TSFS: A Novel Algorithm for Single View Co-training

Wen Zhang Quan Zheng
Department of Automation and Joint Lab of Network Communication System & Control
University of Science & Technology of China
Hefei, 230027, China
green@mail.ustc.edu.cn

Abstract

Co-training has been validated to be effective in various applications. However, it is a challenging task to apply co-training on the data without two independent and "good enough" views. In this paper, we propose a novel subspace feature set splitting algorithm, called Two-view Subspace Feature Splitting (TSFS), to make co-training better usable on single view data. We first project both labeled and unlabeled data into a lower dimensional subspace through Singular Value Decomposition (SVD), in which all features of data are orthogonal to each other. And then a greedy two-view feature selection strategy is proposed for feature set splitting. We introduce the energy function of each view to guarantee the quality of each split feature set. Experimental results well validated the effectiveness of TSFS in contrast to several recent studies on single view co-training.

1. Introduction

It is always highly expensive to acquire sufficient labeled data in various supervised learning tasks [2, 7, 8]. On the contrary, there may be abundant unlabeled data available for learning. The co-training framework [1], which is a semi-supervised learning framework, tries to combine the insufficient labeled data and a large number of unlabeled data to achieve better learning performance. It firstly trains two classifiers from two different views of the labeled data. And then each classifier is reinforced by the learning results in the other view. Through this way, the performance of some classifiers can be improved considerably [1, 10].

However, the co-training framework has two strong assumptions which are hard to be satisfied in real world applications [1]. In details, it requires the training data to have two conditionally independent views and each view is sufficiently good for training. However, in practice, many learning tasks have only one view of the data. This strongly limits the application of co-training framework. Fortunately, some recent studies show that even though the two assumptions are not satisfied, co-training is applicable by randomly splitting the features of training data into two views. This motivates us to explore the problem that, how to effectively split the features of the single view training data into two views which are as independent as possible and each view is sufficiently good for training? In this paper, we propose a novel subspace feature splitting algorithm, which is named as Two-view Subspace Feature Splitting (TSFS) algorithm, to address the problem of single view co-training. The TSFS algorithm consists of two key steps, (1) subspace learning; and (2) multi-view feature selection. In the first stage, we project both labeled and unlabeled data, into a lower dimensional subspace through Singular Value Decomposition (SVD), in which all the newly generated features of data are orthogonal to each other. And then in the second stage, we propose a greedy two-view feature selection strategy to split feature set into two views in the orthogonal subspace. Finally, based on the two views generated by TSFS, we apply co-training with classical classifiers such as NB or K-NN to utilize both labeled and unlabeled data for learning. Experimental results on four UCI data sets and a real text dataset show that TSFS can obtain a better performance in contrast to some recent studies.

2. Related work

Among various semi-supervised learning algorithms, co-training framework is one of the most commonly used strategies [1, 10]. The co-training was firstly proposed by Blum et al. [1] in 1998. They formally formulated it as a PAC-style learning. However, two strong assumptions exist in the proposed framework: (1) there are two naturally independent views of the training data; and (2) each view of the dataset is

978-0-7695-3605-7/09 $25.00 © 2009 IEEE
DOI 10.1109/CSO.2009.251

IEEE computer society

sufficiently good for learning. Unfortunately, these theoretical assumptions are too strong to be satisfied in practice. J.Chan et al. [6] applied co-training on single view email classification by random feature set splitting. Their results showed that the co-training with random feature set splitting outperforms classical supervised learning algorithm. Similarly, Nigam et al. [10] experimentally studied the effect of co-training on datasets that did not satisfy these assumptions. Experimental results demonstrate that co-training is possible to improve the classical supervised learning algorithms by random split. Felix et al. [3] followed the idea of Nigam and Ghani. They proposed a MI based feature splitting algorithm by minimizing the conditional mutual information between two views, which is different from our proposed TSFS. Our experiments show that co-training with TSFS achieves better results than both random splitting and mutual information based approach.

3. Two-view Subspace Feature Split

3.1. Problem formulation

In the remaining part of this paper, the data set $\mathcal{D}$ is represented by a matrix $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{m \times n}$, where each row stands for a data sample and each column denotes a feature of these data. Moreover, data in every row are treated as vectors $\boldsymbol{d}_i$ with n features, so data set $\mathcal{D}$ can also be regarded as a set of vectors $\{\boldsymbol{d}_i\ ; i=1,2,...,m\}$.

Let labeled data set be denoted by $\mathcal{L} = \{\langle \boldsymbol{d}_1, y_1 \rangle, \langle \boldsymbol{d}_2, y_2 \rangle, \dots, \langle \boldsymbol{d}_n, y_n \rangle\}$, y_i is the class label of the corresponding data $\boldsymbol{d}_i$. Meanwhile, we define unlabeled dataset as $\mathcal{U} = \{\boldsymbol{d}'_1,\ \boldsymbol{d}'_2, \dots, \boldsymbol{d}'_m\}$, where $\boldsymbol{d}'_i$ is data sample vector contained by $\mathcal{D}$. However, its label is unknown. Classical supervised learning model aims to learn a classifier c: $\mathcal{D} \rightarrow Y$ from labeled data $\mathcal{L}$, where Y is the set of labels for training data and thus the problem is to use classifier c to predict the class labels of the unlabeled data in $\mathcal{U}$.

Co-training framework was proposed for the dataset which have naturally independent views. Suppose the two views of original data set are $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$. Then, co-training propose to utilize the labeled data sets $\mathcal{L}^{(1)}, \mathcal{L}^{(2)}$ to train two classifiers c_1 and c_2 respectively. After that, each of the two classifiers is reinforced by the learning results on the unlabeled data set $\mathcal{U}^{(1)} \subseteq \mathcal{D}^{(1)}$, $\mathcal{U}^{(2)} \subseteq \mathcal{D}^{(2)}$ in the other view. Blum and Mitchell [1] have proved that co-training algorithms are feasible when certain assumptions are satisfied. The first assumption is that the features in either view are conditionally independent of the features in the other view, given the class of sample. The second one is that the quality of the two views is sufficiently high for classification. However, most data sets do not satisfy these two strong assumptions. This motivates us to find out a method which can split the data set $\mathcal{D}$ into two views $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$ which can satisfy them.

3.2. Subspace learning

To obtain two conditionally independent views, in practice, we first perform orthogonal feature extraction to transform all features into an orthogonal subspace. The orthogonal feature transformation is essentially a dimension reduction procedure, in which a function f: $\mathbb{R}^{\mathrm{n}} \rightarrow \mathbb{R}^{\mathrm{k}}$ $(\mathrm{k} < \mathrm{n})$ is employed to project the data. Hence, through f, every sample $\boldsymbol{d_i} \in \mathbb{R}^{\mathrm{n}}$ in data set is projected into a subspace $\mathbb{R}^{\mathrm{k}}$. We consider only the linear approaches in this paper due to the high time complexity of nonlinear subspace learning in learning f.

In this paper, we propose to utilize *Principle Component Analysis(PCA)* [5] to learn the orthogonal subspace. The goal of the PCA is to project the data into a subspace whose basis vectors correspond to the directions with maximal variances. Suppose $\boldsymbol{M} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1,2,\dots,\mathrm{n}}(\boldsymbol{d_i} - \overline{\boldsymbol{d}})(\boldsymbol{d_i} - \overline{\boldsymbol{d}})^{\boldsymbol{T}}$ is the covariance matrix of $\boldsymbol{\mathcal{D}}$. Then it can be proven that the row vectors of $\boldsymbol{V}$ are the k leading eigenvectors of the covariance matrix $\boldsymbol{M}$. To compute the eigenvectors of $\boldsymbol{M}$, we utilize *Singular Value Decomposition (SVD)* on the covariance matrix $\boldsymbol{M}$. Through *SVD*, we have

$$\boldsymbol{X}^{\boldsymbol{T}}\ \boldsymbol{M}\,\boldsymbol{X} = \lambda \qquad (1)$$

where $\boldsymbol{X}$ is an eigenvector of $\boldsymbol{M}$ and λ is corresponding eigenvalue. Suppose the set of eigenvalues and corresponding eigenvectors are $\boldsymbol{\mathcal{A}} = \{\lambda_1, \lambda_2, \dots, \lambda_{\mathrm{n}}\}$ and $\boldsymbol{\mathcal{B}} = \{\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_{\mathrm{n}}\}$ respectively. For every $\lambda_i \in \boldsymbol{\mathcal{A}}$ and $\mathrm{X}_i \in \boldsymbol{\mathcal{B}}$, $1 \leq i \leq n$, Eqn. (1) is satisfied. Then, the linear projection matrix $\boldsymbol{V} \in \mathbb{R}^{\mathrm{k} \times \mathrm{n}}$ is generated from $\boldsymbol{\mathcal{B}}$ by:

$$\boldsymbol{V} = \left[\boldsymbol{X}_{\boldsymbol{i}_1}\,\boldsymbol{X}_{\boldsymbol{i}_2} \cdots \boldsymbol{X}_{\boldsymbol{i}_{\boldsymbol{k}}}\right]^{\mathbf{T}} \qquad (2)$$

where $\boldsymbol{X}_{\boldsymbol{i}_{\boldsymbol{t}}} \in \boldsymbol{\mathcal{B}}$ are k leading eigenvectors in $\boldsymbol{\mathcal{B}}$, $1 \leq \mathrm{t} \leq \mathrm{k}$. For every $\boldsymbol{d_i}$ in data set $\boldsymbol{\mathcal{D}}$, $\boldsymbol{V}$ can be employed to project it into a subspace $\mathbb{R}^{\mathbf{k}}$. We finally apply PCA to construct a projection T_{PCA} on data set $\mathcal{D}$: T_{PCA} ($\mathcal{D}$) = $\boldsymbol{\mathcal{D}}\boldsymbol{V}^{\mathbf{T}}$. Note that all features in $\widetilde{\boldsymbol{\mathcal{D}}}$ =T_{PCA} ($\boldsymbol{\mathcal{D}}$) are orthogonal to each other. (i.e., linear independent, is considered as near independent)

There is an one to one mapping between features in subspace $\mathbb{R}^{\mathrm{k}}$ and eigenvectors in $\boldsymbol{\mathcal{B}}$ or eigenvalues in $\boldsymbol{\mathcal{A}}$. Thus, in this paper, we define $Q(\lambda_i)$ as the feature corresponding to eigenvalue $\lambda_i \in \boldsymbol{\mathcal{A}}$, while $Q^{-1}(\boldsymbol{e})$ can also be used to represented the eigenvalue corresponding to feature $\boldsymbol{e}$.

3.3. Two-view feature selection

After projecting the data into orthogonal subspace, we aim to find out two high quality views of the data in the orthogonal space. In this section, we try to utilize eigenvalues of features to guide us in splitting the feature set.

Definition 1: A **division** P on feature set $\mathcal{F} = \{f_1, f_2, \dots f_n\}$ is a division that split $\mathcal{F}$ into two sub feature sets $\mathcal{F}_1 \subseteq \mathcal{F}$ and $\mathcal{F}_2 \subseteq \mathcal{F}$ which satisfy $\mathcal{F}_1 \cup \mathcal{F}_2 = \mathcal{F}$

For a given data set matrix $\widetilde{\boldsymbol{\mathcal{D}}}$ projected by T_{PCA}, Let $\mathcal{F}_{\widetilde{\boldsymbol{\mathcal{D}}}}$ be its features set. We try to obtain a division P^* on $\mathcal{F}_{\widetilde{\boldsymbol{\mathcal{D}}}}$ to make a certain target function F(P) = F($\mathcal{F}_1$, $\mathcal{F}_2$) reach a near optimal value. Since features of $\widetilde{\boldsymbol{\mathcal{D}}}$ is as independent as possible with each other, we can focus on searching a division P^* on $\mathcal{F}_{\widetilde{\boldsymbol{\mathcal{D}}}}$ that causes each features set to be sufficient for classification. Hence, the target function only needs to measure the degree of quality of the two-views.

According to the PCA[5], the ratio of eigenvalues can reflect how much information of the original dataset can be maintained by only reserve part of the features. Hence, we introduce energy function $\mathrm{E}(\cdot)$ to measure significance of feature set $\mathcal{S}$ in $\widetilde{\mathcal{D}}$:

Definition 2: The significance of feature set $\boldsymbol{\mathcal{S}} \subseteq \mathcal{F}_{\widetilde{\mathbf{D}}}$ in data set $\widetilde{\boldsymbol{\mathcal{D}}}$ generated from original data set $\boldsymbol{\mathcal{D}}$ by T_{PCA} is defined as below:

$$E(\mathcal{S}) = \frac{\sum_{\boldsymbol{\lambda}_k \in \mathcal{Z}} \boldsymbol{\lambda}_k}{\sum_{\boldsymbol{\lambda}_i \in \mathcal{A}} \boldsymbol{\lambda}_i} \tag{3}$$

where $\mathcal{Z} = \{\lambda = Q^{-1}(s) | s \in \mathcal{S}\}$ and $\mathcal{A} = \{\lambda = Q^{-1}(s) | s \in \mathcal{F}_{\widetilde{\mathcal{D}}}\}$. It is clear that, along with the increase of $\mathrm{E}(\mathcal{S})$, the significance of feature set $\mathcal{S}$ increases. Let $\mathcal{P}_{\mathcal{F}} = \left\{\mathrm{P}_{\mathcal{F}}^{(1)}, \mathrm{P}_{\mathcal{F}}^{(2)}, \dots, \mathrm{P}_{\mathcal{F}}^{(u)}\right\}$ be one set of divisions on feature set $\mathcal{F}$, for every $\mathrm{P}_{\mathcal{F}}^{(i)} \in \mathcal{P}_{\mathcal{F}}$, two subsets of $\mathcal{F}$ are obtained: $\mathcal{S}_i$ and $\mathcal{F} - \mathcal{S}_i$. Because of the orthogonality between features in $\mathcal{F}$, the two disjoint feature sets can be regarded as independent of each other. In such a case, we only need to find a division P^* which makes the two-views be sufficiently good for classification.

Definition 3: we define energy diversity function in division $\mathrm{P}_{\mathcal{F}}^{(i)}$:

$$Div\left(P_{\mathcal{F}}^{(i)}\right) = |E(\mathcal{S}_i) - E(\mathcal{F} - \mathcal{S}_i)| \tag{4}$$

We can transform Eqn. (4) as:

$$\begin{aligned} Div\left(P_{\mathcal{F}}^{(i)}\right) &= |E(\mathcal{S}_i) - E(\mathcal{F} - \mathcal{S}_i)| \\ &= \left|\frac{\sum_{\boldsymbol{\lambda}_k \in \mathcal{Z}_1} \boldsymbol{\lambda}_k - \sum_{\boldsymbol{\lambda}_k \in \mathcal{Z}_2} \boldsymbol{\lambda}_k}{\sum_{\boldsymbol{\lambda}_i \in \mathcal{A}} \boldsymbol{\lambda}_i}\right| \\ &= \frac{1}{A}\left|\sum_{\boldsymbol{\lambda}_k \in \mathcal{Z}_1} \boldsymbol{\lambda}_k - \sum_{\boldsymbol{\lambda}_k \in \mathcal{Z}_2} \boldsymbol{\lambda}_k\right| \end{aligned} \tag{5}$$

where A is utilized to represent $\sum_{\boldsymbol{\lambda}_i \in \mathcal{A}} \boldsymbol{\lambda}_i$. To gain two "good enough" views, we hope that the difference between the two disjoint views' energies is as small as possible because the sum of energy is definite. In other words, an optimal division P^* that satisfies below equation is desired:

$$P^* = argmin_{P_{\mathcal{F}}^{(i)} \in \mathcal{P}_{\mathcal{F}}} \left|\sum_{\boldsymbol{\lambda}_k \in \mathcal{Z}_1} \boldsymbol{\lambda}_k - \sum_{\boldsymbol{\lambda}_k \in \mathcal{Z}_2} \boldsymbol{\lambda}_k\right| \tag{6}$$

subject to the constraint:

$$\begin{cases} E(\mathcal{S}_1) \geq \tau \\ E(\mathcal{F} - \mathcal{S}_1) \geq \tau \end{cases} \tag{7}$$

where τ is the threshold of each view's energy and $\mathcal{F}_1$, $\mathcal{F} - \mathcal{F}_1$ are two disjoint sub feature sets generated by a division on $\mathcal{F}$, we try to find an optimal division P^* under the constraint. In practice, we cannot guarantee it to be satisfied if we want the two views to be fully independent (i.e. strictly disjoint). So, we balancing the independence and view quality by letting the two views share some common features that correspond to a high eigenvalues such that we maximize the views' quality through sharing the minimal number of features.

In this paper, we apply a greedy strategy to obtain an approximate optimal division $\hat{P}^*$. First, the eigenvalues are sorted in a descending order. Then two-views share features corresponding to top s eigenvalues. Begin with (s+1) *th* feature, we put odd features into one view and even features into another view. Algorithm 1 outlines this process. Through it, the feature set $\mathcal{F}_{\widetilde{\mathbf{D}}}$ is divided into two subsets: $\mathcal{F}_1$ and $\mathcal{F}_2$. According to Eqn.(2), We can generate two linear projection matrix $\boldsymbol{V}_1$, $\boldsymbol{V}_2$ based on $\hat{\mathcal{S}}^*$, $\hat{\mathcal{T}}^*$. For every $\boldsymbol{d}_i \in \mathcal{D}$, two vectors $\tilde{d}_i^{(1)}$and $\tilde{d}_i^{(2)}$are obtained. They each represents value vector of feature set $\mathcal{F}_1$ and $\mathcal{F}_2$, forming two views of data $\tilde{\boldsymbol{d}}_i = T_{PCA}$ ($\boldsymbol{d}_i$). Finally, we can apply co-training algorithm on the two views to train two classifiers c_1, c_2.

```
Algorithm 1
Obtain an approximate optimal division approximate_optimal_div( )
Input: eigenvalues set $\boldsymbol{\mathcal{A}} = \{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \dots, \boldsymbol{\lambda}_n\}$
    k: the number of leading eigenvalues.
    s: the number of shared features.
Procedure:
    1. eigenvalue set $\mathcal{A}$ are sorted descending, then k leading
       eigenvalues are selected by order: $\boldsymbol{\mathcal{A}}' = \{\boldsymbol{\lambda}_{i_1}, \boldsymbol{\lambda}_{i_2}, \dots, \boldsymbol{\lambda}_{i_k}\}$
    2. $p \leftarrow 1, T \leftarrow |\boldsymbol{\mathcal{A}}'|, \hat{\mathcal{S}}^* \leftarrow \Omega, \hat{\mathcal{T}}^* \leftarrow \Omega$
    3. if ( $p > T$ ) exit
    4. if ( $p \leq s$ ) $\hat{\mathcal{S}}^* = \hat{\mathcal{S}}^* \cup \boldsymbol{i_p}$ , $\hat{\mathcal{T}}^* = \hat{\mathcal{T}}^* \cup \boldsymbol{i_p}$,
       go to (6)
    5. if ($p$ % 2 == 0) $\hat{\mathcal{S}}^* = \hat{\mathcal{S}}^* \cup \boldsymbol{i_p}$
       else $\hat{\mathcal{T}}^* = \hat{\mathcal{T}}^* \cup \boldsymbol{i_p}$
    6. $p$++, go to (3)
Output: $\hat{\mathcal{S}}^*$, $\hat{\mathcal{T}}^*$
```

4. Experiments

4.1. Data sets

To validate the effect of TSFS, We use three of the UCI benchmark data sets: *Australia, ionosphere, magic gamma telescope* and a semi-artificial dataset constructed from *20newsgroup* [9] for our experiment. Similar to previous works [1, 10], the ratio of positive samples to negative ones in $\mathcal{L}$**,** is equal to the ratio in the entire data set. Moreover, 70% of the samples in $\mathcal{U}$ are randomly selected as unlabeled data, the remaining ones are used for testing. The experiments are repeated 10 times and the reported results are averaged. We try various numbers of overlapped features between two views in TSFS algorithm, which correspond to s in algorithm 1. Then we select s which satisfies constraint (7) from these values. In this paper, we let τ in constraint (7) be 0.5.

To compare TSFS with truly independent split and *maxInd*[6] split, we use the News2×2 dataset proposed by Nigam et al. [10]. We select 4 newsgroups from the 20newsgroups dataset as shown in Table 1. The News2×2 dataset was configured like this: randomly selecting documents from newsgroups 1 and 2 to make positive ones, from newsgroups 3 and 4 to make negative ones. This joining is done in such a way that the words in the first and third newsgroups come from the same vocabulary, while the words in the second and fourth newsgroups come from another vocabulary. Apparently, the News2×2 dataset has a natural feature split. Then, feature selection is applied on the new dataset to choose the important features. Top 200 words for each split with high document frequency are selected. Then each document is represented with the TF-IDF weights of the selected features [4].

Table 1 The News2×2 dataset

Class	Feature Set A	Feature Set B
Pos	1.comp.os.ms.windows.misc	2.talk.politics.misc
Neg	3.comp.sys.ibm.pc.hardware	4.talk.politics.guns

4.2. Experimental results

4.2.1 UCI Data Set. On the UCI Data Set, we show the comparison of the performance of two base learning algorithms (NB and KNN) as well as the performance when we apply the classical co-training with these base learning algorithms, and the performance of TSFS. In Fig.1 we show the experimental results on the three UCI benchmark data sets: *Australia, ionosphere,* and *magic gamma telescope* respectively. In these experiments, the co-training proceeds identically as in Blum et al.[1] except that we run co-training until it gives labels to all the unlabeled samples. On these data sets, we define distance metric as Euclidean distance for KNN.

In Fig.1, x-coordinate represents the number of feature shared between two-views in TSFS algorithm, y-coordinate represents the accuracy of algorithms and the real values near the points in TSFS curve represent the energy of features shared. Apparently the results of KNN and random split (CT-random) are straight lines in Fig.1. Fig.1(a), (b) and (c) are the experiment results when we utilize NB as base classifier, meanwhile Fig.1(d), (e) and (f) are the results of experiment utilizing KNN as base classifier. From the six figures, we notice that when the energy of features shared reaches around 0.5, TSFS achieves its best results: it obtain 16.4%, 5.3%, 12.3%, 6.2%, 4.6%, 3.2% higher accuracy than the second best algorithms respectively. One explanation is that a trade-off between the independence of two views and energy in each view is satisfied by this time. We call the overlap at this time as a **trade-off overlap**. In addition, co-training with random split is even not better than the supervised classifier sometimes (Fig.1 (a), (c), (e) and (f)). A possible reason is that the two-views generated by random split are hard to satisfy the two assumptions in co-training setting, such that it is hard for it to benefit from the procedure of co-training.

4.3.2 News2×2 dataset. We also conduct experiments on News2×2 dataset. We applied 5 algorithms on the News2×2 dataset: co-training with truly independent split (CT-Ind), co-training with *maxInd* (CT-*maxInd*), co-training with random split (CT-random), co-training with TSFS, KNN. Among them, co-training with *maxInd* was introduced before, and achieved good performance on the same dataset (**News2×2**). Due to space limitation, all the co-training algorithms above only utilize KNN as base classifiers. For KNN we utilize cosine similarity to scale distance of two samples. According to preceding experimental results, to achieve the best performance of TSFS, we select a suitable value of overlap which ensures that the energy of features shared reaches around 0.5.

Table 2 Experimental results on News2×2 dataset

Algorithm	TSFS	CT-random	CT-maxInd	CT-Ind	KNN
Accuracy	**0.875**	0.830	0.825	**0.892**	0.684

Experimental results are showed in Table 2. The accuracy rate of TSFS is 5.4%, 5.9%, 25.9% higher than the random split, *maxInd* and KNN respectively, and only 1.3% lower than truly independent split which is the ideal split in co-training.

5. Conclusion and future work

In this paper, we propose a novel feature set splitting method for single view co-training problem, called TSFS. Its general idea is to project the original data set into a subspace in which all features are orthogonal to each other, then apply a greedy two-view feature selection strategy on the subspace data set to gain two high quality views. For measuring the quality of each view, we introduce an energy function of view based on the eigenvalues corresponding to the features in this view. To validate the effectiveness of TSFS, we apply it on several real world datasets. In our experiments, we compared TSFS with some state of the art approaches on the three UCI datasets and news2×2 dataset. It can be seen from the results that the TSFS with **trade-off overlap** can achieve the best performance among the algorithms when there is no natural split on feature set. In the future, we will theoretically study the trade-off between the independence of two views and energy in each view.

Acknowledgement

This work is supported by Hi-Tech Research and Development Program of China (863 program, No.2008AA01Z147)

References

[1] A.Blum, T.Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, Madison, WI, USA, 1998.

[2] D.Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, Massachussets, USA, 1995.

[3] F.Feger, I.Koprinska. Co-training Using RBF Nets and Different Feature Splits. In *Proceeding of International Joint Conference on Neural Networks*, Vancouver, BC, Canada, 2007.

[4] H.Park, M.Jeon and J. Rosen. Lower Dimensional Representation of Text Data Based on Centroids and Least Squares. *BIT Numerical Math*. Vol.43, 427-448, 2003.

[5] I.T.Jolliffe. *Principal Component Analysis*. Springer-Velag, New York, 1986.

[6] J.Chan, I.Koprinska and J.Poon. Co-training with a Single Natural Feature Set Applied to Email Classification. In *proceedings of International Conference on Web Intelligence*. Beijing, China, 2004.

[7] J.Han, M.Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2001.

[8] K.Hiraoka, K.Hidai, M.Hamahira, H.Mizoguchi, T. Mishima and S.Yoshizawa. Successive Learning of Linear Discriminant Analysis: Sanger-Type Algorithm. In *Proceedings of 14th International Conference on Pattern Recognition*. 2000.

[9] K.Lang. NewsWeeder: Learning to Filter Netnews. In *Proceedings of 12th International Conference on Machine Learning*. Tahone City, California, USA, 1995.

[10] K.Nigam, R.Ghani. Analyzing the Effectiveness and Applicability of Co-Training. In *Proceeding of the 9th International Conference on Information and Knowledge Management*. McLean, Virginia, USA, 2000

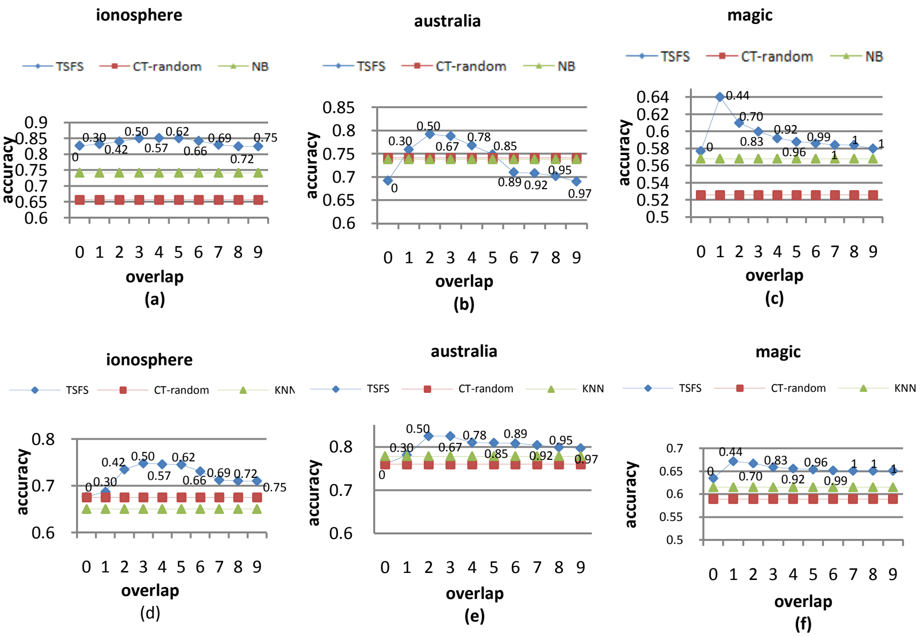

Figure 1. TSFS with different feature overlap on the UCI data set