

Serendipitous Learning: Learning Beyond the Predefined Label Space

Dan Zhang
Computer Science
Department
Purdue University
West Lafayette, IN
zhang168@cs.purdue.edu

Yan Liu
Computer Science
Department
University of Southern
California
Los Angeles, CA
yanliu.cs@usc.edu

Luo Si
Computer Science
Department
Purdue University
West Lafayette, IN
lsi@cs.purdue.edu

ABSTRACT

In machine learning, most supervised learning methods are developed for learning from training examples within a predefined label space and make predictions on which classes (among those predefined labels) each test example belongs to. However, in many real world applications, such as text and image categorization, we are often confronted with the learning scenarios in which the label space needs to be enlarged during testing phase, that is, the test examples may belong to some new classes which have not been defined during the training phase. How to make accurate predictions for examples in existing classes and at the same time identify novel examples from new classes is an extremely challenging problem, which has not been addressed before. This paper explores this novel and practical learning scenario, which is named as Serendipitous Learning (SL). The basic idea is to leverage the knowledge in the labeled examples to help identify the unknown classes. In particular, a maximum margin formulation is proposed to model both the classification loss on the known classes and the clustering performance on the unknown ones. An efficient optimization algorithm is designed based on Constrained Concave-Convex Procedure (CCCP) and the bundle method to solve the corresponding optimization problem. Furthermore, an efficient online learning algorithm is proposed for large-scale applications of the proposed problem with guaranteed bounds on regret. The experimental results on two synthetic datasets and two real world datasets demonstrate the advantages of the proposed method over other baseline algorithms.

Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning—*Knowledge acquisition*

General Terms

Algorithms, Performance, Experimentation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD'11, August 21–24, 2011, San Diego, California, USA.
Copyright 2011 ACM 978-1-4503-0813-7/11/08 ...\$10.00.

Keywords

Serendipitous Learning, Maximum Margin Classification, Maximum Margin Clustering, Label Space

1. INTRODUCTION

One of the basic assumptions in most supervised machine learning algorithms is that the label space is predefined and shared by the training and testing examples [5, 12]. However, this assumption is often violated in many open-domain applications of classification. For example, in online webpage classification, we might be able to provide a list of common classes, such as *politics*, *entertainment* and *sports*. But it is extremely difficult to gather a complete list of all classes as well as their corresponding training examples beforehand since the webpages on any new topics and new classes can appear as the data come; in protein sequence classification, biologists can define protein classification hierarchies based on domain knowledge and literatures, but it is almost impossible to name all classes since many protein families are still new to the biologists and new types of proteins can arise as nature evolves. This type of learning scenario, which is referred to as serendipitous learning (SL), can be summarized as follows: given labeled examples in a predefined label space, our goal is to design learning algorithms which can make accurate predictions for test examples in existing classes, and also identify new classes. Despite of its practical and theoretical importance, to the best of our knowledge, SL has not been fully explored in existing works and therefore is the focus of this paper.

The concept of serendipitous learning is very common in human learning. Many novel scientific findings or new knowledge may occur by chance, or as a by-product of the main task. In these cases, humans need to expand their knowledge base accordingly, and draw the common features of the novel categories.

Several recent research works share similar spirits with our serendipitous learning problem [1, 3, 4, 11, 20]. Unsupervised transfer classification [20] builds the classification model for a target/novel class in the absence of any labeled training example but with given labeled examples belonging to auxiliary classes similar to the target class. However, in this learning scenario, the prior information on the correlations between the target class and auxiliary classes has to be provided, which significantly limits its practical uses. The research on never-ending learning [3, 4] mainly focuses on self-supervised learning from primarily unlabeled data, i.e., how to use knowledge about subset and mutual exclusion relations between classes to couple the semi-supervised learning of these classes in order to improve the accuracy of learning. As an application of machine learning, topic detection and tracking (TDT) [1] aims to

correctly classify the documents in known categories, while identifying the documents that belong to new categories. However, in TDT, these new categories are identified separately by thresholding the similarities between new documents and labeled ones. TDT does not explicitly model the joint set of labeled and unlabeled documents in a unified label space that contains both predefined label space and new label space. Moreover, TDT does not provide theoretical analysis as that in this paper. As a means of incorporating the unlabeled examples to improve the classification performance, semi-supervised learning [25] learns a better model by utilizing the distribution similarities between the labeled and unlabeled examples on the same label space while serendipitous learning needs to find similarities between existing label space and new label space.

This paper proposes a large margin formulation to address the serendipitous learning problem. More specifically, the proposed formulation intends to maximize the margins on both the labeled and unlabeled examples simultaneously, in which the unlabeled set contains both the observed categories in the labeled set and some novel categories that do not appear in the labeled set. From another perspective, since there is no labeled information for the novel categories, we can consider the task of identifying these novel categories as a clustering problem. Meanwhile, classifying the unlabeled examples into the predefined categories can be considered as a supervised problem. These two objectives are both incorporated into the proposed formulation for designing classifiers of classes that have and have not been observed before. Therefore, the proposed formulation can also be considered as an integration of both the supervised and unsupervised problems. To solve this formulation, we propose an efficient and effective way based on the *Constrained Concave-Convex Procedure (CCCP)* [17] and an adaption of the bundle method [13, 15]. In order to handle large-scale data for online learning applications, an online version of the proposed method is developed thereafter. According to Theorem 3.1, the regret of the proposed method, i.e., the difference between the total loss and its optimal value for a fixed solution, is confined within $O(\log(t))$, where t is the number of online examples. Finally, we conduct experiments on two synthetic datasets and two application datasets to demonstrate the advantages of the proposed method.

The rest of the paper is organized as follows: Section 2 discusses some related works. Section 3 describes the proposed methodology, the proposed online learning method, as well as some theoretical analysis. Section 4 presents the experimental results. At the end of this paper, conclusions will be drawn in Section 5.

2. RELATED WORKS

2.1 Never-Ending Learning

Never-Ending Learning [3, 4] is proposed to enable computer agents to learn forever. In particular, each day, the computer agents need to (1) extract information from the web to update a growing structured knowledge base; (2) learn to update the model with the new knowledge base. The basic motivation for this work is that the vast redundancy of information will enable a system with the right learning mechanisms to succeed, and since the system keeps updating itself, the model trained by the right learning methods will be more and more accurate. It is clear that never-ending learning focuses on self-supervised learning from primarily unlabeled data.

Our proposed method can be applied to works in the scenario of Never-Ending Learning, since it does not require the label spaces for the training and testing sets to be the same. The online learning method of SL also enables the efficient processing with sequential data, and discovering novel categories.

2.2 Unsupervised Transfer Classification

In [20], the authors study a novel learning problem as unsupervised transfer classification, which builds classification models for a novel/target class in the absence of any labeled training example for that class. In their proposed learning scenario, they assume that the following side information is available: (1) some labeled examples in other classes, i.e., auxiliary classes; (2) the class information, including the prior for the novel class and the conditional probabilities between this class and the auxiliary classes. With these two assumptions, the authors propose a framework which is based on the generalized maximum entropy model. This model effectively transfers the label information from the auxiliary classes to the novel class. They further give some theoretical analysis, showing that under certain assumptions, the learned classification model converges to the optimal one.

Unsupervised transfer classification is related to the proposed work, since it also contains some labeled examples from some other domains and examples from an unknown category. However, in a lot of real world applications, the class information may not be available. Furthermore, in the proposed serendipitous learning scenario, we need to deal with the case when the number of unknown categories is more than one, and the class information is not available. Although no class information is required in the proposed method, the classifiers of known categories can serve as some prior knowledge for unknown categories, since the classifiers of different categories should be as distinct as possible.

2.3 Topic Detection and Tracking (TDT)

As a means of effectively retrieving and organizing examples into groups of events dynamically, the notion of Topic Detection and Tracking (TDT) [1] is proposed and has received considerable attentions. TDT mainly investigates methods for automatically organizing news stories by the events that they discuss. It includes several specific evaluations, i.e., (1) splitting the stream of news into stories that are about a single topic (stream segmentation); (2) arranging stories into groups so that each group discusses a single topic (link detection); (3) identifying the onset of a new topic (first story detection); (4) exploiting user feedback to monitor the stream of news (story tracking). The main goals of TDT are to monitor a stream of broadcast news stories, and try to identify the relationships between these stories.

The third evaluation of the TDT method can be adapted to address the proposed serendipitous learning scenario by first identifying examples that do not belong to the existing category and then clustering these examples into several different categories. However, detecting the onset of new topics cannot be directly applied to identifying the examples that do not belong to the existing categories. This is because in TDT the novel category examples are identified through setting a single threshold. But this threshold may not be optimal for detecting all of the novel examples. The proposed SL algorithm in this paper explicitly models the joint set of labeled and unlabeled documents in a unified label space that contains both predefined label space and new label space, which promises to substantially improve the performance. This claim can be further verified through the experiments in Section 4.

2.4 Maximum Margin Classification and Clustering

Maximum Margin learning is an important technique for classification and dimensionality reduction [12, 14]. As a typical Maximum Margin classification method, Support Vector Machines (SVM) has received considerable attentions in the past decade. In SVM, the margin is defined as the distance between the classification hy-

perplanes and a set of support examples. In its formulation, this kind of margins should be maximized such that the label assignments for the labeled examples should be as certain/determined as possible.

Recently, the idea of maximum margin learning has also been applied to data clustering, which is usually referred to as Maximum Margin Clustering (MMC). In [18], the authors assign instances to two classes $\{-1, +1\}$ so that the separation between the two classes can be as large as possible. The maximum margin clustering method has a solid theoretical foundation and performs much better than the previous methods. But in [18] it can only deal with the two class separation problem. In the later work [19], it is extended to the multi-class case. However, one of the potential problems in these methods is that they are often very time consuming. In [22, 23, 24], the authors relax the original problem and solve the relaxed problem in a more efficient way.

The previous MMC works have already shown their superior performances in dealing with clustering problems. However, there is no prior work on investigating how to deal with the proposed learning scenario, in which the classification and clustering problems exist simultaneously. In this learning scenario, we need to train some large margin classifiers on the examples that belong to the known classes while a set of large margin clustering hyperplanes should also be trained to identify the categories that have never appeared before.

3. METHODOLOGY

3.1 Problem Statement and Notations

Suppose we are given a training dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\} \in \{\mathcal{X}, \mathcal{Y}\}$, and m unlabeled examples $\{\mathbf{x}_1^U, \dots, \mathbf{x}_m^U\} \in \mathcal{X}$. \mathcal{X} is a d -dimensional space. \mathcal{Y} denotes the label space for the training set and $y_i \in \{1, 2, \dots, l\}$, where l is the number of classes. Different from traditional learning problems, the label space of the unlabeled examples are larger than that of the labeled examples, and is denoted as \mathcal{Y}' , $\mathcal{Y}' \supseteq \mathcal{Y}$. Suppose $\mathcal{Y}' = \{1, 2, \dots, l+k\}$, where k is the number of new categories in the unlabeled examples. The main objective of *Serendipitous Learning (SL)* is to infer the labels of the unlabeled examples, so that the examples belonging to \mathcal{Y} can be correctly classified, and the examples belonging to the novel categories, i.e., $\mathcal{Y}' - \mathcal{Y} = \{l+1, l+2, \dots, l+k\}$, can be clustered to the correct group.

3.2 Formulation

The main challenge for SL is how to integrate the classification task on the known categories and the clustering task on the unknown categories together in a unified framework. The basic motivation of the methodology that we employ here is that the classifiers for different categories should be as distinct as possible, which will serve as a bridge to connect these two types of tasks together.

In particular, the proposed formulation tries to find a set of hyperplanes that can separate examples from different categories as much as possible. To model these hyperplanes, for each classifier and cluster $p \in \{1, 2, \dots, l+k\}$, we define a weight vector \mathbf{w}_p . This is a multi-class problem, where $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l)$ are used to identify the known categories, and $(\mathbf{w}_{l+1}, \dots, \mathbf{w}_{l+k})$ are used for clustering the unknown category examples. For examples belonging to \mathcal{Y} , we want their labels to be as determined as possible, while for examples belonging to $\mathcal{Y}' - \mathcal{Y}$, their cluster margins should be maximized. In particular, the large margin formulation of SL can

be formally given as follows:

$$\begin{aligned} \min_{\mathbf{w}_1, \dots, \mathbf{w}_{l+k}, \mathbf{y}^U} & \frac{1}{2} \sum_{i=1}^{l+k} \|\mathbf{w}_i\|^2 + \frac{C_1}{m+n} \sum_{j=1}^n \xi_j + \frac{C_2}{m+n} \sum_{j=1}^m \xi_j^U \\ \text{s.t.} & \quad \forall j \in \{1, \dots, n\}, r = 1, \dots, l \\ & \quad \mathbf{w}_{y_j}^T \mathbf{x}_j + \delta_{y_j, r} - \mathbf{w}_r^T \mathbf{x}_j \geq 1 - \xi_j \\ & \quad \forall j \in \{1, \dots, m\}, r = 1, \dots, l+k \\ & \quad \mathbf{w}_{y_j^U}^T \mathbf{x}_j^U + \delta_{y_j^U, r} - \mathbf{w}_r^T \mathbf{x}_j^U \geq 1 - \xi_j^U \\ & \quad \forall p \in \{l+1, \dots, l+k\}, \forall q \in \{l+1, \dots, l+k\} \\ & \quad -e \leq \sum_{j=1}^m \mathbf{w}_p^T \mathbf{x}_j^U - \sum_{j=1}^m \mathbf{w}_q^T \mathbf{x}_j^U \leq e, \end{aligned} \quad (1)$$

where \mathbf{y}^U is the class assignments for the unlabeled examples. $\delta_{y_j, r}$ is an indication function. $\delta_{y_j, r} = 1$, if $y_j = r$; and otherwise 0. $\delta_{y_j^U, r}$ is defined likewise. The objective function is composed of three parts. The first part $\frac{1}{2} \sum_{i=1}^{l+k} \|\mathbf{w}_i\|^2$ is a regularizer, which confines the capacity of the classifiers. The second part $\frac{C_1}{m+n} \sum_{j=1}^n \xi_j$ is the loss function defined on the labeled examples, while the third part $\frac{C_2}{m+n} \sum_{j=1}^m \xi_j^U$ penalizes the small margins for the unlabeled examples. C_1 and C_2 are two trade-off parameters, which tune the importance of these three terms. There are three sets of constraints in this formulation. The first set of constraints require that the labeled examples should be correctly classified and their associated labels should be assigned as certain/determined as possible. The second set of constraints mainly focus on the unlabeled set, in which we want the classification and clustering assignments for the unlabeled examples to be as determined as possible. Actually, a similar motivation is given in Transductive Support Vector Machines (TSVM) [7] and Semi-Supervised Support Vector Machines [2], in which the authors require that the classification margins for the unlabeled examples should be maximized so that the distribution of unlabeled examples can serve as the prior knowledge for designing the corresponding classifiers. The third set of constraints suggest that the hyperplanes for the novel categories should be balanced, which can effectively avoid the trivial solution of assigning all of the novel category examples into the same cluster. e is a non-negative parameter that controls this balance. The smaller e is, the more balance the classes should be.

From the perspective of classifier design, this formulation can be considered as a combination of classification and clustering. The classifiers that are used for labeled examples should also be able to give a good classification performance on the unlabeled examples that belong to the known categories, while the hyperplanes for unknown categories cluster the examples in unknown categories into k groups. The classifiers for the known categories should be as far away from the hyperplanes of the unknown categories as possible, and vice versa. This exclusive relationship serves as the prior information in this formulation.

The proposed formulation (1) is reasonable. However, it is non-convex and cannot be solved directly. In the following part, we will suggest a way to solve this formulation efficiently. In particular, we will show how we convert and relax the proposed problem to a form that can be solved through a series of sophisticated optimization strategies, which is a combination of the *Constrained Concave-Convex Procedure (CCCP)* and the bundle method.

First of all, to simplify the formulation, without loss of generality, the following notations are introduced:

$$\tilde{\mathbf{w}} = [\mathbf{w}_1^T, \dots, \mathbf{w}_{l+k}^T]^T, \tilde{\mathbf{x}}_j^{(p)} = [\mathbf{0}, \dots, \mathbf{x}_j^T, \dots, \mathbf{0}]^T. \quad (2)$$

In this transformation, $\mathbf{0}$ is a $1 \times d$ zero vector. In $\tilde{\mathbf{x}}_j^{(p)}$, only the $(p-1)d$ to pd -th elements are nonzero and equals \mathbf{x}_j . It is clear that, with this transformation, $\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{(p)}$ equals $\mathbf{w}_p^T \mathbf{x}_j$. Then, the problem (1) can be equivalently transformed to:

$$\begin{aligned}
\min_{\tilde{\mathbf{w}}} \quad & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \frac{C_1}{m+n} \sum_{j=1}^n \xi_j + \frac{C_2}{m+n} \sum_{j=1}^m \xi_j^U \\
\text{s.t.} \quad & \forall j \in \{1, \dots, n\}, r = \{1, \dots, l\} \setminus y_j \\
& \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{(y_j)} - \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{(r)} \geq 1 - \xi_j \\
& \forall j \in \{1, \dots, m\}, r = 1, \dots, l+k \\
& \max_p \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p)} + \delta_{y_j^U, r} - \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(r)} \geq 1 - \xi_j^U \\
& \forall p \in \{l+1, \dots, l+k\}, \forall q \in \{l+1, \dots, l+k\} \\
& -e \leq \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p)} - \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(q)} \leq e. \quad (3)
\end{aligned}$$

Here, $\tilde{\mathbf{x}}_j^{U(p)}$ is defined similar to $\tilde{\mathbf{x}}_j^{(p)}$, and $y_j^U = \arg \max_p \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p)}$. After this transformation, the number of variables that need to be optimized is reduced by m . Although simplified, this optimization problem is still intractable, due to the non-convexity of $\delta_{y_j^U, r}$ and $\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(r)}$. To solve this problem, we propose to absorb $\delta_{y_j^U, r}$ and relax $\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(r)}$. In particular, we replace $\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(r)}$ with $\text{mean}_q(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(q)})$, and transform the notion of margin from the difference between the largest and second largest outputs to the difference between the largest and the associated mean output with respect to all of hyperplanes. After this relaxation, problem (3) turns to:

$$\begin{aligned}
\min_{\tilde{\mathbf{w}}} \quad & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \frac{C_1}{m+n} \sum_{j=1}^n \xi_j + \frac{C_2}{m+n} \sum_{j=1}^m \xi_j^U \\
\text{s.t.} \quad & \forall j \in \{1, \dots, n\}, r = \{1, \dots, l\} \setminus y_j \\
& \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{y_j} - \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^r \geq 1 - \xi_j \\
& \forall j \in \{1, \dots, m\}, \\
& \max_p \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p)} - \text{mean}_q(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(q)}) \geq 1 - \xi_j^U \\
& \forall p \in \{l+1, \dots, l+k\}, \forall q \in \{l+1, \dots, l+k\} \\
& -e \leq \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p)} - \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(q)} \leq e. \quad (4)
\end{aligned}$$

The second set of constraints can be considered as the difference between two convex functions. Therefore, CCCP, which is an optimization method that deals with the concave convex objective functions and constraints, can be used to solve this formulation.

Given an initial point $\tilde{\mathbf{w}}^{(0)}$, CCCP computes $\tilde{\mathbf{w}}^{(t+1)}$ from $\tilde{\mathbf{w}}^{(t)}$ iteratively by replacing $\max_p \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p)}$ with its first order Taylor expansions at $\tilde{\mathbf{w}}^{(t)}$, and solves the resulting quadratic programming problem, until convergence. For the t -th CCCP iteration, the sub-

¹We use the superscript t to denote that the result is obtained from the t -th CCCP iteration, i.e., $\tilde{\mathbf{w}}^{(t)}$ is optimized from the t -th CCCP iteration step.

Algorithm: Serendipitous Learning (SL)
Input:
1. Labeled examples: $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$. Unlabeled examples: $\{\mathbf{x}_1^U, \dots, \mathbf{x}_m^U\}$
2. Parameters: the trade-off parameters C_1 and C_2 ; the class balance parameter e ; CCCP precision parameter $\delta_1 = 0.001$; the bundle method precision $\delta_2 = 0.001$.
Output: The class assignments \mathbf{y}^U and $\tilde{\mathbf{w}}^{(t)}$
CCCP Iterations:
1. Construct $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}}_i^{(r)}\}, \tilde{\mathcal{X}}^U = \{\tilde{\mathbf{x}}_i^{U(r)}\}$
2. Initialize $\tilde{\mathbf{w}}^0, t=0, \Delta J = 10^3, J^{(-1)} = 10^3$
3. while $\Delta J / J^{(t-1)} > \delta_1$ do
4. Derive problem (5) by updating $p_j^{(t)} = \arg \max_p \tilde{\mathbf{w}}^{(t)T} \tilde{\mathbf{x}}_j^{U(p)}$.
5. $t = t + 1, s = 0$
Bundle Method Iterations:
6. repeat
7. $s = s + 1$
8. Compute the gradient for the empirical loss: $\mathbf{a}^{ts} = \partial_{\tilde{\mathbf{w}}} R_{emp}^{(t)}(\tilde{\mathbf{w}}^{(t_{s-1})})$, and $b^{ts} = R_{emp}^{(t)}(\tilde{\mathbf{w}}^{(t_{s-1})}) - \langle \tilde{\mathbf{w}}^{(t_{s-1})}, \mathbf{a}^{ts} \rangle$.
9. Derive the optimization problem: $R_s^{CP} = \max_{1 \leq i \leq s} \langle \tilde{\mathbf{w}}, \mathbf{a}^{ti} \rangle + b^{ti}$
10. $\tilde{\mathbf{w}}^{ts} = \arg \min_{\tilde{\mathbf{w}}} \frac{1}{2} \ \tilde{\mathbf{w}}\ ^2 + R_s^{CP}$,
s.t., $-e \leq \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p_j^{(t)})} - \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(q)} \leq e$
11. $\epsilon_s = \min_{0 \leq i \leq s} J(\tilde{\mathbf{w}}^{ti}) - J_s(\tilde{\mathbf{w}}^{ts})$
12. until $\epsilon_s \leq \delta_2$
13. $J^{(t)} = \min_{0 \leq i \leq s} J(\tilde{\mathbf{w}}^{ti}), \tilde{\mathbf{w}}^{(t)} = \tilde{\mathbf{w}}^{ts}$
14. $\Delta J = J^{(t-1)} - J^{(t)}$
15. end while
16. Class Assignment:
For \mathbf{x}_i^U , the corresponding class assignment $\mathbf{y}_i^U = \arg \max_{p \in \{1, 2, \dots, l+k\}} (\tilde{\mathbf{w}}^{(t)})^T \tilde{\mathbf{x}}_i^{U(p)}$

Table 1: Algorithm Description: Serendipitous Learning

problem is as follows:

$$\begin{aligned}
\min_{\tilde{\mathbf{w}}} \quad & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \frac{C_1}{m+n} \sum_{j=1}^n \xi_j + \frac{C_2}{m+n} \sum_{j=1}^m \xi_j^U \\
\text{s.t.} \quad & \forall j \in \{1, \dots, n\}, r = \{1, \dots, l\} \setminus y_j \\
& \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{y_j} - \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^r \geq 1 - \xi_j \\
& \forall j \in \{1, \dots, m\}, \\
& \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p_j^{(t)})} - \text{mean}_q(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(q)}) \geq 1 - \xi_j^U \\
& \forall p \in \{l+1, \dots, l+k\}, \forall q \in \{l+1, \dots, l+k\} \\
& -e \leq \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p)} - \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(q)} \leq e, \quad (5)
\end{aligned}$$

where $p_j^{(t)} = \arg \max_p \tilde{\mathbf{w}}^{(t)T} \tilde{\mathbf{x}}_j^{U(p)}$. For each sub-problem, if the number of classes and number of examples are high, the computational cost would be huge. In this paper, we propose an optimization based on the bundle method [13, 15] to solve this subproblem efficiently and effectively.

The basic motivation of the bundle method is to approximate the objective function $J(\tilde{\mathbf{w}})$ through a set of linear functions, where $J(\tilde{\mathbf{w}}) = \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \frac{C_1}{m+n} \sum_{j=1}^n \xi_j + \frac{C_2}{m+n} \sum_{j=1}^m \xi_j^U$. In particular, this objective function is lower bounded as follows:

$$J(\tilde{\mathbf{w}}) \geq \max_{1 \leq i \leq s} \{J(\tilde{\mathbf{w}}^{t_{i-1}}) + \langle \tilde{\mathbf{w}} - \tilde{\mathbf{w}}^{t_{i-1}}, \mathbf{a}_i \rangle\},$$

where $\tilde{\mathbf{w}}^{t_i}$ ($1 \leq i \leq s$) is a set of points picked by the bundle method, and \mathbf{a}_i is the gradient/sub-gradient at point $\tilde{\mathbf{w}}^{t_i}$. The bundle method monotonically decreases the gap between $J(\tilde{\mathbf{w}})$ and $J_s(\tilde{\mathbf{w}}) = \max_{1 \leq i \leq s} \{J(\tilde{\mathbf{w}}^{t_{i-1}}) + \langle \tilde{\mathbf{w}} - \tilde{\mathbf{w}}^{t_{i-1}}, \mathbf{a}_i \rangle\}$ such that the minimal point of $J(\tilde{\mathbf{w}})$ can be approximated by the minimal point of the line segments $J_s(\tilde{\mathbf{w}})$.

The concrete procedure is described in Table 1. Here, $R_{emp}^{(t)}(\tilde{\mathbf{w}}) = \frac{C_1}{m+n} \sum_{j=1}^n \max\{0, \max_{r \setminus y_j} \{1 - \tilde{\mathbf{w}}^T (\tilde{\mathbf{x}}_j^{(y_j)} - \tilde{\mathbf{x}}_j^{(r)})\}\} + \frac{C_2}{m+n} \sum_{j=1}^m$

²Throughout this paper, t_s is used to denote the s -th iteration of the bundle method iteration for solving the problem derived from the t -th CCCP iteration.

$\max\{0, (1 - \tilde{\mathbf{w}}^T (\tilde{\mathbf{x}}_j^{U(p_j^{(t)})} - \text{mean}_q \tilde{\mathbf{x}}_j^{U(q)}))\}$. It is clear that $R_{emp}^{(t)}(\tilde{\mathbf{w}})$ is a non-smooth function. So, its gradient cannot be derived directly. Instead, we can calculate the subgradient as follows:

$$\begin{aligned} \partial_{\tilde{\mathbf{w}}} R_{emp}^{(t)}(\tilde{\mathbf{w}}^{(t_{s-1})}) &= \frac{C_1}{m+n} \sum_{j=1}^n \sum_{u=1}^l I_u(\tilde{\mathbf{x}}_j, \tilde{\mathbf{w}}^{(t_{s-1})})(\tilde{\mathbf{x}}_j^{(u)} - \tilde{\mathbf{x}}_j^{(y_j)}) \\ &+ \frac{C_2}{m+n} \sum_{j=1}^m I(\tilde{\mathbf{x}}_j^U, \tilde{\mathbf{w}}^{(t_{s-1})})(\text{mean}_q \tilde{\mathbf{x}}_j^{U(q)} - \tilde{\mathbf{x}}_j^{U(p_j^{(t)})}) \end{aligned}$$

Here, $I_u(\tilde{\mathbf{x}}_j, \tilde{\mathbf{w}}^{(t_{s-1})})$ equals 1 if $\max_{r \setminus y_j} \{1 - \tilde{\mathbf{w}}^T(\tilde{\mathbf{x}}_j^{(y_j)} - \tilde{\mathbf{x}}_j^{(r)})\} > 0$ and $u = \arg \max_{r \setminus y_j} \{1 - \tilde{\mathbf{w}}^T(\tilde{\mathbf{x}}_j^{(y_j)} - \tilde{\mathbf{x}}_j^{(r)})\}$, and otherwise 0. $I(\tilde{\mathbf{x}}_j^U, \tilde{\mathbf{w}}^{(t_{s-1})})$ equals 1 if $(1 - \tilde{\mathbf{w}}^{(t_{s-1})T}(\tilde{\mathbf{x}}_j^{U(p_j^{(t)})} - \text{mean}_q \tilde{\mathbf{x}}_j^{U(q)})) > 0$ and otherwise 0.

The outer iteration of the proposed method is CCCP. It has already been proved that the CCCP decreases the objective function monotonically [17]. The inner iteration is the bundle method for regularized objective risk minimization. Similar to the proof in [15], we can prove that given δ_2 , this method converges to the optimal solution in $O(1/\delta_2)$ steps. Due to the lack of space, the concrete proof is omitted here.

3.3 Online Learning

In previous sections, we have presented SL method, which can be trained in a batch mode given a set of labeled and unlabeled examples. However, in many real world problems, it is highly possible that the data comes in sequential order. By obtaining data in a continuous data stream, novel categories could be introduced gradually. In this section, an online learning method for SL is proposed. In particular, we focus on the online learning algorithm that takes examples, i.e. $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T\}$, either labeled or unlabeled, sequentially as they become available, and updates the weight vectors. Suppose at time t^3 , the most up-to-date weight $\tilde{\mathbf{w}}_t$ is available. We are trying to minimize the regret defined as follows:

$$\begin{aligned} R_t &= \sum_{\tau=-(m+n-1)}^0 \left(\frac{1}{2} \|\tilde{\mathbf{w}}_0\|^2 + f_\tau(\tilde{\mathbf{w}}_0) \right) + \sum_{\tau=1}^t \left(\frac{1}{2} \|\tilde{\mathbf{w}}_\tau\|^2 + \right. \\ & \left. f_\tau(\tilde{\mathbf{w}}_\tau) - \min_{\tilde{\mathbf{w}} \in K} \sum_{\tau=-(m+n-1)}^t \left(\frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + f_\tau(\tilde{\mathbf{w}}) \right) \right) \\ \text{s.t. } & \tilde{\mathbf{w}}_\tau \in K, \end{aligned} \quad (6)$$

where K denotes the feasible region, defined by $\forall p \in \{l+1, \dots, l+k\}, \forall q \in \{l+1, \dots, l+k\}, -e \leq \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p)} - \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(q)} \leq e$. Without loss of generality, we assume that the previous $m+n$ labeled and unlabeled examples appear in sequential from time $-(m+n-1)$ to time 0. Next, we will elaborate how the cost function $f_\tau(\tilde{\mathbf{w}}_\tau)$ is designed with different kinds of input data. Suppose the example \mathbf{z}_τ is labeled, and its label is \mathbf{y}_τ . Then, $f_\tau(\tilde{\mathbf{w}}_\tau) = C_1 \max\{0, \max_{r \setminus \mathbf{y}_\tau} \{1 - \tilde{\mathbf{w}}_\tau^T(\mathbf{z}_\tau^{(\mathbf{y}_\tau)} - \mathbf{z}_\tau^{(r)})\}\}$. If \mathbf{z}_τ is an unlabeled example, \mathbf{z}_τ is first classified, with the corresponding label assignment $\mathbf{y}_\tau = \arg \max_r \tilde{\mathbf{w}}_\tau^T \mathbf{z}_\tau^{(r)}$. The corresponding $f_\tau(\tilde{\mathbf{w}}_\tau)$ is then defined as $C_2 \max\{0, (1 - \tilde{\mathbf{w}}_\tau^T(\mathbf{z}_\tau^{(\mathbf{y}_\tau)} - \text{mean}_q \mathbf{z}_\tau^{(q)}))\}$. Since $f_\tau(\tilde{\mathbf{w}}_\tau)$ is non-smooth, when calculating the gradients, we can use the sub-gradient instead.

The complete online learning algorithm is described in Table 2, which can be considered as an extension of online gradient descent

³To avoid confusion, t here means the t -th online example, which is different from the previous usage for t -th CCCP iteration.

Algorithm: SL-Online
Input:
1. $\tilde{\mathbf{w}}$ obtained from Table 1.
2. Parameters: C_1, C_2, e from Table 1.
Output: the updated model \mathbf{w}_T
1. Initialize $\tilde{\mathbf{w}}_1 = \tilde{\mathbf{w}}$
2. for $\tau = 1, 2, \dots, T$ do
3. if \mathbf{z}_τ is labeled with \mathbf{y}_τ , then
$f_\tau(\tilde{\mathbf{w}}) = C_1 \max\{0, \max_{r \setminus \mathbf{y}_\tau} \{1 - \tilde{\mathbf{w}}^T(\mathbf{z}_\tau^{(\mathbf{y}_\tau)} - \mathbf{z}_\tau^{(r)})\}\}$
4. if \mathbf{z}_τ is unlabeled, then $\mathbf{y}_\tau = \arg \max_r \tilde{\mathbf{w}}^T \mathbf{z}_\tau^{(r)}$, and
$f_\tau(\tilde{\mathbf{w}}) = C_2 \max\{0, (1 - \tilde{\mathbf{w}}^T(\mathbf{z}_\tau^{(\mathbf{y}_\tau)} - \text{mean}_q \mathbf{z}_\tau^{(q)}))\}$
5. Compute $\eta_{\tau+1} = (m+n+\tau)^{-1}$
6. $\tilde{\mathbf{w}}_{\tau+1}^b = \tilde{\mathbf{w}}_\tau - \eta_{\tau+1} \partial_{\tilde{\mathbf{w}}} f_\tau(\tilde{\mathbf{w}}_\tau)$
7. Solve $\tilde{\mathbf{w}}_{\tau+1} = \arg \min_{\tilde{\mathbf{w}}} \ \tilde{\mathbf{w}} - \tilde{\mathbf{w}}_{\tau+1}^b\ ^2$,
s.t. $\forall p \in \{l+1, \dots, l+k\}, \forall q \in \{l+1, \dots, l+k\}$,
$-e \leq \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(p)} - \sum_{j=1}^m \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_j^{U(q)} \leq e$.
8. end for

Table 2: Algorithm Description: The Online Learning Algorithm for Serendipitous Learning

method [6]. As we shall see later, the regret of the proposed method is guaranteed to be upper bounded by $O(\log(t))$.

3.4 Theoretical Analysis

THEOREM 3.1. *At time t , the regret defined in Eq.(6) will be upper bounded by:*

$$\frac{t \log(t)}{2(m+n+t)} \max_{\tau \in \{1, \dots, t\}} \mathbf{G}_\tau^2 + C, \quad (7)$$

where C is negative and equals $\sum_{\tau=-(m+n-1)}^0 (\frac{1}{2} \|\tilde{\mathbf{w}}_{-(m+n-1):0}^*\|^2 + f_\tau(\tilde{\mathbf{w}}_{-(m+n-1):0}^*)) - \sum_{\tau=-(m+n-1)}^0 (\frac{1}{2} \|\tilde{\mathbf{w}}^*\|^2 + f_\tau(\tilde{\mathbf{w}}^*))$, and $\mathbf{G}_\tau = \|\partial_{\tilde{\mathbf{w}}} f_\tau(\tilde{\mathbf{w}}_\tau)\|$. Here, $\tilde{\mathbf{w}}_{-(m+n-1):0}^*$ is the optimal solution from Table 1 and $\tilde{\mathbf{w}}^* = \arg \min_{\tilde{\mathbf{w}} \in K} \sum_{\tau=-(m+n-1)}^t (\frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + f_\tau(\tilde{\mathbf{w}}))$.

PROOF. This proof is an extension of the convergence proof in [6]. It is clear that if t equals 0, since $\tilde{\mathbf{w}}_0 = \arg \min_{\tilde{\mathbf{w}} \in K} \sum_{\tau=-(m+n-1)}^0 (\frac{1}{2} \|\tilde{\mathbf{w}}_0\|^2 + f_\tau(\tilde{\mathbf{w}}_0))$, the regret R_t equals 0. So, our focus is on the case when t is larger than 1.

To solve this problem, first of all, suppose $\tilde{\mathbf{w}}_{1:t}^* = \arg \min_{\tilde{\mathbf{w}} \in K} \sum_{\tau=1}^t (\frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + f_\tau(\tilde{\mathbf{w}}))$. So,

$$\begin{aligned} R_t &= \sum_{\tau=-(m+n-1)}^0 \left(\frac{1}{2} \|\tilde{\mathbf{w}}_0\|^2 + f_\tau(\tilde{\mathbf{w}}_0) \right) + \sum_{\tau=1}^t \left(\frac{1}{2} \|\tilde{\mathbf{w}}_\tau\|^2 \right. \\ & \left. + f_\tau(\tilde{\mathbf{w}}_\tau) - \sum_{\tau=-(m+n-1)}^t \left(\frac{1}{2} \|\tilde{\mathbf{w}}^*\|^2 + f_\tau(\tilde{\mathbf{w}}^*) \right) \right) \\ & \leq \sum_{\tau=1}^t \left(\frac{1}{2} \|\tilde{\mathbf{w}}_\tau\|^2 + f_\tau(\tilde{\mathbf{w}}_\tau) \right) - \sum_{\tau=1}^t \left(\frac{1}{2} \|\tilde{\mathbf{w}}^*\|^2 + f_\tau(\tilde{\mathbf{w}}^*) \right) + C \\ & \leq \sum_{\tau=1}^t \left(\frac{1}{2} \|\tilde{\mathbf{w}}_\tau\|^2 + f_\tau(\tilde{\mathbf{w}}_\tau) \right) - \sum_{\tau=1}^t \left(\frac{1}{2} \|\tilde{\mathbf{w}}_{1:t}^*\|^2 + f_\tau(\tilde{\mathbf{w}}_{1:t}^*) \right) + C. \end{aligned}$$

Then, we will show that $\sum_{\tau=1}^t (\frac{1}{2} \|\tilde{\mathbf{w}}_\tau\|^2 + f_\tau(\tilde{\mathbf{w}}_\tau)) - \sum_{\tau=1}^t (\frac{1}{2} \|\tilde{\mathbf{w}}_{1:t}^*\|^2 + f_\tau(\tilde{\mathbf{w}}_{1:t}^*))$ is upper bounded. Suppose $\Phi_\tau(\tilde{\mathbf{w}}) = \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + f_\tau(\tilde{\mathbf{w}})$, we have:

$$\Phi_\tau(\tilde{\mathbf{w}}_{1:t}^*) \geq \Phi_\tau(\tilde{\mathbf{w}}_\tau) + \mathbf{G}_\tau^T(\tilde{\mathbf{w}}_{1:t}^* - \tilde{\mathbf{w}}_\tau) + \frac{1}{2} \|\tilde{\mathbf{w}}_{1:t}^* - \tilde{\mathbf{w}}_\tau\|^2,$$

and use the expression for $\tilde{\mathbf{w}}_{\tau+1}$ implies:

$$\begin{aligned} \|\tilde{\mathbf{w}}_{\tau+1} - \tilde{\mathbf{w}}_{1:t}^*\|^2 &\leq \|\tilde{\mathbf{w}}_\tau - \tilde{\mathbf{w}}_{1:t}^*\|^2 - 2\eta_{t+1} \mathbf{G}_\tau^T(\tilde{\mathbf{w}}_\tau - \tilde{\mathbf{w}}_{1:t}^*) \\ & \quad + \eta_{t+1}^2 \|\mathbf{G}_\tau\|^2. \end{aligned}$$

It indicates that

$$2\mathbf{G}_\tau^T(\tilde{\mathbf{w}}_\tau - \tilde{\mathbf{w}}_{1:t}^*) \leq \frac{\|\tilde{\mathbf{w}}_\tau - \tilde{\mathbf{w}}_{1:t}^*\|^2 - \|\tilde{\mathbf{w}}_{\tau+1} - \tilde{\mathbf{w}}_{1:t}^*\|^2}{\eta_{t+1}} + \eta_{t+1}\|\mathbf{G}_\tau\|^2$$

By combining the previous equations together, we can get:

$$2(\Phi_\tau(\tilde{\mathbf{w}}_\tau) - \Phi_\tau(\tilde{\mathbf{w}}_{1:t}^*)) \leq \frac{\|\tilde{\mathbf{w}}_\tau - \tilde{\mathbf{w}}_{1:t}^*\|^2 - \|\tilde{\mathbf{w}}_{\tau+1} - \tilde{\mathbf{w}}_{1:t}^*\|^2}{\eta_{t+1}} + \eta_{t+1}\|\mathbf{G}_\tau\|^2 - \|\tilde{\mathbf{w}}_{1:t}^* - \tilde{\mathbf{w}}_\tau\|^2$$

Summing τ over $\{1, \dots, t\}$:

$$2 \sum_{\tau=1}^t (\Phi_\tau(\tilde{\mathbf{w}}_\tau) - \Phi_\tau(\tilde{\mathbf{w}}_{1:t}^*)) \leq \sum_{\tau=1}^t \|\tilde{\mathbf{w}}_\tau - \tilde{\mathbf{w}}_{1:t}^*\|^2 \left(\frac{1}{\eta_{\tau+1}} - \frac{1}{\eta_\tau} - 1 \right) + \sum_{\tau=1}^t \mathbf{G}_\tau^2 \eta_{\tau+1}$$

In the proposed algorithm, $\frac{1}{\eta_{\tau+1}} - \frac{1}{\eta_\tau} - 1 = (m+n+\tau) - (m+n+\tau-1) - 1 = 0$. Therefore,

$$2 \sum_{\tau=1}^t (\Phi_\tau(\tilde{\mathbf{w}}_\tau) - \Phi_\tau(\tilde{\mathbf{w}}_{1:t}^*)) \leq \sum_{\tau=1}^t \frac{\mathbf{G}_\tau^2}{(m+n+\tau)} \leq \sum_{\tau=1}^t \frac{\mathbf{G}_\tau^2/\tau}{1 + \frac{m+n}{t}} \leq \frac{t \log(t)}{m+n+t} \max_{\tau \in \{1, \dots, t\}} \mathbf{G}_\tau^2.$$

So, for the t -th online example, the increase of the regret incurred by using the method in Table 2 is upper bounded by $O(\log(t))$. It is clear that this upper bound can be significantly decreased given enough labeled and unlabeled examples offline. \square

Dataset	# Dim	Labeled Set		Unlabeled Set	
		# Categories	# inst	#Categories	# inst
Synthetic I	2	3	60	5	700
Synthetic II	2	3	300	5	700
20Newsgroups	6600	5	2350	8	5328
ReutersV1	8625	3	3505	6	10513

Table 3: Dataset Descriptions for Four Datasets as Synthetic I, Synthetic II, 20 Newsgroups, and ReutersV1

4. EXPERIMENTS

A set of experiments on four datasets, i.e., two synthetic datasets, 20Newsgroups, and ReutersV1, are conducted to validate the effectiveness of the proposed method.

4.1 Datasets

Synthetic Datasets: We generate two synthetic datasets, as shown in Fig.1(a) and Fig.2, by independently extracting examples from 5 two-dimensional gaussian distributions, and each gaussian component represents a specific category. Among these five categories, two of them are used as the unknown categories in the unlabeled set, while the other three categories are considered as known categories. The difference between these two synthetic datasets is that Synthetic I is a easy linear separable classification task, which is used for illustrating the performance of the proposed method, while Synthetic II is much more difficult and is mainly designed for quantitative measurements. A detailed description of these two datasets can be found in Table 3.

20Newsgroups: This is a benchmark dataset⁴ for text categorization, which contains 20 categorizes. In the labeled examples,

⁴<http://people.csail.mit.edu/jrennie/20Newsgroups/>

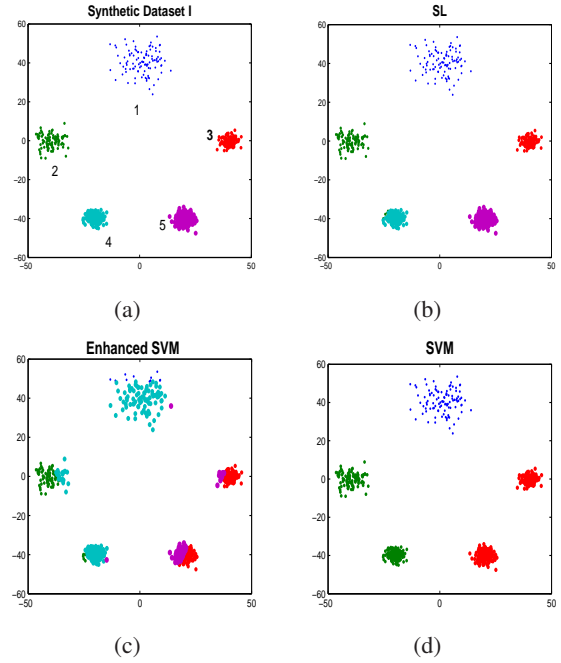


Figure 1: (a) illustrates the unlabeled set of the first synthetic dataset, where each color represents a specific class, and the class ids are indicated. Class 1, 2, and 3 are used as known classes, while 4 and 5 are considered as unknown categories. (b) demonstrates the execution results of SL. (c) is the results of Enhanced SVM, while (d) shows the classification results of traditional SVM. It is clear that on this dataset, the proposed method recovers the original class membership with 99.43% accuracy, and its performance far exceeds that of Enhanced SVM and traditional SVM.

there are in total 5 randomly selected categories, while in the unlabeled set, another 3 categories are used as the novel categories. Please refer to Table 3 for more details.

Reuters-Volume I (ReutersV1): It is an archive of over 800,000 manually categorized newswire stories [8]. A subset of ReutersV1 is used. There are in total 126 categories in this dataset. This dataset is split into a labeled set and an unlabeled set. There are in total 3 categories in the labeled set, and 6 categories in the unlabeled set. Please refer to Table 3 for more details.

4.2 Methods

We compare the proposed algorithm with four baseline methods, i.e., the traditional large margin classification algorithm – SVM [12], which has shown its superior performance compared with Naive Bayes and Logistic Regression in text classification applications, Enhanced SVM as an adapted form of SVM for serendipitous learning as described below; the generative method – Gaussian Mixture Models [9]; as well as the Topic Detection and Tracking method – TDT [21].

In particular, for SVM, we train a set of large margin classifiers on the labeled set, and directly use these classifiers to infer the labels of the unlabeled examples. But by using this method, the novel categories in the unlabeled set cannot be detected. So, we further measure the performance of Enhanced SVM, in which a set of classifiers are first trained by using SVM, and after that the least confident examples given by SVM are considered as the novel examples. Maximum Margin Clustering (MMC) [24] is then

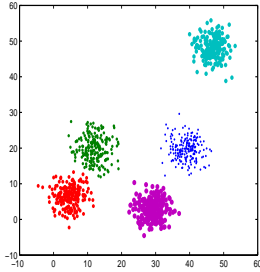


Figure 2: The Illustration of Synthetic Dataset II

used to cluster these examples into k different groups. For both SVM and Enhanced SVM, the parameters for the hinge loss are selected through 5 fold cross validations on the training set. For the MMC step in Enhanced SVM, the clustering parameter is selected through grid search. For the generative method, i.e., GMM, for each iteration, the number of generative models is fixed to be the true number of categories on the unlabeled set. Then, the gaussian models are trained on both the labeled and unlabeled sets together. Different from the traditional GMM, which is mainly used for clustering, to make fair comparisons, we are forcing labeled examples in the same categories to be in the same gaussian model during each iteration of GMM. Since GMM is a non-deterministic method, the final reported results are summarized over 20 independent runs. In TDT, similar to [21], we use the nearest neighbor method as the base classifier and detect the novel examples by measuring the similarities between each unlabeled example and its nearest neighbor. Since there is no novel category in the labeled set, the threshold for detecting the novel examples is chosen to be the best one on the unlabeled set. After the novel examples are identified, MMC is used to categorize them into predefined groups. For the proposed method, it has three parameters, C_1 , C_2 and e . For C_1 , the parameter is tuned through 5 fold cross validation, while the parameters for C_2 and e are set through grid search.

4.3 Evaluation Metric

There are two different types of unlabeled examples, i.e., the ones whose categories have already appeared in the training set, and the ones that fall into novel categories. So, there are also two different types of accuracies – classification accuracy (Classification Acc) and clustering accuracy (Clustering Acc) [16, 18, 24].

The classification accuracy measures the extent to which the examples belonging to the previously appeared categories are correctly classified. The clustering accuracy is used for examples from novel categories. In particular, we first take a set of labeled examples from novel categories, remove the labels of these examples and run the algorithms. Then we relabel these examples using the clustering assignments returned by the algorithms. Finally, we measure the percentage of correct classifications by comparing the true labels and the labels assigned by the clustering algorithms as: $Acc = \frac{1}{n} \sum_{i=1}^n \delta(y_i, map(c_i))$, where, $map(\cdot)$ is a function that maps each cluster index to a class label, which can be found by the Hungarian algorithm [10]. c_i and y_i are the cluster index of \mathbf{x}_i and the true class label. $\delta(a, b)$ is a function that equals 1 when a equals b , and 0 otherwise. It is clear that clustering accuracy discovers the one-to-one relationship between clusters, and measures the extent to which cluster assignments are associated with the corresponding true categories. The greater clustering accuracy is, the better the clustering algorithm performs.

The total accuracy is reported in this paper, which measures whether each example is correctly classified or clustered to the right group. More precisely, $Total Accuracy = Ratio_{Known} \times Classification Acc + Ratio_{Unknown} \times Clustering Acc$, where $Ratio_{Known}$ denotes the true ratio of examples that belong to the known categories, and $Ratio_{Unknown}$ refers to that of examples belonging to unknown categories.

	Synthetic II	20Newsgroup	ReutersV1
Enhanced SVM	31.8	42.7	56.1
SVM	28.6	27.5	32.7
TDT	24.6	37.1	42.2
GMM	×	31.0	54.8
SL	51.7	46.1	63.6

Table 4: Classification Results. GMM is a perfect fit for the Synthetic dataset, since this dataset is generated using GMM models. Therefore, the performance of GMM on the Synthetic dataset II is not reported. As we shall see from the real world datasets, i.e., 20Newsgroup and ReutersV1, the performance of GMM is worse than the proposed method.

4.4 Classification Results

First of all, the demonstration of the advantages of the proposed method on Synthetic Dataset I is illustrated in Fig.1. It is clear that compared with Enhanced SVM and SVM, the proposed method can perfectly identify the original data classes, while SVM and Enhanced SVM performs relatively poor in this dataset.

The average classification results of 20 independent runs on the remaining three datasets are reported in Fig.3 with varying training ratios. In Table 4, we further report the average performance with training ratio being fixed to be 1. It is clear that on all of these three datasets, the proposed method shows the best performances among all algorithms.

The advantages of the proposed method lie in the fact that it tries to model all of the categories in the unlabeled set in a unified framework. Through this approach, the large margin classifiers for the known categories and the unknown categories can better benefit each other, which results in a higher accuracy compared with the other methods.

SVM and Enhanced SVM are two large margin classification methods. For SVM, it is clear that on all of these three datasets, its performance is not directly comparable with that of the others. This is because it only classifies the examples into known categories while neglecting the novel categories. Enhanced SVM contains two steps. In the first step, it trains a set of large margin classifiers. In the second step, the novel examples are identified and clustered. This is a heuristic choice. However, it has two drawbacks. First of all, it cannot identify the novel examples accurately. Secondly, since it is a two step approach, the classifiers for the known and unknown categories are trained independently, although their corresponding classifiers are not independent.

The TDT method used in this paper is originally designed for the first story detection, and is adapted to solve the proposed problem. It can be seen that its performance is not very promising. This is because it identifies the novel category examples by computing its cosine similarity with its nearest neighbor, and comparing this value with a threshold. However, as can also be seen from Fig.2, examples whose cosine similarities are high may actually belong to different categories. Therefore, fixing a single threshold for all of the examples is not optimal for this application. On the contrary, this problem can be effectively avoided by separating examples through

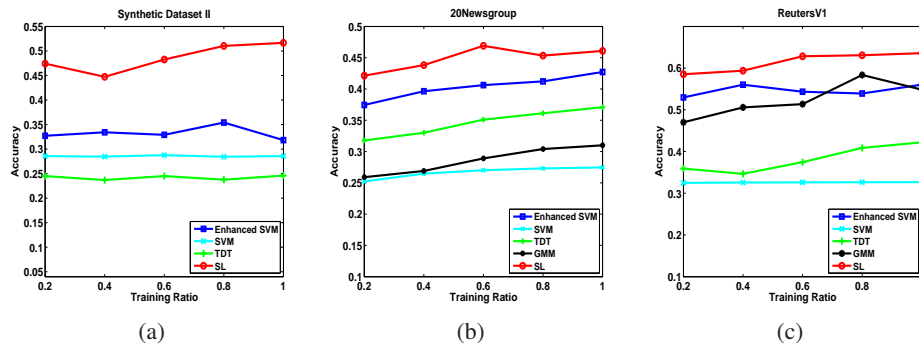


Figure 3: Classification Results with Varying Ratio of Training Examples. It is clear that the proposed method shows the best performance on all of these three datasets.

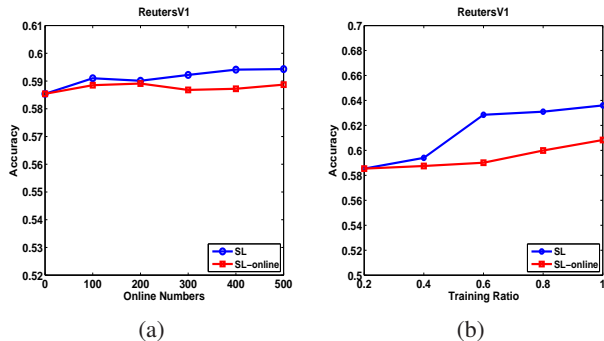


Figure 4: The performance of the online version of the proposed method. Fig.(a) measures the performance change with different number of unlabeled examples, while Fig.(b) measures that of the labeled examples. We compare the performance of online learning with the non-online version. It can be seen that the performances of these two methods are close.

large margin classifiers, and considering different classes together. That is why the proposed method performs much better than TDT.

As a generative method, GMM shows a great advantage in modeling the data distribution through a set of gaussian models. It can be used to model the synthetic dataset perfectly. However, in real world applications, where the gaussian distribution assumption does not hold, its performance is worse than the proposed method.

4.5 Online Learning Results

The performance of the associated online learning method for SL is reported in Fig.4. In particular, we measure the online learning accuracies by increasing the number of labeled and unlabeled examples separately. The originally labeled examples are fixed to be 20 percents of the whole training set. In Fig.4(a), the performance of online learning for unlabeled examples is reported, while the performance of online learning for labeled examples is reported in Fig.4(b). For Fig.4(a), we add extra unlabeled examples to the unlabeled set in sequence, while for Fig.4(b), the labeled examples in the database are added to the labeled set one by one. In both of these two figures, we can conclude that the performance of the proposed online learning is comparable with the non-online algorithm. As can be concluded from Table 2, the computational cost for online learning is very low, and only involves some simple calculations and a quadratic programming problem, which can be solved efficiently through the dual form.

Since for online learning the classification error may be accumulated with the increase of labeled examples, we further investigate at which point the whole set of classifiers need to be retrained. From Fig.4(b), we can see that if the online number is around 1.5 times larger than that of the originally labeled ones, then, the performance difference between the online learning and non-online learning will be large enough for considering to retrain the classifiers.

5. CONCLUSIONS

Most traditional supervised learning methods are developed to learn a model from labeled examples and use this model to classify the unlabeled ones into the same label space predefined by the models. However, in many real world applications, the label spaces for both the labeled/training and unlabeled/testing examples can be different. To solve this problem, this paper proposes a novel notion of Serendipitous Learning (SL), which is defined to address the learning scenarios in which the label space can be enlarged during the testing phase. In particular, a large margin approach is proposed to solve SL. The basic idea is to leverage the knowledge in the labeled examples to help identify novel/unknown classes, and the large margin formulation is proposed to incorporate both the classification loss on the examples within the known categories, as well as the clustering loss on the examples in unknown categories. An efficient optimization algorithm based on CCCP and the bundle method is proposed to solve the optimization problem of the large margin formulation of SL. Moreover, an efficient online learning method is proposed to address the issue of large scale data in online learning scenario, which has been shown to have a guaranteed learning regret. An extensive set of experimental results on two synthetic datasets and two datasets from real world applications demonstrate the advantages of the proposed method over several other baseline algorithms. One limitation of the proposed method is that the number of unknown classes is given in advance. It may be possible to remove this constraint if we model it by using a non-parametric way. We also plan to do experiments on more real world applications in the future.

6. ACKNOWLEDGEMENT

We would like to express our sincere thanks to Dr. Zenglin Xu (Purdue University), Prof. S.V.N. Vishwanathan (Purdue University), Feng Yan (Purdue University), and the anonymous reviewers for their valuable comments and suggestions. This research was partially supported by the NSF research grants IIS-0746830, CNS-1012208, IIS-1017837, CCF- 0939370.

7. REFERENCES

- [1] J. Allan. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers Norwel, 2002.
- [2] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *NIPS*, pages 368–374, 1998.
- [3] J. Betteridge, A. Carlson, S. Hong, and E. Hruschka Jr. Toward Never Ending Language Learning. In *AAAI*, 2009.
- [4] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.
- [5] R. Duda, P. Hart, and D. Stork. *Pattern classification*. Wiley-Interscience, 2001.
- [6] E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *COLT*, pages 499–513, 2006.
- [7] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.
- [8] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.
- [9] X. Liu, Y. Gong, W. Xu, and S. Zhu. Document clustering with cluster refinement and model selection capabilities. In *SIGIR*, pages 191–198, 2002.
- [10] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization : Algorithms and Complexity*. Dover Publications, July 1998.
- [11] D. Preston, C. E. Brodley, R. Khardon, D. Sulla-Menashe, and M. A. Friedl. Redefining class definitions using constraint-based clustering: an application to remote sensing of the earth’s surface. In *KDD*, pages 823–832, 2010.
- [12] B. Scholkopf and A. Smola. *Learning with kernels*. MIT press Cambridge, Mass, 2002.
- [13] A. Smola, S. Vishwanathan, and Q. Le. Bundle methods for machine learning. *NIPS*, 20, 2008.
- [14] Q. Tao, D. Chu, and J. Wang. Recursive support vector machines for dimensionality reduction. *IEEE Transactions on Neural Networks*, 19(1):189–193, 2008.
- [15] C. H. Teo, S. V. N. Vishwanathan, A. J. Smola, and Q. V. Le. Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, 11:311–365, 2010.
- [16] H. Valizadegan and R. Jin. Generalized Maximum Margin Clustering and Unsupervised Kernel Learning. *NIPS*, 2006.
- [17] A. S. Vishwanathan, A. J. Smola, and S. V. N. Vishwanathan. Kernel methods for missing variables. In *AISTAT*, pages 325–332, 2005.
- [18] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. *NIPS*, 17:1537–1544, 2005.
- [19] L. Xu and D. Schuurmans. Unsupervised and semi-supervised multi-class support vector machines. In *AAAI*, pages 904–910, 2005.
- [20] T. Yang, R. Jin, A. K. Jain, Y. Zhou, and W. Tong. Unsupervised transfer classification: application to text categorization. In *KDD*, pages 1159–1168, 2010.
- [21] Y. Yang, J. Z. 0003, J. G. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *KDD*, pages 688–693, 2002.
- [22] K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum margin clustering made practical. *IEEE Transactions on Neural Networks*, 20(4):583–596, 2009.
- [23] B. Zhao, F. Wang, and C. Zhang. Efficient maximum margin clustering via cutting plane algorithm. *SDM*, pages 751–762, 2008.
- [24] B. Zhao, F. Wang, and C. Zhang. Efficient multiclass maximum margin clustering. *ICML*, pages 751–762, 2008.
- [25] X. Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2006.