Signal Processing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Multi-task support vector machines for feature selection with shared knowledge discovery

Sen Wang [a], Xiaojun Chang [a,*], Xue Li [a], Quan Z. Sheng [b], Weitong Chen [a]

[a] The University of Queensland, Australia
[b] The University of Adelaide, Australia

ARTICLE INFO

Article history:
Received 18 August 2014
Received in revised form
9 December 2014
Accepted 10 December 2014

Keywords:
Feature selection
Multi-task learning
Trace norm
Low-rank

ABSTRACT

Feature selection is an effective way to reduce computational cost and improve feature quality for the large-scale multimedia analysis system. In this paper, we propose a novel feature selection method in which the hinge loss function with a $\ell_{2,1}-$norm regularization term is used to learn a sparse feature selection matrix for each learning task. Meanwhile, shared information exploiting across multiple tasks has been also taken into account by imposing a constraint which globally limits the combined feature selection matrices to be low-rank. A convex optimization method is proposed to use in the framework by minimizing the trace norm of a matrix instead of minimizing the rank of a matrix directly. Afterwards, gradient descent is applied to find the global optimum. Extensive experiments have been conducted across eight datasets for different multimedia applications, including action recognition, face recognition, object recognition and scene recognition. Experimental results demonstrate that the proposed method performs better than other compared approaches. Especially, when the shared information across multiple tasks is very beneficial to the multi-task learning, obvious improvements can be observed.

1. Introduction

With the explosion of multimedia data in our daily life, effective and efficient ways managing such big scale data have attracted much research attention over the past few years. Thus, a mechanism, which is able to facilitate indexing, ranking and classification, is in great demand. Most of the existing approaches which improve the system performance, from a hierarchical perspective, can be categorized into two groups. The first group of methods focuses on designing new machine learning algorithms for different applications at a higher level. In contrast, crafting and selecting distinctive features can obtain significant improvement of the performance at a lower level in a system. Comparing with crafting specific features, such as SIFT, STIP, and MFCC, feature selection chooses a subset of variables from the raw features aiming to describe the input data more productively than the original ones. This is achieved by removing noise or irrelevant variables resulting in satisfied system performance, such as higher prediction accuracy.

In the literature, feature selection models can be divided into three major categories: filter methods, wrapper methods and embedded methods. In filter methods [1], all features are firstly ranked and those features with higher ranked scores are then selected as the distinctive features before feeding them to the classifier. In this manner, selection and prediction are separate which means that the feature selection is independent of the classifier. The features which have no influences on the class information should be removed as irrelevant features. In the light of this principal, a suitable ranking criterion is applied to evaluate relevance of the variables to the class labels by comparing with a threshold.

In wrapper methods [2,3], the classifier is bound up with a specific search algorithm aiming to select a subset of

* Corresponding author. Tel.: +61 4 26960518; Fax: +61 7 3365 4999.
E-mail address: x.chang@uq.edu.au (X. Chang).

http://dx.doi.org/10.1016/j.sigpro.2014.12.012
0165-1684/

features yielding the best prediction results. For each subset of the original features, prediction scores can be regarded as evaluation outcomes. The best subset is consequently obtained by optimizing the objective function. In embedded methods [4–7], variable selection is directly integrated into the training process. For instance, Neumann et al. [8] investigate a few SVM-based feature selection methods which are similar with our proposed method in this paper.

During the last decades, feature selection has been pervasively applied in a number of applications due to its advantages mentioned above. In bioinformatics, feature selection has been used to remove redundancy and noise from raw data. In [9], a wrapper-based feature selection algorithm and the SVM classifier have been applied to discover discriminative features and predict the functional classes of proteins. Conilione and Wang [10] perform a correlation-based feature selection method to recognize promoter regions. Also, in [11], features are ranked and selected to improve the performance of the SVM classifier-based system, miTarget, which predicts miRNA target gene. In [12], a semi-supervised feature selection algorithm has been proposed for video and image analysis. Besides, feature selection has been also utilized in other research areas. In the field of multimedia analysis, multimedia data, i.e. images and videos, often employ high dimensional representation models. For example, Bag-of-Visual Word (BoVW) model and its extension (spatial and temporal pyramid BoVW model) are very popular for encoding features in image and video analysis systems over the past few years [13–17]. These representation methods have huge number of features causing heavy computational cost. To improve efficiency and effectiveness, many works [18–21] have paid great attention to tackle this problem.

In this paper, we focus on feature selection when the labeled data are few. Generally speaking, there are two ways to deal with the small number of labeled data. The first one is known as semi-supervised learning. As manifold structure has shown to be effective for data analysis, e.g., retrieval [22] and recognition [23–25], one possible direction is to leverage the local geometry for feature selection [26,27]. In addition, transfer learning and multi-task learning are able to utilize knowledge from related tasks, which are particularly suitable for the cases when the labeled data are few [28–30]. Although multi-task learning and shared structural analysis have attracted much research attention, feature selection using multi-task learning while exploiting shared knowledge across multiple tasks has been largely ignored [31]. It is intuitive that learners absorb new knowledge more quickly and accurately if already known experience which is relevant to the new learning task can be taken into account. In [32–34], learning relatedness of tasks simultaneously is preferable to independently learning each task. Particularly, when there are only a few labeled data per task, discovering shared knowledge across multiple tasks can significantly improve the performance for multimedia analysis [29]. Similarly, when labeled data are scarce, feature selection can also benefit from relatedness learning across multiple learning tasks. In fact, most of the existing feature selection algorithms [35–38] only focus on uncovering distinctive subsets of features for each independent task. Shared information among multiple tasks is rarely considered when selecting features. In [39], a discriminative model that exploits shared structures across multiple tasks is proposed to select features. In this paper, we propose a multi-task SVM-based feature selection algorithm. The main contributions of this work are as follows:

- The proposed method efficiently selects discriminative features for each task by restricting feature selection matrix for each task to be sparse. More specifically, for the l-th learning task, $\ell_{2,1}-$norm penalty term makes the feature selection matrix be sparse in rows. A small number of non-zero rows will be selected as the most common and distinctive features across different classes. Correlations among different classes are thus exploited which might be beneficial to perform feature selection for each learning task.
- Relatedness of features among multiple tasks has been also taken into account by simultaneously mining the shared structure across all tasks. A global constraint is imposed to a feature selection matrix, which combines feature selection matrices from all tasks, to be low-rank. In this way, common patterns or shared information across multiple tasks can be discovered.
- Instead of directly minimizing the rank of a matrix, which is a solution to shared information discovery but non-convex, we propose to minimize trace norm of a matrix, which is a convex optimization problem. By leveraging each task with the entire learning objective, our proposed work is superior to all the compared methods according to the experiment results. Notably, our method outperforms single task feature selection counterparts when labeled training samples for each task are limited.

The rest of paper is organized as follows: Section 2 elaborates our proposed method followed by a brief illustration of optimization in Section 3. Experimental setup and results will be demonstrated and reported in Section 4. Section 5 will conclude this paper.

2. Multi-task feature selection with shared knowledge exploiting

In a problem of multi-task learning, given t tasks and c classes, training samples for each task are denoted as $\{(x_l^i, y_l^i)\}|_{i=1}^{n_l}$, $y_l^i \in \{1, \ldots, c_l\}$. n_l and c_l are the number of training samples and the number of classes in the l-th task, respectively. Feature selection for each task is equal to find a set of feature selection functions $\{f_l\}|_{l=1}^{t}$. From the perspective of supervision, supervised and unsupervised feature selection algorithms and related variants [40–43] have flourished in the past decade. As demonstrated in [19,44,37], exploiting correlations among features using sparsity-based methods are beneficial to feature selection. In [19,37,45], $\ell_{2,1}-$norm regularization is used as an effective model to uncover feature correlations for a single task learning problem. $\ell_{2,1}-$norm of an arbitrary matrix $M \in \mathbb{R}^{d \times c}$ is defined as

$$\|M\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} M_{ij}^2} \tag{1}$$

Comparing with some sparsity-based norm functions [46], the $\ell_{2,1}-$norm, from the definition, computes the ℓ_2-norm of

Please cite this article as: S. Wang, et al., Multi-task support vector machines for feature selection with shared knowledge discovery, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2014.12.012

each row, $\|M_{i*}\|_2$, followed by the summation of the resulting vectors, $\sum_{i=1}^{d}\|M_{i*}\|_2$, which is actually the ℓ_1-norm of the results. By minimizing the $\ell_{2,1}-$norm regularizer, entries in each row are similar which favors uniformity, while some rows are shrinking to zero rows. In this way, the most discriminative features can be selected by choosing those non-zero rows, if M is used as a feature coefficient matrix. In a problem of multi-task feature selection, the framework can be written as

$$\min_{f_l}\sum_{l=1}^{t}\left(\sum_{i=1}^{m_l}\mathrm{loss}(f_l(x_l^i),y_l^i)+\alpha g(\cdot)\right), \tag{2}$$

where $f_l(x_l)=w_l^T\cdot x_l$ and $g(\cdot)=\|w_l\|_{2,1}$ is the $\ell_{2,1}-$norm regularizer selecting the most distinctive features for the l-th task by leveraging the sparsity of each feature coefficient matrix with the promising performance. α is the regularization parameter. Note that $w_l=[w_0,w_1,\ldots,w_d]^T\in\mathbb{R}^{d+1}$ and $x_l=[1,x_1,\ldots,x_d]^T\in\mathbb{R}^{d+1}$. The biased term for $f_l(x_l)$ is therefore w_0.

As mentioned before, it is assumed that if multiple tasks contain shared knowledge from which can be borrowed, it might be beneficial to a single task learning. In other words, exploiting shared information among multiple tasks is helpful to improve the performance of multi-task learning. In [32,47], relatedness across multiple tasks is exploited by learning a shared low-dimensional subspace. Refs. [28,31,48] have demonstrated that shared information across multiple tasks can be uncovered by imposing a low-rank constraint to the matrix $W=[W_1,\ldots,W_t]$ which is stacked by feature selection matrix of each task all together. $W_l|_{l=1}^t=[w_{l1},\ldots,w_{lc_l}]\in\mathbb{R}^{(d+1)\times c_l}$. Similarly, low-rank constraint has been also applied to uncovering shared information between positive and related training data in [49]. Inspired by these works, we propose our framework for multi-task feature selection as

$$\min_{W_l}\sum_{l=1}^{t}\left(\mathrm{loss}(W_l^TX_l,Y_l)+\alpha\|W_l\|_{2,1}\right)+\beta\,\mathrm{Rank}(W), \tag{3}$$

Because minimizing a rank of a matrix is non-convex, we propose to minimize the trace norm of W which is a convex optimization problem. For an arbitrary matrix A, its trace norm is defined as

$$\|A\|_*=\mathrm{Tr}\left((AA^T)^{1/2}\right) \tag{4}$$

$\mathrm{Tr}(\cdot)$ is the trace operation. To step further, we propose to use hinge loss function, $\max(0,1-y\cdot w^Tx)$, in the objective function. Note that comparisons and discussions on various loss functions are beyond the scope of this paper. In this paper, we just use hinge loss due to its simplicity. After integrating the hinge loss and trace norm regularizer in (4), the objective function is re-written as

$$\min_{w_{lk}}\left(\sum_{l=1}^{t}\left(\sum_{k=1}^{c_l}\sum_{i=1}^{n_k}(1-(w_{lk}^T\cdot x_{lk}^{(i)})y_{lk}^{(i)})_+\right)+\alpha\|W_l\|_{2,1}\right)+\beta\|W\|_*\Big), \tag{5}$$

where operator $(\boldsymbol{z})_+=\mathbf{max}(0,\boldsymbol{z})$. $W=[W_1,\ldots,W_t]$ and $\|\cdot\|_*$ is the trace norm. c_l and n_k is the number of classes for the l-th task and the number of training samples with respect to the k-th class. α and β are regularization parameters. This objective function is convex and can reach the global optimum. We name this algorithm as *M*ulti-task *F*eature *S*election with *S*hared *K*nowledge *D*iscovery (*MFSSKD*).

3. Optimization

In this section, we will elaborate the corresponding optimization and algorithm of the framework. According to [31], the objective function in (5) can be re-formulated as

$$\min_{w_{lk}}\left(\sum_{l=1}^{t}\sum_{k=1}^{c_l}\sum_{i=1}^{n_k}(1-(w_{lk}^Tx_{lk}^{(i)})y_{lk}^{(i)})_+ +\alpha\sum_{l=1}^{t}\mathrm{Tr}(W_l^TD_lW_l)+\beta\mathrm{Tr}(W^TDW)\right), \tag{6}$$

D_l is a diagonal matrix which is defined as

$$D_l=\begin{bmatrix}\frac{1}{2\|z_l^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|z_l^d\|_2}\end{bmatrix} \tag{7}$$

and D is equal to

$$D=\frac{1}{2}(WW^T)^{-1/2} \tag{8}$$

Because $W_l=[w_{l1},\ldots,w_{lc_l}]$ and $W=[W_1,\ldots,W_t]$, the second term in (6) can be formulated as

$$\mathrm{Tr}(W_l^TD_lW_l)=\sum_{k=1}^{c_l}w_{lk}^TD_lw_{lk} \tag{9}$$

Similarly, the last term in (6) can be formulated as

$$\mathrm{Tr}(W^TDW)=\sum_{l=1}^{t}\sum_{k=1}^{c_l}w_{lk}^TDw_{lk} \tag{10}$$

Taking (9) and (10) into (6), the objective function arrives at

$$\min_{w_{lk}}J(w_{lk})=\min_{w_{lk}}\sum_{l=1}^{t}\sum_{k=1}^{c_l}\left(\sum_{i=1}^{n_k}(1-w_{lk}^Tx_{lk}^{(i)}y_{lk}^{(i)})_+ +w_{lk}^T(\alpha D_l+\beta D)w_{lk}\right) \tag{11}$$

Since we use hinge loss in (11), we split the entire cost function into two parts:

$$J(w_{lk})=J_1(w_{lk})+J_2(w_{lk}),$$

where $J_1(w_{lk})=\sum_{l=1}^{t}\sum_{k=1}^{c_l}\sum_{i=1}^{n_k}(1-w_{lk}^Tx_{lk}^{(i)}y_{lk}^{(i)})_+$ and $J_2(w_{lk})=\sum_{l=1}^{t}\sum_{k=1}^{c_l}w_{lk}^T(\alpha D_l+\beta D)w_{lk}$. By setting derivative of (11) with respect to w_{lk}, we have

$$\begin{aligned}\frac{\partial J(w_{lk})}{\partial w_{lk}}&=\frac{\partial J_1(w_{lk})}{\partial w_{lk}}+\frac{\partial J_2(w_{lk})}{\partial w_{lk}}\\ &=\sum_{i=1}^{n_k'}-x_{lk}^{(i)}y_{lk}^{(i)}+2(\alpha D_l+\beta D)w_{lk},\end{aligned} \tag{12}$$

where n_k' is the number of training samples, for the k-th class, which satisfy with $w_{lk}^Tx_{lk}^{(i)}y_{lk}^{(i)}<1$, $i=1,\ldots,n_k$ and $n_k'\le n_k$. Consequently, we use gradient descent in Algorithm 1 to iteratively update w_{lk} until convergence of objective function has reached. The most time consuming operation is the calculation of D according to (8). Thus, computational complexity of Algorithm 1 is $O(d^3)$, where d is the feature dimensionality.

Algorithm 1. *M*ulti-task *F*eature *S*election with *S*hared *K*nowledge *D*iscovery.

Input:

The l-th task training data from c_l classes: $X_l|_{l=1}^t$;

Please cite this article as: S. Wang, et al., Multi-task support vector machines for feature selection with shared knowledge discovery, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2014.12.012

The l-th task training data labels $Y_l|_{l=1}^{t}$;
Parameters α, β.

Output:

Feature selection matrix $W_l|_{l=1}^{t}$.

Initialize $W_l=[w_{l1},\ldots,w_{lc_l}]\in\mathbb{R}^{(d+1)\times c_l}$ randomly;
Compute $W=[W_1,\ldots,W_t]\in\mathbb{R}^{(d+1)\times\sum_{i=1}^{t}c_i}$;
Compute D_l of W_l according to (7);
Compute D of W according to (8);
repeat
  for $l\leftarrow 1$ **to** t **do**
    for $k\leftarrow 1:c_l$ **do**
      repeat
        Compute $w_{lk}^{r+1}=w_{lk}^{r}-\lambda\frac{\partial J(w_{lk})}{\partial w_{lk}}$ according to (12)
      Update W_l, W, D_l and D
      until *Convergence*;
until *Convergence*;
Return $W_l|_{l=1}^{t}$.

4. Experiments

In this section, we will demonstrate our experimental settings followed by a brief introduction of datasets and compared methods in this paper. Consequently, the experimental results will be reported, including the overall performance comparison, studies of parameter sensitivity and convergence demonstrations of Algorithm 1.

4.1. Experimental settings

To evaluate our algorithm, we have conducted extensive experiments and made the comparisons on a number of datasets regarding different multimedia applications:

- *ADL-RFID* [50]: It contains 10 housekeeping activities (vacuuming, ironing, dusting, brooming, mopping, cleaning windows, making bed, watering plants, washing dishes, and setting the table). The total length of the dataset is 240 min. The data was recorded by 12 subjects in a controlled lab environment. Only RFID tag data are used in this experiment. There are 191 objects in the lab that are associated with RFID tags.
- *COIL-20*: It contains 1440 gray-scale images of 20 objects (72 images per object) under various poses. The objects are rotated through 360° and taken at the interval of 5 degrees.
- *COREL-5K*: This dataset consists of 5000 images categorized into 50 classes, such as *beach*, *bird*, *jewelry*, *and sunset*, and each class has 100 images.
- *HumanEva* [51]: We downsample 10,000 3D motion/pose data for two objects from HumanEva motion dataset in this experiment as the same as in [52]. Each object has 5000 samples in five action classes: *boxing, gesturing, jogging, throw-catch* and *walking*. In this dataset, we consider each action performed by each person as an independent action learning task.
- *KTH actions dataset*: KTH dataset records 6 categories of actions: *walking*, *jogging*, *running*, *boxing*, *hand-waving* and *hand-clapping*. Each action is performed by 25 subjects under 4 different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. In total, KTH contains 599 video clips (2391 sequences) with the resolution of 160×120 pixels.
- *MIML dataset* [53]: It consists of 2000 natural scene images belonging to the classes *desert*, *mountains*, *sea*, *sunset*, *and trees*. Images are collected from both the COREL image collection and Internet. Meanwhile there are over 22% images with multi-class labels.
- *SCENE* [54]: This dataset consists of 2400 natural scene images which can be categorized into six primary classes: *Beach*, *Sunset*, *Fall foliage*, *Field*, *Mountain* and *Urban*.
- *UMIST*: The UMIST, which is also known as the Sheffield Face Database, consists of 564 images of 20 individuals. Each individual is shown in a variety of poses from profile to frontal views.

For those image-based datasets (COIL20, COREL5K, MIML, SCENE and UMIST), raw pixels have been used as input features. For those video datasets (KTH, HumanEva), a number of man-crafted features have been proposed. However, we only use the most typical feature to evaluate the performance. For example, STIP feature has been extracted from KTH data. Then, a dictionary with 1000 visual words has been learned from a sampled STIP feature subset by KMeans clustering. The video data in KTH dataset are then represented as 1000 dimensional feature vectors by bag-of-words model. Table 1 shows details of datasets. We compare our proposed method with the following methods:

1. *All variables*: All original variables are preserved as the baseline in the experiments.
2. *Max variance*: Features are ranked according to the variance magnitude of each feature in a descending

Table 1
Dataset description.

Datasets	Feature dimension	Number of classes	Number of samples
ADL-RFID	191	10	61
COIL-20	1024	20	1440
Corel5K	423	50	5000
HumanEva	168	10	10,000
KTH	1000	6	2391
MIML	423	5	2000
UMIST	644	20	575
SCENE	294	6	2407

Table 2
Performance comparison (MAP %) across all datasets.

Datasets	Ours	All features	Max variance	Fisher score	ST-SVM	FSSI
ADL-RFID	**29.85**	23.49	23.51	27.34	24.41	28.18
COIL-20	**94.85**	94.55	94.51	94.34	93.41	94.18
Corel5K	**21.91**	21.67	21.63	21.58	19.88	21.78
HumanEva	**91.79**	89.11	88.66	90.53	88.80	91.47
KTH	**76.06**	74.91	74.87	75.15	74.62	75.36
MIML	**39.91**	33.97	33.95	39.17	35.67	37.71
UMIST	**98.39**	98.01	98.01	98.26	98.05	98.18
SCENE	**36.08**	32.40	32.44	34.13	32.82	35.89

Please cite this article as: S. Wang, et al., Multi-task support vector machines for feature selection with shared knowledge discovery, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2014.12.012

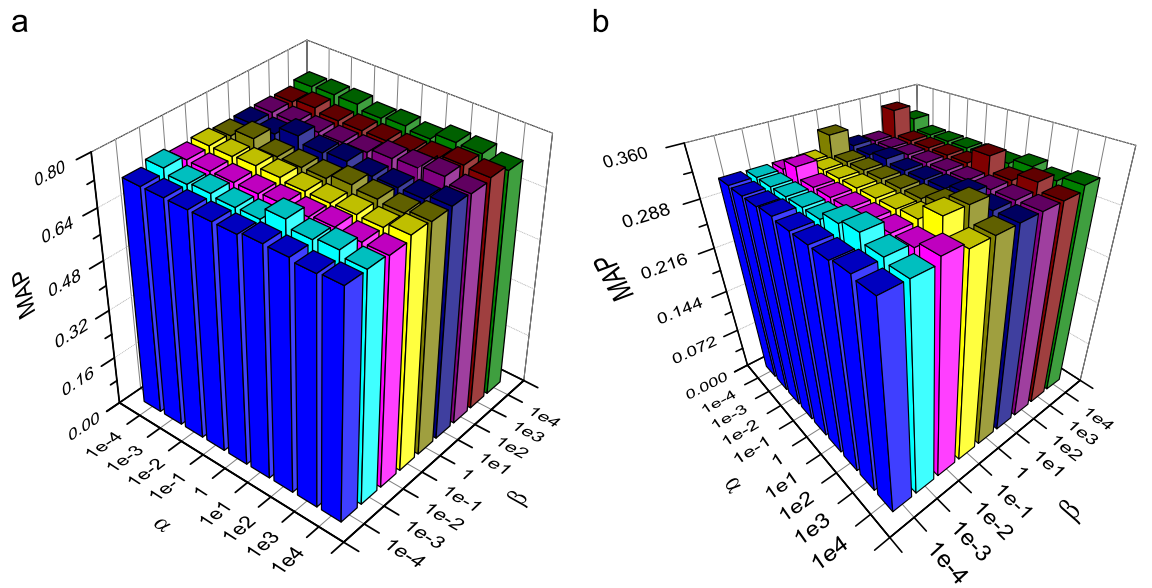

Fig. 1. Performance variations w.r.t different combinations of α and β. (a) HumanEva. (b) Scene.

order. The highest ranked features are selected.

3. *Fisher score* [40]: It is a classical filter method in which the discriminative power of each feature is ranked according to some univariate metric.
4. *Single task SVM* (*ST-SVM*): To validate the benefit from shared knowledge mining in the proposed method, SVM-based feature selection using $\ell_{2,1}-$norm for single task is also compared in the experiments. This method is equal to our proposed method without the low-rank constraint to the stacked feature selection matrices.
5. *Feature selection with shared information* (*FSSI*) [31]: It employs the least squared loss function and $\ell_{2,1}-$norm regularization for each single task, and a joint trace norm minimization to exploit shared information among multiple tasks.

For all the compared methods, corresponding parameters are all tuned in the same range of $\{10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}\}$. We treat each class of the dataset as a learning task. To evaluate feature selection methods, we train a linear SVM classifier for each learning task. The parameter of SVM, which controls the trade off between allowing training errors and forcing rigid margins, is also tuned in the aforementioned range. For all experiments, we randomly select 10 labeled data per class as the training sets and use the rest of data as the testing sets.[1] Average Precision and Mean Average Precision are used as metrics for each learning task and the holistic performance, respectively. Learning rate λ is initialized to 0.1 and adapts after each iteration step by dividing the number of iteration.

[1] Due to the size of ADL-RFID dataset, we only randomly select 2 labeled data per class.

4.2. Experimental results

We have compared five feature selection algorithms over eight datasets involving a number of multimedia applications: action recognition (ADL-RFID, KTH, HumanEva), face recognition(UMIST), object recognition (COIL-20, COREL-5K) and scene recognition (MIML, SCENE). In Table 2, the performance comparison across all the datasets measured by MAP is reported. From the outcome, it is observed that the proposed method has been consistently better than the other compared algorithms. Notably, our algorithm and FSSI consistently perform fairly better on ADL-RFID, HumanEva, KTH, MIML and SCENE against those methods which neglect the shared knowledge across multiple learning tasks (All Features, Max Variance, Fisher Score, ST-SVM). Meanwhile our algorithm and FSSI are slightly better than the counterparts on COIL-20, COREL-5K and UMIST. The reason for this outcome might be because the shared information across multiple learning tasks in ADL-RFID, HumanEva, KTH, MIML and SCENE is more helpful than the one in COIL-20, COREL-5K and UMIST. Even though exploiting shared information across multi-tasks is employed, only slight improvements have been observed in the latter three datasets. Besides, our approach is marginally better than FSSI over most of the datasets except for MIML (2.2% improvement).

We study the sensitivity of parameters in two experiments. In the first one, we select HumanEva and SCENE as examples to demonstrate the performance variations of the proposed feature selection approach with respect to different combinations of parameters α and β in Fig. 1. In the second experiment, we fix both $\alpha = 1$ and the percentage of selected features. Performance variations of two tasks in KTH, jogging and handclapping, are reported under different β. When β is close to zero, the contribution of shared information exploiting in (5) is removed which means little shared information has been taken into account in this case. In Fig. 2, for both two tasks, performance initializes at a lower value and then

Please cite this article as: S. Wang, et al., Multi-task support vector machines for feature selection with shared knowledge discovery, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2014.12.012

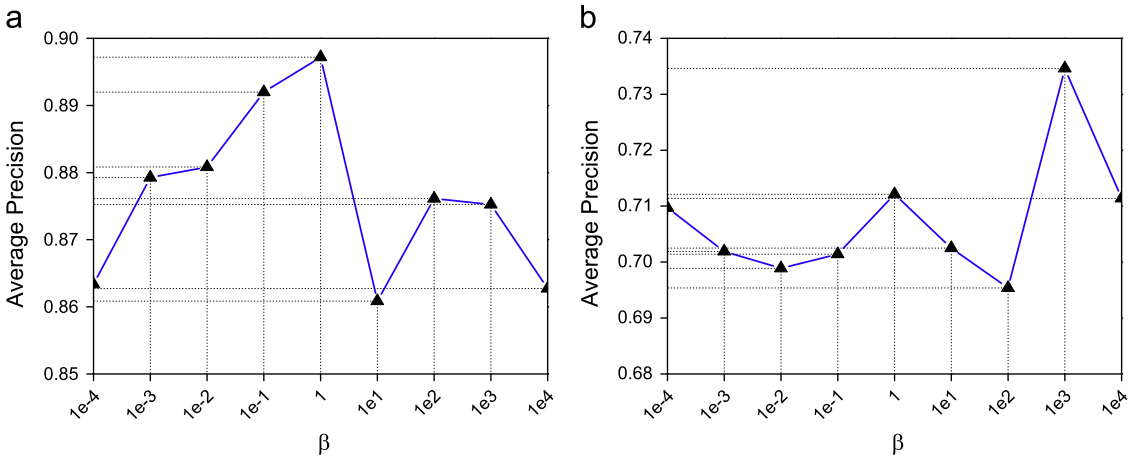

Fig. 2. Performance variations of jogging and handclapping with different β on KTH dataset. (a) Jogging. (b) Handclapping.

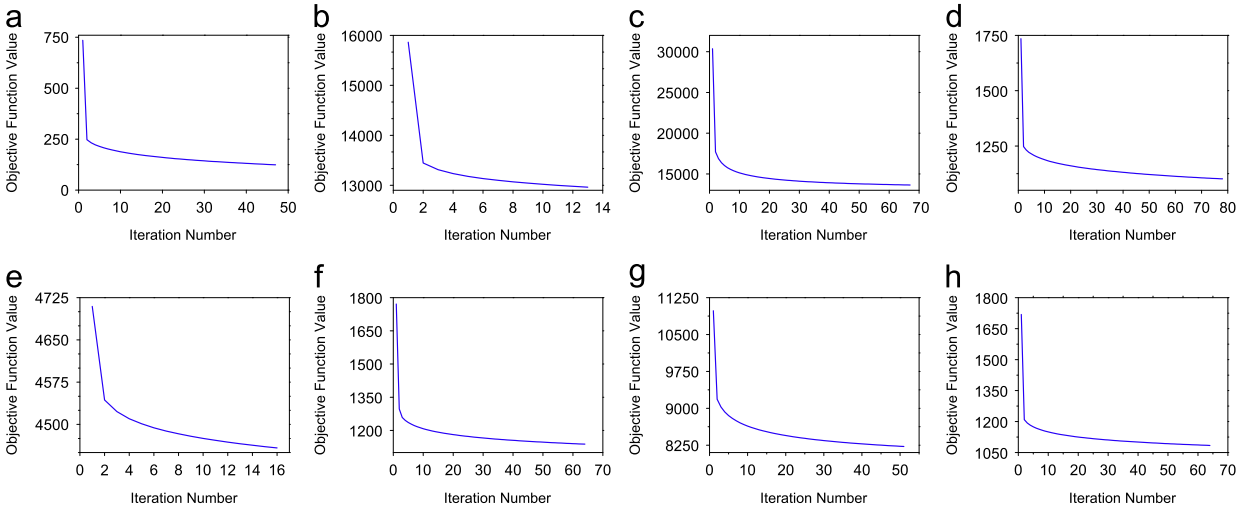

Fig. 3. The objective function values across all datasets when $\alpha = 1$ and $\beta = 1$. (a) ADL-RFID. (b) COIL-20. (c) COREL-5K. (d) HumanEva (e) KTH. (f) MIML. (g) UMIST. (h) SCENE.

peaks at the highest point when β is set to a proper value. More specifically, the classification performance of jogging starts with a small proportion of shared information when $\beta = 10^{-4}$. When $\beta = 1$, a 3.4% improvement is observed. We also have similar observations for other tasks across all the datasets. This validates that the shared information can be beneficial to learning of single task.

In Fig. 3, the objective function value of (5) at each iteration step for each dataset is plotted when all parameters are set to 1. It is observed that the objective function monotonically decreases until convergence. However, the pace of convergence varies for different datasets. For instance, our algorithm converges much faster on COIL-20 and KTH than on the others when both α and β are set to 1.

5. Conclusions and future work

In this paper, we propose a feature selection algorithm in which both information from each learning task and shared structures among multiple tasks are jointly considered. Specifically, we propose to use hinge loss function with the $\ell_{2,1}-$norm regularization to learn feature selection matrix for each task. Sparsity of the feature selection matrix helps us to discover the correlations within each task while choosing distinctive features. Meanwhile, we also impose a global constraint to exploit shared information across multiple tasks. This is achieved by minimizing the trace norm of $W = [W_1, \ldots, W_t]$. Gradient descent is applied to reach the global optimum. Extensive experiments on a variety of datasets have demonstrated that the proposed method improves the performance particularly when the shared information is rich and helpful on some multimedia datasets.

Acknowledgment

This work was supported by the Australian Research Council Discover Project under Grant no. DP 130104614.

Please cite this article as: S. Wang, et al., Multi-task support vector machines for feature selection with shared knowledge discovery, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2014.12.012

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Australian Research Council.

References

[1] P. Langley, et al., Selection of Relevant Features in Machine Learning, Defense Technical Information Center, 1994.
[2] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1) (1997) 273–324.
[3] P.M. Narendra, K. Fukunaga, A branch and bound algorithm for feature subset selection, IEEE Trans. Comput. 100 (9) (1977) 917–922.
[4] P.A. Mundra, J.C. Rajapakse, Svm-rfe with mrmr filter for gene selection, IEEE Trans. Nanobiosci. 9 (1) (2010) 31–37.
[5] O. Chapelle, S.S. Keerthi, Multi-class feature selection with support vector machines, in: Proceedings of the American Statistical Association, 2008.
[6] R. Archibald, G. Fann, Feature selection and classification of hyperspectral images with support vector machines, IEEE Geosci. Remote Sens. Lett. 4 (4) (2007) 674–677.
[7] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, Q. Tian, Discovering discriminative graphlets for aerial image categories recognition, IEEE Trans. Image Process. 22 (12) (2013) 5071–5084.
[8] J. Neumann, C. Schnörr, G. Steidl, Combined svm-based feature selection and classification, Mach. Learn. 61 (1–3) (2005) 129–150.
[9] A. Al-Shahib, R. Breitling, D. Gilbert, Feature selection and the class imbalance problem in predicting protein function from sequence, Appl. Bioinform. 4 (3) (2005) 195–203.
[10] P. Conilione, D. Wang, et al., A comparative study on feature selection for e. coli promoter recognition, Int. J. Inf. Technol. 11 (2005) 54–66.
[11] S.-K. Kim, J.-W. Nam, J.-K. Rhee, W.-J. Lee, B.-T. Zhang, mitarget: microrna target gene prediction using a support vector machine, BMC Bioinform. 7 (1) (2006) 411.
[12] Y. Yan, H. Shen, L. Liu, Z. Ma, C. Gao, N. Sebe, Glocal tells you more: coupling glocal structural for feature selection with sparsity for image and video classification, Comput. Vis. Image Underst. 123 (2014) 99–109.
[13] Z. Ma, Y. Yang, N. Sebe, K. Zheng, A. Hauptmann, Multimedia event detection using a classifier-specific intermediate representation, IEEE Trans. Multimed. 15 (7) (2013) 1628–1637.
[14] Y. Yan, G. Liu, S. Wang, J. Zhang, K. Zheng, Graph-based clustering and ranking for diversified image search, Multimed. Syst. http://dx.doi.org/10.1007/s00530-014-0419-4.
[15] X. Zhang, X. Zhao, Z. Li, J. Xia, R. Jain, W. Chao, Social image tagging using graph-based reinforcement on multi-type interrelated objects, Signal Process. 93 (8) (2013) 2178–2189.
[16] J.A. Nasiri, N.M. Charkari, K. Mozafari, Energy-based model of least squares twin support vector machines for human action recognition, Signal Process. 104 (2014) 248–257.
[17] L. Zhang, Y. Gao, Y. Xia, Q. Dai, X. Li, A fine-grained image categorization system by cellet-encoded spatial pyramid modeling, IEEE Trans. Ind. Electron. PP(99) (2014) 1–1. http://dx.doi.org/10.1109/TIE.2014.2327558.
[18] Feature selection for fast speech emotion recognition, in: Proceedings of the 17th International Conference on Multimedia 2009, Vancouver, British Columbia, Canada, October 19–24, 2009.
[19] Z. Ma, F. Nie, Y. Yang, J.R. Uijlings, N. Sebe, Web image annotation via subspace-sparsity collaborated feature selection, IEEE Trans. Multimed. 14 (4) (2012) 1021–1030.
[20] X. Chang, F. Nie, Y. Yang, H. Huang, A convex formulation for semi-supervised multi-label feature selection, in: AAAI Conference on Artificial Intelligence, 2014.
[21] L. Zhang, Y. Gao, C. Hong, Y. Feng, J. Zhu, D. Cai, Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition, IEEE Trans. Cybern. 44 (8) (2014) 1408–1419, http://dx.doi.org/10.1109/TCYB.2013.2285219.
[22] Y. Yang, Y. Zhuang, F. Wu, Y. Pan, Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval, IEEE Trans. Multimed. 10 (3) (2008) 437–466.
[23] S. Wang, Z. Ma, Y. Yang, X. Li, C. Pang, A. Hauptmann, Semi-supervised multiple feature analysis for action recognition, IEEE Trans. Multimed. 16 (2) (2014) 289–298.
[24] H. Shen, Y. Yan, S. Xu, N. Ballas, W. Chen, Evaluation of semi-supervised learning method on action recognition, Multimedia Tools and Applications, Springer, US, 2014 http://dx.doi.org/10.1007/s11042-014-1936-z.
[25] Y. Han, Z. Xu, Z. Ma, Z. Huang, Image classification with manifold learning for out-of-sample data, Signal Process. 93 (8) (2013) 2169–2177.
[26] L. Zhang, M. Song, Y. Yang, Q. Zhao, C. Zhao, N. Sebe, Weakly supervised photo cropping, IEEE Trans. Multimed. 16 (1) (2014) 94–107.
[27] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, X. Zhou, Semi-supervised feature selection via spline regression for video semantic recognition, IEEE Transactions on Neural Networks and Learning Systems (IEEE T-NNLS), http://dx.doi.org/10.1109/TNNLS.2014.2314123.
[28] Z. Ma, Y. Yang, F. Nie, N. Sebe, S. Yan, A.G. Hauptmann, Harnessing lab knowledge for real-world action recognition, Int. J. Comput. Vis. 109 (1–2) (2014) 60–73.
[29] Z. Ma, Y. Yang, N. Sebe, A.G. Hauptmann, Knowledge adaptation with partially shared features for event detection using few exemplars, IEEE Trans. Pattern Anal. Mach. Intell. 36 (9) (2014) 1789–1802.
[30] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S.L. Risacher, A.J. Saykin, L. Shen, High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction, in: Neural Information Processing Systems, 2012, pp. 1286–1294.
[31] Y. Yang, Z. Ma, A.G. Hauptmann, N. Sebe, Feature selection for multimedia analysis by sharing information among multiple tasks, IEEE Trans. Multimed. 15 (3) (2013) 661–669.
[32] R.K. Ando, T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, J. Mach. Learn. Res. 6 (2005) 1817–1853.
[33] A. Evgeniou, M. Pontil, Multi-task feature learning, in: Neural Information Processing Systems, vol. 19, 2007, p. 41.
[34] R. Caruana, Multitask learning, Springer, US, 1998.
[35] Feature selection for high-dimensional data: A fast correlation-based filter solution: Proceedings of the Twentieth International Conference on Machine Learning, AAAI press, Washington, DC, USA, August 21–24, 2003.
[36] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[37] L2,1-norm regularized discriminative feature selection for unsupervised learning. in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16–22, 2011. AAAI press.
[38] X. Chang, H. Shen, S. Wang, J. Liu, X. Li, Semi-supervised feature analysis for multimedia annotation by mining label correlation, in: Pacific–Asia Conference on Knowledge Discovery and Data Mining, 2014.
[39] Y. Han, J. Zhang, Z. Xu, S.-I. Yu, Discriminative multi-task feature selection, in: AAAI (Late-Breaking Developments), 2013.
[40] Pattern Classification: R. O DUDA, P. E. HART, Pattern Classification. John Wiley and Sons, Inc., New York, USA.
[41] M.H. Law, M.A. Figueiredo, A.K. Jain, Simultaneous feature selection and clustering using mixture models, IEEE Trans. Pattern Anal. Mach. Intell. 26 (9) (2004) 1154–1166.
[42] Estimating attributes: analysis and extensions of relief: Springer, Berlin, Heidelberg, 1994.
[43] Z. Zhao, L. Wang, H. Liu, Efficient spectral feature selection with minimum redundancy, in: AAAI Conference on Artificial Intelligence, 2010.
[44] B. Krishnapuram, A. Harternink, L. Carin, M.A. Figueiredo, A bayesian approach to joint feature selection and classifier design, IEEE Trans. Pattern Anal. Mach. Intell. 26 (9) (2004) 1105–1111.
[45] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint l21-norms minimization, in: Neural Information Processing Systems, 2010.
[46] D. Meng, Q. Zhao, Z. Xu, Improve robustness of sparse pca by l1-norm maximization, Pattern Recognit. 45 (1) (2012) 487–497.
[47] A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, Mach. Learn. 73 (3) (2008) 243–272.
[48] F. Nie, H. Huang, C.H.Q. Ding, Low-rank matrix recovery via efficient schatten p-norm minimization, in: AAAIConference on Artificial Intelligence, 2012.
[49] How related exemplars help complex event detection in web videos, in: IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1–8, 2013.
[50] M. Stikic, T. Huynh, K. Van Laerhoven, B. Schiele, Adl recognition based on the combination of rfid and accelerometer sensing, in: IEEE Second

Please cite this article as: S. Wang, et al., Multi-task support vector machines for feature selection with shared knowledge discovery, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2014.12.012

International Conference on Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008, 2008, pp. 258–263.

[51] L. Sigal, A.O. Balan, M.J. Black, Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, Int. J. Comput. Vis. 87 (1–2) (2010) 4–27.

[52] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 723–742.

[53] Z.-H. Zhou, M.-L. Zhang, Multi-instance multi-label learning with application to scene classification, in: Neural Information Processing Systems, 2006, pp. 1609–1616.

[54] M.R. Boutell, J. Luo, X. Shen, C.M. Brown, Learning multi-label scene classification, Pattern Recognit. 37 (9) (2004) 1757–1771.

Please cite this article as: S. Wang, et al., Multi-task support vector machines for feature selection with shared knowledge discovery, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2014.12.012