IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 37, NO. 1, JANUARY 2015

Towards Making Unlabeled Data Never Hurt

Yu-Feng Li and Zhi-Hua Zhou, Fellow, IEEE

制作无标签数据永远不会受到伤害

李玉峰，周志华，IEEE院士

Abstract—It is usually expected that learning performance can be improved by exploiting unlabeled data, particularly when the number of labeled data is limited. However, it has been reported that, in some cases existing semi-supervised learning approaches perform even worse than supervised ones which only use labeled data. For this reason, it is desirable to develop safe semi-supervised learning approaches that will not significantly reduce learning performance when unlabeled data are used. This paper focuses on improving the safeness of semi-supervised support vector machines (S3VMs). First, the S3VM-us approach is proposed. It employs a conservative strategy and uses only the unlabeled instances that are very likely to be helpful, while avoiding the use of highly risky ones. This approach improves safeness but its performance improvement using unlabeled data is often much smaller than S3VMs. In order to develop a safe and well-performing approach, we examine the fundamental assumption of S3VMs, i.e., low-density separation. Based on the observation that multiple good candidate low-density separators may be identified from training data, safe semi-supervised support vector machines (S4VMs) are here proposed. This approach uses multiple low-density separators to approximate the ground-truth decision boundary and maximizes the improvement in performance of inductive SVMs for any candidate separator. Under the assumption employed by S3VMs, it is here shown that S4VMs are provably safe and that the performance improvement using unlabeled data can be maximized. An out-of-sample extension of S4VMs is also presented. This extension allows S4VMs to make predictions on unseen instances. Our empirical study on a broad range of data shows that the overall performance of S4VMs is highly competitive with S3VMs, whereas in contrast to S3VMs which hurt performance significantly in many cases, S4VMs rarely perform worse than inductive SVMs.

摘要 - 通常期望通过利用未标记的数据来改善学习性能，特别是当标记数据的数量有限时。然而，据报道，在某些情况下，现有的半监督学习方法比仅使用标记数据的监督学习方法表现得更差。出于这个原因，期望开发安全的半监督学习方法，其在使用未标记数据时不会显着降低学习性能。本文着重于提高半监督支持向量机（S3VMs）的安全性。首先，提出了S3VM-us方法。它采用保守策略，仅使用非常有用的未标记实例，同时避免使用高风险实例。这种方法提高了安全性，但使用未标记数据的性能提升通常比S3VM小得多。为了开发安全且性能良好的方法，我们研究了S3VM的基本假设，即低密度分离。基于观察到可以从训练数据中识别出多个好的候选低密度分离器，这里提出了安全的半监督支持向量机（S4VM）。该方法使用多个低密度分离器来近似地面实况决策边界，并最大化任何候选分离器的归纳SVM的性能改进。在S3VM采用的假设下，这里显示S4VM可证明是安全的，并且使用未标记数据的性能改进可以最大化。还介绍了S4VM的样本外扩展。此扩展允许S4VM对看不见的实例进行预测。我们对广泛数据的实证研究表明，S4VM的整体性能与S3VM竞争激烈，而与在许多情况下显着损害性能的S3VM相比，S4VM很少比感应式SVM表现更差。

Index Terms—Unlabeled data, semi-supervised learning, safe, S3VMs, S4VMs

索引术语 - 未标记数据，半监督学习，安全，S3VM，S4VM

Ç

# 1 INTRODUCTION

T

RADITIONAL supervised learning often assumes that large numbers of labeled data are readily available for training. In many practical applications, however, the acquisition of class labels is expensive because the labeling process requires human effort and expertise. For example, in computer-aided medical diagnosis, large numbers of X-ray images can be obtained from routine examinations, but it is costly and difficult for physicians to mark all focuses in all images. In this case, training with only labeled data may not lead to a good performance. It is possible to employ semi-supervised learning [10], [34], [51], [52] that exploits the wide availability of unlabeled data to improve performance. During the past decade, semi-supervised learning has attracted significant attention. It has been found useful in many applications, including text categorization [23], image retrieval [42], bioinformatics [24], and natural language processing [19].

介绍

传统的监督学习通常假设大量标记数据可用于培训。 然而，在许多实际应用中，类标签的获取是昂贵的，因为标签过程需要人力和专业知识。 例如，在计算机辅助医学诊断中，可以从常规检查中获得大量的X射线图像，但是医生在所有图像中标记所有焦点是昂贵且困难的。 在这种情况下，仅使用标记数据进行培训可能无法获得良好的性能。 有可能采用半监督学习[10]，[34]，[51]，[52]，利用未标记数据的广泛可用性来提高性能。 在过去的十年中，半监督学习引起了人们的极大关注。 它被发现在许多应用中都很有用，包括文本分类[23]，图像检索[42]，生物信息学[24]和自然语言处理[19]。

Existing semi-supervised approaches can be roughly grouped into four categories. The first category is generative methods, e.g., [35], [36]. These methods extend supervised generative models by incorporating unlabeled data, and estimate model parameters and labels using techniques such as the EM algorithm [17]. The second category is graph-based methods, e.g., [2], [7], [34], [48], [53]. These

|  |
| --- |
| The authors are with the National Key Laboratory for Novel Software Tech-nology, Nanjing University and Collaborative Innovation Center of Novel  Software Technology and Industrialization, Nanjing 210023, China. E-mail: {liyf, zhouzh}@lamda.nju.edu.cn.  Manuscript received 9 Jan. 2013; revised 11 Dec. 2013; accepted 27 Dec. 2013.  Date of publication 13 Jan. 2014; date of current version 5 Dec. 2014.  Recommended for acceptance by K. Borgwardt.  For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2014.2299812 |

methods encode both the labeled and unlabeled instances in a graph and then assign class labels to the unlabeled data such that their inconsistencies with both the labeled data and the underlying graph are minimized. The third category is disagreement-based methods, e.g., [8], [50]. These methods typically involve multiple learners and improve them through the exploitation of disagreement among the learners. The fourth category is semi-supervised support vector machines (S3VMs), e.g., [4], [23]. They use unlabeled data to regularize the decision boundary so that it can pass through low-density regions [12].

现有的半监督方法可以大致分为四类。第一类是生成方法，例如[35]，[36]。这些方法通过合并未标记的数据来扩展监督的生成模型，并使用诸如EM算法的技术来估计模型参数和标签[17]。第二类是基于图的方法，例如[2]，[7]，[34]，[48]，[53]。这些方法在图中编码已标记和未标记的实例，然后将类标签分配给未标记的数据，以使它们与标记数据和基础图的不一致性最小化。第三类是基于分歧的方法，例如[8]，[50]。这些方法通常涉及多个学习者，并通过利用学习者之间的分歧来改进它们。第四类是半监督支持向量机（S3VM），例如，[4]，[23]。他们使用未标记的数据来规范决策边界，使其可以通过低密度区域[12]。

|  |
| --- |
| 0162-8828 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications\_standards/publications/rights/index.html for more information. |

It is generally accepted that by using unlabeled data, semi-supervised learning can help improve the performance, particularly when the number of labeled data is limited. Many empirical studies, however, show that there are cases in which the use of unlabeled data decreases the performance [7], [11], [13], [14], [16], [20], [36], [47], [50]. Such phenomena undeniably encumber the deployment of semisupervised learning in real applications, especially tasks requiring high reliability, because users usually require that new techniques (such as semi-supervised learning) should perform at least as well as existing techniques (such as pure supervised learning). For this reason, it is desirable to have safe semi-supervised learning approaches which never reduce learning performance significantly when using unlabeled data. This is a challenging task, and only a few authors have explicitly tried to reduce the chance of performance degeneration [14], [27], even though there are already many studies on semi-supervised learning. Safe, here means that the generalization performance is never statistically significantly worse than methods using only labeled data. It is meaningless to talk about a single trial, because for a single trial, even exploiting more labeled data might result in a worse performance.

通常认为，通过使用未标记的数据，半监督学习可以帮助改善性能，特别是当标记数据的数量有限时。然而，许多实证研究表明，有些情况下使用未标记的数据会降低性能[7]，[11]，[13]，[14]，[16]，[20]，[36]， 47]，[50]。这种现象无可否认地妨碍了在实际应用中部署半监督学习，尤其是需要高可靠性的任务，因为用户通常要求新技术（如半监督学习）至少应该与现有技术（如纯监督学习）一样好。 。出于这个原因，期望具有安全的半监督学习方法，其在使用未标记数据时从不显着降低学习性能。这是一项具有挑战性的任务，只有少数作者明确尝试降低性能退化的机会[14]，[27]，尽管已经有很多关于半监督学习的研究。安全，这意味着泛化性能在统计上永远不会比仅使用标记数据的方法差。谈论单个试验是没有意义的，因为对于单个试验，即使利用更多标记的数据也可能导致更差的性能。

Cozman et al. [16] discussed the reason why unlabeled data can increase classification error for generative methods. They conjectured that the performance degeneration is caused by incorrect model assumptions, because fitting unlabeled data based on an incorrect model assumption will mislead the learning process. However, it is very difficult to make a correct model assumption without sufficient domain knowledge. For graph-based methods, researchers realized that graph construction is the crucial problem. However, developing a good graph in general situations remains an open problem. Disagreement-based methods usually use pseudo-labels of unlabeled data provided by multiple learners to enhance the labeled data set. In this way, incorrect pseudo-labels may disrupt the learning process. One possible solution is to use data editing techniques to examine data that may have been pseudo-labeled [27]. However, such solutions work well only on dense data. This is because data editing techniques usually rely on the data neighboring information. With S3VMs, the correctness of the optimization objective has been studied on very small data sets [11]. However, there is no clear solution that can be used to prevent performance from degeneration when using unlabeled data. There are also some general discussions on the usefulness of unlabeled data from a theoretical perspective [1], [3], [38]. In particular, in [1], the authors showed that when unlabeled data provide a good regularizer, a purely inductive supervised SVM on labeled data using such a regularizer guarantee a good generalization. Deriving such a good regularizer, however, remains an open problem.

Cozman等人。 [16]讨论了未标记数据可能增加生成方法的分类错误的原因。他们猜测性能退化是由不正确的模型假设引起的，因为根据不正确的模型假设拟合未标记的数据会误导学习过程。但是，如果没有足够的领域知识，很难做出正确的模型假设。对于基于图的方法，研究人员意识到图形构造是关键问题。但是，在一般情况下开发一个好图仍然是一个悬而未决的问题。基于分歧的方法通常使用由多个学习者提供的未标记数据的伪标签来增强标记数据集。这样，不正确的伪标签可能会破坏学习过程。一种可能的解决方案是使用数据编辑技术来检查可能已经伪标记的数据[27]。但是，此类解决方案仅适用于密集数据。这是因为数据编辑技术通常依赖于数据邻近信息。对于S3VM，已经在非常小的数据集上研究了优化目标的正确性[11]。但是，当使用未标记的数据时，没有明确的解决方案可用于防止性能退化。从理论的角度来看，还有一些关于未标记数据有用性的一般性讨论[1]，[3]，[38]。特别地，在[1]中，作者表明，当未标记的数据提供良好的正则化器时，使用这种正则化器对标记数据进行纯粹的归纳监督SVM保证了良好的泛化。然而，推导出如此优秀的正规则仍然是一个悬而未决的问题。

Particularly, S3VMs have been widely applied to many tasks [10], and their representative algorithm, TSVM [23], has won the Ten-Year Best Paper Award for machine learning in 2009. Most research efforts on S3VMs address its complexity [11], [15], [23], [28], with little effort on its safeness, although many empirical studies have shown that S3VMs also reduce performance, sometimes even seriously [10],

[42], [47].

特别是，S3VM已被广泛应用于许多任务[10]，其代表性算法TSVM [23]在2009年获得了机器学习十年最佳论文奖。大多数关于S3VM的研究工作都解决了其复杂性[11] ，[15]，[23]，[28]虽然很少努力保证其安全性，尽管许多实证研究表明S3VMs也会降低性能，有时甚至会严重[10]，

[42]，[47]。

This paper focuses on improving the safeness of S3VMs. First, because the main use of unlabeled data is to determine data distribution, it is here conjectured that the degradation of the performance degeneration of S3VMs is caused by unlabeled instances that are obscure or misleading for the discovery of the underlying distribution. For this reason, the S3VM with unlabeled data selection (S3VM-us) approach is here proposed. It uses hierarchical clustering to estimate the reliability of unlabeled instances and then removes the ones with the lowest reliability.

本文着重于提高S3VM的安全性。 首先，因为未标记数据的主要用途是确定数据分布，所以推测S3VM性能退化的退化是由于未标记的实例引起的，这些实例对于底层分布的发现来说是模糊或误导的。 出于这个原因，这里提出了具有未标记数据选择（S3VM-us）方法的S3VM。 它使用层次聚类来估计未标记实例的可靠性，然后删除具有最低可靠性的实例。

Our empirical studies show that S3VM-us improves the safeness of S3VMs. However, its improvement in performance using unlabeled data is not as considerable as S3VMs. To develop a safe and well-performing approach, we then examine the fundamental assumption of S3VMs, i.e., low-density separation (LDS), and get another conjecture on the reason of performance degeneration. Given a few labeled data and many more unlabeled data, there is usually more than one large-margin low-density separator. However, it is hard to determine which one is optimal based on the limited labeled data. Although these low-density separators are all consistent with the limited labeled data, they can be very diverse with respect to the instance space.

我们的实证研究表明，S3VM-us提高了S3VM的安全性。 但是，使用未标记数据的性能提升并不像S3VM那么大。 为了开发一种安全且性能良好的方法，我们研究了S3VM的基本假设，即低密度分离（LDS），并对性能退化的原因进行了另一种猜想。 给定一些标记数据和更多未标记数据，通常有多个大边距低密度分离器。 然而，基于有限的标记数据很难确定哪一个是最佳的。 尽管这些低密度分离器都与有限的标记数据一致，但它们在实例空间方面可以非常多样化。

In this way, incorrect selection may result in a reduced performance. Based on this observation, the S4VMs (Safe S3VMs) approach, the main contribution of this paper, is proposed. S4VMs use multiple low-density separators to approximate the ground-truth decision boundary and maximize the improvement in performance against inductive SVMs for any candidate separator. S4VMs are shown to be safe and to achieve the maximal performance improvement under the low-density assumption of S3VMs. An out-ofsample extension of S4VMs is also presented so that S4VMs can make predictions on unseen instances. Our empirical studies performed on a broad range of data sets show that S4VMs perform highly competitive with S3VMs. More importantly, unlike S3VMs which significantly reduce performance in many cases, S4VMs are rarely inferior to inductive SVMs.

这样，不正确的选择可能导致性能降低。 基于这一观察，提出了S4VMs（Safe S3VMs）方法，这是本文的主要贡献。 S4VM使用多个低密度分离器来近似地面实况决策边界，并最大化针对任何候选分离器的感应SVM的性能改进。 S4VM被证明是安全的，并且在S3VM的低密度假设下实现了最大的性能提升。 还提供了S4VM的示例扩展，以便S4VM可以对看不见的实例进行预测。 我们对广泛的数据集进行的实证研究表明，S4VM与S3VM的竞争非常激烈。 更重要的是，与在许多情况下显着降低性能的S3VM不同，S4VM很少低于归纳SVM。

The rest of this paper is organized as follows. S3VMs are briefly introduced in Section 2. The S3VM-us and S4VMs are introduced in Sections 3 and 4. Empirical results are report in Section 5. Conclusions are presented in Section 6.

本文的其余部分安排如下。 S3VM简要介绍在第2节中.S3VM-us和S4VM在第3节和第4节中介绍。实验结果在第5节中报告。结论见第6节。

# 2 BRIEF INTRODUCTION TO S3VMS

Inspired by the success of the large-margin principle [40], S3VMs extend inductive supervised SVMs to semi-supervised learning. They simultaneously learn the optimal decision function and the labels of unlabeled instances such that the decision boundary has a large margin on both the labeled and unlabeled data. It was discovered that S3VMs realize the low-density assumption [12] which states that the decision boundary will go across low-density regions.

S3VMS简介

受大边际原则[40]的成功启发，S3VM将归纳监督SVM扩展到半监督学习。 他们同时学习最佳决策函数和未标记实例的标签，使得决策边界在标记和未标记数据上都有很大的余量。 人们发现S3VM实现了低密度假设[12]，该假设表明决策边界将跨越低密度区域。

Formally, we consider binary classification here. Let X be the input space and Y ¼ f1g be the label space. Given al set of l labeled instancesl u fxi;yigi 1 and u unlabeled instanf jgjþl 1, S3VMs aim to ¼find a decision function ces x

¼ þ f : X ! f1g and a label assignment on unlabeled instances y ¼ fylþ1;...;ylþug 2 B such that the following functional is minimized:

在形式上，我们在这里考虑二元分类。 设X为输入空间，Y = f1g为标签空间。 给定l个标记的实例l y fxi; yigi 1和u unlabeled instanfjgjþl1，S3VMs旨在找到决策函数ces x

¼þf：X！ f1g和未标记实例上的标签分配y =fylþ1; ...;ylþug2B使得以下功能最小化：

1 2 l lþu

fmin2H 2kfkH þ C1Xi¼1 ‘ðyi;fðxiÞÞ þ C2 jX¼lþ1‘ðyj;fðxjÞÞ: (1) y2B

Here B is a set of label assignments obtained from domain knowledge. For example, when the class proportion of unlabeled data is closely related to that of labeled data (also refer to as balance constraint [11], [23]), we can set

这里B是从领域知识获得的一组标签分配。 例如，当未标记数据的类比例与标记数据的比例密切相关时（也称为平衡约束[11]，[23]），我们可以设置

P

lþu 1 y y

B ¼ fy 2 f1guj b j¼lþ j Pil¼1 i bg; u l

where b is a small constant controlling the inconsistency of class proportions. H is the Reproducing Kernel Hilbert Space (RKHS) induced by a kernel function k. ‘ðy;fðxÞÞ ¼ maxf0;1 yfðxÞg is the hinge loss used in SVMs. C1 and

C2 are two regularization parameters trading off model complexity and empirical losses on the labeled and unlabeled data, respectively.

其中b是控制类比例不一致的小常数。 H是由核函数k诱导的再生核希尔伯特空间（RKHS）。 'ðy;fðxÞÞ¼maxf0;1yfðxÞg是SVM中使用的铰链损耗。 C1和

C2是两个正则化参数，分别在标记和未标记数据上折衷模型复杂性和经验损失。

Similar to supervised SVMs, S3VMs favor the decision boundary having a large margin on all training data.

According to [12], they inherently favor the decision boundary going through low-density regions. Otherwise a large loss will occur with respect to the objective of S3VMs [12].

与监督的SVM类似，S3VM倾向于在所有训练数据上具有较大余量的决策边界。

根据[12]，他们固有地倾向于通过低密度区域的决策边界。 否则，就S3VMs的目标而言会出现大的损失[12]。

Unlike supervised SVMs where the training labels are complete, S3VMs need to infer the integer-value labels of the unlabeled instances, resulting in a difficult mixedinteger programming problem. Great efforts have been devoted to coping with the high complexity of S3VMs. Roughly speaking, they can be grouped into four categories. The first kind of approaches is based on global combinatorial optimization. Examples include branch-andbound methods [4], [11], which solve S3VMs globally and obtain good performance on small data sets. The second kind of approaches is based on global heuristic search, which gradually increases the difficulty of solving the non-convex part in Eq. (1). Examples include TSVM [23] which gradually increases the influence of unlabeled data (i.e., the value of C2), the deterministic annealing approach [37] which gradually increases the temperature of an entropy function in optimization, and the continuation method [9] which first introduces a surrogate smooth function and then gradually decreases the smoothness of the surrogate function to approach the objective in Eq. (1). The third kind of approaches is based on convex relaxation, which transforms Eq. (1) into a relaxed convex problem. Examples include the semi-definite programming (SDP) relaxation [6], [43], and the minimax relaxation [28], [29], [30] which is tighter and more scalable than the SDP relaxation. The fourth kind of approaches is based on efficient non-convex optimization techniques. Examples include UniverSVM [15] which employs concaveconvex procedure (CCCP) [44], and meanS3VM [28] which employs alternating optimization [5].

与训练标签完整的监督SVM不同，S3VM需要推断未标记实例的整数值标签，从而导致困难的混合整数编程问题。已经付出了巨大的努力来应对S3VM的高度复杂性。粗略地说，它们可以分为四类。第一种方法基于全局组合优化。示例包括分支和绑定方法[4]，[11]，它们全局解决S3VM并在小数据集上获得良好性能。第二种方法是基于全局启发式搜索，这逐渐增加了求解方程中非凸部分的难度。 （1）。例子包括TSVM [23]逐渐增加未标记数据的影响（即C2的值），确定性退火方法[37]逐渐增加优化中熵函数的温度，以及延续方法[9]首先介绍一个代理平滑函数，然后逐渐降低代理函数的平滑度，以接近方程式中的目标。 （1）。第三种方法是基于凸松弛，它改变了方程。 （1）进入一个松弛的凸问题。例子包括半定规划（SDP）松弛[6]，[43]和极小极大弛豫[28]，[29]，[30]，它比SDP弛豫更紧凑，更具可扩展性。第四种方法基于有效的非凸优化技术。例子包括UniverSVM [15]，它采用凹凸过程（CCCP）[44]，而meanS3VM [28]采用交替优化[5]。

Because S3VMs involve a complicated optimization task, most previous efforts were devoted to handling the high complexity, whereas few literatures have explicitly studied the safeness of S3VMs.

由于S3VM涉及复杂的优化任务，以前的大多数努力都致力于处理高复杂性，而很少有文献明确研究过S3VM的安全性。

# 3 S3VM-US

It is generally accepted that the major utility of unlabeled data is to disclose useful information about the underlying data distribution [10]. When some unlabeled instances are obscure or misleading for the discovery of the underlying distribution, learning performance may be reduced by using those data. Based on this observation, S3VM-us, which tries to exclude highly risky unlabeled instances, is proposed.

3 S3VM美

人们普遍认为，未标记数据的主要用途是披露有关基础数据分布的有用信息[10]。 当一些未标记的实例对于发现底层分布而言模糊或误导时，可以通过使用这些数据来减少学习性能。 基于这一观察结果，提出了试图排除高风险未标记实例的S3VM-us。

In the following, two simple approaches to exclude highly risky unlabeled instances, i.e., the S3VM-c and S3VM-p approaches, are first introduced and by examining the deficiencies of S3VM-c and S3VM-p, S3VM-us is then presented. For the simplicity of notations, the training set is denoted as D ¼ ffxi;yigli¼1;fxjgljþ¼ulþ1g. The predicted labels for x by inductive SVM (using labeled data only) and S3VM are denoted as ysvmðxÞ and ys3vmðxÞ, respectively.

The transpose of a vector is denoted by the superscript 0.

在下文中，首先介绍两种排除高风险未标记实例的简单方法，即S3VM-c和S3VM-p方法，然后通过检查S3VM-c和S3VM-p的缺陷，然后介绍S3VM-us。 为了简化符号，训练集表示为D = ffxi; yigli = 1;fxjgljþ¼ulþ1g。 通过归纳SVM（仅使用标记数据）和S3VM的x的预测标记分别表示为ysvmðxÞ和ys3vmðxÞ。

矢量的转置由上标0表示。

## 3.1 Two Simple Approaches

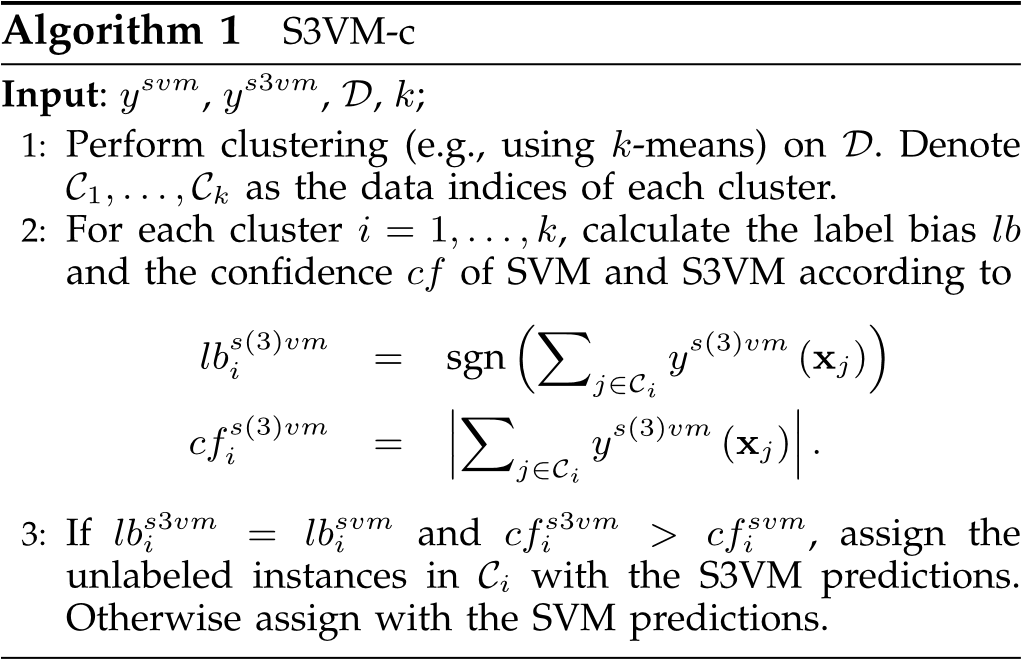
### 3.1.1 S3VM-c

The first simple approach S3VM-c is motivated by [38]. It suggests that unlabeled data will be helpful when the component density sets are discernible, where component density sets refer to regions of data distribution with non-zero probability density. To implement this idea, in S3VM-c, the component density sets are simulated by clusters obtained with a clustering algorithm, and the discernibility is simulated by a disagreement between S3VM and inductive SVM based on bias and confidence. It is noteworthy that other simulations are also possible. As Algorithm 1 shows, we rely on the prediction of S3VM if S3VM obtains the same bias but enhances the confidence of the inductive SVM. Otherwise we will rely on the prediction of the inductive SVM.

3.1 两个简单的方法

3.1.1 S3VM-C

第一个简单的方法S3VM-c的动机是[38]。 它表明，当组件密度集是可辨别的时，未标记的数据将是有用的，其中组件密度集指的是具有非零概率密度的数据分布区域。 为了实现这个想法，在S3VM-c中，通过使用聚类算法获得的聚类来模拟组件密度集，并且通过基于偏差和置信度的S3VM和归纳SVM之间的不一致来模拟可辨别性。 值得注意的是，其他模拟也是可能的。 正如算法1所示，如果S3VM获得相同的偏差但是增强了归纳SVM的置信度，我们依赖于S3VM的预测。 否则，我们将依赖于归纳SVM的预测。



### 3.1.2 S3VM-p

The second simple approach S3VM-p is motivated by the confidence estimation in label propagation methods [48], [53], where the confidence can be naturally regarded as a measurement of the reliability of unlabeled data.

Formally, to estimate the confidence of unlabeled data, let yl ¼ ½y1;...;yl0 2 f1gl1 and Fl ¼ ½ðyl þ 1Þ=2;ð1 ylÞ=2 2 f0;1gl2 be the vector- and matrix-form of labeled data, respectively. Let W ¼ ½wij 2 RðlþuÞðlþuÞ be the similarity matrix of training data, and L ¼ D W the Laplacian matrix of W, where D is a diagonal matrix with entries di ¼ Pljþ¼u1 wij, i ¼ 1;...;luþ u. According to [53], the predictions of unlabeled data F are derived as,

3.1.2 S3VM-P

第二种简单方法S3VM-p的动机是标签传播方法中的置信度估计[48]，[53]，其中置信度可以自然地被视为未标记数据可靠性的度量。

形式上，为了估计未标记数据的置信度，让yl =½½1; ...; yl0 2 f1gl1和Fl =½ðylþ1Þ= 2;ð1ylÞ= 2 2 f0; 1gl2是标记数据的向量和矩阵形式 ， 分别。 设W =½wij2RðlþuÞðlþuÞ为训练数据的相似矩阵，L = D W W的拉普拉斯矩阵，其中D为对角矩阵，条目为di¼Pljþ¼u1wij，i = 1; ...;luþu。 根据[53]，未标记数据F的预测推导为，

Fu ¼ Lu;u1Wu;lFl; (2)

where Lu;u refers to a sub-matrix of L on the block of unlabeled data, Wu;l refers to a sub-matrix of W on the block between labeled and unlabeled data, and Lu;u1 refers to the inverse matrix of Lu;u. In Fu, note that the two entries of each row refer to the confidence estimations belonging to two different classes. We then assign each unlabeled instance xj with the label ylpðxjÞ ¼ sgnðFujl;1  Fujl;2Þ, and the confidence hjl ¼ jFjul;1

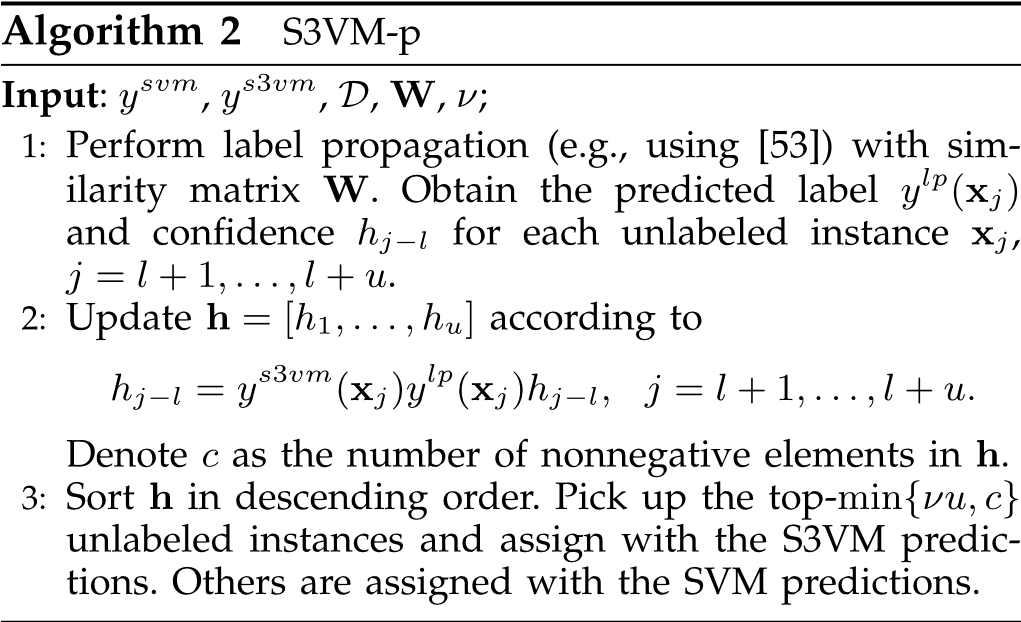
u

Fjl;2j, j ¼ l þ 1;...;l þ u. As Algorithm 2 shows, after confidence estimation, similar to S3VM-c, we consider the risk of unlabeled data by bias and confidence. If S3VM obtains the same bias of label propagation and the confidence is high, we use the S3VM prediction. Otherwise we use the inductive SVM prediction instead.

其中Lu; u指的是未标记数据块上的L的子矩阵，Wu; l指的是标记和未标记数据之间的块上的W的子矩阵，Lu; u1指的是Lu的逆矩阵;ü。 在Fu中，请注意每行的两个条目是指属于两个不同类的置信度估计。 然后，我们为每个未标记的实例xj分配标签ylpðxjÞ¼sgnðFujl; 1 Fujl;2Þ，以及置信度hjl = jFjul; 1

ü

Fjl; 2j，j = lþ1; ...; lþu。 正如算法2所示，在置信度估计之后，类似于S3VM-c，我们通过偏差和置信度来考虑未标记数据的风险。 如果S3VM获得相同的标签传播偏差并且置信度很高，我们使用S3VM预测。 否则，我们使用归纳SVM预测



## 3.2 S3VM-us

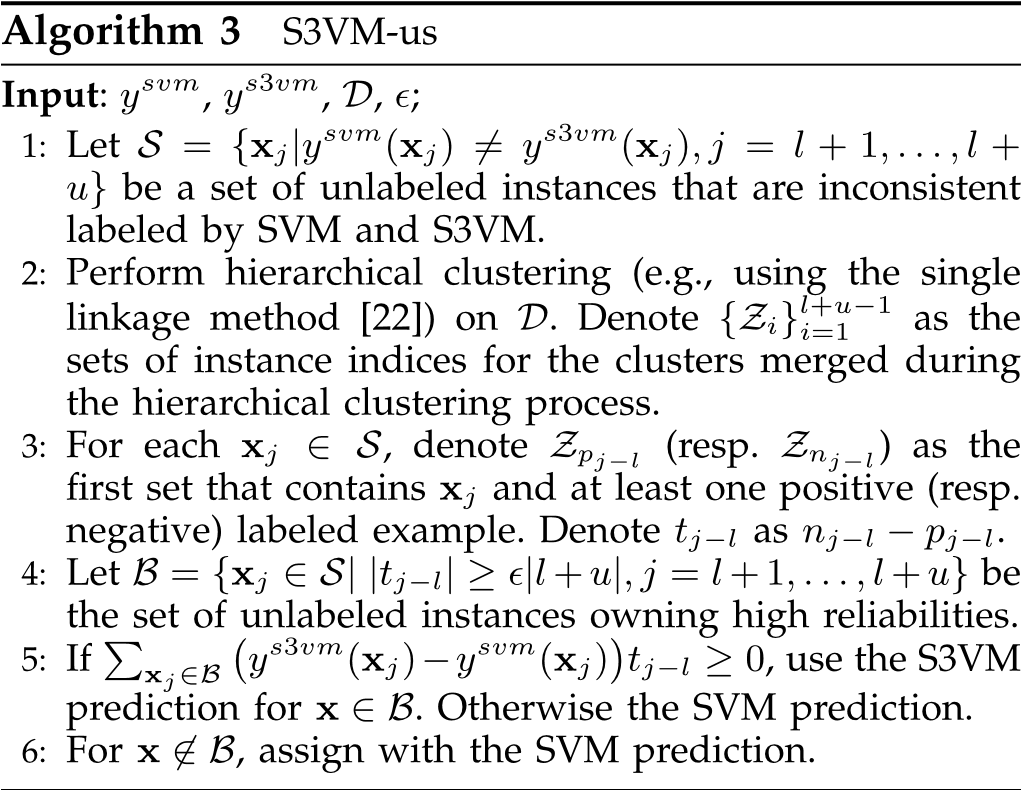
S3VM-c and S3VM-p have not been reported before. Our empirical studies show that they are capable of reducing the chances of performance degeneration. However, they both suffer from some deficiencies. S3VM-c works in a local manner and the relations between clusters are never considered. In S3VM-p, as stated in [41], the confidence estimated with label propagation methods might be incorrect if the label initialization is highly imbalanced. Moreover, both S3VM-c and S3VM-p heavily rely on S3VM predictions. This might be risky when S3VM suffers from a serious reduced performance.

The examination of the deficiencies of S3VM-c and S3VM-p suggests us to exploit the relations between clusters and reduce the sensitivity to the label initialization. This motivates our S3VM-us approach.

3.2 S3VM美

之前没有报道过S3VM-c和S3VM-p。 我们的实证研究表明，它们能够降低性能退化的可能性。 但是，他们都有一些不足之处。 S3VM-c以本地方式工作，从不考虑集群之间的关系。 在S3VM-p中，如[41]中所述，如果标签初始化高度不平衡，则使用标签传播方法估计的置信度可能不正确。 此外，S3VM-c和S3VM-p都严重依赖于S3VM预测。 当S3VM性能严重下降时，这可能会有风险。

检查S3VM-c和S3VM-p的缺陷表明我们要利用集群之间的关系并降低对标签初始化的敏感性。 这激发了我们的S3VM-us方法。



As Algorithm 3 shows, S3VM-us employs hierarchical clustering [22]. It first initializes each single instance as a cluster and then merges two of the clusters with the shortest distance. This process repeats until all the instances are merged into one cluster. It is not hard to validate that hierarchical clustering considers the between-cluster relations. Moreover, since hierarchical clustering is an unsupervised method, it does not suffer from the label initialization problem.

正如算法3所示，S3VM-us采用了层次聚类[22]。 它首先将每个单个实例初始化为一个簇，然后合并两个具有最短距离的簇。 重复此过程，直到所有实例合并到一个群集中。 验证层次聚类是否考虑了簇间关系并不难。 此外，由于分层聚类是无监督方法，因此不会遇到标签初始化问题。

To estimate the reliability on unlabeled instances, let pjl and njl denote the lengths of paths from an unlabeled

为了估计未标记实例的可靠性，让pjl和njl表示来自未标记的路径的长度

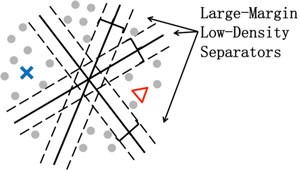


Fig. 1. There are multiple large-margin low-density separators coinciding well with labeled data (cross and triangle).

图1.有多个大边距低密度分离器与标记数据（交叉和三角形）很好地吻合。

instance xj to its nearest positive and negative labeled instances, respectively. The difference between pil and nil is simply taken as an estimation of reliability. Intuitively, the larger the difference between pjl and njl, the higher the reliability on labeling xj.

Our empirical studies in Section 5 show that S3VM-us effectively improves the safeness of S3VMs. However, its improvement in performance is often marginal when compared with existing S3VMs. To develop safe and well-performing methods, it might be insufficient to purely rely on the selection of unlabeled instances. This motivates us to develop the S4VM approach presented in the next section.

实例xj分别为其最接近的正面和负面标记实例。 pil和nil之间的差异简单地被视为可靠性的估计。 直观地，pjl和njl之间的差异越大，标记xj的可靠性越高。

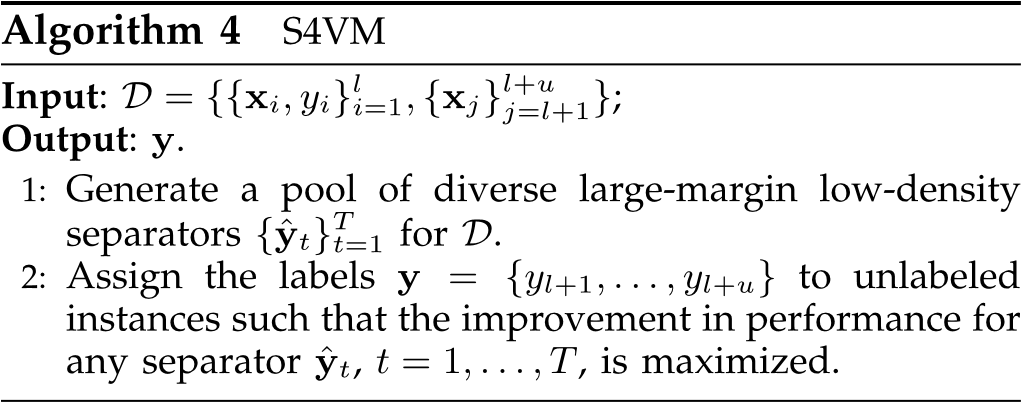
我们在第5节中的实证研究表明，S3VM-us有效地提高了S3VM的安全性。 但是，与现有的S3VM相比，它在性能方面的改进通常很小。 为了开发安全且性能良好的方法，纯粹依赖于未标记实例的选择可能是不够的。 这促使我们开发下一节中介绍的S4VM方法。

# 4 S4VMS

As previously mentioned, the underlying assumption of S3VMs is low-density separation. That is, the ground-truth is realized by a large-margin low-density separator. However, as illustrated in Fig. 1, given limited labeled data and many more unlabeled data, there usually exist multiple large-margin low-density separators. Although these separators all coincide well with the labeled data, they could be quite diverse with respect to the feature space, and thus an inadequate selection may lead to a serious performance reduction. This observation incites us the design of S4VMs. Specifically, S4VMs first generate a pool of diverse largemargin low-density separators, and then try to maximize the improvement in performance for any separator. The pseudo-code of S4VM is summarized in Algorithm 4.

4个S4VMS

如前所述，S3VM的基本假设是低密度分离。 也就是说，通过大范围的低密度分离器实现了基本事实。 然而，如图1所示，给定有限的标记数据和许多未标记的数据，通常存在多个大边缘低密度分离器。 尽管这些分隔符与标记数据完全一致，但它们在特征空间方面可能非常多样化，因此选择不当可能导致严重的性能降低。 这种观察激发了我们S4VM的设计。 具体而言，S4VM首先生成一系列不同的大型低密度分离器，然后尝试最大化任何分离器的性能改进。 算法4总结了S4VM的伪代码。



In the following, we will first introduce how to build S4VMs given a pool of diverse large-margin low-density separators, and then present two different implementations for generating the pool.

在下文中，我们将首先介绍如何在给定多个大边距低密度分隔符池的情况下构建S4VM，然后提供两种不同的实现来生成池。

4.1 Building S4VMs from a Pool of Separators Let ybe the ground-truth label assignment and ysvm be the predictive labels of inductive SVM on unlabeled instances. For any label assignment of unlabeled instances y ¼ fylþ1;...;ylþug, denote gainðy;y;ysvmÞ and lossðy;y; ysvmÞ as the gained and lost accuracies compared to the inductive SVM. Our goal is to learn a label assignment y such that the improved performance against the inductive

SVM is maximized, maxy2f1gu gainðy;y;ysvmÞ lossðy;y;ysvmÞ; (3)

where is a parameter for trading-off how much risk the user would like to undertake. In the sequel, we will denote gainðy;y^;ysvmÞ lossðy;y^;ysvmÞ as Jðy;y^;ysvmÞ, for the simplicity of notations.

4.1从分离器池构建S4VM让y成为真实的标签分配，ysvm是未标记实例上的归纳SVM的预测标签。 对于未标记实例的任何标签分配y =fylþ1; ...;ylþug，表示gainðy; y;ysvmÞ和lossðy; y; 与归纳SVM相比，ysvmÞ作为获得和丢失的准确度。 我们的目标是学习标签分配，以便提高针对归纳的性能

SVM最大化，maxy2f1guinjy; y;ysvmÞinfeðy; y;ysvmÞ;（3）

哪个是用于权衡用户想要承担多少风险的参数。 在续集中，我们将表示增益; y ^;ysvmÞdossðy; y ^;ysvmÞ为Jðy; y ^;ysvmÞ，表示符号的简单性。

The difficulty in solving Eq. (3) lies in the fact that the ground-truth yis unknown. Otherwise it is trivial to output y ¼ yas the optimal solution. Given a pool of T lowhere we assume that the ground-truthdensity separators fy^tgTt¼1, as employed by existing S3VMs,y is realized by a low-density separator, i.e., y2 M, fy^tgTt¼1. Without further domain knowledge in distinguishing these separators, we then maximize the worst-case improvement over inductive SVM (Eq. (4)), and denote ~~y~~ as the optimal solution

解决Eq的困难。 （3）在于事实真相是未知的。 否则输出y = y作为最优解是微不足道的。 给定T low lowhere，我们假设现有S3VM使用的地面实况密度分离器fy ^ tgTt = 1由低密度分离器实现，即y 2 M，fy ^ tgTt = 1。 如果没有进一步的区域知识来区分这些分离器，那么我们最大化了对归纳SVM的最坏情况改进（方程（4）），并将y表示为最优解

~~y~~ ¼ arg max min Jðy;y^;ysvmÞ: (4) y2f1gu y^2M

The following theorem shows that by taking the low-density assumption as typical S3VMs, i.e., y2 fy^tgTt¼1, S4VMs are provably safe.

以下定理表明，通过将低密度假设作为典型的S3VM，即y 2 fy ^ tgTt = 1，S4VM可证明是安全的。

Theorem 1. If y2 fy^tgTt¼1 and 1, the accuracy of ~~y~~ is never worse than that of ysvm.

Proof. Note that ~~y~~ is the optimal solution and Jðysvm; y^;ysvmÞ is zero for any y^, we have min Jð~~y~~;y^;ysvmÞ min Jðysvm;y^;ysvmÞ ¼ 0: (5)

y^2M y^2M

Further note that y2 M, we have

Jð~~y~~;y;ysvmÞ miny^2MJð~~y~~;y^;ysvmÞ: (6)

From Eqs. (5) and (6), Jð~~y~~;y;ysvmÞ 0, i.e., gainð~~y~~;y; ysvmÞ lossð~~y~~;y;ysvmÞ. Recall that 1, we then have gainð~~y~~;y;ysvmÞ lossð~~y~~;y;ysvmÞ and thus the theorem is proved. tu

According to Theorem 1, it is easy to get the following proposition.

Proposition 1.satisfyingsvm. minIfy^y2M 2Jðfyy^;ty^g;Tt¼y1svmandÞ 0 , is never worse than that1, the accuracy of any y of y

Simply outputting the predictive results of the inductive SVM would be also safe but evidently not useful. Thus, it is important to study the performance improvement of S4VMs. The following proposition shows that S4VMs achieve the maximal performance improvement in the worst cases.

定理1.如果y 2 fy ^ tgTt1和1，则y的精度永远不会比ysvm的精度差。

证明。 注意y是最优解和Jðysvm; y ^;ysvmÞ对于任何y ^都为零，我们有minJðy; y ^;ysvmÞminJðysvm; y ^;ysvmÞ= 0：（5）

Ÿ^2M Y 12M

进一步注意y 2 M，我们有

JDY; Y;ysvmÞMINY^2MJðy; Y^;ysvmÞ：（6）

来自Eqs。 （5）和（6），Jðy; y;ysvmÞ0，即gainðy; y; ysvmÞvalueðy; y;ysvmÞ。 回想一下，然后我们得到了收益; y;ysvmÞlossðy; y;ysvmÞ因此证明了该定理。TU

根据定理1，很容易得到以下命题。

命题1.satisfyingsvm。 minIfy ^y2M2Jðfyy^; ty ^ g;Tt¼y1svmandÞ0，永远不会比那个y，y的任何y的精度都差

简单地输出归纳SVM的预测结果也是安全的，但显然没有用。 因此，研究S4VM的性能改进非常重要。 以下命题表明，在最坏的情况下，S4VM可以实现最大的性能提升。

Proposition 2. If y2 fy^tgTt¼1 and ¼ 1, the accuracy of ~~y~~ achieves the maximal performance improvement over that of ysvm in the worst cases.

It is noteworthy that S4VMs are somewhat relevant to ensemble methods [49], and the spirit of S4VMs is not specific to S3VMs, which may also be extended to other semisupervised learning methods.

In the following, we will present the optimization of Eq. (4) and an out-of-sample extension of S4VMs in

Sections 4.1.1 and 4.1.2, respectively.

命题2.如果y 2 fy ^ tgTt = 1和¼1，在最坏的情况下，y的精度达到了ysvm的最大性能改善。

值得注意的是，S4VM与集合方法[49]有些相关，而S4VM的精神并不是特定于S3VM的，它们也可以扩展到其他半监督学习方法。

在下文中，我们将介绍方程的优化。 （4）和S4VMs的样本外扩展

第4.1.1和4.1.2节分别。

## 4.1.1 Optimization

Note that the gainðy;y^;ysvmÞ and lossðy;y^;ysvmÞ are linear functions with respect to y, i.e.,

4.1.1优化

注意，gainðy; y ^;ysvmÞ和lossðy; y ^;ysvmÞ是关于y的线性函数，即

svm lþu svm

gainðy;y^;y Þ ¼ X Iðyj ¼ y^jÞIy^j 6¼ yj

j¼lþ1

¼ lþu 1 þ yjy^j 1 ysvmj y^j

; j¼lþ1 2 2

X

svm lþu svm

lossðy;y^;y Þ ¼ X Iðyj 6¼ y^jÞIy^j ¼ yj

j¼lþ1

¼ lþu 1 yjy^j 1 þ ysvmj y^j

: j¼lþ1 2 2

X

Hence, Jðy;y^t;ysvmÞ is also linear to y and can be cast as c0ty þ dt, where ct ¼ 14½ð1 þ Þy^t þ ð 1Þysvm and dt ¼ 14½ð1 þ Þy^0tysvm þ ð1 Þ.

因此，Jðy; y ^ t;ysvmÞ也与y呈线性关系，可以转换为c0tyþdt，其中ct =14½ð1þÞy^tþð1Þysvm和dt¼14½ð1þÞy^ 0tysvmþð1Þ。

By introducing an additional variable t, the inner minimization in Eq. (4) can be reformulated as a maximization problem, and Eq. (4) becomes

通过引入另一个变量t，方程式中的内部最小化。 （4）可以重新表述为最大化问题，而Eq。 （4）变成

maxy maxt t

s. t. t c0ty þ dt;8t ¼ 1;...;T; y 2 f 1gu:

(7)

Though Eq. (7) is still a difficult mixed-integer linear programming problem, according to Proposition 1, optimal solutions are not necessary for achieving safeness. A simple method is then presented. Specifically, we first relax the integer-form of constraint f1gu into its convex hull ½1;1u, and obtain the optimal solution of the resultant convex linear programming problem. We then project it back to an integer solution with the minimum distance. If the objective value of the resultant integer solution is smaller than zero, ysvm is output as the final solution. It is not hard to verify that our solution satisfies Proposition 1.

虽然Eq。 （7）仍然是一个困难的混合整数线性规划问题，根据命题1，最优解不是实现安全性所必需的。 然后介绍一种简单的方法。 具体来说，我们首先将约束f1gu的整数形式放宽到其凸壳½1; 1u，并获得所得凸线性规划问题的最优解。 然后我们将其投影回具有最小距离的整数解。 如果得到的整数解的目标值小于零，则输出ysvm作为最终解。 验证我们的解决方案是否满足命题1并不难。

It is notable that prior knowledge on low-density separators can be easily incorporated into our framework. Specifically, by introducing the dual variables a ¼ ½a1;...;aT 0 0 for the constraints in Eq. (7), one can have the Lagrangian of Eq. (7) as

值得注意的是，关于低密度分离器的先验知识可以很容易地纳入我们的框架中。 具体而言，通过引入双变量a =½a1; ...; aT 0 0来表示方程式中的约束。 （7），可以得到方程的拉格朗日量。 （7）as

T

Lðt;y; aÞ ¼ t Xatðt c0ty dtÞ: (8)

t¼1

Setting the partial derivation w.r.t. t to zero, we have

设置部分派生w.r.t. t为零，我们有

T

@L=@t ¼ 1 Xat ¼ 0: (9)

t¼1

With Eq. (9), the inner maximization of Eq. (7) can be replaced by its dual and Eq. (7) becomes

随着Eq。 （9），方程的内在最大化。 （7）可以用它的双重和Eq代替。 （7）变成

T

maxu min Xatðc0ty þ dtÞ: (10)

y2f1g PTt¼1at¼1;a 0 t¼1

Here at can be interpreted as a probability that y^t discloses the ground-truth solution. Hence, if prior knowledge about the probabilities a is available, one can readily learn the optimal y with respect to the target in Eq. (10) using the known a.

这里可以解释为y ^ t公开地面实况解的概率。 因此，如果关于概率a的先验知识可用，则可以容易地获得关于等式1中的目标的最优y。 （10）使用已知的a。

## 4.1.2 Out-of-Sample Extension

Eq. (4) works in the transductive setting [40] which could not make predictions on unseen instances. To overcome this, an out-of-sample extension (also named as induction extension [52]) of S4VMs is presented.

4.1.2样本外扩展

式。 （4）在转换设置[40]中工作，无法对看不见的实例进行预测。 为了克服这个问题，提出了S4VM的样本外扩展（也称为感应扩展[52]）。

One common practice to achieve this is to freeze the transductive setting on the set of both testing and unlabeled instances [52]. Formally, for any given testing instancelow-density separators, andz, let fy^zt gTt¼1 be the predictive labels of multipleysvm;z be the predictive label of the inductive SVM. One need to learn a label assignment for both testing and unlabeled instances such that the objective of S4VM is maximized max t

实现这一目标的一种常见做法是冻结测试和未标记实例集上的转换设置[52]。 形式上，对于任何给定的测试实例低密度分离器，andz，让fy ^ ztgTt¼1成为multipleysvm的预测标签; z是归纳SVM的预测标签。 需要为测试和未标记的实例学习标签分配，以使S4VM的目标最大化

y2f1gu; yz2f1g; t (11)

s.t. t ½ct;cz0½y;yz þ dzt ;8t ¼ 1;...;T;

where cz ¼ 14½ð1 þ Þy^zt þ ð 1Þysvm;z and dzt ¼ dt ð1þ

Þy^zt ysvm;z. This, however, will be computationally prohibitive especially when there are a large number of instances for testing.

其中cz¼14½ð1þÞy^ ztþð1Þysvm; z和dzt¼dtð1þ

Þy^ zt ysvm; z。 然而，这在计算上是禁止的，特别是当存在大量测试实例时。

To alleviate the computational load, we present an efficient algorithm for approximate solutions. Specifically, note that when yz is fixed to ysvm;z, Eq. (11) is equivalent to transductive S4VM, i.e., Eq. (7), and thus the solution of Eq. (7) (denoted by y) provides a quite good approximation to Eq. (11). This observation motivates us to solve the following much simpler problem instead of the complicated one in Eq. (11),

为了减轻计算量，我们提出了一种有效的近似解算法。 具体来说，请注意当yz固定为ysvm时; z，Eq。 （11）相当于转导S4VM，即Eq。 （7），因而解决了方程式 （7）（用y表示）提供了对方程的非常好的近似。（11）。 这种观察促使我们解决以下更简单的问题，而不是方程式中的复杂问题。（11），

max t

yz2f1g; t (12)

s.t. t ½ct;cz0½y;yz þ dzt ;8t ¼ 1;...;T:

It is efficient to derive the optimal solution of Eq. (12). We just need to enumerate the two possible values of yz and then pick up the one with the smaller objective value. As will be validated empirically in Section 5.2, our approximation is quite effective.

导出方程的最优解是有效的。（12）。 我们只需要枚举yz的两个可能值，然后选择具有较小目标值的值。 正如将在5.2节中凭经验验证的那样，我们的近似非常有效。

4.2 Generating the Pool of Diverse Separators Denote hðf;y^Þ as the objective function of S3VMs in Eq. (1) for the sake of simplicity,

2 l lþu hðf;y^Þ ¼kfk þ C1X‘ðyi;fðxiÞÞ þ C2 X ‘ðy^j;fðxjÞÞ:

H i¼1 j¼lþ1

To generate a pool of diverse separatorsT fftgTt¼1 and their corresponding label assignments fy^tgt¼1, in this paper we consider to minimize the following function:

T

fft;y^t2BgtT¼1 Xt¼1 ð Þ þ f g ¼1: (13) min h ft;y^t MV y^t Tt

Here V refers to a penalty reflecting the diversity of separators, i.e., the larger the diversity, the smaller the penalty. M is a large constant (e.g., 105 in our experiments) enforcing large diversity. It is easy to realize that minimizing Eq. (13) favors the separators with large margins as well as large diversities.

4.2生成不同分隔符池表示hðf; y ^Þ作为方程式中S3VM的目标函数。 （1）为简单起见，

2llþuhðf; y ^Þ¼kfkþC1X'ðyi;fðxiÞÞþC2X'ðy^ j;fðxjÞÞ：

^ hi¼1j¼lþ1

为了生成一个不同的分隔符池fftgTt¼1及其相应的标签分配fy ^tgt¼1，在本文中我们考虑最小化以下函数：

Ť

fft; y ^t2BgtT¼1Xt¼1ðÞþfg¼1：（13）min h ft; y ^ t MV y ^ t Tt

这里V指的是反映分隔符多样性的惩罚，即多样性越大，惩罚越小。 M是一个大的常数（例如，在我们的实验中为105），强大的多样性。 很容易意识到最小化Eq。 （13）支持利润率高且分散性大的分离器。

We consider the penalty as a sum of pairwise terms, i.e., Vðfy^tgTt¼1Þ ¼ P1t¼6 t~2T d½ðy^0tuy^t~ 1 &Þ where d is the indicator function and & 0;1 is a constant (e.g., 0:5 in our experiments). It is notable that other penalty quantities can be also applicable.

Recall that fðxÞ ¼ w0fðxÞ þ b is a linear model in S3VMs, where fðxÞ is a feature mapping induced by the kernel k, i.e., kðx;x^Þ ¼ fðxÞ0fðx^Þ and b is a bias term. Eq. (13) then becomes

我们将惩罚视为成对项的总和，即Vðfy^tgTt¼1Þ¼P1t¼6t~2Td½ðy^ 0tuy ^ t~1＆Þ其中d是指标函数，＆0; 1是常数（例如，0：5 in 我们的实验）。 值得注意的是，其他惩罚数量也可适用。

回想一下，在S3VMs中，fðxÞ是一个线性模型，其中fðxÞ是由核k引起的特征映射，即kðx; x ^Þ¼fðxÞ0fðx^Þ和b是一个偏差项。式。 （13）然后成为

T l

t tmint t 1 X1kwtk2 þ C1X‘ðyi;w0tfðxiÞ þ btÞ

fw ;b ;y^ 2BgT¼ t¼1 2 i¼1

lþu

þ C2 X ‘ðy^t;j;w0tfðxjÞ þ btÞ (14)

j¼lþ1 y^0ty^t~ þ M d 1 &:

u

1t6¼t~T

X

To address Eq. (14), in the sequel, two implementations are presented. One is based on a global simulated annealing (SA) search while the other is based on an efficient sampling strategy.

It is notable that exhaustively searching all possible large-margin low-density separators is prohibitive. Fortunately, according to Theorem 1, generating a large-margin low-density separator to realize the ground-truth is only a sufficient rather than necessary condition to have safe S3VMs. As will be validated in our empirical studies, even on many cases in which the ground-truth is not realized by any of the generated large-margin low-density separators, S4VMs still work quite well.

解决方程式 （14），在续集中，提出了两种实现方式。 一种是基于全局模拟退火（SA）搜索，而另一种是基于有效的采样策略。

值得注意的是，彻底搜索所有可能的大边缘低密度分离器是令人望而却步的。 幸运的是，根据定理1，生成一个大边距的低密度分离器来实现真实性只是拥有安全S3VM的充分而非必要的条件。 正如我们的实证研究中所证实的那样，即使在许多情况下，任何生成的大边缘低密度分离器都没有实现真实性，S4VM仍然可以很好地工作。

## 4.2.1 Global Simulated Annealing Search

Our first implementation to address Eq. (14) is based on global search, e.g., simulated annealing search [25]. SA is a probabilistic method for approaching global solutions of objective functions which suffer from multiple local minima. Specifically, at each step, SA replaces the current solution by a random nearby solution with a probability. The probability depends on two factors, i.e., the value difference between their corresponding function targets, and a global parameter, i.e., the temperature P, which gradually decreases during the process. When P is large, the current solution almost changes randomly. While as P approaches zero, the changes are increasingly “downhill”. In theory, the probability that SA converges to the global solution approaches to 1 as SA procedure is continued [26].

4.2.1全局模拟退火搜索

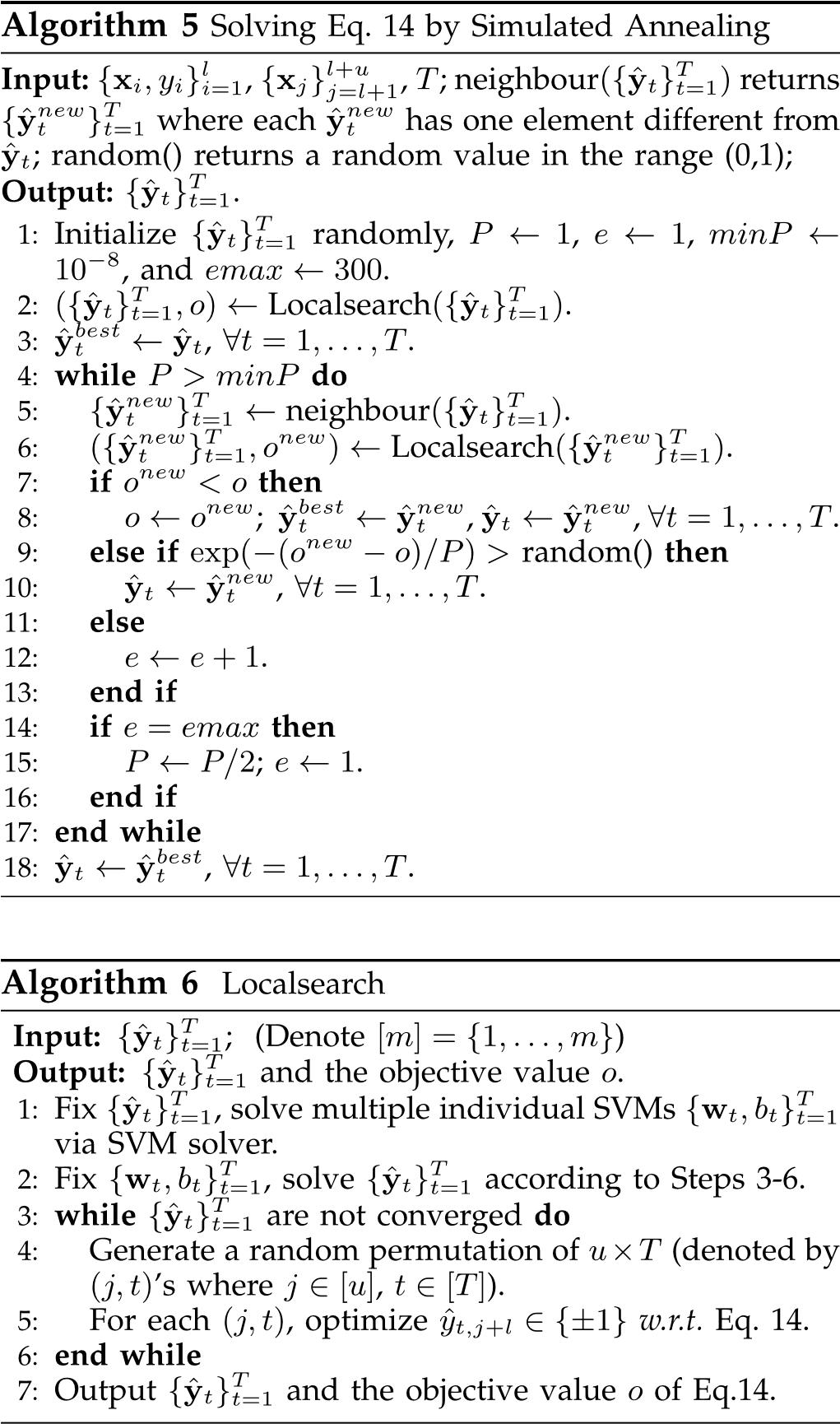
我们的第一个实现，以解决方程式 （14）基于全局搜索，例如模拟退火搜索[25]。 SA是用于接近具有多个局部最小值的目标函数的全局解的概率方法。 具体而言，在每个步骤中，SA用概率随机附近解决方案替换当前解决方案。 概率取决于两个因素，即它们相应的功能目标之间的值差异，以及在该过程期间逐渐减小的全局参数，即温度P. 当P很大时，当前的解决方案几乎随机变化。 当P接近零时，变化越来越“下坡”。 理论上，随着SA程序的继续，SA收敛到全局解的概率接近1 [26]。

To alleviate the low convergence rate of standard SA, inspired by [37], a deterministic local search scheme is used. Specifically, when fy^tgTt¼1 are fixed, fwt;btgTt¼1 are solvedT via multiple individual SVM subroutines. WhenT fwt;btgt¼1 are fixed, fy^tgt¼1 are updated based on local binary search.

Algorithm 5 presents the pseudo-code of our simulated annealing approach for Eq. (14), where the local search subroutine is given in Algorithm 6.

为了减轻标准SA的低收敛速度，受[37]的启发，使用了确定性局部搜索方案。 具体来说，当fy ^tgTt¼1固定时，fwt;btgTt¼1通过多个单独的SVM子程序求解。 当T fwt;btgt¼1固定时，基于本地二进制搜索更新fy ^tgt¼1。

算法5给出了我们的模拟退火方法的伪代码。 （14），其中局部搜索子程序在算法6中给出。



## 4.2.2 Representative Sampling

To further alleviate the computational burden, our second implementation is based on heuristic representative sampling. Recall that the goal of Eq. (13) can be realized by finding multiple large-margin low-density separators and then keeping only representative ones with large diversity. This motivates us to have a two-stage method, a) search for multiple large-margin low-density separators at first and then b) select the representative separators. Algorithm 7 presents the pseudo-code of our second implementation.

As Algorithm 7 shows, multiple candidate largemargin low-density separators are first obtained by [46]. A clustering algorithm is then applied to identify the representative separators. This approach is simple. As will be validated empirically in Section 5, it is also efficient and effective.

We call our S4VM using simulated annealing as S4VMa, and the one using sampling as S4VMs.

4.2.2代表性抽样

为了进一步减轻计算负担，我们的第二个实现基于启发式代表性采样。 回想一下Eq的目标。 （13）可以通过找到多个大边距低密度分离器，然后只保留具有大量多样性的代表性分离器来实现。 这促使我们采用两阶段方法，a）首先搜索多个大边距低密度分离器，然后b）选择代表性分离器。 算法7给出了我们第二种实现的伪代码。

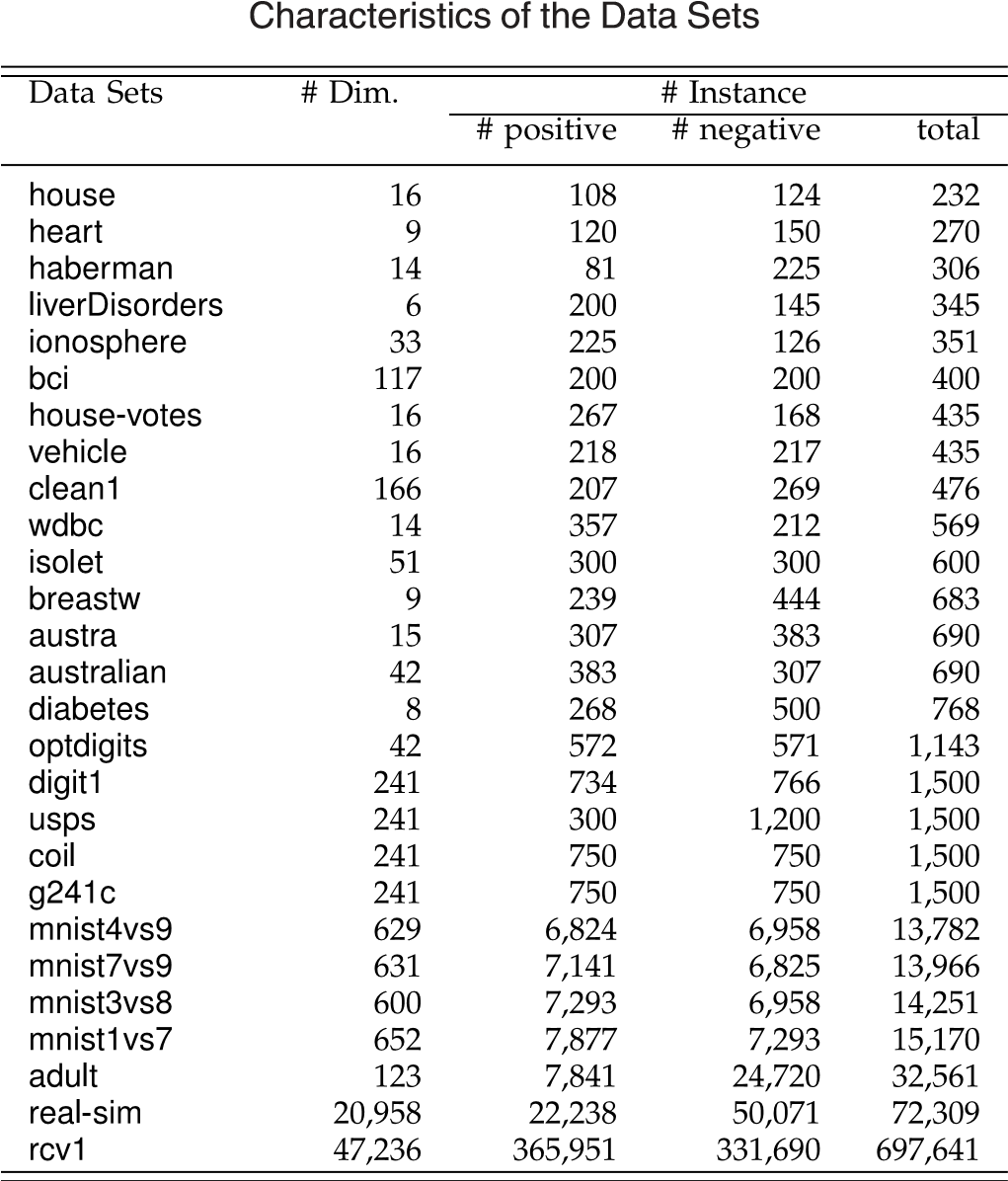
如算法7所示，首先通过[46]获得多个候选大型低密度分离器。 然后应用聚类算法来识别代表性分隔符。 这种方法很简单。 正如将在第5节中凭经验验证的那样，它也是高效和有效的。

我们将使用模拟退火的S4VM称为S4VMa，将采样作为S4VM使用。

## 5 EMPIRICAL STUDY

In this section, the proposed approaches are evaluated on a broad range of tasks including five semi-supervised

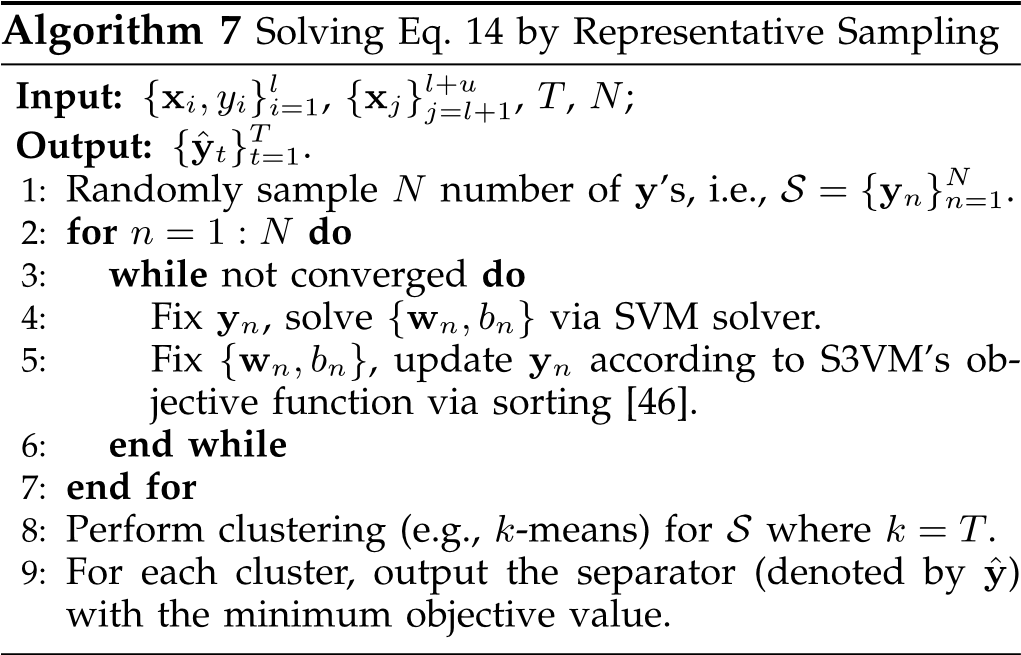
TABLE 1



benchmark data sets,1 digit1, USPS, BCI, g241c, COIL, and 15 UCI data sets2 and four large scale data sets, adult, mnist, real-sim, rcv1. The size of data ranges from 232 to more than 600;000, and the dimensionality ranges from 6 to more than 40;000. mnist has 45 pairs of binary classification problems, and we focus on its four most difficult pairs [46]. Table 1 summarizes the characteristics of the data sets.

5实证研究

在本节中，所提出的方法在广泛的任务上进行评估，包括5个半监督基准数据集，1个数字1，USPS，BCI，g241c，COIL和15个UCI数据集2和4个大规模数据集，成人，mnist ，real-sim，rcv1。 数据的大小范围从232到超过600;并且维度范围从6到超过40,000。 mnist有45对二元分类问题，我们关注其中最困难的四对[46]。 表1总结了数据集的特征。



To satisfy the balance constraint required by S3VMs, for each data set, we randomly select 10 instances whose class proportion is closely related to the whole data set, to be served as labeled instances. The remaining data are served as the unlabeled instances. The experiments repeat for 30 times. The average performance and standard deviation are recorded.

为了满足S3VM所需的平衡约束，对于每个数据集，我们随机选择10个类比例与整个数据集密切相关的实例作为标记实例。 其余数据用作未标记的实例。 实验重复30次。 记录平均性能和标准偏差。

1. http://www.kyb.tuebingen.mpg.de/ssl-book/.
2. http://archive.ics.uci.edu/ml/datasets.html.

Inductive SVM and S3VM serve as the two baseline approaches. For small and medium scale data sets, LIBSVM3 [18] and TSVM4 [23] are employed. For large scale data sets, due to the high computational load of LIBSVM and TSVM, efficient LIBLINEAR5 [21] and UniverSVM6 [15] serve as baselines instead. Both the linear and RBF kernels are used for small and medium scale data sets, and linear kernel is always used for large scale data sets.

对于中小规模数据集，使用LIBSVM3 [18]和TSVM4 [23]。 对于大规模数据集，由于LIBSVM和TSVM的高计算负荷，高效的LIBLINEAR5 [21]和UniverSVM6 [15]代替了基线。 线性和RBF内核都用于中小规模数据集，线性内核总是用于大规模数据集。

Three S3VM variants using multiple low-density separators are also compared. Specifically, S3VMbest presents the best performance among the multiple candidate separators (note that this method is impractical). S3VMmin selects the low-density separator with minimum objective value. S3VMcom combines the candidate separators using uniform weights.

还比较了使用多个低密度分离器的三种S3VM变体。 具体来说，S3VMbest在多个候选分隔符中呈现最佳性能（请注意，此方法不切实际）。 S3VMmin选择具有最小目标值的低密度分隔符。 S3VMcom使用统一权重组合候选分隔符。

The parameters are set as follows. Following the setups in [10], the regularization parameter C is fixed to 100 and the width of RBF kernel is set to the average distance between instances for inductive SVM. The regularization parameters C1, C2 and b in the balance constraint are fixed to 100, 0:1 and 0:1 for all S3VMs and S4VMs. For S3VM-c, the cluster number k is fixed to 50. For S3VM-p, the parameter h is fixed to 0.1 and the similarity matrix is constructed via Gaussian distance where the width is set to the average distance between instances. For S3VM-us, the parameter is fixed to 0.1. For S4VMa, the number of separators T and the risk parameter are both fixed to 3. For S4VMs, the sampling size N, the number of separators T, and the risk parameter are fixed to 100, 10 and 3, respectively. The linear program in S4VMs is conducted using the linprog function in MATLAB.

参数设置如下。 在[10]中的设置之后，正则化参数C被固定为100并且RBF内核的宽度被设置为用于归纳SVM的实例之间的平均距离。 对于所有S3VM和S4VM，平衡约束中的正则化参数C1，C2和b固定为100,0：1和0：1。 对于S3VM-c，簇号k固定为50.对于S3VM-p，参数h固定为0.1，并且相似性矩阵通过高斯距离构造，其中宽度设置为实例之间的平均距离。 对于S3VM-us，参数固定为0.1。 对于S4VMa，分隔符数T和风险参数都固定为3.对于S4VM，采样大小N，分隔符数T和风险参数分别固定为100,10和3。 S4VM中的线性程序使用MATLAB中的linprog函数进行。

### 5.1 Comparison Results

Intensive comparison results are shown in Table 2. Although simulated annealing was used to improve the efficiency of S3VMs [37], it still involves high computational load. Table 2 only reports the performance of S4VMa on 11 small UCI data sets.

5.1比较结果

密集的比较结果如表2所示。虽然模拟退火用于提高S3VM的效率[37]，但它仍然涉及高计算负荷。 表2仅报告了11个小型UCI数据集上S4VMa的性能。

Table 2 shows that S4VMa performs highly competitive with S3VM. Specifically, S3VM significantly outperforms inductive SVM on 5 of the 11 cases with linear kernel, and 7 of the 11 cases with RBF kernel; while S4VM significantly outperforms inductive SVM on seven cases for both the linear and RBF kernels.

表2显示S4VMa与S3VM的竞争非常激烈。 具体来说，S3VM在线性内核11例中有5例明显优于归纳SVM，11例RBF内核中有7例; 而对于线性和RBF内核，S4VM在七种情况下明显优于归纳SVM。

More importantly, unlike S3VM which causes significant degeneration of the performance on one case with linear kernel and two cases with RBF kernel, S4VMa is never inferior to inductive SVM. The Wilcoxon sign tests at 95 percent significance level confirm that S4VMa is significantly better than inductive SVM with both linear and RBF kernels, but S3VM does not show such a significance.

更重要的是，与使用线性内核和两个RBF内核的情况导致性能显着退化的S3VM不同，S4VMa从不逊色于归纳SVM。 Wilcoxon符号测试的显着性水平为95％，证实S4VMa明显优于线性和RBF内核的归纳SVM，但S3VM没有显示出如此重要的意义。

Table 2 also shows the highly competitive performance of S4VMs and S3VM-us compared with S3VM. Specifically, in terms of pairwise comparison, S4VMs is found to be superior to S3VM on 16 of the 27 cases with linear kernel, and 11 of the 20 cases with RBF kernel. S3VM-us is superior to S3VM on nine and eight of the 20 cases with linear and RBF kernel, respectively. In terms of wins, with linear kernel, S3VM outperforms inductive SVM on 44 percent (12/27) of the cases; while S4VMs and S3VM-us outperform inductive SVM on 59 percent (16/27) and 45 percent (9/20), respectively. Similar observations can be found for RBF kernel. On 55, 55 and 50 percent of the cases, S3VM, S4VMs and S3VM-us significantly outperform inductive SVM, which are also competitive.

表2还显示了S4VM和S3VM-us与S3VM相比具有高度竞争力的性能。 具体来说，在成对比较方面，27个线性内核中的16个发现S4VM优于S3VM，而20个RBF内核中有11个发现S4VM优于S3VM。 S3VM-us在线性和RBF内核的20个案例中分别有9个和8个优于S3VM。 在胜利方面，对于线性内核，S3VM在44％（12/27）的情况下优于归纳SVM; 而S4VM和S3VM-us的表现分别优于感应式SVM，分别为59％（16/27）和45％（9/20）。 对于RBF内核可以找到类似的观察结果。 在55％，55％和50％的案例中，S3VM，S4VM和S3VM-us明显优于归纳SVM，后者也具有竞争力。

Unlike S3VM whose performance is found to decrease significantly on three cases with linear kernel and six cases with RBF kernel, S3VM-us shows decreased performance on only one case, and S4VMs never show decreased performance. Both S3VM-c and S3VM-p are capable of reducing the chance of performance degeneration, but they do not perform as well as S3VM-us. S3VMmins and S3VMcoms still show significantly reduced performance in many cases. The Wilcoxon sign tests at 95 percent significance level validate S4VMs and S3VM-us to be significantly better than inductive SVM with both linear and RBF kernels, but other semi-supervised methods, such as S3VM, S3VM-c, S3VM-p, S3VMmins and S3VMcoms , do not obtain significance.

与S3VM不同，S3VM的性能在线性内核和6个RBF内核的情况下显着下降，S3VM-us仅在一种情况下表现出性能下降，而S4VM从未表现出性能下降。 S3VM-c和S3VM-p都能够降低性能退化的可能性，但它们的性能不如S3VM-us。 在许多情况下，S3VMmins和S3VMcoms仍然显示出显着降低的性能。 具有95％显着性水平的Wilcoxon符号测试验证了S4VM和S3VM-us明显优于线性和RBF内核的归纳SVM，但其他半监督方法，如S3VM，S3VM-c，S3VM-p，S3VMmins和 S3VMcoms，没有意义。

Although S3VM-us is found to be safer than S3VM, it employs a conservative strategy and its improvement is often much smaller than that of S3VM. In contrast, S4VMs takes the improvement in performance into account and performs much better. Specifically, in terms of average performance, S4VMs is superior to S3VM-us. It reaches 75.91 percent versus S3VM-us’s 74.97 percent on the 40 cases of S3VM-us reported in Table 2. The paired t-tests at 95 percent significance level show that S4VMs performed significantly better than S3VM-us. These comparisons confirm that S4VMs is better than S3VM-us.

虽然发现S3VM-us比S3VM更安全，但它采用了保守的策略，其改进通常远小于S3VM。 相比之下，S4VM将性能方面的改进考虑在内并且表现更好。 具体而言，就平均性能而言，S4VM优于S3VM-us。 在表2中报告的40个S3VM-us案例中，它与S3VM-us的74.97％相比达到了75.91％。显着性水平为95％的配对t检验表明，S4VM的表现明显优于S3VM-us。 这些比较证实S4VMs优于S3VM-us。

The condition of Theorem 1 is already weaker than the traditional low-density assumption in S3VMs, the theorem may not always hold in practice. That is, the ground-truth may not reside among the low-density separators (cf. the performance of S3VMbests ). Even in such cases, S4VMs still work well. That might be because i) Theorem 1 only presents a sufficient rather than necessary condition for safeness, and ii) the analysis of the diversity among lowdensity separators [39], provides an explanation to S4VMs’ superiority to single separator.

定理1的条件已经弱于S3VM中的传统低密度假设，该定理在实践中可能并不总是成立。 也就是说，真实性可能不属于低密度分离器（参见S3VMbests的性能）。 即使在这种情况下，S4VM仍能正常运行。 这可能是因为i）定理1仅提供了足够而非必要的安全条件，以及ii）低密度分离器之间的多样性分析[39]，为S4VMs对单一分离器的优越性提供了解释。

### 5.2 Out-of-Sample Extension

Table 3 shows the performance of S4VMs with out-of-sample extension on small and medium scale data sets. For each data set, 75 percent of instances are used for training, among which 10 are served as labeled data and required to be satisfied by the balance constraint. The remaining instances are used for testing. Experiment repeats for 30 times. The average performance and standard deviation are recorded.

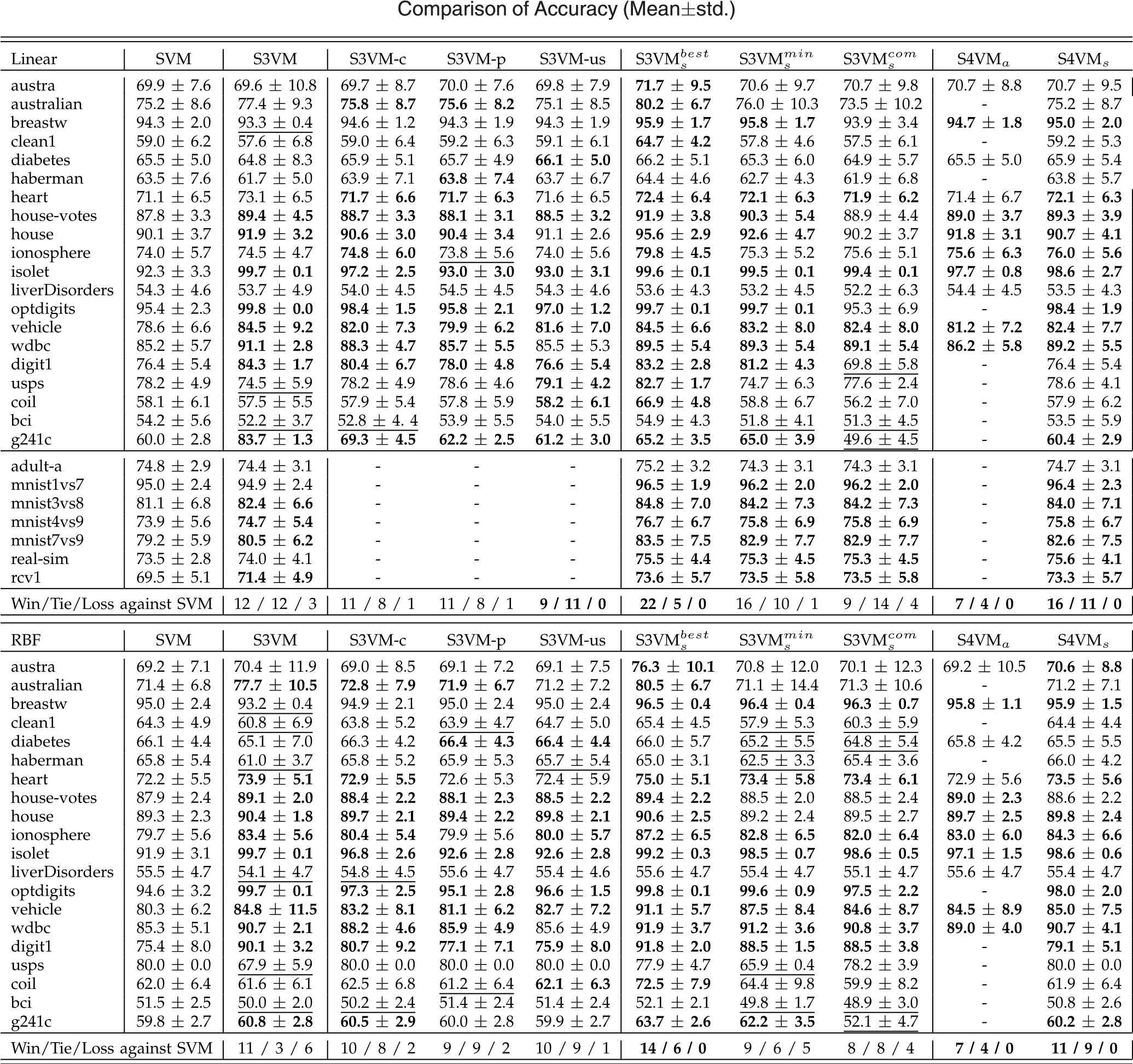
5.2样本外扩展

表3显示了S4VM在中小规模数据集上具有样本外扩展的性能。 对于每个数据集，75％的实例用于培训，其中10个用作标记数据，并且需要通过平衡约束来满足。 其余实例用于测试。 实验重复30次。 记录平均性能和标准偏差。

|  |  |
| --- | --- |
|  | of wins, S4VMs performs the best in comparison with the |
| 1. http://www.csie.ntu.edu.tw/~cjlin/libsvm/. 2. http://svmlight.joachims.org/. 3. http://www.csie.ntu.edu.tw/~cjlin/linlinear/. 4. http://mloss.org/software/view/19/. | other three S3VMs. More importantly, unlike the other S3VMs, such as S3VM, S3VMmins and S3VMcoms , which show significant performance reductions in many cases,  从表3可以看出，S4VM在样本外扩展方面运行良好。 具体来说，就胜利而言，S4VM与之相比表现最佳  其他三个S3VM。 更重要的是，与其他S3VM不同，例如S3VM，S3VMmins和S3VMcoms，它们在许多情况下显示出显着的性能降低， |

As can be seen from Table 3 that S4VMs works quite well with out-of-sample extension. Specifically, in terms

TABLE 2



Entries of semi-supervised methods (S3VM, S3VM-c, S3VM-p, S3VM-us, S3VMbests , S3VMmins , S3VMcoms , S4VMa and S4VMs) are bolded/underlined if they are significantly better/worse than SVM (paired t-tests at 95 percent significance level). ‘-’ marks cases suffering from high computational cost

or memory overhead.

S4VMs is never inferior to inductive SVM. The Wilcoxon sign tests at 95 percent significance level confirm that S4VMs is significantly better than inductive SVM with both linear and RBF kernels, and the other three S3VMs do not achieve significance.

S4VM从不逊色于归纳SVM。 Wilcoxon符号测试的显着性水平为95％，证实S4VM明显优于线性和RBF内核的归纳SVM，其他三个S3VM没有达到显着性。

### 5.3 Influence of the Number of Labeled Data

Table 4 shows the performance of S4VMs under different numbers of labeled examples. As can be seen from Table 4 that S4VMs is found to be highly competitive with S3VM for each number of labeled examples. Specifically, in terms of wins, S3VM obtains significance on 19/20/20 of the 40 cases for 20, 50 and 100 labeled examples, respectively; while S4VMs outperforms on 20/20/17 cases accordingly.

5.3标记数据的影响

表4显示了在不同数量的标记示例下S4VM的性能。 从表4可以看出，对于每个标记的示例，发现S4VM与S3VM具有高度竞争性。 具体来说，就胜利而言，S3VM分别在20个，50个和100个标签示例的40个案例的19/20/20中获得重要性; 而S4VM在相应的20/20/17案例中表现优异。

In terms of pairwise comparison (suppose win, tie and loss stand for scores of 1, 0 and 1 for each data set), S4VMs outscores S3VM on seven data sets, scores the same as S3VM on seven data sets, and lower on six data sets.

在成对比较方面（假设win，tie和loss代表每个数据集的分数为1,0和1），S4VM在7个数据集上超过S3VM，在7个数据集上得分与S3VM相同，在6个数据上得分更低集。

|  |
| --- |
| TABLE 3  Comparison of Accuracy (Meanstd.) with Out-of-Sample Extension  表3  精确度（meanstd。）与样本外扩展的比较 |

More importantly, in contrast to S3VM that significantly reduces performance on 17 cases, S4VMs only shows decreased performance on three cases which all happen on liverDiscorders with linear kernel. The might be because, in that setting, even the S3VMbests approach (which always selects the best candidate separator) cannot achieve a comparable performance against the inductive SVM (the accuracies of S3VMbests are 56.9, 61.2 and 64.5 for 20, 50 and 100 labeled examples, which are all significantly inferior to the inductive SVM). The Wilcoxon sign tests at 95% significance level confirm that S4VMs is significantly better than the inductive SVM on each number of label examples, whereas S3VM does not show significance.

更重要的是，与在17个案例中显着降低性能的S3VM相比，S4VM仅显示三个案例的性能下降，这些都发生在具有线性内核的liverDiscorders上。 可能是因为，在该设置中，即使是S3VMbests方法（总是选择最佳候选分隔符）也无法实现与归纳SVM相当的性能（对于20,50和100个标记示例，S3VMbests的准确度为56.9,61.2和64.5 ，这些都明显不如归纳SVM）。 在95％显着性水平的Wilcoxon符号测试证实，在每个标签示例数量上，S4VM明显优于感应SVM，而S3VM没有显示出显着性。

5.4 Influence of the Number of Unlabeled Data Table 5 shows the performance of S4VMs with different numbers of unlabeled instances. As can be seen, similar to the cases in Section 5.3, S4VMs still performs highly competitive with S3VM, both in terms of the wins as well as the pairwise comparison. Furthermore, unlike S3VM which significantly hurts performance on 23 cases, S4VMs never shows decreased performance. The Wilcoxon sign tests at 95 percent significance level still conform that S4VMs is significantly better than inductive SVM on each number of unlabeled instances, and S3VM does not show such a significance.

5.4未标记数据的数量的影响表5显示了具有不同数量的未标记实例的S4VM的性能。 可以看出，类似于5.3节中的情况，S4VM在胜利和成对比较方面仍然与S3VM竞争激烈。 此外，与23VM严重损害性能的S3VM不同，S4VM从未表现出性能下降。 在95％显着性水平的Wilcoxon符号测试仍然符合S4VM在每个未标记实例上明显优于归纳SVM，并且S3VM没有显示出如此重要性。

### 5.5 Influence of the Balance Constraint

One piece of prior knowledge of S3VMs is the balance constraint. Although the balance constraint is often a mild assumption, it might still be violated in some cases. To study the influence of the balance constraint, 10 labeled examples whose class proportion is substantially different from that of remaining unlabeled data, are randomly selected, and the balance constraint is still required for S3VMs and S4VM. Experiments are repeated for 30 times. The average performance and standard deviation on UCI data sets with linear kernel are reported in Table 6.

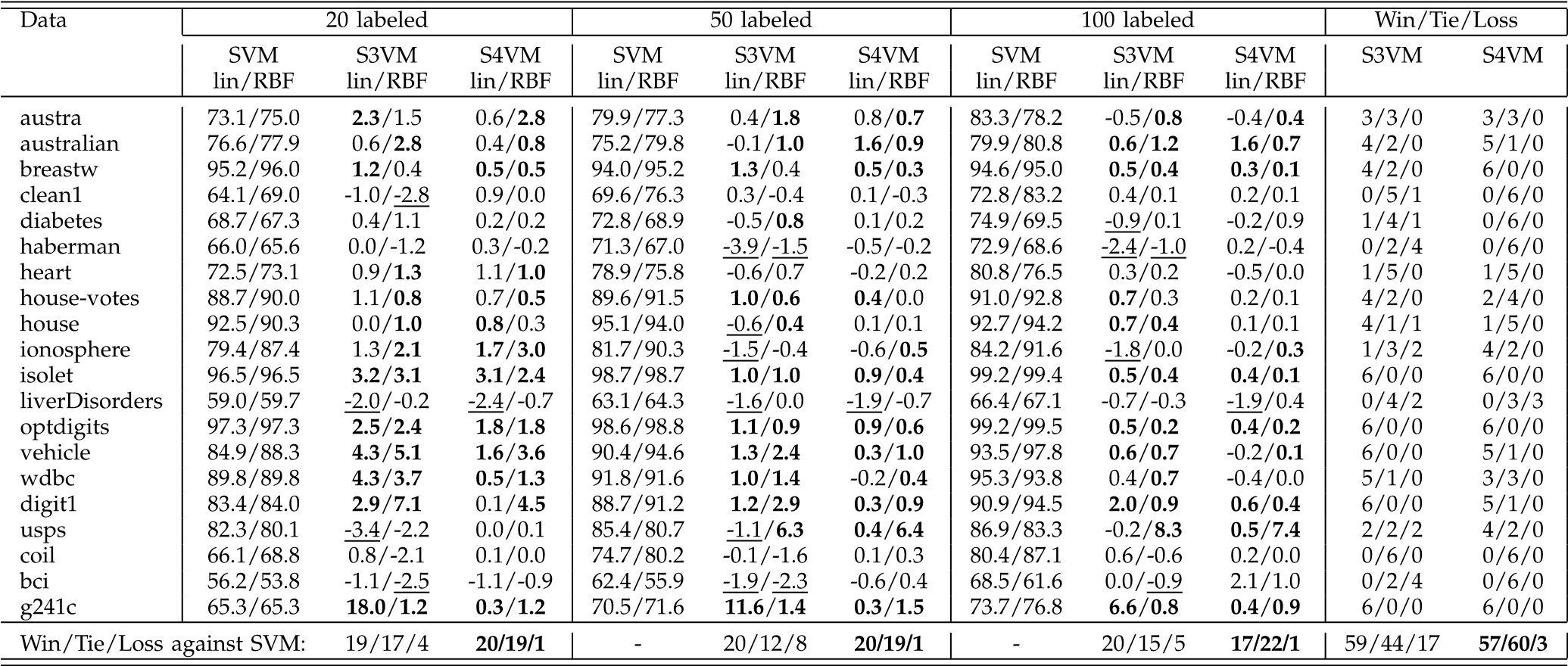
The results show that both the S4VMs and S3VM perform much worse than those without the violation of the balance constraint (cf. results in Table 2). Moreover, although S4VMs has already substantially improved the safeness of S3VM, it still shows significant decrease performance on two cases. This suggests that, in the cases in

TABLE 4

Accuracy of SVM and Accuracy Improvements of S4VMs and S3VM against SVM on Different Numbers of Labeled Data

表4

SVM的准确性和S4VM和S3VM对不同标记数据的SVM精度的改进



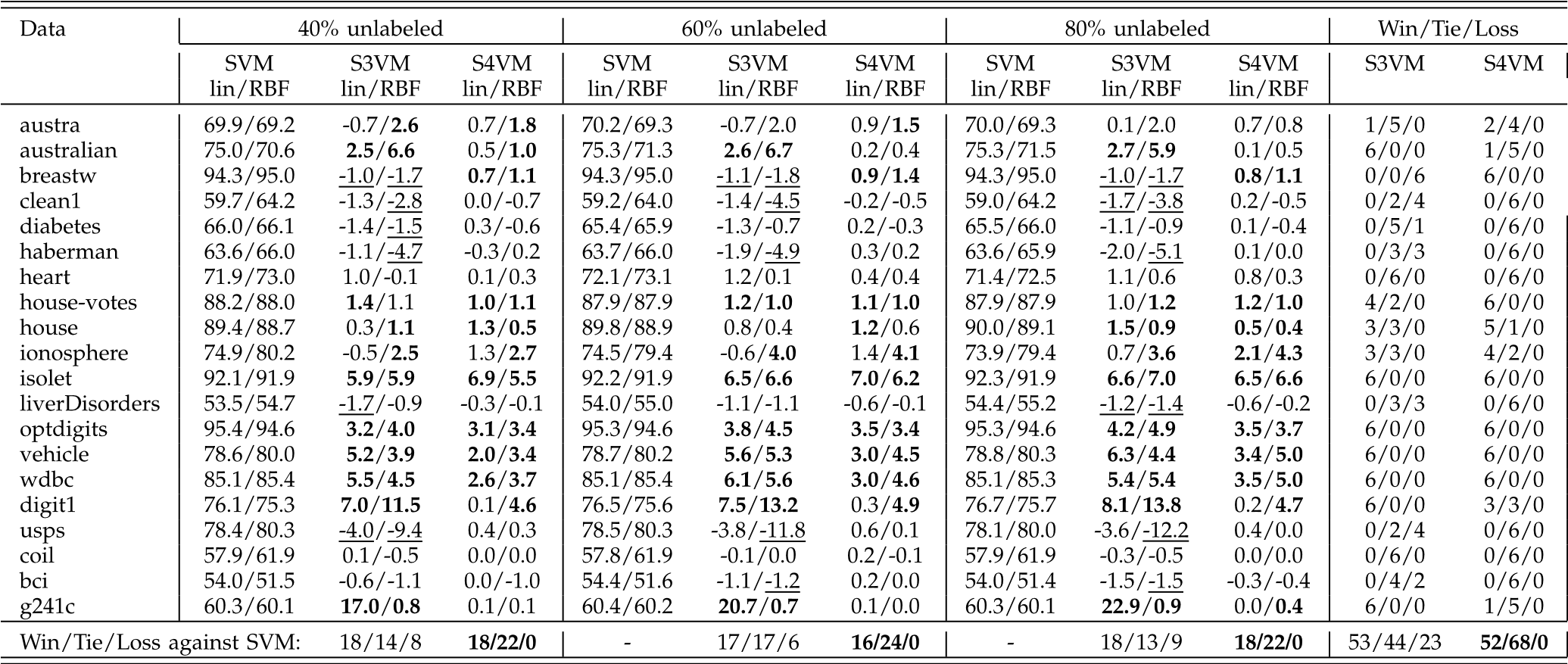
The accuracy Improvement of algo against SVM is calculated by ðaccalgo  accsvmÞ. ‘lin’ stands for the linear kernel.

TABLE 5

Accuracy of SVM and Accuracy Improvements of S4VMs and S3VM on Different Numbers of Unlabeled Data

表5

SVM的准确性和S4VM和S3VM对不同数据无标签数据的准确性改进



which the class proportion of unlabeled instances cannot be estimated using existing labeled examples, it is still challenging to have safe S3VMs.

5.5平衡约束的影响

S3VM的一个先验知识是平衡约束。虽然平衡约束通常是一个温和的假设，但在某些情况下可能仍会违反。为了研究平衡约束的影响，随机选择10个标记的例子，其类比例与剩余的未标记数据的比例大不相同，并且S3VM和S4VM仍然需要平衡约束。实验重复30次。表6中报告了具有线性内核的UCI数据集的平均性能和标准偏差。

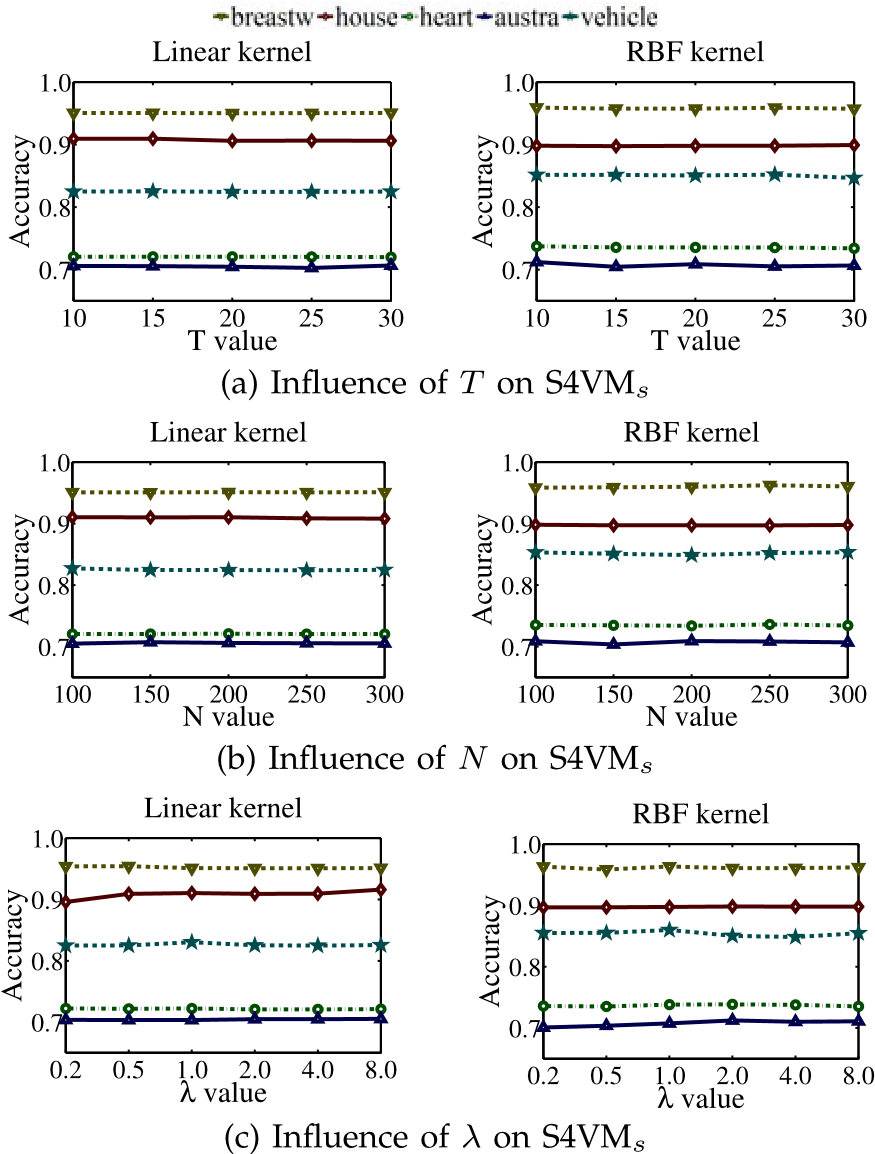
结果表明，S4VM和S3VM的性能都比没有违反平衡约束的情况差得多（参见表2中的结果）。此外，尽管S4VM已经大大提高了S3VM的安全性，但它仍然在两种情况下显示出显着的降低性能。这表明，在使用现有标记示例无法估计未标记实例的类别比例的情况下，拥有安全S3VM仍然具有挑战性。

### 5.6 Influence of Parameters

S4VMs has four parameters, i.e., sampling size N, cluster number T, risk parameter and the kernel type to set. In previous empirical studies, N, T and are set as default values, i.e., 100, 10 and 3. Fig. 2 further studies the influence of N, T and with linear and RBF kernels on five representative data sets (the results on other data sets are similar) with 10 labeled examples by fixing other examples is too small to afford a reliable model selection. Moreover, paired t-tests at 95 percent significance level confirm that S4VMs does not reduce performance on all the cases in Figs. 2a, 2b and 2c when 1.

### 5.7 Running Time

Following the setup in Section 5.2, Fig. 3 gives the training and testing time of S3VM and S4VMs with linear kernel on UCI data sets. S4VMs runs approximately 10 times of S3VM. That is because S4VMs needs to generate T low-density separators, where T is usually a small constant (such as

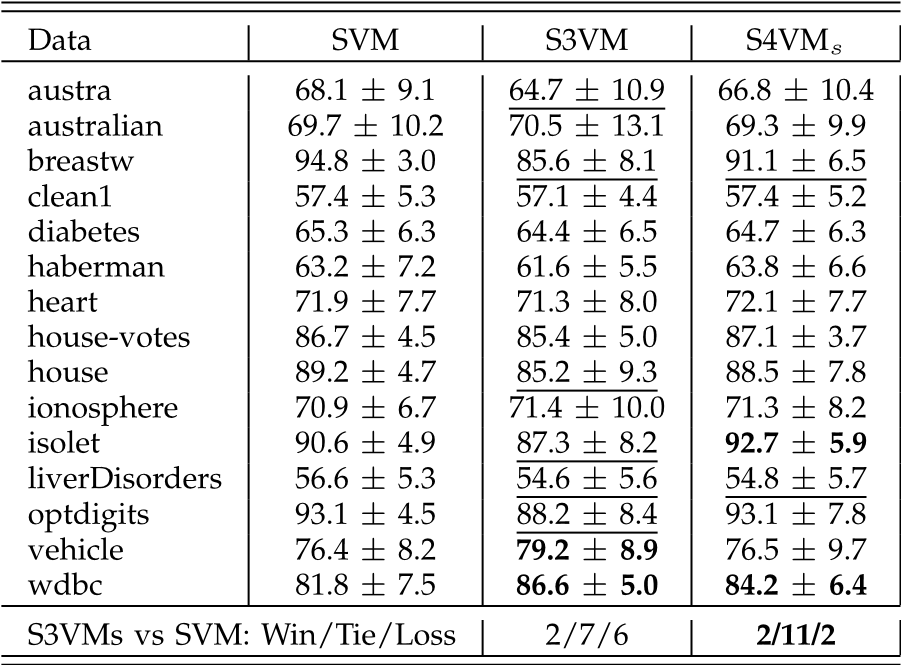
parameters as default values.

It can be seen that, though the number of labeled examples is small, the performance of S4VMs is quite insensitive to the setting of the parameters. One possible reason is that, rather than simply picking one low-density separator, S4VMs optimize the assignment of labels in the worst cases. This property makes S4VMs even more attractive, especially when the number of labeled

TABLE 6

Comparison of Accuracy (Meanstd.) when the Balance

Constraint Is Violated

 Fig. 2. Parameter influence with 10 labeled examples.

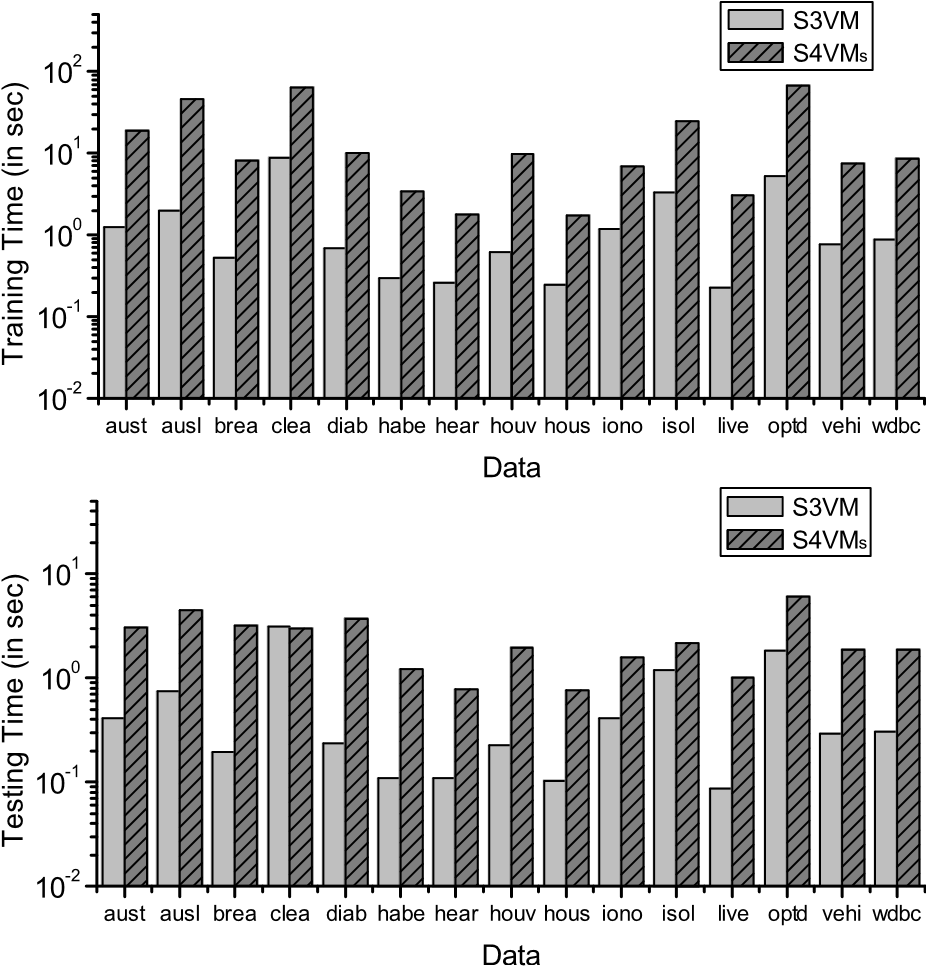


Fig. 3. Training and testing time (in seconds) of S3VM and S4VMs on UCI data sets with linear kernel.

10 in our experiments). It is notable that the implementation of S4VMs is inherently parallelizable, and thus S4VMs can be accelerated by parallel implementations or by using more efficient S3VM solvers.

### 5.8 Comparison with Other S3VMs

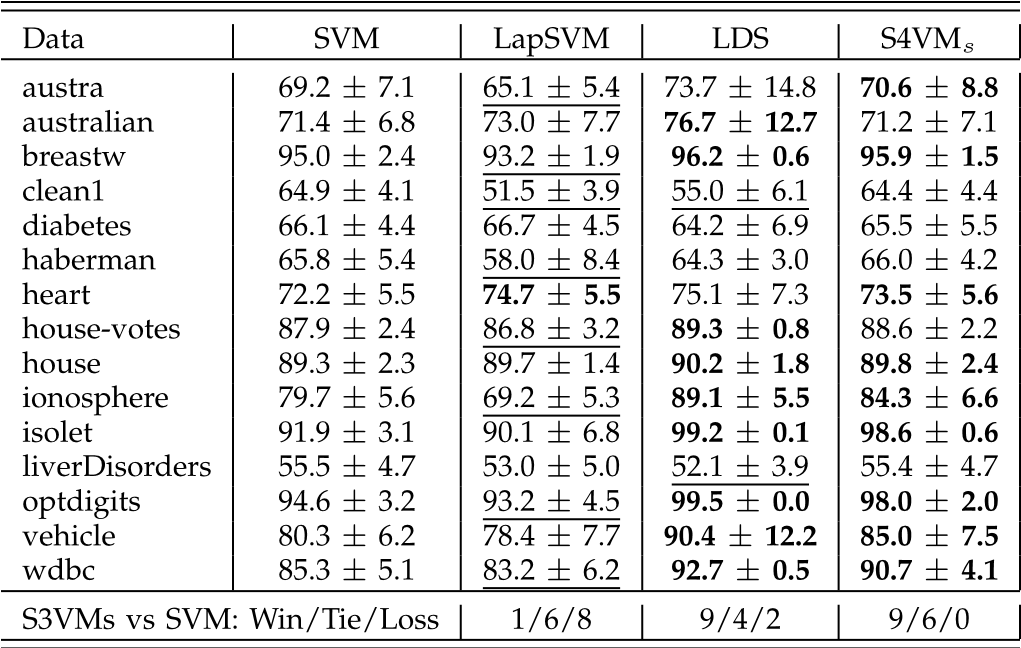
Table 7 shows the accuracy of other S3VM implementations. Specifically, Laplacian SVM (LapSVM) [2]7 which incorporates manifold assumption into S3VMs, and low density separation [12]8 which first introduces a graph-based distance for instances and then optimizes the objective of S3VM with the gradient descent method, are compared with inductive SVM. The parameters gA, gI of LapSVM are set to the same as the parameters C1 and C2 in S3VMs and S4VMs (i.e., 100 and 0:1). The r in LDS is set to 4 which achieves the best performance reported in the paper. Since LDS is based on RBF kernel, RBF kernel is used for inductive SVM and LapSVM. The other parameters are with the default settings recommended by the paper. As shown in Table 7, similar to TSVM [23], other S3VM implementations like LapSVM and LDS also decrease the performance significantly in some cases.

## 6 CONCLUSION

The purpose of this paper is to develop safe semi-supervised support vector machines (S3VMs) which never perform significantly inferior to inductive SVMs that only use labeled data. Based on our preliminary works in [31], [32], this paper first proposes the S3VM-us approach. This approach uses only the unlabeled instances that are very likely to be helpful, and thus avoids the use of highly risky unlabeled instances. Our empirical studies show

1. http://manifold.cs.uchicago.edu/manifold\_regularization/software.
2. http://olivier.chapelle.cc/lds/.

TABLE 7 Accuracy of Other S3VMs (Meanstd.)



that this approach improves safeness but only improves the performance slightly, usually much less than S3VMs. To develop a safe and well-performing approach, we reexamine the fundamental assumption of S3VMs, i.e., lowdensity separation. Based on the observation that multiple low-density separators can be identified from training data, S4VMs (Safe S3VMs) approach, the main contribution of this paper, is proposed. This approach attempts to avoid the risk of using a poor separator. Under the lowdensity assumption used by S3VMs, S4VMs are found to be provably safe and to achieve the maximum improvement in performance. An out-of-sample extension of S4VMs is also presented so that S4VMs can make predictions on unseen instances. Our empirical studies on a broad range of data sets show that the overall performance of S4VMs is highly competitive with S3VMs, but unlike S3VMs which show significant reduced performance in many cases, S4VMs are rarely inferior to inductive SVMs.

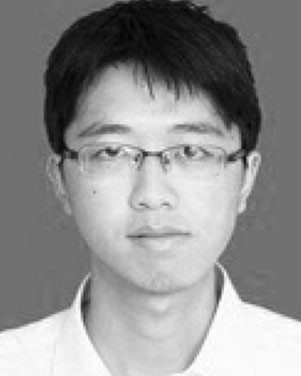
Our empirical studies in Table 2 reveal that even when low-density assumption does not hold, S4VMs still work well. We conjecture that this is because S4VMs exploit multiple separators rather than relying on a single separator. In this way, its robustness benefits from an inherent ensemble learning mechanism [49]. Further study on this issue is an interesting future work. It is also possible to combine the advantages of S3VM-us and S4VMs to develop approaches that are even stronger than the current S4VMs. Moreover, extending the spirit of S4VMs to graph-based semi-supervised methods [2], [33], [45], [53], as well as connecting the safeness to the generalization are worth studying in the future.

# ACKNOWLEDGMENTS

The authors want to thank the associate editor and reviewers for helpful comments and suggestions. They thank Teng Zhang for help in some experiments. They also thank Jianxin Wu and Cam-Tu Nguyen for proofreading the paper. This research was partially supported by the National Fundamental Research Program of China (2014CB340501), the National Science Foundation of China (61403186, 61333014, 61021062), Jiangsu Science Foundation (BK20140613) and Baidu fund.

# REFERENCES

1. M.F. Balcan and A. Blum, “A Discriminative Model for SemiSupervised Learning,” J. ACM, vol. 57, no. 3, article 19, 2010.
2. M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples,” J. Machine Learning Research, vol. 7, pp. 2399-2434, 2006.
3. S. Ben David, T. Lu, and D. Pal, “Does Unlabeled Data Provably Help? Worst-Case Analysis of the Sample Complexity of SemiSupervised Learning,” Proc. 21st Ann. Conf. Learning Theory, pp. 33-44, 2008.
4. K. Bennett and A. Demiriz, “Semi-Supervised Support Vector Machines,” Proc. Advances in Neural Information Processing Systems, vol. 11, pp. 368-374, 1999.
5. J.C. Bezdek and R.J. Hathaway, “Convergence of Alternating Optimization,” Neural, Parallel & Scientific Computations, vol. 11, no. 4, pp. 351-368, 2003.
6. T. De Bie and N. Cristianini, “Convex Methods for Transduction,” Proc. Advances in Neural Information Processing Systems, vol. 16, pp. 73-80, 2004.
7. A. Blum and S. Chawla, “Learning from Labeled and Unlabeled Data Using Graph Mincuts,” Proc. Eight Int’l Conf. Machine Learning, pp. 19-26, 2001.
8. A. Blum and T. Mitchell, “Combining Labeled and Unlabeled Data with Co-Training,” Proc. Seventh Ann. Conf. Computational Learning Theory, 1998.
9. O. Chapelle, M. Chi, and A. Zien, “A Continuation Method for Semi-Supervised SVMs,” Proc. 23rd Int’l Conf. Machine Learning, pp. 185-192, 2006.
10. Semi-Supervised Learning, O. Chapelle, B. Scholkopf, and A. Zien,€ eds., MIT Press, 2006.
11. O. Chapelle, V. Sindhwani, and S.S. Keerthi, “Optimization Techniques for Semi-Supervised Support Vector Machines,” J. Machine Learning Research, vol. 9, pp. 203-233, 2008.
12. O. Chapelle and A. Zien, “Semi-Supervised Classification by Low Density Separation,” Proc. 10th Int’l Workshop Artificial Intelligence and Statistics, pp. 57-64, 2005.
13. N. Chawla and G. Karakoulas, “Learning from Labeled and Unlabeled Data: An Empirical Study across Techniques and Domains,” J. Artificial Intelligence Research, vol. 23, pp. 331-366, 2005.
14. K. Chen and S. Wang, “Semi-Supervised Learning via Regularized Boosting Working on Multiple Semi-Supervised Assumptions,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 129-143, Jan. 2011.
15. R. Collobert, F. Sinz, J. Weston, and L. Bottou, “Large Scale Transductive SVMs,” J. Machine Learning Research, vol. 7, pp. 1687-1712, 2006.
16. F.G. Cozman, I. Cohen, and M.C. Cirelo, “Semi-Supervised Learning of Mixture Models,” Proc. 20th Int’l Conf. Machine Learning, pp. 99-106, 2003.
17. A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” J. Royal Statistical Soc. B, vol. 39, no. 1, pp. 1-38, 1977.
18. R.E. Fan, P.H. Chen, and C.J. Lin, “Working Set Selection Using Second Order Information for Training Support Vector Machines,” J. Machine Learning Research, vol. 6, pp. 1889-1918, 2005.
19. C. Goutte, H. Dejean, E. Gaussier, N. Cancedda, and J.M. Renders, “Combining Labelled and Unlabelled Data: A Case Study on Fisher Kernels and Transductive Inference for Biological Entity Recognition,” Proc. Sixth Conf. Natural Language Learning, pp. 1-7, 2002.
20. Y. Grandvalet and Y. Bengio, “Semi-Supervised Learning by Entropy Minimization,” Proc. Advances in Neural Information Processing Systems, vol. 17, 2004.
21. C.J. Hsieh, K.W. Chang, C.J. Lin, S.S. Keerthi, and S. Sundararajan, “A Dual Coordinate Descent Method for Large-Scale Linear SVM,” Proc. 25th Int’l Conf. Machine Learning, pp. 408-415, 2008.
22. A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1988.
23. T. Joachims, “Transductive Inference for Text Classification Using Support Vector Machines,” Proc. 16th Int’l Conf. Machine Learning, pp. 200-209, 1999.
24. N. Kasabov and S. Pang, “Transductive Support Vector Machines and Applications in Bioinformatics for Promoter Recognition,” Proc. Int’l Conf. Neural Networks and Signal Processing, pp. 1-6, 2003.
25. S. Kirkpatrick, “Optimization by Simulated Annealing: Quantitative Studies,” J. Statistical Physics, vol. 34, no. 5, pp. 975-986, 1984.
26. P.J.M. Laarhoven and E.H.L. Aarts, Simulated Annealing: Theory and Applications. Springer, 1987.
27. M. Li and Z.-H. Zhou, “SETRED: Self-Training with Editing,” Proc. Ninth Pacific-Asia Conf. Knowledge Discovery and Data Mining, pp. 611-621, 2005.
28. Y.-F. Li, J.T. Kwok, and Z.-H. Zhou, “Semi-Supervised Learning Using Label Mean,” Proc. 26th Int’l Conf. Machine learning, pp. 633640, 2009.
29. Y.-F. Li, I.W. Tsang, J.T. Kwok, and Z.-H. Zhou, “Tighter and Convex Maximum Margin Clustering,” Proc. 12th Int’l Conf. Artificial Intelligence and Statistics, pp. 344-351, 2009.
30. Y.-F. Li, I.W. Tsang, J.T. Kwok, and Z.-H. Zhou, “Convex and Scalable Weakly Labeled SVMS,” J. Machine Learning Research, vol. 14, pp. 2151-2188, 2013.
31. Y.-F. Li and Z.-H. Zhou, “Improving Semi-Supervised Support Vector Machines through Unlabeled Instances Selection,” Proc. 25th AAAI Conf. Artificial Intelligence, pp. 386-391, 2011.
32. Y.-F. Li and Z.-H. Zhou, “Towards Making Unlabeled Data Never Hurt,” Proc. 28th Int’l Conf. Machine learning, pp. 10811088, 2011.
33. W. Liu, J.-F. He, and S.-F. Chang, “Large Graph Construction for Scalable Semi-Supervised Learning,” Proc. 27th Int’l Conf. Machine Learning, pp. 679-686, 2010.
34. W. Liu, J. Wang, and S.-F. Chang, “Robust and Scalable GraphBased Semisupervised Learning,” Proc. IEEE, vol. 100, no. 9, pp. 2624-2638, Sep. 2012.
35. D.J. Miller and H.S. Uyar, “A Mixture of Experts Classifier with Learning Based on Both Labelled and Unlabelled Data,” Proc. Advances in Neural Information Processing Systems, vol. 9, pp. 571577, 1997.
36. K. Nigam, A.K. McCallum, S. Thrun, and T. Mitchell, “Text Classification from Labeled and Unlabeled Documents Using EM,” Machine Learning, vol. 39, no. 2, pp. 103-134, 2000.
37. V. Sindhwani, S.S. Keerthi, and O. Chapelle, “Deterministic Annealing for Semi-Supervised Kernel Machines,” Proc. 23rd Int’l Conf. Machine Learning, pp. 841-848, 2006.
38. A. Singh, R. Nowak, and X. Zhu, “Unlabeled Data: Now It Helps, Now It Doesn’t,” Proc. Advances in Neural Information Processing Systems, vol. 21, pp. 1513-1520, 2009.
39. E.K. Tang, P.N. Suganthan, and X. Yao, “An Analysis of Diversity Measures,” Machine Learning, vol. 65, no. 1, pp. 247-271, 2006.
40. V.N. Vapnik, Statistical Learning Theory. Wiley, 1998.
41. J. Wang, T. Jebara, and S.-F. Chang, “Graph Transduction via Alternating Minimization,” Proc. 25th Int’l Conf. Machine Learning, pp. 1144-1151, 2008.
42. L. Wang, K. Chan, and Z. Zhang, “Bootstrapping SVM Active Learning by Incorporating Unlabelled Images for Image Retrieval,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 629-634, 2003.
43. L. Xu and D. Schuurmans, “Unsupervised and Semi-Supervised Multi-Class Support Vector Machines,” Proc. 20th Nat’l Conf. Artificial Intelligence, pp. 904-910, 2005.
44. A.L. Yuille and A. Rangarajan, “The Concave-Convex Procedure,” Neural Computation, vol. 15, no. 4, pp. 915-936, 2003.
45. K. Zhang, J.T. Kwok, and B. Parvin, “Prototype Vector Machine for Large Scale Semi-Supervised Learning,” Proc. 26th Int’l Conf. Machine Learning, pp. 1233-1240, 2009.
46. K. Zhang, I.W. Tsang, and J.T. Kwok, “Maximum Margin Clustering Made Practical,” Proc. 24th Int’l Conf. Machine Learning, pp. 1119-1126, 2007.
47. T. Zhang and F. Oles, “The Value of Unlabeled Data for Classification Problems,” Proc. 17th Int’l Conf. Machine Learning, pp. 11911198, 2000.
48. D. Zhou, O. Bousquet, T.Navin Lal, J. Weston, and B. Scholkopf,€ “Learning with Local and Global Consistency,” Proc. Advances in Neural Information Processing Systems, vol. 16, pp. 595-602, 2004.
49. Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Chapman & Hall, 2012.
50. Z.-H. Zhou and M. Li, “Tri-Training: Exploiting Unlabeled Data Using Three Classifiers,” IEEE Trans. Knowledge and Data Eng., vol. 17, no. 11, pp. 1529-1541, Nov. 2005.
51. Z.-H. Zhou and M. Li, “Semi-Supervised Learning by Disagreement,” Knowledge and Information Systems, vol. 24, no. 3, pp. 415-439, 2010.
52. X. Zhu, “Semi-Supervised Learning Literature Survey,” technical report, Univ. of Wisconsin-Madison, 2007.
53. X. Zhu, Z. Ghahramani, and J. Lafferty, “Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions,” Proc. 20th Int’l Conf. Machine Learning, pp. 912-919, 2003.

Yu-Feng Li received the BSc and PhD degrees in computer science from Nanjing University, China, in 2006 and 2013, respectively. He joined the Department of Computer Science & Technology at Nanjing University as an assistant researcher in 2013. His main research interests include machine learning and data mining. He has won the Microsoft Fellowship Award (2009) and the Excellent Doctoral Dissertation Award of Chinese Computer Federation (2013). He has been a Program Committee member of several

conferences including IJCAI’11, ICDM’11, AAAI’12 (special track on AI and the web), IJCNN’13 and ICML’14.

Zhi-Hua Zhou (S’00-M’01-SM’06) received the BSc, MSc, and PhD degrees in computer science from Nanjing University, China, in 1996, 1998, and 2000, respectively, all with the highest honors. He joined the Department of Computer Science and Technology at Nanjing University as an assistant professor in 2001, and is currently a professor and director of the LAMDA group. His research interests are mainly in artificial intelligence, machine learning, data mining, pattern recognition and multimedia information retrieval. In these areas, he has published more than 100 papers in leading international journals or conference proceedings, and he holds 12 patents. He has received various awards/honors including the IEEE CIS Outstanding Early Career Award, the National Science & Technology Award for Young Scholars of China, the Fok Ying Tung Young Professorship Award, the Microsoft Young Professorship Award, and nine international journals/conferences paper or competition awards. He is an associate editor-in-chief of the Chinese Science Bulletin, associate editor of the ACM Transactions on Intelligent Systems and Technology and on the editorial boards of various other journals. He is the founder and Steering Committee Chair of ACML, and Steering Committee member of PAKDD and PRICAI. He serves/served as general chair/co-chair of ACML’12, ADMA’12 and PCM’13, program chair/co-chair for PAKDD’07, PRICAI’08, ACML’09, SDM’13, etc., workshop chair of KDD’12, program vice chair or area chair of various conferences, and chaired many domestic conferences in China. He is the chair of the Machine Learning Technical Committee of the Chinese Association of Artificial Intelligence, chair of the Artificial Intelligence & Pattern Recognition Technical Committee of the China Computer Federation, vice chair of the Data Mining Technical Committee of the IEEE Computational Intelligence Society, and the chair of the IEEE Computer Society Nanjing Chapter. He is a fellow of the IAPR, the IEEE, and the IET/IEE.

" For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.