**Improving Semi-Supervised Support Vector Machines**

**Through Unlabeled Instances Selection**

Yu-Feng Li, Zhi-Hua Zhou[[1]](#footnote-1)

*National Key Laboratory for Novel Software Technology*

*Nanjing University, Nanjing 210093, China*

改进半监督支持向量机

通过未标记的实例选择

李玉峰，周志华

新型软件技术国家重点实验室

南京大学，南京210093



# Abstract

抽象

Semi-supervised support vector machines (S3VMs) are a kind of popular approaches which try to improve learning performance by exploiting unlabeled data. Though S3VMs have been found helpful in many situations, they may degenerate performance and the resultant generalization ability may be even worse than using the labeled data only. In this paper, we try to reduce the chance of performance degeneration of S3VMs. Our basic idea is that, rather than exploiting all unlabeled data, the unlabeled instances should be selected such that only the ones which are very likely to be helpful are exploited, while some highly risky unlabeled instances are avoided. We propose the S3VM-*us* method by using hierarchical clustering to select the unlabeled instances. Experiments on a broad range of data sets over eighty-eight different settings show that the chance of performance degeneration of S3VM-*us* is much smaller than that of existing S3VMs.

[1](http://arxiv.org/abs/1005.1545v2)

[arXiv:1005.1545v2 [cs.LG] 9 May 201](http://arxiv.org/abs/1005.1545v2)

*Key words:* unlabeled data, performance degeneration, semi-supervised support vector machine

半监督支持向量机（S3VM）是一种流行的方法，试图通过利用未标记的数据来提高学习性能。尽管已经发现S3VM在许多情况下是有用的，但它们可能会降低性能，并且由此产生的泛化能力甚至可能比仅使用标记数据更糟糕。在本文中，我们尝试降低S3VM性能退化的可能性。我们的基本思想是，应该选择未标记的实例，而不是利用所有未标记的数据，以便仅利用非常可能有用的实例，而避免一些高风险的未标记实例。我们通过使用分层聚类来选择未标记的实例来提出S3VM-us方法。对超过八十八种不同设置的各种数据集进行的实验表明，S3VM-us的性能退化机会远小于现有S3VM。

关键词：未标记数据，性能退化，半监督支持向量机



# Introduction

介绍

In many real situations there are plentiful unlabeled training data while the acquisition of class labels is costly and difficult. Semi-supervised learning tries to exploit unlabeled data to help improve learning performance, particularly when there are limited labeled training examples. During the past decade, semi-supervised learning has received significant attention and many approaches have been developed [6, 29, 28].

在许多实际情况中，存在大量未标记的训练数据，而获得类别标签是昂贵且困难的。 半监督学习试图利用未标记的数据来帮助提高学习成绩，特别是在标签训练示例有限的情况下。 在过去的十年中，半监督学习受到了极大的关注，并且已经开发了许多方法[6,29,28]。

Among the many semi-supervised learning approaches, S3VMs (semi-supervised support vector machines) [3, 15] are popular and have solid theoretical foundation. However, though the performances of S3VMs are promising in many tasks, it has been found that there are cases where, by using unlabeled data, the performances of S3VMs are even worse than SVMs simply using the labeled data [25, 6, 7]. To enable S3VMs to be accepted by more users in more application areas, it is desirable to reduce the chances of performance degeneration by using unlabeled data.

在众多半监督学习方法中，S3VM（半监督支持向量机）[3,15]很受欢迎，具有坚实的理论基础。 然而，尽管S3VM的性能在许多任务中都很有前途，但已经发现，通过使用未标记的数据，有些情况下，S3VM的性能甚至比仅使用标记数据的SVM更差[25,6,7]。 为了使更多用户能够接受更多应用程序区域的S3VM，希望通过使用未标记的数据来降低性能退化的可能性。

In this paper, we focus on transductive learning and present the S3VM-*us* (S3VM with Unlabeled instances Selection) method. Our basic idea is that, given a set of unlabeled data, it may be not adequate to use all of them without any sanity check; instead, it may be better to use only the unlabeled instances which are very likely to be helpful while avoiding unlabeled instances which are with high risk. To exclude highly risky unlabeled instances, we first introduce two baselines, where the first baseline uses standard clustering technique motivated by the discernibility of density set [21] while the other one uses label propagation technique motivated by confidence estimation. Then, based on the analysis of the deficiencies of the two baseline approaches, we propose the S3VM-*us* method, which employs hierarchical clustering to help select unlabeled instances. Comprehensive experiments on a broad range of data sets over eightyeight different settings show that, the chance of performance degeneration of S3VM-*us* is much smaller than that of TSVM [15], while the overall performance of S3VM-*us* is competitive with TSVM.

在本文中，我们专注于转换学习并呈现S3VM-us（S3VM与未标记实例选择）方法。我们的基本想法是，给定一组未标记的数据，在没有任何健全性检查的情况下使用所有这些数据可能是不够的;相反，最好只使用很可能有用的未标记实例，同时避免高风险的未标记实例。为了排除高风险的未标记实例，我们首先引入两个基线，其中第一个基线使用由密度集的可辨别性驱动的标准聚类技术[21]，而另一个使用由置信度估计驱动的标签传播技术。然后，基于对两种基线方法的不足之处的分析，我们提出了S3VM-us方法，该方法采用分层聚类来帮助选择未标记的实例。对八种不同设置的各种数据集进行全面的实验表明，S3VM-us的性能退化机会远小于TSVM [15]，而S3VM-us的整体性能与TSVM相比具有竞争力。

The rest of this paper is organized as follows. Section 2 briefly reviews some related work. Section 3 introduces two baseline approaches. Section 4 presents our S3VM-*us* method. Experimental results are reported in Section 5. The last section concludes this paper.

本文的其余部分安排如下。 第2节简要回顾了一些相关的工作。 第3节介绍了两种基线方法。 第4节介绍了我们的S3VM-us方法。 实验结果在第5节中报告。最后一节总结了本文。

# Related Work

相关工作

Roughly speaking, existing semi-supervised learning approaches mainly fall into four categories. The first category is generative methods, e.g., [19, 20], which extend supervised generative models by exploiting unlabeled data in parameter estimation and label estimation using techniques such as the EM method. The second category is graph-based methods, e.g., [4, 30, 26], which encode both the labeled and unlabeled instances in a graph and then perform label propagation on the graph. The third category is disagreementbased methods, e.g., [5, 27], which employ multiple learners and improve the learners through labeling the unlabeled data based on the exploitation of disagreement among the learners. The fourth category is S3VMs, e.g., [3, 15], which use unlabeled data to regularize the decision boundary to go through low density regions [8].

粗略地说，现有的半监督学习方法主要分为四类。 第一类是生成方法，例如[19,20]，其通过利用诸如EM方法之类的技术在参数估计和标记估计中利用未标记数据来扩展监督生成模型。 第二类是基于图的方法，例如[4,30,26]，其在图中编码标记和未标记的实例，然后在图上执行标记传播。 第三类是基于分歧的方法，例如[5,27]，其采用多个学习者并通过基于对学习者之间的分歧的利用来标记未标记的数据来改进学习者。 第四类是S3VM，例如[3,15]，它们使用未标记的数据来规范决策边界以通过低密度区域[8]。

Though semi-supervised learning approaches have shown promising performances in many situations, it has been indicated by many authors that using unlabeled data may hurt the performance [20, 25, 11, 27, 9, 16, 2, 21]. In some application areas, especially the ones which require high reliability, users might be reluctant to use semi-supervised learning approaches due to the worry of obtaining a performance worse than simply neglecting unlabeled data. As typical semi-supervised learning approaches, S3VMs also suffer from this deficiency.

尽管半监督学习方法在许多情况下表现出了很好的表现，但许多作者已经指出，使用未标记的数据可能会损害性能[20,25,11,27,9,16,2,21]。 在某些应用领域，特别是那些需要高可靠性的应用领域，用户可能不愿意使用半监督学习方法，因为担心获得的性能比简单地忽略未标记的数据更糟糕。 作为典型的半监督学习方法，S3VM也存在这种缺陷。

The usefulness of unlabeled data has been discussed theoretically [16, 2, 21] and validated empirically

[9]. Many literatures indicated that unlabeled data should be used carefully. For generative methods, Cozman et al. [11] showed that unlabeled data can increase error even in situations where additional labeled data would decrease error. One main conjecture on the performance degeneration is attributed to the difficulties of making a right model assumption which prevents the performance from degenerated by fitting with unlabeled data. For graph-based methods, more and more researchers recognize that graph construction is more crucial than how the labels are propagated, and some attempts have been devoted to using domain knowledge or constructing robust graphs [1, 14]. As for disagreement-based method, the generalization ability has been studied with plentiful theoretical results based on different assumptions

[5, 12, 23, 24]. As for S3VMs, the correctness of the S3VM objective has been studied on small data sets

[7].

已经在理论上讨论了未标记数据的有用性[16,2,21]，并在经验上进行了验证

[9]。许多文献表明应该仔细使用未标记的数据。对于生成方法，Cozman等。 [11]表明，即使在额外的标记数据会减少错误的情况下，未标记的数据也会增加错误。关于性能退化的一个主要猜想归因于制定正确的模型假设的困难，该假设通过拟合未标记的数据来防止性能退化。对于基于图的方法，越来越多的研究人员认识到图形构造比标签的传播方式更为重要，并且一些尝试致力于使用领域知识或构建鲁棒图[1,14]。对于基于分歧的方法，基于不同的假设，对泛化能力进行了充分的理论研究

[5,12,23,24]。对于S3VM，已经在小数据集上研究了S3VM目标的正确性

[7]。

It is noteworthy that though there are many work devoted to cope with the high complexity of S3VMs [15, 10, 7, 18], there was no proposal on how to reduce the chance of performance degeneration by using unlabeled data. There was a relevant work which uses data editing techniques in semi-supervised learning [17]; however, it tries to remove or fix suspicious unlabeled data during training process, while our proposal tries to select unlabeled instances for S3VM and SVM predictions after the S3VM and SVM have already been trained.

值得注意的是，尽管有许多工作致力于应对S3VM的高复杂性[15,10,7,18]，但没有提出如何通过使用未标记数据来降低性能退化的可能性。 有一项相关工作在半监督学习中使用数据编辑技术[17]; 但是，它试图在训练过程中删除或修复可疑的未标记数据，而我们的提议尝试在S3VM和SVM已经过训练后为S3VM和SVM预测选择未标记的实例。

# Two Baseline Approaches

两种基线方法

As mentioned, our intuition is to use only the unlabeled data which are very likely to help improve the performance and keep the unlabeled data which are with high risk to be unexploited. In this way, the chance of performance degeneration may be significantly reduced. Current S3VMs can be regarded as an extreme case which believes that all unlabeled data are with low risk and therefore all of them should be used; while inductive SVMs which use labeled data only can be regarded as another extreme case which believes that all the unlabeled data are high risky and therefore only labeled data are used.

如上所述，我们的直觉是仅使用未标记的数据，这些数据很可能有助于提高性能并保持未标记的高风险数据未被开发。 以这种方式，可以显着降低性能退化的可能性。 当前的S3VM可以被视为一种极端情况，它认为所有未标记的数据都具有低风险，因此应该使用所有这些数据; 而仅使用标记数据的归纳SVM可以被认为是另一种极端情况，它认为所有未标记的数据都是高风险的，因此仅使用标记数据。

Specifically, we consider the following problem: Once we have obtained the predictions of inductive SVM and S3VM, how to remove risky predictions of S3VM such that the resultant performance could be often better and rarely worse than that of inductive SVM?

具体来说，我们考虑以下问题：一旦我们获得了归纳SVM和S3VM的预测，如何去除S3VM的风险预测，使得最终的性能往往更好，并且很少比归纳SVM更差？

There are two simple ideas that are easy to be worked out to address the above problem, leading to two baseline approaches, namely S3VM-*c* and S3VM-*p*.

有两个简单的想法很容易解决上述问题，导致两种基线方法，即S3VM-c和S3VM-p。

In the sequel, suppose we are given a training data set D = LSU where L = {(**x**1*,y*1)*,...,*(**x***l,yl*)} denotes the set of labeled data and U = {**x***l*+1*,...,***x***l*+*u*} denotes the set of unlabeled data. Here **x** ∈ X is an instance and *y* ∈ {+1*,*−1} is the label. We further let *ySV M*(**x**) and *yS*3*V M*(**x**) denote the predicted labels on **x** by inductive SVM and S3VM, respectively.

在下文中，假设我们被给予训练数据集D = LSU，其中L = {（x1，y1），...，（x1，y1）}表示标记数据的集合，并且U = {x1 + 1，。 ..，xl + u}表示未标记数据的集合。 这里x∈X是一个实例，y∈{+ 1，-1}是标签。 我们进一步让ySV M（x）和yS3V M（x）分别通过归纳SVM和S3VM表示x上的预测标签。

## S3VM-c

The first baseline approach is motivated by the analysis in [21] which suggests that unlabeled data help when the component density sets are discernable. Here, one can simulate the component density sets by clusters and discernibility by a condition of disagreements between S3VM and inductive SVM. We consider the disagreement using two factors, i.e., *bias* and *confidence*. When S3VM obtains the same bias as inductive SVM and enhances the confidence of inductive SVM, one should use the results of S3VM; otherwise it may be risky if we totally trust the prediction of S3VM.

第一种基线方法是由[21]中的分析推动的，这表明未标记的数据有助于何时可以辨别组分密度集。 在这里，可以通过S3VM和归纳SVM之间的不一致条件，通过聚类和可辨别性来模拟组件密度集。 我们考虑使用两个因素的分歧，即偏见和信心。 当S3VM获得与归纳SVM相同的偏差并增强归纳SVM的置信度时，应该使用S3VM的结果; 否则，如果我们完全信任S3VM的预测可能会有风险。

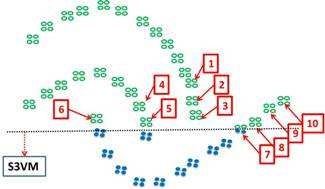
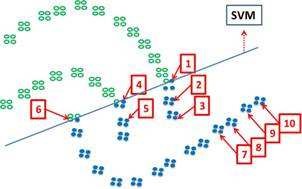
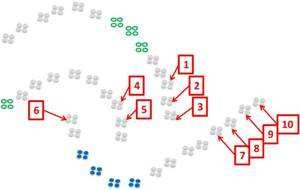
Algorithm 1 gives the S3VM-*c* method and Figure 1(d) illustrates the intuition of S3VM-*c*. As can be seen, S3VM-*c* inherits the correct predictions of S3VM on groups {1*,*4} while avoids the wrong predictions of S3VM on groups {7*,*8*,*9*,*10}.

算法1给出了S3VM-c方法，图1（d）给出了S3VM-c的直觉。 可以看出，S3VM-c在组{1,4}上继承了S3VM的正确预测，同时避免了对组{7,8,9,10}的S3VM的错误预测。

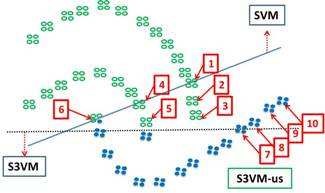
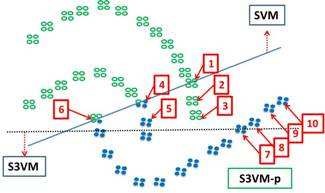
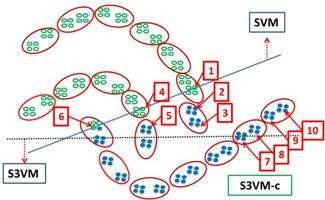
## S3VM-p

The second baseline approach is motivated by confidence estimation in graph-based methods, e.g., [30], where the confidence can be naturally regarded as a risk measurement of unlabeled data.

第二种基线方法的动机是基于图的方法中的置信度估计，例如[30]，其中置信度可以自然地被视为未标记数据的风险度量。



(a) (b) (c)



(d) (e) (f)

Figure 1: Illustration with artificial three-moon data. (a) Labeled data (empty and filled circles) and unlabeled data (gray points). The blocked numbers highlight groups of four unlabeled instances. Classification results of (b) Inductive SVM (using labeled data only); (c) S3VM; (d) S3VM-*c*, where each circle presents a cluster; (e) S3VM-*p*; (f) Our proposed S3VM-*us*.

图1：人工三月数据的插图。 （a）标记数据（空白和实心圆圈）和未标记数据（灰点）。 被阻止的数字突出显示了四个未标记实例的组。 （b）归纳SVM的分类结果（仅使用标记数据）; （c）S3VM; （d）S3VM-c，其中每个圆圈呈现一个簇; （e）S3VM-p; （f）我们建议的S3VM-us。

Formally, to estimate the confidence of unlabeled data, let **F***l* = [(**y***l* + 1)*/*2*,*(1 − **y***l*)*/*2] ∈ {0*,*1}*l*×2 be the label matrix for labeled data where **y***l* = [*y*1*,...,yl*]′ ∈ {±1}*l*×1 is the label vector. Let **W** = [*wij*] ∈ R(*l*+*u*)×(*l*+*u*) be the weight matrix of training data and **Λ** is the laplacian of **W**, i.e., **Λ** = **D** − **W** where **D** = *diag*(*di*) is a diagonal matrix with entries *di* = P*j wij*. Then, the predictions of unlabeled data can be obtained by [30]

**F***,*

形式上，为了估计未标记数据的置信度，令F1 = [（yl + 1）/ 2，（1 - y1）/ 2]∈{0,1} l×2为标记数据的标签矩阵，其中yl = [ y1，...，yl]'∈{±1} l×1是标签矢量。 令W = [wij]∈R（l + u）×（l + u）为训练数据的权重矩阵，Λ为W的拉普拉斯，即Λ= D-W其中D = diag（di）为a 条目di = Pj wij的对角矩阵。 然后，可以通过[30]获得未标记数据的预测

F ，

where **Λ***u,u* is the sub-matrix of **Λ** with respect to the block of unlabeled data, while **W***u,l* is the submatrix of **W** with respect to the block between labeled and unlabeled data. Then, assign each point **x***i*

with the label  and the confidence . After

confidence estimation, similar to S3VM-*c*, we consider the risk of unlabeled data by two factors, i.e., *bias* and *confidence*. If S3VM obtains the same bias of label propagation and the confidence is high enough, we use the S3VM prediction, and otherwise we take SVM prediction.

其中，u，u是关于未标记数据块的Λ的子矩阵，而Wu，l是关于标记和未标记数据之间的块的W的子矩阵。 然后，分配每个点xi

有了标签和信心。 后

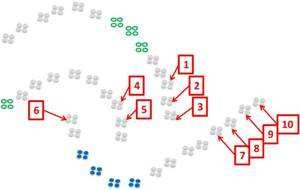
置信度估计，类似于S3VM-c，我们考虑两个因素的无标签数据的风险，即偏差和置信度。 如果S3VM获得相同的标签传播偏差并且置信度足够高，我们使用S3VM预测，否则我们采用SVM预测。

Algorithm 2 gives the S3VM-*p* method and Figure 1(e) illustrates the intuition of S3VM-*p*. As can be seen, the correct predictions of S3VM on groups {2*,*3} are inherited by S3VM-*p*, while the wrong predictions of S3VM on groups {7*,*8*,*9*,*10} are avoided.

算法2给出了S3VM-p方法，图1（e）说明了S3VM-p的直觉。 可以看出，S3VM-p继承了对组{2,3}的S3VM的正确预测，同时避免了对组{7,8,9,10}的S3VM的错误预测。

# Algorithm 1 S3VM-*c*

算法1 **S3VM-*c***



**Input**: *ySV M*, *yS*3*VM*, D and parameter *k*

输入：ySV M，yS3VM，D和参数k

1: Perform partitional clustering (e.g., *k*means) on D. Denote C1*,...,*C*k* as the data indices of each cluster re-

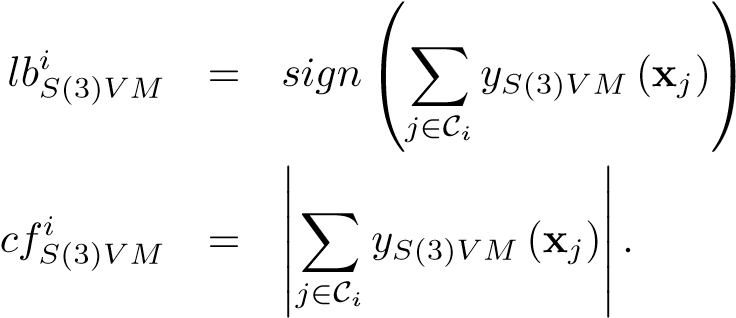
spectively.

2: For each cluster *i* = 1*,...,k*, calculate the label bias *lb* and confidence *cf* of SVM and S3VM according to:

1：在D上执行分区聚类（例如，kmeans）。将C1，...，Ck表示为每个聚类的数据索引。

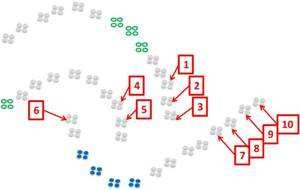
spectively。

2：对于每个簇i = 1，...，k，根据以下公式计算SVM和S3VM的标签偏差lb和置信度cf：



3: If *lbiSVM* = *lbiS*3*VM* & *cfSi*3*VM > cfSV Mi* , use the prediction of S3VM; otherwise use the prediction of SVM.

3：如果lbiSVM = lbiS3VM＆cfSi3VM> cfSV Mi，则使用S3VM的预测; 否则使用SVM的预测。



|  |  |
| --- | --- |
| **Algorithm 2** | S3VM-*p*  算法2 S3VM-*p* |

**Input**: *ySV M*, *yS*3*VM*, D, **W** and parameter *η*

输入：ySV M，yS3VM，D，W和参数η

1: Perform label propagation (e.g., [30]) with **W**, obtain the predicted label *ylp*(**x***i*) and confidence *hi* for each unlabeled instance **x***i*, *i* = *l* + 1*,...,l* + *u*.

2: Update **h** according to *hi* = *yS*3*VM*(**x***i*)*ylp*(**x***i*)*hi,i* = *l* + 1*,...,l* + *u.*

Let *c* denote the number of nonnegative entries in **h**.

3: Sort **h**, pick up the top-min{*ηu,c*} values and use the predictions of S3VM for the corresponding unlabeled instances, otherwise use the predictions of SVM.

1：用W执行标记传播（例如，[30]），获得每个未标记实例xi的预测标记ylp（xi）和置信度hi，i = 1 + 1，...，l + u。

2：根据hi = yS3VM（xi）ylp（xi）hi，i = 1 + 1，...，l + u更新h。

设c表示h中非负数的条目数。

3：对h进行排序，选取最高{ηu，c}值，并使用S3VM的预测作为相应的未标记实例，否则使用SVM的预测。



# Our Proposed Method

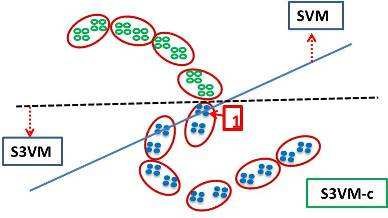
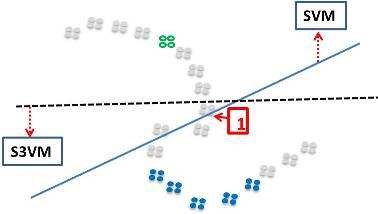
4.我们的建议方法

## Deficiencies of S3VM-c and S3VM-p

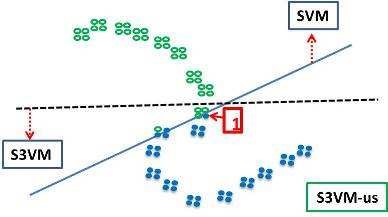
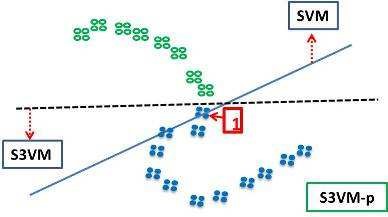
4.1。 S3VM-c和S3VM-p的缺陷

S3VM-*c* and S3VM-*p* are capable of reducing the chances of performance degeneration by using unlabeled data, however, they both suffer from some deficiencies. For S3VM-*c*, it works in a local manner and the relation between clusters are never considered, leading to the unexploitation of some helpful unlabeled instances, e.g., unlabeled instances in groups {2*,*3} in Figure 2(d). For S3VM-*p*, as stated in [22], the confidence estimated by label propagation approach might be incorrect if the label initialization is highly imbalanced, leading to the unexploitation of some useful unlabeled instances, e.g., groups {4*,*5} in Figure 2(e).

S3VM-c和S3VM-p能够通过使用未标记的数据来降低性能退化的可能性，但是，它们都存在一些缺陷。 对于S3VM-c，它以本地方式工作，并且从不考虑簇之间的关系，导致一些有用的未标记实例的未开发，例如图2（d）中组{2,3}中的未标记实例。 对于S3VM-p，如[22]中所述，如果标签初始化高度不平衡，标签传播方法估计的置信度可能不正确，导致一些有用的未标记实例的未开发，例如图中的组{4,5}图2（e）。



(a) (b)



(c) (d)

Figure 2: Illustration with artificial two-moon data when S3VM degenerates performance. (a) Labeled data (empty and filled circles) and unlabeled data (gray points). The blocked number highlight a group of four unlabeled instances. Classification results of (b) S3VM-*c*, where each circle presents a cluster; (c) S3VM-*p*; (d) Our proposed S3VM-*us*.

图2：S3VM退化性能时带有人工双月数据的插图。 （a）标记数据（空白和实心圆圈）和未标记数据（灰点）。 被阻止的数字突出显示一组四个未标记的实例。 （b）S3VM-c的分类结果，其中每个圆圈呈现一个簇; （c）S3VM-p; （d）我们建议的S3VM-us。

Moreover, both S3VM-*c* and S3VM-*p* heavily rely on the predictions of S3VM, which might become a serious issue especially when S3VM obtains degenerated performance. Figures 2(b) and 2(c) illustrate the behaviors of S3VM-*c* and S3VM-*p* when S3VM degenerates performance. Both S3VM-*c* and S3VM-*p* erroneously inherit the wrong predictions of S3VM of group 1.

此外，S3VM-c和S3VM-p都严重依赖于S3VM的预测，这可能成为一个严重的问题，特别是当S3VM获得退化的性能时。 图2（b）和2（c）说明了当S3VM退化性能时S3VM-c和S3VM-p的行为。 S3VM-c和S3VM-p都错误地继承了组1的S3VM的错误预测。

## S3VM-us

The deficiencies of S3VM-*c* and S3VM-*p* suggest to take into account of cluster relation and make the method insensitive to label initialization. This motivates us to use hierarchical clustering [13], leading to our proposed method S3VM-*us*.

S3VM-c和S3VM-p的缺陷建议考虑群集关系并使该方法对标签初始化不敏感。 这促使我们使用层次聚类[13]，导致我们提出的方法S3VM-us。

Hierarchical clustering works in a greedy and iterative manner. It first initials each singe instance as a cluster and then at each step, it merges two clusters with the shortest distance among all pairs of clusters. In this step, the cluster relation is considered and moreover, since hierarchical clustering works in an unsupervised setting, it does not suffer from the label initialization problem.

分层聚类以贪婪和迭代的方式工作。 它首先将每个单个实例初始化为一个簇，然后在每个步骤中，它合并两个簇，所有簇对之间的距离最短。 在该步骤中，考虑了聚类关系，而且，由于分层聚类在无监督设置中工作，因此不会遇到标签初始化问题。

Suppose *pi* and *ni* are the lengths of paths from the instance **x***i* to its nearest positive and negative labeled instances, respectively, in hierarchical clustering. We simply take the difference between *pi* and *ni* as an estimation of the confidence on the unlabeled instance **x***i*. Intuitively, the larger the difference between *pi* and *ni*, the higher the confidence on labeling **x***i*.

假设pi和ni是分层聚类中从实例xi到其最近的正和负标记实例的路径长度。 我们简单地将pi和ni之间的差异作为对未标记实例xi的置信度的估计。 直观地，pi和ni之间的差异越大，对标记xi的置信度越高。

# Algorithm 3 S3VM-*us*

**算法3 S3VM-*us***



**Input**: *ySV M*, *yS*3*VM*, D and parameter *ǫ*

输入：ySV M，yS3VM，D和参数ǫ

1: Let S be a set of the unlabeled data **x** such that *ySVM*(**x**) 6= *yS*3*V M*(**x**).

2: Perform hierarchical clustering, e.g., single linkage method [13].

3: For each unlabeled instance **x***i* ∈ S, calculate *pi* and *ni*, that is, the length of the paths from **x***i* to its nearest positive and negative labeled instances, respectively. Denote *ti* = (*ni* − *pi*).

4: Let B be the set of unlabeled instances **x***i* in S satisfying |*ti*| ≥ *ǫ*|*l* + *u*|.

5: If P**x***i*∈B *yS*3*VM*(**x***i*)*ti* ≥ P**x***i*∈B *ySVM*(**x***i*)*ti*, predict the unlabeled instances in B by S3VM and otherwise by SVM.

6: Predict the unlabeled data **x** 6∈ B by SVM.

1：令S是一组未标记数据x，使得ySVM（x）6 = yS3V M（x）。

2：执行分层聚类，例如，单链接方法[13]。

3：对于每个未标记的实例xi∈S，计算pi和ni，即分别从xi到其最接近的正和负标记实例的路径长度。 表示ti =（ni-pi）。

4：设B是满足| ti |的S中的未标记实例xi的集合 ≥ǫ| l + u |。

5：如果Pxi∈ByS3VM（xi）ti≥Pxi∈BySVM（xi）ti，则通过S3VM预测B中的未标记实例，否则通过SVM预测。

6：通过SVM预测未标记数据x6∈B。



Algorithm 3 gives the S3VM-*us* method and Figures 1(f) and 2 illustrate the intuition of S3VM-*us*. As can be seen, the wrong predictions of S3VM on groups {7*,*8*,*9*,*10} are avoided by S3VM-*us*, the correct predictions of S3VM on groups {2*,*3*,*4*,*5} are inherited, and S3VM-*us* does not erroneously inherit the wrong predictions of S3VM on group 1 in Figure 2.

算法3给出了S3VM-us方法，图1（f）和2说明了S3VM-us的直觉。 可以看出，S3VM-us避免了对组{7,8,9,10}的S3VM的错误预测，继承了组{2,3,4,5}上S3VM的正确预测，并且S3VM- 我们不会错误地继承图2中组1的S3VM的错误预测。

# Experiments

5.实验

## Settings

5.1。设置

We evaluate S3VM-*us* on a broad range of data sets including the semi-supervised learning benchmark data sets in [6] and sixteen UCI data sets[[2]](#footnote-2). The benchmark data sets are g241c, g241d, Digit1, USPS, TEXT and BCI. For each data, the archive2 provides two data sets with one using 10 labeled examples and the other using 100 labeled examples. As for UCI data sets, we randomly select 10 and 100 examples to be used as labeled examples, respectively, and use the remaining data as unlabeled data. The experiments are repeated for 30 times and the average accuracies and standard deviations are recorded. It is worth noting that in semi-supervised learning, labeled examples are often too few to afford a valid cross validation, and therefore hold-out tests are usually used for the evaluation.

我们在广泛的数据集上评估S3VM-us，包括[6]中的半监督学习基准数据集和16个UCI数据集。 基准数据集是g241c，g241d，Digit1，USPS，TEXT和BCI。 对于每个数据，archive2提供两个数据集，一个使用10个标记的示例，另一个使用100个标记的示例。 对于UCI数据集，我们分别随机选择10和100个示例作为标记示例，并将剩余数据用作未标记数据。 重复实验30次，记录平均精度和标准偏差。 值得注意的是，在半监督学习中，标记的示例通常太少而无法提供有效的交叉验证，因此通常使用保持测试进行评估。

In addition to S3VM-*c* and S3VM-*p*, we compare with inductive SVM and TSVM3 [15]. Both linear and

Gaussian kernels are used. For the benchmark data sets, we follow the setup in [6]. Specifically, for the case of 10 labeled examples, the parameter *C* for SVM is fixed to where *m* = *l* + *u* is

the size of data set and the Gaussian kernel width is set to *δ*, i.e., the average distance between instances. For the case of 100 labeled examples, *C* is fixed to 100 and the Gaussian kernel width is selected from {0*.*25*δ,*0*.*5*δ,δ,*2*δ,*4*δ*} by cross validation. On UCI data sets, the parameter *C* is fixed to 1 and the Gaussian kernel width is set to *δ* for 10 labeled examples. For 100 label examples, the parameter *C* is selected from {0*.*1*,*1*,*10*,*100} and the Gaussian kernel width is selected from {0*.*25*δ,*0*.*5*δ,δ,*2*δ,*4*δ*} by cross validation.

除了S3VM-c和S3VM-p之外，我们还与归纳SVM和TSVM3进行了比较[15]。 线性和

使用高斯核。 对于基准数据集，我们遵循[6]中的设置。 具体地，对于10个标记示例的情况，SVM的参数C固定为m = 1 + u的情况

数据集的大小和高斯核宽度设置为δ，即实例之间的平均距离。 对于100个标记示例的情况，C被固定为100并且通过交叉验证从{0.25δ，0.5δ，δ，2δ，4δ}中选择高斯核宽度。 在UCI数据集上，参数C固定为1，高斯核宽度设置为δ，用于10个标记示例。 对于100个标签示例，参数C选自{0.1,1,10,100}，并且通过交叉验证从{0.25δ，0.5δ，δ，2δ，4δ}中选择高斯核宽度。

For S3VM-*c*, the cluster number *k* is fixed to 50; for S3VM-*p*, the weighted matrix is constructed via Gaussian distance and the parameter *η* is fixed to 0.1; for S3VM-*us*, the parameter *ǫ* is fixed to 0.1.

对于S3VM-c，簇号k固定为50; 对于S3VM-p，加权矩阵通过高斯距离构造，参数η固定为0.1; 对于S3VM-us，参数ǫ固定为0.1。

## Results

5.2。结果

The results are shown in Tables 1 and 2. As can be seen, the performance of S3VM-*us* is competitive with TSVM. In terms of average accuracy, TSVM performs slightly better (worse) than S3VM-*us* on the case of 10 (100) labeled examples. In terms of pairwise comparison, S3VM-*us* performs better than TSVM on 13/12 and 14/16 cases with linear/Gaussian kernel for 10 and 100 labeled examples, respectively. Note that in a number of cases, TSVM has large performance improvement against inductive SVM, while the improvement of S3VM-*us* is smaller. This is not a surprise since S3VM-*us* tries to improve performance with the caution of avoiding performance degeneration.

结果显示在表1和表2中。可以看出，S3VM-us的性能与TSVM竞争。 就平均准确性而言，在10（100）个标记示例的情况下，TSVM比S3VM-us表现稍好（差）。 在成对比较方面，S3VM-us在13/12和14/16情况下的性能优于TSVM，线性/高斯内核分别用于10和100个标记示例。 请注意，在许多情况下，TSVM对归纳SVM的性能提升很大，而S3VM-us的改进较小。 这并不奇怪，因为S3VM-us试图通过避免性能退化来提高性能。

Though TSVM has large improvement in a number of cases, it also has large performance degeneration in cases. Indeed, as can be seen from Tables 1 and 2, TSVM is significantly inferior to inductive SVM on 8/44, 19/44 cases for 10 and 100 labeled examples, respectively. Both S3VM-*c* and S3VM-*p* are capable to reduce the times of significant performance degeneration, while S3VM-*us* does not significantly degenerate performance in the experiments.

虽然TSVM在许多情况下有很大的改进，但在某些情况下它也会有很大的性能退化。 实际上，从表1和表2中可以看出，TSVM在8 / 44,19 / 44个案例中分别对于10和100个标记的例子明显不如诱导SVM。 S3VM-c和S3VM-p都能够减少显着性能退化的时间，而S3VM-us不会显着降低实验中的性能。

## Parameter Influence

5.3。 参数影响

S3VM-*us* has a parameter *ǫ*. To study the influence of *ǫ*, we run experiments by setting *ǫ* to different values (0.1, 0.2 and 0.3) with 10 labeled examples. The results are plotted in Figure 3. It can be seen that the setting of *ǫ* has influence on the improvement of S3VM-*us* against inductive SVM. Whatever linear kernel or gaussian kernel is used, the larger the value of *ǫ*, the closer the performance of S3VM-*us* to

Table 1: Accuracy (mean ± std.) on 10 labeled examples. ‘SVM’ denotes inductive SVM which uses labeled data only. For the semi-supervised methods (TSVM, S3VM-*c*, S3VM-*p* and S3VM-*us*), if the performance is significantly better/worse than SVM, the corresponding entries are bolded/underlined (paired *t*-tests at 95% significance level). The win/tie/loss counts with the fewest losses are bolded.

S3VM-us有一个参数ǫ。 为了研究influence的影响，我们通过将ǫ设置为具有10个标记实例的不同值（0.1,0.2和0.3）来进行实验。 结果如图3所示。可以看出，setting的设置对S3VM-us对归纳SVM的改进有影响。 无论使用线性内核还是高斯内核，ǫ的值越大，S3VM-us的性能越接近

表1：10个标记实施例的准确度（平均值±标准）。 'SVM'表示仅使用标记数据的归纳SVM。 对于半监督方法（TSVM，S3VM-c，S3VM-p和S3VM-us），如果性能明显优于/差于SVM，则相应的条目用粗体/下划线表示（配对t检验，显着性水平为95％））。 赢得/平局/损失计数最少的损失是粗体。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data | SVM | TSVM | S3VM-*c* | S3VM-*p* | S3VM-*us* |
|  | ( linear / gaussian ) | ( linear / gaussian ) | ( linear / gaussian ) | ( linear / gaussian ) | ( linear / gaussian ) |
| BCI | 50.7±1.5 / 52.7±2.7 | 49.3±2.8 / 51.4±2.7 | 50.2±2.0 / 52.2±2.6 | 50.6±1.6 / 52.6±2.7 | 50.9±1.6 / 52.6±2.7 |
| g241c | 53.2±4.8 / 53.0±4.5 | **78.9**±**4.7** / **78.5**±**5.0** | 55.2±8.3 / 55.3±8.8 | **53.9**±**5.8** / **53.6**±**5.3** | 53.5±4.8 / 53.2±4.5 |
| g241d | 54.4±5.4 / 54.5±5.2 | 53.6±7.8 / 53.2±6.5 | 53.8±5.4 / 53.6±5.0 | 54.1±5.3 / 54.0±5.2 | 54.4±5.3 / 54.4±5.2 |
| digit1 | 55.4±10.9 / 75.0±7.9 | **79.4**±**1.1** / **81.5**±**3.1** | 56.1±12.2 / **77.3**±**8.2** | 56.2±12.2 / 75.0±8.1 | **58.1**±**9.6** / 75.1±7.8 |
| USPS | 80.0±0.1 / 80.7±1.8 | 69.4±1.2 / 73.0±2.6 | 80.0±0.1 / 80.4±2.5 | 80.0±0.1 / 80.5±2.1 | 80.0±0.1 / 80.7±1.8 |
| Text | 54.7±6.3 / 54.6±6.3 | **71.4**±**11.7** / **71.2**±**11.4** | **56.8**±**8.8** / **56.5**±**8.7** | **55.3**±**6.6** / **55.2**±**6.8** | 58.0±9.0 / 57.8±8.9 |
| house | 90.0±6.0 / 84.8±11.8 | 6.9 | 89.8±6.2 / 84.8±11.9 | 89.5 | 90.1±6.1 / 85.4±11.4 |
| heart | 58.8±10.5 / 63.9±11.6 | **72.4**±**12.6** / **72.6**±**10.4** | **59.0**±**10.8** / **64.4**±**11.6** | 58.6±10.6 / 63.8±11.7 | **61.9**±**9.7** / 65.1±11.0 |
| heart-statlo | g 74.6±4.8 / 69.9±10.1 | 74.9±6.6 / **73.9**±**5.9** | 74.5±5.2 / 70.1±10.2 | 74.5±4.9 / 70.0±10.2 | 74.2±5.4 / 71.7±6.9 |
| ionosphere | 70.4±8.7 / 65.8±9.8 | 72.0±10.5 / **76.1**±**8.2** | **70.9**±**9.0** / **66.1**±**9.9** | 70.4±8.7 / 66.0±9.7 | 70.7±8.3 / 67.4±6.7 |
| vehicle | 73.2±8.9 / 58.3±9.5 | 72.1±9.4 / **63.2**±**7.8** | 73.5±9.4 / 58.4±9.6 | 72.6±9.1 / 58.0±9.5 | **74.5**±**9.3** / **64.2**±**9.1** |
| house-votes | 85.5±7.0 / 79.7±10.7 | 83.8±6.1 / **84.0**±**5.3** | 85.7±7.0 / 80.1±10.6 | 85.3±6.9 / 79.7±10.7 | 86.0±5.7 / **84.3**±**6.1** |
| wdbc | 65.6±7.5 / 73.8±10.3 | **90.0**±**6.1** / **88.9**±**3.7** | 65.7±7.8 / **74.9**±**10.9** | **66.1**±**8.0** / 73.9±10.5 | **65.8**±**7.5** / **73.9**±**10.3** |
| clean1 | 58.2±4.2 / 53.5±6.2 | 57.0±5.1 / 53.3±4.8 | 57.8±4.4 / 53.3±6.2 | **58.5**±**4.2** / 53.3±6.3 | 58.2±4.2 / **55.0**±**8.1** |
| isolet | 93.8±4.3 / 82.0±15.7 | 84.2±10.9 / **86.7**±**9.5** | **94.5**±**5.1** / **83.2**±**16.0** | 93.0±4.7 / 81.7±15.7 | 93.7±4.3 / **84.1**±**12.6** |
| breastw | 93.9±4.8 / 92.3±10.1 | 89.2±8.6 / 88.9±8.8 | 94.2±4.9 / 92.4±10.0 | 93.9±4.9 / 92.2±10.0 | 93.6±5.4 / 92.4±9.9 |
| australian | 70.4±9.2 / 60.3±8.4 | 69.6±11.9 / **68.6**±**11.4** | 70.1±9.8 / 60.4±8.3 | 70.5±9.4 / 60.5±8.8 | 70.3±9.2 / 60.8±7.9 |
| diabetes | 63.3±6.9 / 66.3±3.5 | 63.4±7.6 / 65.8±4.6 | 63.2±6.8 / 65.9±3.0 | 63.4±6.6 / 66.2±3.4 | 63.3±6.9 / 66.3±3.5 |
| german | 65.2±4.9 / 65.1±12.0 | 63.7±5.6 / 63.5±5.1 | 65.6±4.7 / 65.1±11.8 | **65.6**±**4.8** / 65.1±11.9 | 65.2±5.0 / 65.3±11.6 |
| optdigits | 96.1±3.2 / 92.8±9.6 |  | **96.6**±**3.1** / **93.6**±**9.9** |  | 96.9±2.5 / 94.9±5.8 |
| ethn | 56.5±8.8 / 58.5±10.2 | **64.2**±**13.5** / **68.1**±**14.5** | 56.5±8.6 / **59.4**±**11.6** | 56.8±9.1 / 58.6±10.7 | **59.8**±**10.7** / **61.8**±**11.3** |
| sat | 95.8±4.1 / 87.5±10.9 |  | **96.3**±**4.1** / 87.7±11.2 | 94.8 |  |
| Aver. Acc. | 70.9 / 69.3 | 73.5 / 73.8 | 71.2 / 69.8 | 70.9 / 69.3 | 71.6 / 70.8 |
| SVM vs. Semi-Supervised: W/T/L | | 18/18/8 | 14/29/1 | 7/25/12 | **12/32/0** |



Table 2: Accuracy (mean ± std.) on 100 labeled examples. ‘SVM’ denotes inductive SVM which uses labeled data only. For the semi-supervised methods (TSVM, S3VM-*c*, S3VM-*p* and S3VM-*us*), if the performance is significantly better/worse than SVM, the corresponding entries are bolded/underlined (paired *t*-tests at 95% significance level). The win/tie/loss counts with the fewest losses are bolded.

表2：100个标记实施例的准确度（平均值±标准）。 'SVM'表示仅使用标记数据的归纳SVM。 对于半监督方法（TSVM，S3VM-c，S3VM-p和S3VM-us），如果性能明显优于/差于SVM，则相应的条目用粗体/下划线表示（配对t检验，显着性水平为95％））。 赢得/平局/损失计数最少的损失是粗体。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data | SVM | TSVM | S3VM-*c* | S3VM-*p* | S3VM-*us* |
|  | ( linear / gaussian ) | ( linear / gaussian ) | ( linear / gaussian ) | ( linear / gaussian ) | ( linear / gaussian ) |
| BCI | 61.1±2.6 / 65.9±3.1 | 56.4±2.8 / 65.6±2.5 | 58.3±2.6 / 65.6±3.0 | 60.3±2.5 / 65.8±3.0 | 61.0±2.7 / 65.8±3.1 |
| g241c | 76.3±2.0 / 76.6±2.1 | **81.7**±**1.6** / **82.1**±**1.2** | **79.3**±**1.7** / **79.6**±**1.8** | **77.2**±**2.1** / **77.1**±**2.0** | 76.3±2.0 / 76.6±2.1 |
| g241d | 74.2±1.9 / 75.4±1.8 | 76.1±8.5 / 77.9±7.4 | **77.4**±**3.5** / **78.5**±**3.3** | **74.8**±**2.3** / 75.7±2.2 | 74.2±1.9 / 75.4±1.8 |
| digit1 | 50.3±1.2 / 94.0±1.4 | **81.9**±**3.0** / 94.0±2.0 | 50.3±1.2 / **95.0**±**1.5** | 50.3±1.2 / 94.1±1.4 | **67.9**±**1.3** / 94.1±1.4 |
| USPS | 80.0±0.2 / 91.7±1.1 | 78.8±2.0 / 90.9±1.4 | 80.0±0.2 / **92.5**±**1.0** | 80.0±0.2 / 91.6±1.2 | 80.1±0.4 / 91.8±1.1 |
| Text | 73.8±3.3 / 73.7±3.6 | **77.7**±**1.6** / **77.7**±**1.7** | **75.3**±**3.4** / **75.2**±**3.6** | **73.9**±**3.4** / **73.8**±**3.7** | 74.1±3.1 / **74.2**±**3.3** |
| house | 95.7±2.0 / 95.6±1.6 | 94.4±2.5 / 94.8±2.6 | 95.5±1.8 / 95.4±1.8 | 95.6±2.0 / 95.5±1.7 | 95.6±2.0 / 95.6±1.6 |
| heart | 81.5±2.5 / 80.1±2.4 | 80.7±3.1 / 79.5±2.9 | 81.1±3.0 / 79.8±2.5 | 81.5±2.5 / 80.2±2.5 | 81.5±2.6 / 80.1±2.4 |
| heart-statlo | g 81.5±2.4 / 81.4±2.7 | 81.6±2.7 / 79.0±4.5 | 81.2±2.2 / 80.7±3.0 | 81.5±2.4 / 81.2±2.7 | 81.5±2.4 / 81.3±2.7 |
| ionosphere | 87.1±1.5 / 93.2±1.6 | 85.6±2.1 / 92.1±2.3 | **88.7**±**1.3** / 93.4±1.5 | 87.1±1.5 / 93.2±1.6 | 87.1±1.5 / 93.2±1.6 |
| vehicle | 92.9±1.7 / 95.4±1.4 | 91.6±2.5 / 95.4±2.3 | **93.3**±**1.6** / **95.9**±**1.3** | 92.8±1.7 / 95.2±1.5 | **93.0**±**1.7** / **95.5**±**1.4** |
| house-votes | 92.3±1.3 / 92.8±1.2 | 92.0±1.8 / 93.0±1.4 | 92.6±1.2 / 92.9±1.2 | 92.3±1.3 / 92.8±1.2 | 92.3±1.3 / 92.8±1.2 |
| clean1 | 73.0±2.7 / 80.6±3.0 | 73.2±3.1 / 79.1±3.4 | **73.7**±**2.9** / 79.9±2.9 | **73.2**±**2.6** / 80.4±3.2 | 73.1±2.7 / 80.7±3.0 |
| wdbc | 95.6±0.8 / 94.7±0.9 | 94.3±2.3 / 94.1±2.4 | **95.8**±**0.7** / 94.9±0.9 | 95.6±0.8 / 94.7±0.9 | **95.6**±**0.8** / **94.8**±**0.9** |
| isolet | 99.2±0.4 / 99.0±0.6 | 95.9±3.1 / 98.2±2.3 | 99.2±0.4 / **99.2**±**0.5** | 99.0±0.4 / 98.9±0.6 | 99.2±0.4 / **99.1**±**0.5** |
| breastw | 96.4±0.4 / 96.7±0.4 | 96.9±1.9 / **97.1**±**0.5** | **96.6**±**0.4** / **96.9**±**0.4** | 96.3±0.4 / 96.7±0.4 | 96.4±0.4 / 96.7±0.4 |
| australian | 83.8±1.6 / 84.9±1.7 | 82.5±2.6 / 84.6±2.7 | 83.8±1.7 / 85.0±1.6 | 83.9±1.7 / 85.0±1.8 | 83.8±1.7 / 85.0±1.7 |
| diabetes | 75.2±1.7 / 74.7±1.9 | 72.3±2.3 / 71.8±1.8 | 74.9±1.7 / 74.2±2.2 | 75.3±1.6 / 74.7±1.9 | 75.2±1.8 / 74.7±1.9 |
| german | 67.1±2.4 / 72.0±1.5 | 66.1±2.1 / 65.9±3.4 | 67.1±2.2 / 71.6±1.5 | **67.6**±**2.3** / 72.1±1.4 | 67.1±2.4 / 72.1±1.5 |
| optdigits | 99.4±0.3 / 99.4±0.3 | 95.9±3.7 / 97.4±3.1 | 99.5±0.4 / **99.5**±**0.3** | 99.2±0.4 / 99.2±0.4 | 99.5±0.3 / 99.4±0.3 |
| ethn | 91.6±1.6 / 93.4±1.2 | **92.6**±**2.3** / 93.4±3.0 | **93.9**±**1.6** / **95.0**±**1.2** | **91.9**±**1.5** / 93.3±1.2 | **91.7**±**1.5** / 93.4±1.2 |
| sat | 99.7±0.2 / 99.7±0.1 | 96.4±2.8 / 97.6±2.7 | **99.7**±**0.2** / **99.8**±**0.1** | 99.5±0.3 / 99.5±0.3 | 99.7±0.2 / 99.7±0.1 |
| Aver. Acc. | 83.0 / 86.8 | 83.9 / 86.4 | 83.5 / 87.3 | 83.1 / 86.8 | 83.9 / 86.9 |
| SVM vs. Semi-Supervised: W/T/L | | 7/18/19 | 21/16/7 | 8/25/11 | **8/36/0** |



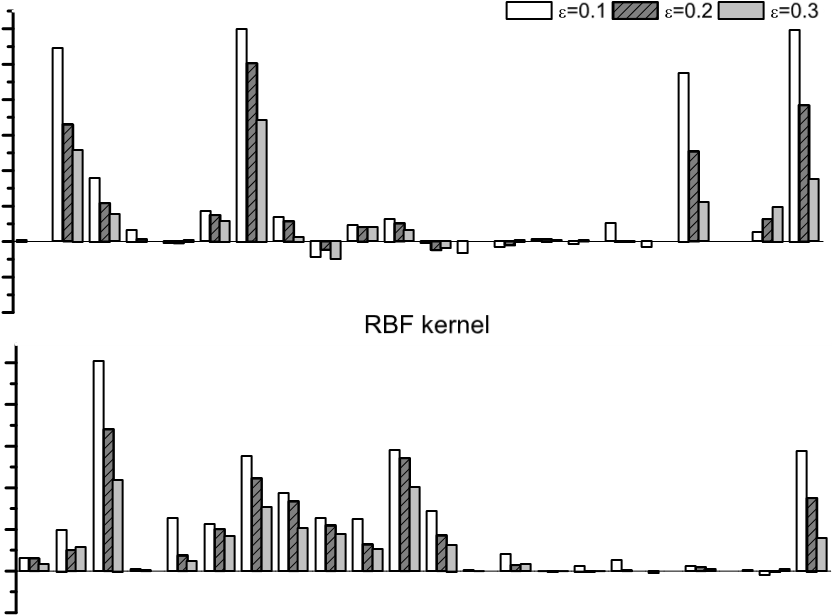


Figure 3: Influence of the parameter *ǫ* on the improvement of S3VM-*us* against inductive SVM.

图3：参数Influence对S3VM-us改进归纳SVM的影响。

SVM. It may be possible to increase the performance improvement by setting a smaller *ǫ*, however, this may increase the risk of performance degeneration.

SVM。 通过设置较小的ǫ可以提高性能，但是，这可能会增加性能退化的风险。

# Conclusion

6，结论

In this paper we propose the S3VM-*us* method. Rather than simply predicting all unlabeled instances by semi-supervised learner, S3VM-*us* uses hierarchical clustering to help select unlabeled instances to be predicted by semi-supervised learner and predict the remaining unlabeled instances by inductive learner.

In this way, the risk of performance degeneration by using unlabeled data is reduced. The effectiveness of S3VM-*us* is validated by empirical study.

在本文中，我们提出了S3VM-us方法。 S3VM-us不是简单地通过半监督学习者预测所有未标记的实例，而是使用层次聚类来帮助选择由半监督学习者预测的未标记实例，并通过归纳学习者预测剩余的未标记实例。

以这种方式，通过使用未标记的数据降低了性能退化的风险。 S3VM-us的有效性通过实证研究得到验证。

The proposal in this paper is based on heuristics and theoretical analysis is future work. It is worth noting that, along with reducing the chance of performance degeneration, S3VM-*us* also reduces the possible performance gains from unlabeled data. In the future it is desirable to develop really *safe* semisupervised learning approaches which are able to improve performance significantly but never degenerate performance by using unlabeled data.

本文的提议是基于启发式和理论分析是未来的工作。 值得注意的是，除了降低性能退化的可能性之外，S3VM-us还降低了未标记数据可能带来的性能提升。 在未来，需要开发真正安全的半监督学习方法，这些方法能够显着提高性能，但绝不会通过使用未标记的数据来降低性能。

# References

1. M. F. Balcan, A. Blum, P. P. Choi, J. Lafferty, B. Pantano, M. R. Rwebangira, and X. Zhu. Person identification in webcam images: An application of semi-supervised learning. In *ICML Workshop on Learning with Partially*

*Classified Training Data*, 2005.

1. S. Ben-David, T. Lu, and D. Pa´l. Does unlabeled data provably help? Worst-case analysis of the sample

complexity of semi-supervised learning. In *COLT*, pages 33–44, 2008.

1. K. Bennett and A. Demiriz. Semi-supervised support vector machines. In *NIPS 11*, pages 368–374. 1999.
2. A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *ICML*, pages

19–26, 2001.

1. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100,

1998.

1. O. Chapelle, B. Scho¨lkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA,

2006.

1. O. Chapelle, V. Sindhwani, and S. S. Keerthi. Optimization techniques for semi-supervised support vector machines. *J. Mach. Learn. Res.*, 9:203–233, 2008.
2. O. Chapelle and A. Zien. Semi-supervised learning by low density separation. In *AISTATS*, pages 57–64,

2005.

1. N. V. Chawla and G. Karakoulas. Learning from labeled and unlabeled data: An empirical study across techniques and domains. *J. Artif. Intell. Res.*, 23:331–366, 2005.
2. R. Collobert, F. Sinz, J. Weston, and L. Bottou. Large scale transductive SVMs. *J. Mach. Learn. Res.*, 7:1687–

1712, 2006.

1. F. G. Cozman, I. Cohen, and M. C. Cirelo. Semi-supervised learning of mixture models. In *ICML*, pages

99–106, 2003.

1. S. Dasgupta, M. L. Littman, and D. McAllester. PAC generalization bounds for co-training. In *NIPS 14*, pages

375–382. 2002.

1. A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ., 1988.
2. T. Jebara, J. Wang, and S. F. Chang. Graph construction and b-matching for semi-supervised learning. In

*ICML*, pages 441–448, 2009.

1. T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages

200–209, 1999.

1. J. Lafferty and L. Wasserman. Statistical analysis of semi-supervised regression. In *NIPS 20*, pages 801–808.

2008.

1. M. Li and Z. H. Zhou. SETRED: Self-training with editing. In *PAKDD*, pages 611–621, 2005.
2. Y.-F. Li, J. T. Kwok, and Z.-H. Zhou. Semi-supervised learning using label mean. In *ICML*, pages 633–640,

2009.

1. D. J. Miller and H. S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *NIPS 9*, pages 571–577. 1997.
2. K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled docu-

ments using EM. *Mach. Learn.*, 39(2):103–134, 2000.

1. A. Singh, R. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn’t. In *NIPS 21*, pages 1513–1520. 2009.
2. J. Wang, T. Jebara, and S. F. Chang. Graph transduction via alternating minimization. In *ICML*, pages 1144–

1151, 2008.

1. W. Wang and Z.-H. Zhou. Analyzing co-training style algorithms. In *ECML*, pages 454–465, 2007.
2. W. Wang and Z.-H. Zhou. A new analysis of co-training. In *ICML*, pages 1135–1142, 2010.
3. T. Zhang and F. Oles. The value of unlabeled data for classification problems. In *ICML*, pages 1191–1198,

2000.

1. D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In

*NIPS 16*, pages 595–602. 2004.

1. Z.-H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl.*

*Data Eng.*, 17(11):1529–1541,2005.

1. Z.-H. Zhou and M. Li. Semi-supervised learning by disagreement. *Knowl. Inf. Syst.*, 24(3):415–439, 2010.
2. X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Dept. Comp. Sci., Univ. Wisconsin-Madison, 2006.
3. X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.

1. Corresponding author. Email: zhouzh@nju.edu.cn

   *Preprint submitted for review October 30, 2018* [↑](#footnote-ref-1)
2. http://archive.ics.uci.edu/ml/ 2

   http://www.kyb.tuebingen.mpg.de/ssl-book/ 3 http://svmlight. joachims.org/ [↑](#footnote-ref-2)