D]. Pittsburgh, PA, USA: Carnegie Mellon University, 2005.

[15] JOACHIMS T. Transductive inference for text classification using support vector machines[C]// Proceedings of the Sixteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999: 200-209.

[16] CHAPELLE O, CHI M, ZIEN A. A continuation

method for semi-supervised SVMs[C]// Proceedings of the 23rd International Conference on Machine Learning. New York, NY, USA: ACM, 2006: 185-192.

[17] SINDHWANI V, KEERTHI S S, CHAPELLE O. Deterministic annealing for semi-supervised kernel machines [C]// Proceedings of the Twenty-Third International Conference on Machine Learning. New York, NY, USA: ACM, 2006: 841- 848.

[18] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples[J]. Journal of Machine Learning Research, 2006,7: 2399-2434.

[19] LI Y F, KWOK J T, ZHOU Z H. Semi-supervised learning using label mean[C]// Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA:ACM, 2009:633-640.

[20] ROSE K. Deterministic annealing for clustering, compression, classifi- cation, regression, and related optimization problems[J]. Proceedings of the IEEE, 1998, 86(11): 2210-2239.

[21] BENNETT K P, DEMIRIZ A. Semi-supervised support vector machin-ES[C]// Proceedings of the 11th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 1998: 368-374.

[22] CHAPELLE O, ZIEN A. Semi-supervised classification by low density separation [C]// Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, 2005:57-64.

[23] BERTSIMAS D, TSITSIKLIS J. Simulated annealing [J]. Statistical Science, 1993, 8(1): 10-15.

[24] KEERTHI S S, DECOSTE D. A modified finite newton method for fast solution of large scale linear SVMs[J]. Journal of Machine Learning Research, 2005, 6: 341-361.

[25] CHAPELLE O, SINDHWANI V, KEERTHI S S. Optimization techniques for semi-supervised support vector machines[J]. Journal of Machine Learning Research, 2008, 9: 203-233.

[26] CHAPELLE O, SCHOLKOPF B, ZIEN A. Semi-supervised learning[J]. IEEE Transactions on Neural Networks, 2009, 20(3): 542-542.

[27] SINDHWANI V, KEERTHI S S. Large scale semi-supervised linear SVMs[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM, 2006: 477-484.

[28] MANGASARIAN O L. A finite newton method for classification[J]. Optimization Methods and Software, 2002, 17(5): 913-929.

[29] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1):1-38.

[30] ZHU X, GHAHRAMANI Z, LAFFERTY J. Semi-supervised learning using Gaussian fields and harmonic functions[C]// Twentieth International Conference on International Conference on Machine Learning. AAAI Press, 2003:912-919.