文章编号:1000-6788(2007)07-0105-06

# SVM 方法及其在客户流失预测中的应用研究

应维云[1],覃　正[2],赵　宇[3],李　兵[3],李　秀[3]
(1. 西安交通大学 管理学院,西安 710049;2. 上海财经大学 信息管理与工程学院,上海 200433;
3. 清华大学 国家 CIMS 工程研究中心,北京 100084)

**摘要**:　客户流失分析与预测是客户关系管理的重要内容.针对客户流失问题,建立了支持向量机预测模型.针对实际客户流失数据中正负样本数量不平衡而且数据量大的特点,提出带有不同类权重参数的支持向量机算法 CW-SVM,通过调整类权重参数改变分类面位置,提高算法分类准确性;将标准支持向量机训练问题转化为运算效率更高的核向量机问题,提出处理不平衡海量数据集的 CWC-SVM 算法.通过实际银行信贷客户数据集测试,该算法与传统预测算法比较,更适合解决大数据集和不平衡数据,取得较好的客户流失预测效果.

**关键词**:　客户流失;支持向量机;客户关系管理;预测
**中图分类号**:　TP18;F270　　**文献标志码**:　A

# Support Vector Machine and Its Application in Customer Churn Prediction

YING Wei-yun[1], QIN Zheng[2], ZHAO Yu[3], LI Bing[3], LI Xiu[3]
(1. School of Management, Xi'an Jiaotong University, Xi'an 710049, China; 2. The School of Information Management & Engineering, Shanghai University of Finance & Economics, Shanghai 200433, China; 3. National CIMS Engineering & Research Center, Tsinghua University, Beijing 100084, China)

**Abstract**: Customer churn analysis and prediction play an important role in customer relationship management and improve benefit of enterprise. A Support Vector Machine model is established to predict customer churn. Customer churn characteristic is presented in this paper. According to the churn data which is large scale and imbalance, this paper presents a two-class model based on improved SVM to predict customer churn. The class weighted SVM model CW-SVM is presented, and the accuracy is improved by adjusting the class weight and the position of boundary. The efficiency is improved by translating the SVM to the Core Vector Machine and a new algorithms CWC-SVM is presented. The arithmetic performance is better than others based on the test of real credit debt data set in the commercial bank.

**Key words**: customer churn; support vector machine; customer relationship management; prediction

## 0　引言

客户流失分析和客户保持是 CRM 的重要组成部分.在很多行业,例如出版业、投资服务业、保险业、电子设备行业、医疗保健行业、信用卡、银行业、互联网服务行业、电话电信行业等,客户保持对公司的利润底线有着惊人的影响,远远超过公司规模、市场份额、单位成本和其它许多通常认为与竞争优势有关的因素的影响[1].一个小的客户保持率的提高都能导致利润可观的改善.对美国 9 个行业的调查数据表明,客户保持率增加 5%,行业平均利润增加幅度在 25%～85%之间.保持现有客户比获取新客户的成本低得多,一般可节约 4～6 倍.客户流失预测已成为公司成功最至关重要的目标[2].

国外已经有学者针对客户流失问题,应用决策树、神经元网络、进化算法等建立了预测模型,并取得了一定应用效果[3,4].但是这些模型的准确率不尽理想,其处理大数据集的能力也有待提高.针对客户流失数据的特点,本文建立支持向量机(Support Vector Machine,SVM)模型来进行预测.通过分析客户流失问题

**收稿日期**:2006-03-08
**资助项目**:国家自然科学基金(70671059)
**作者简介**:应维云(1971-),男,西安交通大学管理学院博士研究生,主要从事商业智能和决策支持系统的研究,E-mail: yingwy@china.com.

的特点,对 SVM 方法进行改进,提出了处理不平衡数据集的 Class Weighted SVM(CW-SVM)方法和处理不平衡大数据集的 Class Weighted Core vector machine SVM(CWC-SVM)方法. 通过对银行实际客户数据分析与试验,取得了比其它算法更好的预测效果. 同时,不平衡大数据集在现实世界中广泛存在,因此本文提出的方法有广泛的实际应用意义.

## 1 客户流失分析问题

### 1.1 客户流失与客户挽留

客户流失是指企业原来的客户中止继续购买企业商品或接受企业服务,转而接受竞争对手商品或服务. 而预测客户流失,采取相应的挽留手段,则称作客户保持.

在实际行业中,有两种基本手段避免客户流失. 无特定目标方法和有特定目标方法. 无特定目标方法通过用最好的产品、大量的广泛的广告来提高客户满意度和忠诚度,从而避免流失. 一个典型的例子就是 AOL 美国在线网站通过升级自身软件和提供更丰富精彩的内容来降低客户流失[5]. 有目标方法通过预测哪些客户有可能流失,并通过提供激励措施或特定服务来避免客户流失. 例如电信行业或银行业对易流失客户提供"套餐"服务或一定折扣来避免客户流失. 为了避免给那些本来不会流失的客户提供不必要的特定服务和折扣,同时避免在通知这些客户上花费的代价,需要根据历史数据,训练一个预测模型,来预测哪些客户可能流失,从而对这些客户采取措施.

### 1.2 客户流失预测分析框架

考虑到客户流失对企业危害的严重性,企业的一个重要任务就是识别哪些用户可能是流失者(churner),哪些客户是忠诚客户,而数据挖掘技术可以辅助实现该功能. 通过建立预测模型,企业可以针对有流失可能的用户,及时采取措施,尽可能的实现客户挽留,降低客户流失率. 同时,企业可以通过预测模型,识别出哪些因素可能是导致客户流失的主要原因. 图 1 是客户流失预测分析框架图,图中的 1、2、3、4 表示预测工作的执行顺序. 训练数据和试验数据中包括多个属性,这些属性可以分为客户静态属性和客户动态交易属性两大类. 预测模型一般采用决策树、回归模型、神经元网络模型等,而本文提出用支持向量机作为预测模型. 训练数据输入预测模型后,经过训练,可以得到模型的实例. 试验数据输入模型实例,就可以得到最终的预测结果. 根据预测需要,可以预测客户是否会流失(If)、什么时候流失(When)以及为什么会流失(Why);本文重点研究其中的一部分内容,即建立基于改进 SVM 的预测模型,利用客户静态属性数据,来预测客户是否会流失,为决策者进一步分析提供依据.

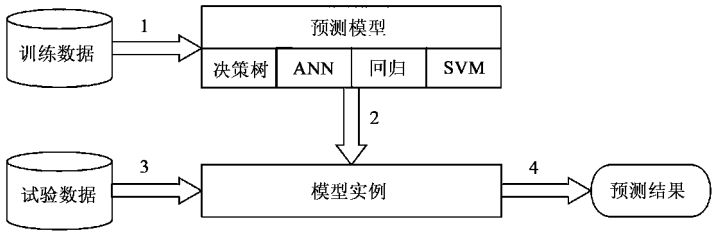


图 1 客户流失预测分析框架图

### 1.3 客户流失问题的特点

客户流失预测一般采用分类模型. 客户流失预测问题有其自身特点:首先,它是一个二分类问题;第二,分类结果要有较高的准确性. 有文献报道,对于某电信公司的 150 万用户,每提高 1 %预测准确性,就可以为企业增收 54 万美元. 但是,已有方法的准确性不能满足实际需要. 第三,客户数据集极端不平衡,流失客户和非流失客户样本数目相差几十到上百倍,传统算法对于这种不平衡数据集的应用效果不尽理想;第四,数据集样本量大,是海量数据集,且样本维数相对较高,神经元网络方法在处理这样大规模数据时需要过长的训练时间. 总之,针对解决客户流失问题,以往的方法还有不足.

## 2 改进 SVM 方法

### 2.1　SVM 简介

支持向量机方法建立在计算学习理论的结构风险最小化原则上,其主要思想是针对二分类问题,在高维空间中寻找一个超平面作为两类的分割,以保证最小的分类误差. SVM 方法非常适合于解决高维、非线性的二分类问题,同时由于其原理上的优势,它的分类准确性要好于传统的决策树等分类方法. 目前,国际上对 SVM 的讨论和研究逐渐广泛,已在模式识别、函数逼近、数据挖掘和非线性系统控制中都得到了很好的应用. 近来,该方法也应用到金融领域,主要是对时间序列的预测与分类,也有学者用 SVM 方法进行了信用评级方面的研究[7]. 而将 SVM 应用于客户流失分析的研究文献极少,主要原因如下:

现实世界中普遍存在非平衡数据,比如基因图谱、医疗诊断、信用卡欺诈检测、客户流失分析等(正负样本数目相差几十倍以上的数据集). 这些数据包含相对极少量的正样本,但是检测出它们又十分必要. 当处理这类数据时,SVM 分类面受到样本分布不均的影响,偏向于正样本,造成 SVM 算法有效性大大下降[6].

SVM 算法复杂度是很大的. 如果训练样本数为 $m$,则 SVM 算法的训练时间复杂度为 $O(m^3)$,空间复杂度至少为 $O(m^2)$. 而现实应用中都是大数据集. 为了降低算法时间和空间复杂度,一些改进算法相继提出. 例如:提高 SVM 的计算速度,以便于处理大规模问题,如分解算法的代表"序列最小化算法"等;利用最优化技术改进支持向量机形式,简化计算过程,如线性 SVM、LS-SVM 等[7];减少训练数据集,例如 boosting 算法和 SVM 算法结合[8]. 但针对非线性大数据集,这些算法运算效率还有待提高.

文献[9]提出了一种核向量机(Core Vector Machine,简称 CVM)方法,它在本质上属于分解算法. 传统的分解算法在训练时,训练向量会多次进出工作向量集合. 而 CVM 方法找到了最小闭包球(minimum enclosing ball,简称 MEB)问题与 SVM 问题的等价途径,通过 MEB 方法来解决 SVM 问题,使得训练向量不会反复进出工作向量集合,从而使时间、空间算法复杂度大大降低. 但是,这个算法不能有效处理非平衡数据.

### 2.2　处理不平衡数据的 CW-SVM 算法原理

Vapnik[10]提出了第一种支持向量机,也称为 C-SVM 或标准支持向量机,但 C-SVM 没有考虑样本数不平衡问题. 本节提出一种扩展的 C-SVM ——分类加权支持向量机(Class Weighted Support Vector Machine,简写为 CW-SVM). CW-SVM 通过设置不同类的权重,可以调整分类面位置,在一定程度上克服 C-SVM 对不平衡数据分类误差大的缺点.

CW-SVM 的原始问题定义为

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_i\sum_{i=1}^{m}\xi_i,$$

$$\text{s.t}\quad y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, i = 1, \cdots, m.$$

其中,$\mathbf{w}^T\phi(\mathbf{x}_i) + b = 0$ 为所要求解的超平面,$\mathbf{w}$ 是超平面的法向量,$b$ 是超平面的偏移量;$C_i$ 是第 $i$ 个类的权重,用于控制训练样本类的重要程度;$\xi_i$ 是松驰变量,表示错分样本的惩罚程度.

CW-SVM 的对偶 Lagrange 表达式为

$$\min_{\alpha} \frac{1}{2}\alpha^T Q\alpha - \mathbf{e}^T\alpha$$

$$0 \leq \alpha_i \leq C_i, i = 1, \cdots, m$$

$$\text{s.t}\quad y^T\alpha = 0,$$

如果对于二分类问题,则用 $C_+$ 表示正类权重,$C_-$ 表示负类权重. 如果 $C_+ = C_- = C$,则 CW-SVM 与 C-SVM 完全相同. 因此,C-SVM 是 CW-SVM 的一种特例. 可以通过设置不同的 $C_+$ 和 $C_-$ 来影响 CW-SVM 对正类和负类的分类准确度.

下面进行仿真试验. 为了便于观察,使用随机产生的二维数据. 正类样本数为 83,负类样本数为 8 个. 三次试验结果如表 1 所示. 图 2 至图 4 显示了当负类和正类权重变化时,分类面移动情况. 随着负类权重 $C_-$ 提高,分类面向正类移动,负类支持向量在减少,负类样本错分数降低;而正类的支持向量数增加,正类

样本错分数增加. 因此,通过调整类权重参数就可以改变所关注类的分类精度.

表 1 正、负类不同权重试验结果

| 试验组别 | $C_-$ | $C_+$ | $C_-/C_+$ | 负类支持向量数 | 负类错分样本数 | 正类支持向量数 | 正类错分样本数 |
|---|---|---|---|---|---|---|---|
| 1 | 100 | 100 | 1:1 | 6 | 2 | 6 | 0 |
| 2 | 500 | 100 | 5:1 | 3 | 1 | 14 | 2 |
| 3 | 1000 | 100 | 10:1 | 2 | 0 | 17 | 6 |

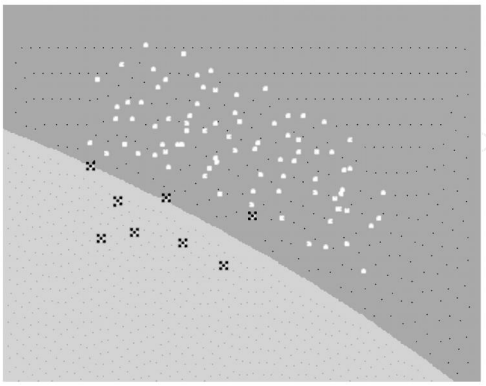

图 2 $C_-/C_+ = 1:1$ 时分类示意图

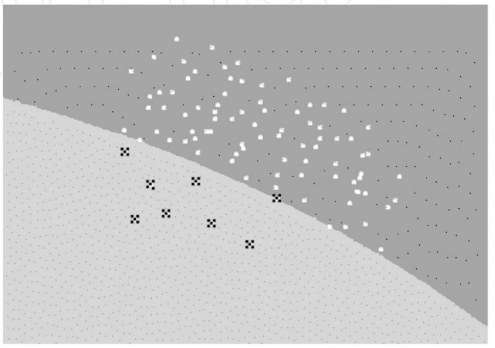

图 3 $C_-/C_+ = 5:1$ 时分类示意图

### 2.3 CWC-SVM 算法描述

文献[9]中指出二分类 SVM 可以表示为一个最小闭包球(Minimum Enclosing Ball, MEB)问题,而 MEB 问题适合处理大数据集. 如果将 SVM 问题转化为 MEB 问题,就可以解决 SVM 不能有效处理大数据集的问题.

本文构造对正类和负类带不同权重的 SVM 算法,并将其等价为 MEB 问题,从而解决 SVM 处理不平衡大数据集问题. 对于正类和负类带不同权重的 SVM 算法,原问题可以表示为

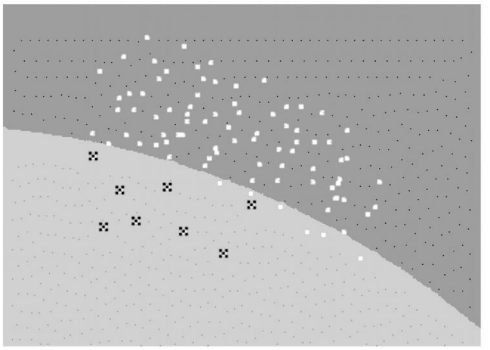

图 4 $C_-/C_+ = 10:1$ 时分类示意图

$$\min_{w,b,\rho,\xi_i} \|w\|^2 + b^2 - 2\rho + C^+ \sum_{\{i|y_i=1\}}^{n_+} \xi_i + C^- \sum_{\{j|y_j=-1\}}^{n_-} \xi_j \tag{1}$$

$$\text{s.t.} \quad y_i(\mathbf{w}\phi(\mathbf{x}_i) + b) \geq \rho - \xi_i, i = 1, \cdots, m$$

其中 $(\mathbf{w}\phi(\mathbf{x}_i) + b) = \rho$ 定义了分隔的超平面, $C^+$ 和 $C^-$ 分别是正类和负类的惩罚参数, $\xi = (\xi_1, \cdots, \xi_m)$ 是松弛向量.

(1)式的对偶问题是

$$\max_{\alpha} -\alpha'(\mathbf{K} \odot \mathbf{yy}' + \mathbf{yy}' + \frac{1}{\widetilde{C}}\mathbf{I})\alpha \tag{2}$$

$$\alpha \geq 0, \alpha' 1 = 1$$

其中 $\alpha = [\alpha_1, \cdots, \alpha_m]$ 是 Lagrange 乘子, $K = [k(x_i, x_j)]$ 是核矩阵,"$\odot$"表示 Hadamard 内积, $y = [y_1, \cdots, y_m]'$, $\widetilde{C}$ 对正类和负类分别取 $C^+$ 和 $C^-$. (2)式可以写成 $\max_{\alpha} -\alpha'\widetilde{K}\alpha, \alpha \geq 0, \alpha' 1 = 1$, 其中 $\widetilde{K} = [\widetilde{k}(z_i, z_j)]$, $\widetilde{k}(z_i, z_j) = y_i y_j k(x_i, x_j) + y_i y_j + \frac{\delta_{ij}}{\widetilde{C}}$. 根据文献[9]定理,带有不同权重的 SVM 问题可以等价于 MEB 问题. 且与文献[9]类似的证明,该算法收敛到最优解. 应用文献[11]中提出的快速近似算法,可以求解此 MEB 问题. 算法的时间和空间复杂度与样本数量成线性关系. 这样,将处理不平衡数据 SVM 问题转化为求解对应的 MEB 问题,充分利用 MEB 问题处理大数据集的能力,从而最终解决不平衡大数据集的分类问题.

## 3　算法试验和结果分析

### 3.1　数据来源

本节的试验数据来自深圳市某银行个人信贷部的客户信贷数据. 数据集中包括顾客数据 12 万条,每条记录有 16 个属性,包括数值属性和文本属性. 数据中有 10 %左右是容易中止业务并造成银行损失的客户. 这一类容易流失的客户样本定义为负类样本,其它客户样本定义为正类样本. 数据集中正类和负类的样本数量比例约为 9 ∶1,属于不平衡数据集. 数据集已经对正类和负类的样本进行了标注. 训练数据集中包括 10 万条数据,测试数据集中包括 2 万条数据.

### 3.2　模型和参数选取

客户流失问题是二分类问题,给定 $m$ 个训练样本$\{z_i = (x_i, y_i)\}_{i=1}^{m}$, $m = 100,000$,其中 $x_i$ 为 16 维样本输入, $y_i \in \{-1, +1\}$为样本输出, -1 表示客户流失, +1 表示客户不流失. 针对这个问题建立支持向量机模型. 因为客户流失数据是非线性数据,因此在 CWC-SVM 和普通 C-SVM 方法中,核函数都采用高斯核 $k(x, y) = \exp(-\|x - y\|^2/\beta)$,其中 $\beta = \frac{1}{m^2}\sum_{i,j=1}^{m}\|x_i - x_j\|^2$. CWC-SVM 中,选择 $C^+ = \frac{l_-}{l_+ + l_-}C$, $C^- = \frac{l_+}{l_+ + l_-}C$, $l_+$ 表示正样本个数, $l_-$ 表示负样本个数, $C$ 根据经验和试验选取 200. CWC-SVM 是基于工具包 libsvm[12] 实现的. 为了进行对比试验,选用普通 C-SVM、决策树、神经元网络模型做对比试验. 决策树模型选用改进 C4.5 模型. 神经元网络模型选用非线性模型,含有一个隐含层,停止条件是误差陷入局部极小.

### 3.3　试验结果和分析

数据挖掘中比较普遍地使用的提升度(Lift)指标来评价模型性能. 在客户关系管理中,一般用"lift"值来衡量预测准确性. Lift 指的是用模型和不用模型相比,预测能力提高的倍数. 比如,使用某个预测模型,在取 10 %总人数时,对应抓到 30 %的流失人数,显然预测效果比随机情况好. 此时模型性能提升: (30 %) / (10 %) = 3(即 lift = 3). 一开始有比较高的 lift 值,然后逐渐平稳下降到 1 的 lift 图代表了一个性能良好的数据挖掘模型.

图 5 是 CWC-SVM, C-SVM, C4.5, ANN 四种算法模型 lift 图.

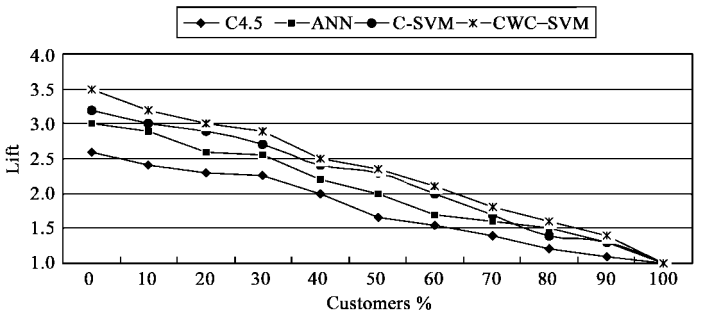


图 5　四种算法(CWC-SVM, C-SVM, C4.5, ANN) lift 图

从结果来看,CWC-SVM 和 C-SVM 方法都要好于神经元网络和决策树方法. 这一点印证了 SVM 方法可以有效解决二分类问题,且准确性相对较高. 如果选择合适的核函数(如本文选择的高斯核),就可以有效处理非线性数据. 而 CWC-SVM 对正类和负类样本选择不同的权重参数,取得了更高的准确率.

图 6 是 CWC-SVM 算法的 CPU 运行时间. 横轴是数据项个数,纵轴是 CPU 运行时间. 为了表示清楚,采用对数单位. 我们试验是在 1.6 GHz Pentium-4 计算机上运行. C-SVM 算法运行时间随着训练数据集规模扩大而显著增加,而 CWC-SVM 算法则大体上和训练数据集规模成线性关系. 可见,CWC-SVM 算法比 C-SVM 算法更适合处理海量数据集.

## 4 结束语

本文重点分析了客户流失问题的特点,针对该问题的特点和现有算法的不足,构建了 SVM 模型预测客户流失.针对客户流失数据集数据量大、数据分布不平衡的特点,本文提出了 CW-SVM 和 CWC-SVM 两种算法,给出了算法主要内容和算法复杂度.通过银行信贷客户数据验证了该算法比其他算法的预测效果更好,同时算法复杂度适合处理大数据集.

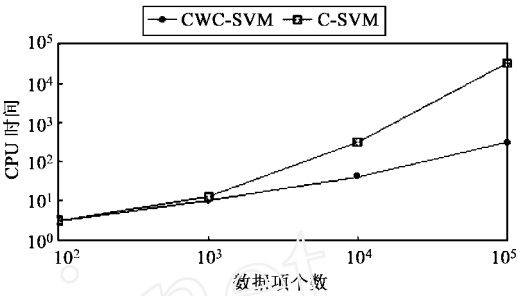


图 6 CWC-SVM 算法和 C-SVM 算法 CPU 运行时间比较

未来的工作包括两方面:一是针对不同数据集,如何有效选择合适的核函数和参数,并结合重抽样等技术,进一步提高预测结果的准确性.二是对分类预测的结果进行解释,从中抽取对决策有用的规则.

### 参考文献

[1] Reichheld Frederick F, Sasser EarlW. Zero defections: Quality comes to services[J]. Harvard Business Review, 1990, 105-111.

[2] Jones Thomas O, Sasser W Earl. Why satisfied customers defect[J]. Harvard Business Review, 1999, 73(6).

[3] Mozer M C, Wolniewicz R, Grimes D B, et al. Churn reduction in the wireless industry[J]. Advances in Neural Information Processing Systems, 2000(12), 935-941.

[4] Lemmens A, Croux C. Bagging and Boosting Classification Trees to predict churn[R], DTEW Research Report 0361, 2003, 40.

[5] Catherine Yang. AOL: Scrambling to Halt the Exodus. Business Week, 2003(8), Issue 3844, 62.

[6] Wu G, Chang E. Class-Boundary Alignment for Imbalanced Dataset Learning[C]//ICML Workshop on Learning from Imbalanced Data Sets II, Washington DC. 2003.

[7] 许建华, 张学工, 李衍达. 支持向量机的新发展[J]. 控制与决策, 2004, 19(5): 481-493.
Xu Jianhua, Zhang Xuegong, Li Yanda. Advances in support vector machines[J]. Control and Decision, 2004, 19(5): 481-493.

[8] Pavlov D, Chudova D, Smyth P. Towards scalable support vector machines using squashing[C]//Proc 6th ACM SIGKDD International Conf on Knowledge Discovery and Data Mining, Boston, Massachusetts, USA, 2000. 295-299.

[9] IvorW. Tsang, James T. Kwok, and Pak-Ming Cheung. Core vector machines: Fast SVM training on very large data sets[J]. Journal of Machine Learning Research, 2005(6): 363-392.

[10] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995

[11] Kumar P, Mitchell J, Yildirim A. Approximate minimum enclosing balls in high dimensions using core-sets. ACM Journal of Experimental Algorithmics, 8, January 2003.

[12] Chih-Chand, Chih-Jen Lin. LibSVM: a library for support vector machine [CP/OL]. http://www.csie.ntu.tw/~cjlin/libsvm, 2001