

[Applied Computing and Informatics 15 (2019) 172–181](https://doi.org/10.1016/j.aci.2017.10.004)

Contents lists available at [ScienceDirect](http://www.sciencedirect.com/science/journal/22108327)

Applied Computing and Informatics

journal homepage: [www.sciencedirect.com](http://www.sciencedirect.com)

|  |  |
| --- | --- |
| Semi-supervised soft margin consistency based multi-view maximum entropy discrimination |  |

Changming Zhua,⇑, Zhe Wangb

a College of Information Engineering, Shanghai Maritime University, Shanghai 201306, PR China   
b College of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, PR China

|  |  |  |
| --- | --- | --- |
| a r t i c l e | i n f o | a b s t r a c t |
| Article history:  Received 16 August 2017  Revised 6 October 2017  Accepted 20 October 2017  Available online 14 August 2018 | | Multi-view maximum entropy discrimination (MVMED) and alternative MVMED (AMVMED) are pro-posed as extensions of maximum entropy discrimination (MED). In MVMED and AMVMED, they use hard margin consistency principle that the decision of margin parameter is related to classifier parameter directly. While the decision always be indirectly in practice, thus soft margin consistency based multi-view maximum entropy discrimination (SMVMED) has been proposed. But it is found that SMVMED is |
| Keywords:  Semi-supervised multi-view learning Multi-view maximum entropy  discrimination  Soft margin consistency | | only adaptive to supervised problems. In this paper, we extend the model of SMVMED to the semi-supervised problems and develop a semi-supervised SMVMED (SSMVMED). Related experiments on multi-view data sets from different aspects have validated the effectiveness of SSMVMED theoretically and empirically. From the experiments, it is found that (1) compared with SMVMED, the average test accuracy of SSMVMED has a 2% enhancement; (2) SSMVMED costs more training time than SMVMED and the extra time is not more than 10%; (3) in terms of the generation of additional unlabeled instances, |

‘mid’ strategy has a better test accuracy than ‘self’ and taking all instances to get the center brings a better test accuracy as well; (4) with SSMVMED, the applications to estimation problem and regression problem will be more feasible.

� 2018 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

|  |  |
| --- | --- |
| 1. Introduction | multi-view data sets, many multi-view learning machines are pro- |

posed as below.

1.1. Background   
 (1) pre-fusion methods: multiple kernel learning (MKL) [1], cen-

Multi-view data set is composed of instances with multiple views and each view consists of multiple features. A corresponding feature group is made up of these features. Take a video data set as an example, suppose this data set consists of multiple videos and each video appears in multiple different forms including visual, audio, and text. Then we treat each form as a view of a video. More-over, each view has several features, for example, text view can be

tered alignment-based MKL algorithms (CABMKL) [2], sim-ple MKL method (SMKL) [3], group Lasso-based MKL method (GLMKL) [4], localized MKL (LMKL) [5].

(2) late-fusion methods: robust late fusion method (RLF) [6]. (3) subspace approaches: multi-view linear discriminant analy- sis (MV-LDA) [7], multi-view canonical correlation analysis (MV-CCA) [8], multi-view locality preserving projections

described by text color, text size, text content and color, size, con- (MV-LPP) [9].

tent form a feature group of text view. In order to process those

(4) disagreement-based methods: co-training [10], confident co-training with data editing (CoTrade) [11], co-regularized

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ⇑ Corresponding author.  E-mail addresses: [cmzhu@shmtu.edu.cn](mailto:cmzhu@shmtu.edu.cn) | (C. | Zhu), | [wangzhe@ecust.edu.cn](mailto:wangzhe@ecust.edu.cn) | Laplacian SVM (Co-Lap) [12]. |
| But these learning machines always neglect to consider uncer- |
| (Z. Wang). | | |
| tainties over model parameters and then maximum entropy dis- |
| Peer review under responsibility of King Saud University. | | |
| crimination (MED) [13] has been developed to consider this issue |

and learn a discriminative classifier. In MED, it learns a distribution

**Production and hosting by Elsevier**

<https://doi.org/10.1016/j.aci.2017.10.004>

pðHÞ over classifier parameter H and this is contrast to the tradi-tional learning machines. In terms of traditional learning machines,

2210-8327/� 2018 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

C. Zhu, Z. Wang / Applied Computing and Informatics 15 (2019) 172–181 173

they only find a single classifier parameter pðHÞ of the discriminant function LðHÞ (e.g., LðXtjHÞ ¼ hTXt þ b; H ¼ fh; bg). While MED obtains a joint distribution pðH;cÞ over H and margin parameters c by minimizing its relative entropy with respect to some prior tar-get distribution p0ðH;cÞ under certain large margin constraints, MED marginalizes out c to obtain pðHÞ [14]. Although MED considers the uncertainties over model parame-ters, its applicable scope limits to single-view problem. Thus multi-view maximum entropy discrimination (MVMED) [15] and alter-native MVMED (AMVMED) [16] are developed to extend the model of MED to multi-view problems. MVMED and AMVMED exploit multiple views in a different style called margin consistency and enforce the margins from two views to be identical. In other words, they adopt the hard margin consistency while have no ability to process large data sets. So soft margin consistency based multi-view maximum entropy discrimination (SMVMED) [17] has been

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| proposed. | Different | from | MVMED | and | AMVMED, | SMVMED |

achieves ‘soft’ margin consistency by utilizing the sum of two KL divergences KLðpðcÞjjqðcÞÞ and KLðqðcÞjjpðcÞÞ in the objective func-tion, where pðcÞ and qðcÞ are the posteriors of two view margins, respectively. By balancing all involved terms in the objective func-tion, SMVMED is more flexible.

While SMVMED is only feasible for supervised problems, i.e., the used instances are labeled. Indeed, most real-world multi-view data sets consist of labeled and unlabeled instances and they are named semi-supervised data sets. In order to process semi-supervised data sets, semi-supervised learning machines have been developed and introduced in many applications [18–24]. In terms of the applications of MED, there also exist some related learning machines for semi-supervised problems, for example, semi-supervised multi-sensor classification via consensus-based

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| multi-view | maximum | entropy | discrimination | [25], | semi- |

supervised learning via generalized maximum entropy [26], and semi-supervised multi-task learning via self-training and maxi-mum entropy discrimination [27]. But these learning machines do not consider soft margin and we find that soft margin has its special physical meaning. As [17] said, if one view corresponds to one special sub-learning machine, its parameters include classifier parameter H and margin parameter c. If the decision of c is related to H directly, we define it as hard margin consistency while if the decision is not directly, we call it as soft margin consistency. For soft margin consistency, [17] has also validated that the decision of c is more flexible and the performance and the conduction of a learning machine is much better and fast.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Universum | learning, | it | aims | to | create | additional | unlabeled |

instances (also called Universum instances or Universum set) and incorporates priori knowledge which is introduced in the form of additional unlabeled instances into the learning process [29]. By Universum learning, the performance of a traditional learning machine can be boosted. Now Universum learning has been grad-ually spread into different applications [30–34] and some related methods are also developed including Universum support vector machine (U-SVM) [35] and self-Universum support vector machine (SUSVM) [36]. So for our SSMVMED, we also adopt Universum learning to add useful discriminant information.

But for these Universum-based learning machines, there still exists two key problems. First one is that when generating Univer-sum set, the weights of views and features which play different discriminant roles are always neglected. Second one is that when

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| generating | the | additional | unlabeled | instances, | traditional |

Universum-based learning machines only adopt labeled or unla-beled instances for generation. In order to solve the first problem, we adopt weighted multi-view clustering (WMVC) [37] which is a multi-view clustering method and can find the optimal cluster assignment. With WMVC, the weights of views and features can be gotten. For the second problem, we try to design some schemes and adopt both labeled and unlabeled instances to generate the additional unlabeled instances.

1.3. Novelty and practical applications

The novelty of the proposed SSMVMED is given below.

First, compared with some semi-supervised learning machines with MED, SSMVMED adopts the soft margin and inherits the advantages of soft margin consistency which makes the decision of margin parameters be more flexible and the performance and the conduction of a learning machine be much better and faster.

Second, SSMVMED extends the model of SMVMED to semi-supervised problems and this enlarges the applicable scope of MED.

Third, during the procedure of SSMVMED, it improves the Universum learning methods, and when generating the additional unlabeled instances, the weights of views and features will be con-sidered and both labeled and unlabeled instances are used for the generation.

The practical applications of SSMVMED can include estimation problems, regression problem, classification problems and so on. In our work, we will adopt some related experiments to show the effectiveness of our proposed SSMVMED in these applications.

1.2. Proposal and trouble 1.4. Framework

Thus, in this paper, we still adopt soft margin consistency and apply it to semi-supervised problems and then propose the semi-supervised soft margin consistency based multi-view maximum entropy discrimination (SSMVMED). But during the process of SSMVMED, there is a potential trouble when the data sets consist

The rest of this paper is organized as below. Related work about MED is given in Section 2. Description of SSMVMED is given in Sec-tion 3. Experiments are given in Section 4. The conclusions are given in Section 5.

|  |  |
| --- | --- |
| of few labeled instances and many unlabeled ones. As we know, compared with unlabeled instances, labeled ones can provide more | 2. Related work |

useful discriminant information while in real-world applications, most data sets consist of few labeled instances and many unlabeled ones and labeling instances is a high-cost task. Thus, for traditional semi-supervised problems, the performances of learning machines are sensitive to the data sets. In order to enhance the performance

Since our SSMVMED is the extension of SMVMED and SMVMED is different from MVMED and AMVMED, thus we will review MVMED, AMVMED, and SMVMED here.

of a learning machine, a widely used and feasible method is gener- 2.1. MVMED

ating additional unlabeled instances with the original labeled or

unlabeled ones and combining all instances together. These addi- MVMED was proposed as an extension of MED to the multi-

tional unlabeled instances will possess some discriminant informa- view learning setting and it considers a joint distribution

tion derived from the original labeled and unlabeled instances. Here, Universum learning [28] is such a kind of method. For

pðH1; H2Þ over the view 1 classifier parameter H1 and view 2 clas-sifier parameter H2. Using the augmented joint distribution

174 C. Zhu, Z. Wang / Applied Computing and Informatics 15 (2019) 172–181

pðH1; H2;cÞ, the model of MVMED is given in Eq. (1) [15] where 1 6 t 6 N. In this model, L1ðX1 tjH1Þ and L2ðX2 tjH2Þ are discriminant functions from two views, respectively. yt is the class label of tth labeled instance Xt and its margin parameter is ct. X1 tand X2 tare the representations of Xt in the first and second views, respectively. With MVMED, multi-view feature selection, multi-view multi-task

learning, multi-view structure learning, and some other problems

can be solved well.

the additional unlabeled instances. Third, apply the original

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| labeled, | unlabeled, | and | the | generated | additional | unlabeled |

instances into the model of SSMVMED.

3.1. Obtain the weights of views and features

In order to obtain the weights of views and features, we adopt WMVC for help. Suppose there is a multi-view data set consisting

|  |  |  |
| --- | --- | --- |
| minpðH1;H2;cÞKLðpðH1; H2;cÞjjp0ðH1; H2;cÞÞ  s:t: ( R R pðH1; H2;cÞ½ytL1ðX1 pðH1; H2;cÞ½ytL2ðX2 tjH1Þ � ct�dH1dH2dc P 0 tjH2Þ � ct�dH1dH2dc P 0 | ð1Þ | of N instances represented by V views, i.e.,  v ¼ fX1 1; X2 1; . . . ; XV 1; . . . ; X1 N; X2 N; . . . ; XV Ng where Xv i2 Rdv is the repre-  sentation of ith instance Xi in the vth view and dvis the dimension |
| of vth view. Here, we let Xv¼ fXv 1; Xv 2; . . . ; Xv Ng represent the vth view  and v can be represented as v ¼ fX1; X2; . . . ; Xv; . . . ; XV�1; XVg. Then according to the notion of WMVC, we try to obtain the weights of |
| 2.2. AMVMED |

views and the weights of features for each view. Let weight of each

Different from MVMED, AMVMED considers two separate dis-tributions pðH1Þ over H1 and pðH2Þ over H2 and balances KL diver-gences of their augmented distributions with respect to the corresponding prior distributions. The model of AMVMED is given in Eq. (2) [16] and 1 6 t 6 N;q is a coefficient.

minp1ðH1;cÞ;p2ðH2;cÞqKLðp1ðH1;cÞjjp0ðH1;cÞÞ ð2Þ þ ð1 � qÞKLðp2ðH2;cÞjjp0ðH2;cÞÞ

s:t: 8 < : As [17] said, for both MVMED and AMVMED, they exploit the R R R pðH1;cÞ½ytL1ðX1 pðH2;cÞ½ytL2ðX2 pðH1;cÞdH1 ¼ tjH1Þ � ct�dH1dc P 0 tjH2Þ � ct�dH2dc P 0 R pðH2;cÞdH2

multiple views in a different style called margin consistency which indicate the margins from two views are enforced to be identical. Moreover, the margins are related to the parameters H1 and H2 directly and those margins are named as hard margins. Although MVMED and AMVMED have provided state-of-the-art multi-view learning performance, hard margin requirement is somewhat too strong to fulfill in many cases. For example, all positive margins can lead to the same label prediction in binary classifications.

2.3. SMVMED

SMVMED is different from MVMED and AMVMED due to the margin of SMVMED is soft. SMVMED achieves margin consistency by minimizing the KL-divergence between the posteriors of margin parameters from two views. Then a trade-off parameter balancing large margin and margin consistency is also introduced to make the model more flexible. The model of SMVMED is given below. Here, pðHÞ or qðHÞ is a distribution over H and pðH1; H2Þ is a joint distribution over H1 and H2. The parameter a is a parameter play-ing the trade-off role of balancing large margin and soft margin consistency. Compared with MVMED and AMVMED, SMVMED is more flexible and the performance and the conduction of a learn-ing machine with SMVMED is much better and faster.

|  |  |
| --- | --- |
| minpðH1;cÞ;qðH2;cÞKLðpðH1Þjjp0ðH1ÞÞ þ KLðqðH2Þjjq0ðH2ÞÞ þ ð1 � aÞKLðpðcÞjjp0ðcÞÞ þ ð1 � aÞKLðqðcÞjjq0ðcÞÞ þ aKLðpðcÞjjqðcÞÞ þ aKLðqðcÞjjpðcÞÞ  s:t: ( R R pðH1;cÞ½ytL1ðX1 qðH2;cÞ½ytL2ðX2 tjH2Þ � ct�dH2dc P 0 tjH1Þ � ct�dH1dc P 0 | ð3Þ |

3. Semi-supervised Soft Margin Consistency based Multi-view Maximum Entropy Discrimination (SSMVMED)

Our proposed SSMVMED consists of three steps. First, compute the weights of views and features by WMVC [37]. Second, generate

|  |
| --- |
| view be xv where v ¼ 1; 2; . . . ; V and weight for the lth feature of  vth view is sv lwhere l ¼ 1; 2; . . . ; dv. Here, each weight should not  be less than zero. Furthermore,PV Xv;Pdv multi-view data set should be divided into several clusters. Let the l¼1sv l¼ 1. Then according to the notion of WMVC, the whole v¼1xv ¼ 1 and for each view  number of clusters be M; k denote the index of clusters, and dik  denote the belonging of the instance Xi, if instance Xi belongs to  kth cluster, then dik ¼ 1, otherwise, dik ¼ 0. For any instance  Xi;PM k¼1dik ¼ 1. The objective function of WMVC is given in Eq. (4).  minfdikgM k¼1;fxvgV v¼1;fsvgV v¼1 eH ð4Þ  s:t:  bPV diagonal matrix where other elements in this matrix are zeros. 8 < >  In this function, eH ¼PV v¼1jjsvjj2;sv ¼ fsv   PM PV Pdv k¼1dik ¼ 1;  v¼1xv ¼ 1;  l¼1sv l¼ 1; 8v;   8i;  xv P 0  2; . . . ;sv   dik 2 f0; 1g  sv  v¼1ðxvÞpPN  lP 0  dv g, and i¼1 PM k¼1dikjjdiagðsvÞðXv  diagðsvÞ represents   i�mv kÞjj2þ  the  Moreover, mv  vth view. Furthermore, bPV of the feature weight vectors sv; 8v so as to avoid the situation that only a few features are selected in getting a very small but mean-k¼ PN PN i¼1dikXv  i¼1dik   is the cluster center of kth cluster in the  v¼1jjsvjj2 is used to control the sparsity |

ingless objective value. The parameters p and b are the exponential and balancing parameters, which are selected according to the pri-ori knowledge of data so as to help controlling the sparsity of the

|  |  |  |  |
| --- | --- | --- | --- |
| view weight vector x ¼ fx1;x2; . . . ;xVg and the feature weight  vectors sv; 8v ¼ 1; 2; . . . ; V respectively. Then with WMVC, xv can be updated by Eq. (5) and svcan be  updated by Eq. (6) until the computations of xv and svbe conver-  gent or the iteration times is up to a maximum number. In these  equations,  ðxvÞpPN i¼1 PM Dv ¼PN k¼1dikðXv  i¼1  i� mv PM k¼1dik diagðsvÞðXv  kÞ  2  lwhere ðHÞlrepresents the lth element���� i� mv kÞ����2  and Bv l¼ bþ | | | |
| xv ¼ | 1   p > 1 PV�1;v ¼ argminuDu 0; u¼1�  otherwise  Du�1=ðp�1Þ  Dv |  | ð5Þ |
| xv ¼ | p ¼ 1 |
| 1  sv l¼ Bv l 8l ð6Þ  Finally, we can get the optimal or final x and svand the weights Pdv m¼1 Bv m  of views and features are also gotten. | | | |

C. Zhu, Z. Wang / Applied Computing and Informatics 15 (2019) 172–181 175

3.2. Approaches of generating additional unlabeled instances

After we get the weights of views and features, we will adopt Universum learning to generate the additional unlabeled instances (i.e., Universum instances or Universum set). In our paper, the gen-eration approaches used are given in Table 1. From this table, it is found that each approach has a code with the form ‘A-B-C’. ‘A’ has three choices, ‘all’, ‘unlabeled’, ‘labeled’. ‘B’ has three choices, ‘all’,‘near’, ‘far’. ‘C’ has two choices, ‘mid’ and ‘self’. For ‘A’, ‘all’ (‘unla-beled’, ‘labeled’) represents that one computes the midpoint of all (unlabeled, labeled) instances as a center. For ‘B’, ‘all’, ‘near’, and ‘far’ represent that one uses all instances, K instances which locates nearest from the center, and K instances which locates far-thest from the center as selected instances respectively. For ‘C’,‘mid’ represents one takes the midpoint of a selected instance and the center to construct Universum set while ‘self’ represents Universum set consists of the selected instances themselves. We

take\_U1�2 as the example. In this approach, we first to compute the mean of all instances as a center, then we select K instances which locates nearest from this center, finally take the midpoint of a selected instance and the center to construct Universum set.

For all approaches used here, when we compute midpoint or distance, the weights of views and features should be used. Con-

|  |
| --- |
| cretely speaking, if there are two instances, X1 ¼ fX1 1; X2 1; . . . ; XV 1g  and X2 ¼ fX1 2; X2 2; . . . ; XV 2g. The weights of views are x1;x2; . . . ;xV  and the weights of features are svwhere v ¼ 1; 2; . . . ; V and  sv¼ fsv 1;sv 2; . . . ;sv dv g. dv is the dimension of vth view and sv lrepre-  sents weight of the lth feature of vth view. Then the midpoint of X1  and X2 is PV v¼1xvPdv l¼1ðsv lðXv 1þXv 2ÞlÞ  2   and the distance between X1 and X2  isPV ment of ðHÞ. v¼1xv Pdv l¼1ðsv lðXv 1� Xv 2Þ lÞ where ðHÞlrepresents the lth ele- |

3.3. Solution of SSMVMED

Once we generate additional unlabeled instances, we will apply them along with the original labeled and unlabeled instances into the model of SSMVMED. For convenience, we adopt a binary-view data set for example. For a data set with more than two views, we can divide it into several binary-view problems, and for each binary-view problem, a sub model of SSMVMED is gotten and they can be integrated together and get the final model for multi-view data sets.

Suppose for a binary-view data sets, there are N (here N is dif-ferent from the N in Section 3.1 which denotes the number of all instances) labeled instances fX1 t; X2 t; ytg and L unlabeled instances (including the generated additional ones) fU1 i; U2 ig. Here, X1 t(X2 t) represents the first (or second) view of the tth labeled instance Xt and yt is its class label. For U1 i(U2 i), it represents the first (or sec-ond) view of ith unlabeled instance Ui. Different from Eq. (3), the model of SSMVMED adds the constraint of unlabeled instances and the parameter c ¼ fcjg where j ¼ 1; 2; . . . ; N; N þ 1; . . . ; N þ L. Namely, SSMVMED considers the margin constraint of each labeled

Table 1   
The codes of used Universum set construction ways.

or unlabeled instance. Eq. (7) shows the model of SSMVMED. In

this model, 1 6 t 6 N; 1 6 i 6 L; pðHÞ (qðHÞ) is a distribution over H and pðH1; H2Þ (qðH1; H2Þ) is a joint distribution over H1 and H2. a is still a parameter playing the trade-off role of balancing

large margin and soft margin consistency. H1 (H2) is the view 1

(2) classifier parameter. c is the margin parameter.

minpðH1;cÞ;qðH2;cÞKLðpðH1Þjjp0ðH1ÞÞ þ KLðqðH2Þjjq0ðH2ÞÞ ð7Þ

þ ð1 � aÞKLðpðcÞjjp0ðcÞÞ þ ð1 � aÞKLðqðcÞjjq0ðcÞÞ

þ aKLðpðcÞjjqðcÞÞ þ aKLðqðcÞjjpðcÞÞ

s:t: 8 >

: In order to optimize Eq. (7), we use an iterative scheme for find-R R R R pðH1;cÞ½ytL1ðX1

qðH2;cÞ½ytL2ðX2

pðH1;cÞ½L1ðU1

qðH2;cÞ½L2ðU2   
 ijH1Þ � ci�dH1dc P 0

ijH2Þ � ci�dH2dc P 0   
tjH2Þ � ct�dH2dc P 0   
tjH1Þ � ct�dH1dc P 0

ing a solution to Eq. (7) which is similar with the one given in [17].

In the mth iteration, we successively update pðmÞðH1;cÞ and qðmÞðH2;cÞ by solving the following two problems (Eqs. (8) and (9)) and before the solution, we choose some initial value for

qð0ÞðH2;cÞ with q0ðH2;cÞ and this makes Eq. (8) be a standard MED problem.

|  |  |
| --- | --- |
| pðmÞðH1;cÞ ¼ argminpðmÞðH1;cÞKLðpðmÞðH1Þjjp0ðH1ÞÞ   þ ð1 � aÞKLðpðmÞðcÞjjp0ðcÞÞ þ aKLðpðmÞðcÞjjqðm�1ÞðcÞÞ s:t: ( R R pðmÞðH1;cÞ½ytL1ðX1 pðmÞðH1;cÞ½L1ðU1 ijH1Þ � ci�dH1dc P 0 tjH1Þ � ct�dH1dc P 0  qðmÞðH2;cÞ ¼ argminqðmÞðH2;cÞKLðqðmÞðH2Þjjq0ðH2ÞÞ   þ ð1 � aÞKLðqðmÞðcÞjjq0ðcÞÞ þ aKLðqðmÞðcÞjjpðmÞðcÞÞ s:t: (  The Lagrangian of Eq. (8) can be written as R R qðmÞðH2;cÞ½ytL2ðX2 qðmÞðH2;cÞ½L2ðU2 ijH2Þ � ci�dH2dc P 0 tjH2Þ � ct�dH2dc P 0 | ð8Þ  ð9Þ |
| L ¼ Z pðmÞðH1Þlog pðmÞðH1Þ p0ðH1ÞdH1 ð10Þ  þ ð1 � aÞ Z pðmÞðcÞlog pðmÞðcÞ p0ðcÞdc þ a Z pðmÞðcÞlog pðmÞðcÞ qðm�1ÞðcÞdc  �X Z pðmÞðH1;cÞkðmÞ 1;t½ytL1ðX1 tjH1Þ � ct�dH1dc  �X Z pðmÞðH1;cÞ/ðmÞ 1;i½L1ðU1 ijH1Þ � ci�dH1dc  ¼ Z pðmÞðH1;cÞlog p0ðH1Þ½p0ðcÞ�1�a½qðm�1ÞðcÞ�a pðmÞðH1;cÞ  �X Z pðmÞðH1;cÞkðmÞ 1;t½ytL1ðX1 tjH1Þ � ct�dH1dc  �X Z pðmÞðH1;cÞ/ðmÞ 1;i½L1ðU1 ijH1Þ � ci�dH1dc | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Code | Way | Code | Way | Code | Way |
| \_U1�1  \_U1�2  \_U1�3  \_U1�4  \_U1�5 | all-all-mid | \_U2�1  \_U2�2  \_U2�3  \_U2�4  \_U2�5 | unlabeled-all-mid | \_U3�1  \_U3�2  \_U3�3  \_U3�4  \_U3�5 | labeled-all-mid |
| all-near-mid | unlabeled-near-mid | labeled-near-mid |
| all-far-mid | unlabeled-far-mid | labeled-far-mid |
| all-near-self | unlabeled-near-self | labeled-near-self |
| all-far-self | unlabeled-far-self | labeled-far-self |

176 C. Zhu, Z. Wang / Applied Computing and Informatics 15 (2019) 172–181

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| where kðmÞ 1 ¼ fkðmÞ 1;tg and /ðmÞ ¼ f/ðmÞ 1;ig are sets of nonnegative  Lagrange multipliers, one for each classification constraint. Then | | | | | | | | | | | |
| we | take | the | partial | | derivative | of | Eq. | (10) | with | respect | to |
| pðmÞðH1;cÞ, set it to be zero and get the solution of pðmÞðH1;cÞ as below.  1 a  pðmÞðH1;cÞ ¼ ZðmÞ 1ðkðmÞ 1Þ  p0ðH1Þ½p0ðcÞ�1�a½qm�1ðcÞ� ð11Þ  e PN t¼1kðmÞ 1;t½ytL1ðX1 tjH1Þ�ct�þPL i¼1/ðmÞ 1;i½L1ðU1 ijH1Þ�ci�  where ZðmÞ 1ðkðmÞ 1Þ is the normalization constant and e is the exponen-  tial operation. According to [17] said, kðmÞ 1 is set by finding the  unique maximum of the following concave objective function. | | | | | | | | | | | |
| JðmÞ 1ðkðmÞ 1Þ ¼ �logZ | | | | 1ðkðmÞ 1Þ | | ð12Þ | | | | | |

Then for Eq. (9), we adopt the same analysis and obtain the

|  |  |  |
| --- | --- | --- |
| solution of qðmÞðH2;cÞ as below.  1  qðmÞðH2;cÞ ¼ ZðmÞ 2ðkðmÞ 2Þ q0ðH2Þ½q0ðcÞ�1�a½pmðcÞ�a ð13Þ  e PN t¼1kðmÞ 2;t½ytL2ðX2 tjH2Þ�ct�þPL i¼1/ðmÞ 2;i½L2ðU2 ijH2Þ�ci�  where kðmÞ 2 ¼ fkðmÞ 2;tg and /ðmÞ ¼ f/ðmÞ 2;ig are another sets of nonnega-  tive Lagrange multipliers. Like kðmÞ 1; kðmÞ is set by finding the unique  maximum of the following concave objective function. | | |
| JðmÞ 2ðkðmÞ 2Þ ¼ �logZ | 2ðkðmÞ 2Þ | ð14Þ |

As [17] said, after each iteration, we calculate the relative error

between values of Eq. (12) from two successively iterations and

that of Eq. (14), respectively, and utilize them for determining con-

vergence. When the relative errors

|  |  |  |
| --- | --- | --- |
| JðmÞ 1ðkðmÞ 1Þ � Jðm�1Þ ðkðm�1Þ  Jðm�1Þ 1 ðkðm�1Þ Þ | Þ | ð15Þ |

and

|  |  |  |
| --- | --- | --- |
| JðmÞ 2ðkðmÞ 2Þ � Jðm�1Þ ðkðm�1Þ  Jðm�1Þ 2 ðkðm�1Þ Þ | Þ | ð16Þ |

are both less than some tolerance �, the iteration ends and finally we

can get the optimal pðH1Þ and qðH2Þ. Then for a test instance

Xr ¼ fX1

the class r; X2 rg, we can use ^y1¼ sign R pðH1ÞL1ðX1 first rjH1ÞdH1

view�to get

while

^y2 ¼ signRpðH2ÞL2ðX2 rjH2ÞdH2

second�is used to compute the class label

view. Finally, we can use

^y ¼ sign x1

get the class label of Xr in the whole sample space without consider-� R pðH1ÞL1ðX1 rjH1ÞdH1 þ x2 R pðH2ÞL2ðX2 rjH2ÞdH2� to

ing the view spaces where x1 and x2 are the weights of views

respectively.

cretely speaking, (a) for each data set, we run the compared learn-ing machine for 10 times and 70% instances for each data set are chosen in random for training and the remaining are chosen for test. In the training set, we randomly choose 30% as the labeled instances while the left 70% are treated as unlabeled instances; (b) for SSMVMED, (b-1) when we compute the weights of views and features, the setting can be referred to [37], i.e., exponential parameter p is selected from the set f1; 2; . . . ; 30g and balancing parameter b is initialized to be 0:1, the cluster number equals to be the class number, the maximum iteration times is 300;xv ¼1 V, and sv l¼ ~~1 dv~~ ; 8l ¼ 1; 2; . . . ; dv; 8v ¼ 1; 2; . . . ; V; (b-2) when we gener-ate the additional unlabeled instances, the used approaches can refer to Table 1. In terms of the number of instances which locates farthest or nearest from the center K is selected from the set f1; ::; Ne�maxg where Ne�max ¼ Nt � Nmax. Nt is the total number of training instances and Nmax is the number of instances from largest training class. For example, a training data set consists of three classes, one has 100 instances, another has 120 instances, and

|  |
| --- |
| the third has 140 instances, then Ne�max ¼ 220; (b-3) since we have  given the solution of pðH1Þ and qðH2Þ and furthermore, we also have given the way to predict a test instance with the optimal  pðH1Þ and qðH2Þ, thus it is found that the key of solution is the  expressions of L1ðX1 tjH1Þ and L2ðX2 tjH2Þ. In our practical experi-  ments, we adopt linear classifier assumptions, i.e.,  L1ðX1 tjH1Þ ¼ hT 1X1 tþ b1 and L2ðX2 tjH2Þ ¼ hT 2X2 tþ b2. Furthermore, in  our experiments, we refer to [17] and suppose that  p0ðH1;cÞ ¼ p0ðH1Þp0ðcÞ ¼ p0ðh1Þp0ðb1Þp0ðcÞ and q0ðH2;cÞ ¼  q0ðH2Þq0ðcÞ ¼ q0ðh2Þq0ðb2Þq0ðcÞ where p0ðh1Þ and q0ðh2Þ are satis-fied with Gaussian distributions with mean 0 and standard devia-  tion I; p0ðb1Þ and q0ðb2Þ are set to non-informative Gaussian  distributions, and p0ðcÞ ¼Q p0ðctÞQ p0ðciÞ and q0ðcÞ ¼Q q0ðctÞ  i¼1  p2p   q0ðciÞ. Here, p0ðctÞ ¼ q0ðctÞ ¼  e�c2 2ð1�ciÞ2which are Gaussian priors with mean 1 that encour-p2p cffiffiffiffi e�c2 2ð1�ctÞ2and p0ðciÞ ¼ q0ðciÞ ¼  ages cffiffiffiffi large margins where c is selected from the set  f21; 22; . . . ; 215g. Finally, for SSMVMED, the tolerance � is initialized to be0:001. |

4.1. Comparison for test accuracy

First, we will show the effectiveness of the proposed SSMVMED on test accuracy. Since related experiments have been validated that SMVMED outperforms MED, MVMED, and AMVMED [17] and the effectiveness of MED-related learning machines have also been validated compared with the traditional multi-view learning machines due to MED-related learning machines consider the uncertainties over model parameters [13–16], thus the used com-pared learning machine is SMVMED here. Experimental setting of SMVMED can be referred to the one of SSMVMED (see (b-3)) since the parameters of SSMVMED include the ones of SMVMED. More-over, the used data sets are six multi-view data sets Course, Cite-

4. Experiments seer, Cora, WebKB, NewsGroup, and Reuters. Information of them

is summarized in Table 2 where C represents the class number.

In order to validate that SSMVMED has a better performance, we conduct our experiments on five parts. They are (1) comparison for test accuracy, (2) comparison for training time and computa-tional complexity, (3) comparison between different additional unlabeled instances generation approaches in terms of test accu-racy and training time, (4) application to estimation problem, and (5) application to regression problem.

The used data sets and compared learning machines are given in related experimental contents and we will show the common experimental settings and the setting for our SSMVMED. Con-

(1) Course data set [10] is used to describe web pages and we want to predict whether the given web page is a course page or not; (2) Citeseer and Cora data sets both consist of 4 views and we choose view content and cites here [38]; (3) WebKB data set consists of web pages collected from four universities: Cornell, Texas, Wiscon-sin and Washington which have 5 categories, i.e., student, project, course, stuff and faculty. Data in WebKB are described with two views: content and citation. We treat WebKB in four separate data sets grouped by universities [39]; (4) NewsGroup data set [40] is of six groups extracted from the 20-Newsgroup dataset, i.e., M2, M5,

C. Zhu, Z. Wang / Applied Computing and Informatics 15 (2019) 172–181 177

Table 2

Brief data set description.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data set | C | N | V | dv(v ¼ 1; 2; . . . ; V) |
| Course | 2 | 1051 | 2 | 66, 5 |
| Citeseer | 6 | 3264 | 2 | 3703, 3264 |
| Cora | 7 | 2708 | 2 | 1433, 2708 |
| Cornell | 5 | 195 | 2 | 1703, 195 |
| Texas | 5 | 185 | 2 | 1703, 185 |
| Washington | 5 | 217 | 2 | 1703, 217 |
| Wisconsin | 5 | 262 | 2 | 1703, 262 |
| News-M2 | 2 | 1200 | 3 | 2000, 2000, 2000 |
| News-M5 | 5 | 500 | 3 | 2000, 2000, 2000 |
| News-M10 | 10 | 500 | 3 | 2000, 2000, 2000 |
| News-NG1 | 2 | 500 | 3 | 2000, 2000, 2000 |
| News-NG2 | 5 | 400 | 3 | 2000, 2000, 2000 |
| News-NG3 | 8 | 1000 | 3 | 2000, 2000, 2000 |
| Reuters | 6 | 1600 | 5 | 2000, 2000, 2000, 2000, 2000 |

M10, NG1, NG2, NG3. Every group contains 10 data sets, and we choose the first set for all six groups in our experiments (see

enhancement. Moreover, from the standard deviation values, it is found that the performance of SSMVMED is more stable than

Table 2). For each data set, there are three views, Partitioning SMVMED.

Around Methods, Supervised Mutual Information, and Unsuper-vised Mutual Information; (5) In terms of Reuters, it is the abbre-viation of Reuters RCV1/RCV2 Multilingual and this data set consists of machine translated documents which are written in five different languages [41,42]. These five languages are English (EN), French (FR), German (GR), Italian (IT), and Spanish (SP). Each lan-guage is treated as a view of this Reuters data set and each docu-ment can be translated from one language to another language. For this data set, the documents are also categorized into six differ-ent topics, i.e., six classes. They are C15, CCAT, E21, ECAT, GCAT,

Furthermore, in order to validate the difference between SSMVMED and SMVMED is significant, we adopt paired t-test [45] (paired t-test is different from t-test) and Nemenyi statistical test [46] for quantitative evaluation analysis. Paired t-test is used to analyze if the differences between two compared learning machines on one data set are significant or not. Then for Nemenyi statistical test, it is used to analyze if the differences between two compared learning machines on multiple data sets are significant or not. Nemenyi is different from another famous test, i.e., Fried-man statistical test which is used to analyze if the differences

M11. between all compared learning machines on multiple data sets

Table 3 shows the related experimental results about the com-parison for test accuracy between SSMVMED and SMVMED. From this table, it is found that the proposed SSMVMED outperforms SMVMED in average since on 12 data sets, SSMVMED has a better performance. The average accuracy of SSMVMED has a 2%

Table 3   
Test accuracies (average value ± std.) compared with SMVMED. Last row list the win/ tie/lose counts of SSMVMED on all data sets with t-test against SMVMED at significance level 95%. The best performance on each data set is in bold. (H) indicates the sig-value with paired t-test.

|  |  |  |
| --- | --- | --- |
| Data set | SSMVMED | SMVMED |
| Course | 0.957 ± 0.008 | 0.943 ± 0.011 (0.032) |
| Citeseer | 0.737 ± 0.008 | 0.714 ± 0.008 (0.041) |
| Cora | 0.827 ± 0.002 | 0.815 ± 0.010 (0.047) |
| Cornell | 0.788 ± 0.021 | 0.760 ± 0.041 (0.023) |

are significant or not. Since the number of compared learning machines here is two, thus we adopt Nemenyi statistical test rather than Friedman statistical test. In generally, the differences always indicate the ones in test accuracy. Thus, here we conduct quantita-tive evaluation analysis in terms of test accuracy.

(A) For paired t-test [45], we use sig-value to represent the sig- nificant differences of test accuracy. When sig-value is less than 0.05, it indicates that the compared two learning machines have a significant difference in the test accuracy on one data set. Furthermore, the difference is more signifi- cant when the sig-value is smaller. According to Table 3, we use between SSMVMED and SMVMED. From the table, we find ðHÞ indicates the sig-value with the comparison

that the difference between SSMVMED and SMVMED is sig-nificant in average.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Texas | 0.797 ± 0.014 | | 0.785 ± 0.037 (0.031) | | (B) For Nemenyi statistical test, the performance of two learning | | | | | | |
| Washington | 0.826 ± 0.019 | | 0.816 ± 0.039 (0.026) | | machines on all data sets is significantly different if the cor- | | | | | | |
| Wisconsin | 0.884 ± 0.011 | | 0.868 ± 0.023 (0.043) | |
| responding average ranks differ by at least the critical | | | | | | |
| News-M2 | 0.967 ± 0.003 | | 0.972 ± 0.010 (0.057) | |
| difference | | | | | | |
| News-M5 | 0.989 ± 0.011 | | 0.965 ± 0.015 (0.037) | |
| News-M10 | 0.855 ± 0.003 | | 0.829 ± 0.017 (0.033) | | CD ¼ qa | | r ffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffi | | ð17Þ | | |
| News-NG1 | 0.929 ± 0.001 | | 0.931 ± 0.025 (0.049) | |
| News-NG2 | 0.953 ± 0.006 | | 0.912 ± 0.009 (0.012) | |
| News-NG3 | 0.915 ± 0.006 | | 0.901 ± 0.009 (0.016) | |
| Reuters | 0.783 ± 0.001 | | 0.753 ± 0.016 (0.033) | |
| Average | 0.872 ± 0.008 | | 0.855 ± 0.019 | | where critical value qa is given in Table 4, N is the number of data sets, k is the number of learning machines. | | | | | | |
| W/T/L | SSMVMED vs. SMVMED | | 13/1/0 | |
| Table 4 | | | | | | | | p   ffiffiffi | . | 9 | 10 |
| Critical values for the two-tailed Nemenyi test. Each critical value qa is based on the studentized range statistic [46] divided by | | | | | | | |
| No. learning machines | | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| q0:05 | 1.960 | | 2.343 | 2.569 | 2.728 | 2.850 | 2.949 | | 3.031 | 3.102 | 3.164 |
| q0:10 | 1.645 | | 2.052 | 2.291 | 2.459 | 2.589 | 2.693 | | 2.780 | 2.855 | 2.920 |

178 C. Zhu, Z. Wang / Applied Computing and Informatics 15 (2019) 172–181

According to Table 3, the average rank of SSMVMED 1:1429 is while the one of SMVMED is 1:8571. Then for Nemenyi statistical

|  |
| --- |
| test, if a ¼ 0:05, the critical value q0:05 is 1:960 (see Table 4) and the corresponding CD is 1:960  since  between the average rank of SSMVMED and the one of SMVMED 1:1429 þ 0:5238 ¼ 1:6667 < 1:8571, q ffiffiffiffiffiffiffiffiffiffiffiffi ¼ 0:5238. Under such a case, so the difference  is significant. When a ¼ 0:10, the critical value q0:10 is 1:645 (see Table 4) and the corresponding CD is 1:645  such a case, since 1:1429 þ 0:4396 ¼ 1:5825 < 1:8571, so we get the similar conclusion. q ffiffiffiffiffiffiffiffiffiffiffiffi ¼ 0:4396. Under |

In other words, with paired t-test and Nemenyi statistical test, we can validate the effectiveness of SSMVMED from the quantita-tive evaluation analysis aspect.

4.2. Comparison for training time and computational complexity

Here, we still adopt SMVMED for comparison about the training time and computational complexity. As we know, compared with SMVMED, our proposed SSMVMED has to generate additional unlabeled instances. Thus SSMVMED has to cost more training time theoretically and Table 5 validates that. From this table, we can find that SSMVMED costs more training time than SMVMED. But the extra time is not more than 10% which is acceptable for us. In other words, we can achieve better test accuracy with only a little extra time with SSMVMED adopted.

In order to validate the higher training time, we give the com-putational complexity of them theoretically. As we know, in SSMVMED, it consists of three steps. So here, we will discuss the computational complexities for different steps. For convenience, we let the number of labeled instances be N, the number of original unlabeled instances be L, the number of additional unlabeled instances be U.

For the first step, the computation focuses on xv and sv l. In terms of xv, its computational complexity depends on Dv and the computational complexity of Dv is OðMðN þ LÞððdvÞ 2 þ 2dvÞÞ. In terms of sv l, its computational complexity depends on Bv lwhich computational complexity is OðMðN þ LÞÞ. Thus, for the first step, the computational complexity is OðVMðN þ LÞððdvÞ 2þ of v¼1dv; M is the number of clus- views. Here, for dv in 2dvÞÞ þ OðdMðN þ LÞÞ where d ¼PV

OðVMðN þ LÞððdvÞ 2 þ 2dvÞÞ þ OðdMðN þ LÞÞ, we can select the lar-gest dvand in fact, this selection won’t influence too much.

For the second step, the computational complexity consists of three parts. For ‘A’, if we select ‘all’, the computational complexity is OððN þ LÞdVÞ; if we select ‘unlabeled’, the computational com-plexity is OðLdVÞ; if we select ‘labeled’, the computational com-

Table 5   
Comparison about average training time (in seconds).

plexity is OðNdVÞ. For ‘B’, if we select ‘near’ or ‘far’, the

computational complexity is OðNðN � 1ÞdV=2Þ; if we select ‘all’,

we can omit the related computational complexity. For ‘C’, if we

select ‘mid’, the computational complexity is OðKdVÞ or

OððN þ LÞdVÞ which depends on the selection of ‘B’ and K is the number of selected instances; if we select ‘self’, we can also omit

the related computational complexity. As a result, the total compu-

tational complexity of the second step is ½minfOðLdVÞ; OðNdVÞg;

maxfOððN þ LÞdVÞ þ OðNðN � 1ÞdV=2Þ þ OðKdVÞ; 2OððN þ LÞdVÞg�. For the third step, the computational complexity is similar with

the one of SMVMED. In this step, the computational complexity of

SSMVMED is OððN þ UÞ2Þ while the one of SMVMED is OððN þ LÞ2Þ. Totally speaking, the computational complexity of SSMVMED is

OðVMðN þ LÞððdvÞ 2 þ 2dvÞÞ þ OðdMðN þ LÞÞ þ ½minfOðLdVÞ; OðNdVÞg;

maxfOððN þ LÞdVÞ þ OðNðN � 1ÞdV=2Þ þ OðKdVÞ; 2OððN þ LÞdVÞg�þ

OððN þ UÞ2Þ and compared with SMVMED, the extra computational

complexity is OðVMðN þ LÞððdvÞ 2 þ 2dvÞÞ þ OðdMðN þ LÞÞþ

½minfOðLdVÞ; OðNdVÞg; maxfOððN þ LÞdVÞ þ OðNðN � 1ÞdV=2Þþ

OðKdVÞ; 2OððN þ LÞdVÞg� þ OððN þ UÞ2Þ� OððN þ LÞ2Þ. From this result, it looks like that our proposed SSMVMED seems to cost more

training time, but compared with OððN þ UÞ2Þ and

OððNþLÞ2Þ;OðVMðNþLÞððdvÞ 2 þ2dvÞÞþOðdMðNþLÞÞþ½minfOðLdVÞ;

OðNdVÞg;maxfOððN þ LÞdVÞ þ OðNðN � 1ÞdV=2Þ þ OðKdVÞ;2OððN þ LÞ

dVÞg� won’t influence too much due to N þ U and N þ L is always

much larger than d;V;M;K.

Now according to the theoretical analysis about computational

complexity, we can also validate the conclusion derived from this

experimental item that with SSMVMED adopted, we can achieve

better test accuracy with only a little extra time.

4.3. Comparison between different additional unlabeled instances

generation approaches in terms of test accuracy and training time

Here, we will discuss the difference between additional unla-

beled instances generation approaches which are given in Table 1

in terms of test accuracy and training time. Figs. 1 and 2 show the

test accuracy and training time of different approaches given in

Table 1 on the used data sets. In these figures, orders in

‘Approaches in Table 1’ represent the approaches in Table 1, i.e.,

1-\_U1�1, 2-\_U1�2, 3-\_U1�3, 4-\_U1�4, 5-\_U1�5, 6-\_U2�1, 7-\_U2�2, 8-\_U2�3,

9-\_U2�4, 10-\_U2�5, 11-\_U3�1, 12-\_U3�2, 13-\_U3�3, 14-\_U3�4, and 15-\_U3�5.

Then from these figures, it is found that different additional unla-

beled instances generation approaches bring different perfor-

mances. We find that for\_U1�x ðx ¼ 1; 2; 3; 4; 5Þ approaches, they

can get better test accuracies compared with the\_Uy�x approaches

where x ¼ 1; 2; 3; 4; 5 and y ¼ 2; 3 in average. Moreover, we find that take ‘mid’ strategy for experiments has a better test accuracy

than ‘self’. In terms of training time, approaches with taking all

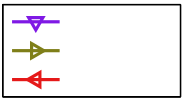
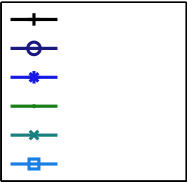
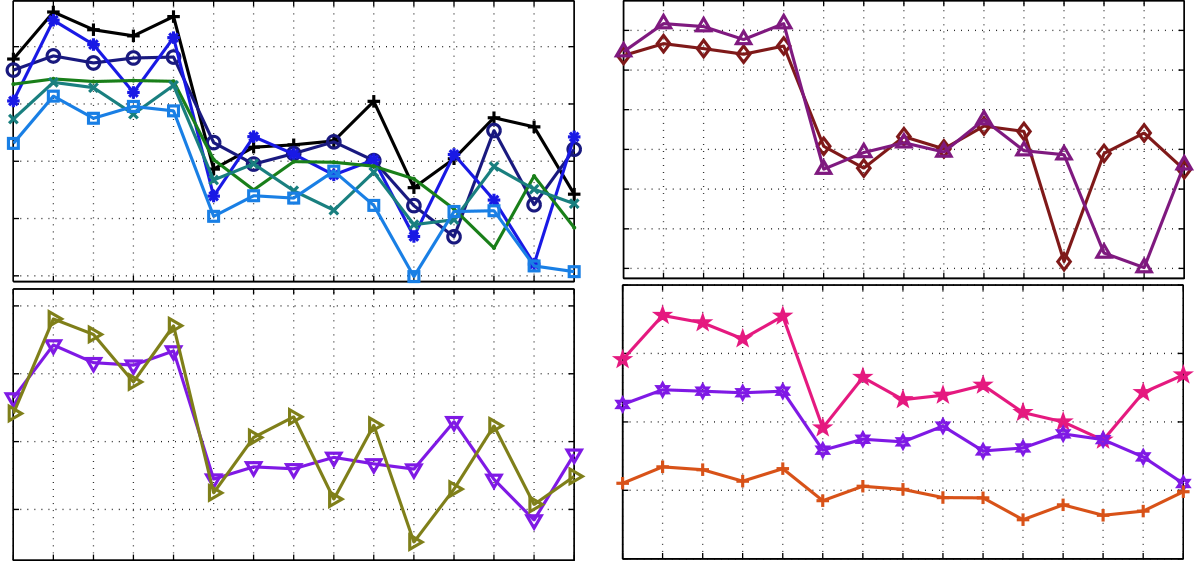
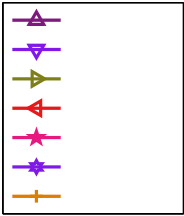
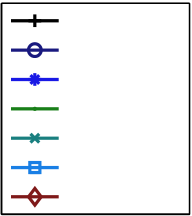
instances as the center bring longer training time due to all

instances rather than labeled or unlabeled instances are used to

|  |  |  |  |
| --- | --- | --- | --- |
| Data set | SSMVMED | SMVMED | compute the center. |
| Course | 2.249 | 2.170 | 4.4. Application to estimation problem |
| Citeseer | 290.545 | 277.717 |
| Cora | 212.896 | 201.727 |
| Cornell | 91.870 | 87.120 | Here, we apply our SSMVMED to estimation problem so as to |
| Texas | 87.832 | 84.980 |
| Washington | 2.933 | 2.677 | validate its effectiveness. The estimation problem discussed here |
| Wisconsin | 1.099 | 1.010 | aims to forecast the demographic trends which has also been dis- |
| News-M2 | 30.396 | 29.913 |
| cussed in [43]. For the forecast, USs Census population data is used. |
| News-M5 | 258.455 | 241.120 |
| In the experiments, we predict the demographic distribution in the |
| News-M10 | 275.072 | 263.533 |
| News-NG1 | 199.226 | 183.350 | year 2010 based on the historical data in years 2000 and 2006. The |

|  |  |  |
| --- | --- | --- |
| News-NG2 | 176.051 | 173.270 |
| News-NG3 | 259.336 | 259.153 |
| Reuters | 234.180 | 222.730 |

prediction is then compared with the actual Census data in the year 2010. This setting is same as the one in [43]. In [43], the authors conducted the estimation experiments with three stages



|  |  |  |  |
| --- | --- | --- | --- |
| Course | test accuracy | C. Zhu, Z. Wang / Applied Computing and Informatics 15 (2019) 172–181 | 179 |
| test accuracy comparison  1  0.7  0.6  0.5  0.4  0.9  0.8  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15  1  0.7  0.6  0.5  0.9  0.8  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 |
| Citeseer |
| Cora |
| Cornell |
| Texas |
| Washington |
| Wisconsin |
| News−M2 |
| News−M5 |
| News−M10 |
| News−NG1 |
| News−NG2 |
| News−NG3 |
| Reuters |

Approaches in Table 1

Fig. 1. Comparison between different additional unlabeled instances generation approaches in terms of test accuracy.

training time comparison

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Course | training time (s) | 300 | 1 | 2 | 3 | 4 | 200 | | | | | | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 11 12 13 14 |
| Citeseer |
| 180 | | | | | | |
| Cora |
| Cornell | 250 | 160 | | | | | | |
| Texas | 200 | 140 | | | | | | |
| Washington |
| 120 | | | | | | |
| Wisconsin | 150 | 100 | | | | | | |
| News−M2 | 100 |
| 5 | 6 | 7 | 8 | 9 | 10 11 12 13 14 | 80  4 |
| News−NG2 | 120 |
| News−NG3 | 2 | 3 | 4 | 3 | | | | | | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 11 12 13 14 15 |
| Reuters | 100 |
| News−M5 | 2 | | | | | | |
| 80 |
| News−M10 |
| News−NG1 | 1 | | | | | | |
| 60 |
| 5 | 6 | 7 | 8 | 9 | 10 11 12 13 14 15 | 0  1 |
| Approaches in Table 1 | | | | | | | Approaches in Table 1 | | | | | |

Fig. 2. Comparison between different additional unlabeled instances generation approaches in terms of training time.

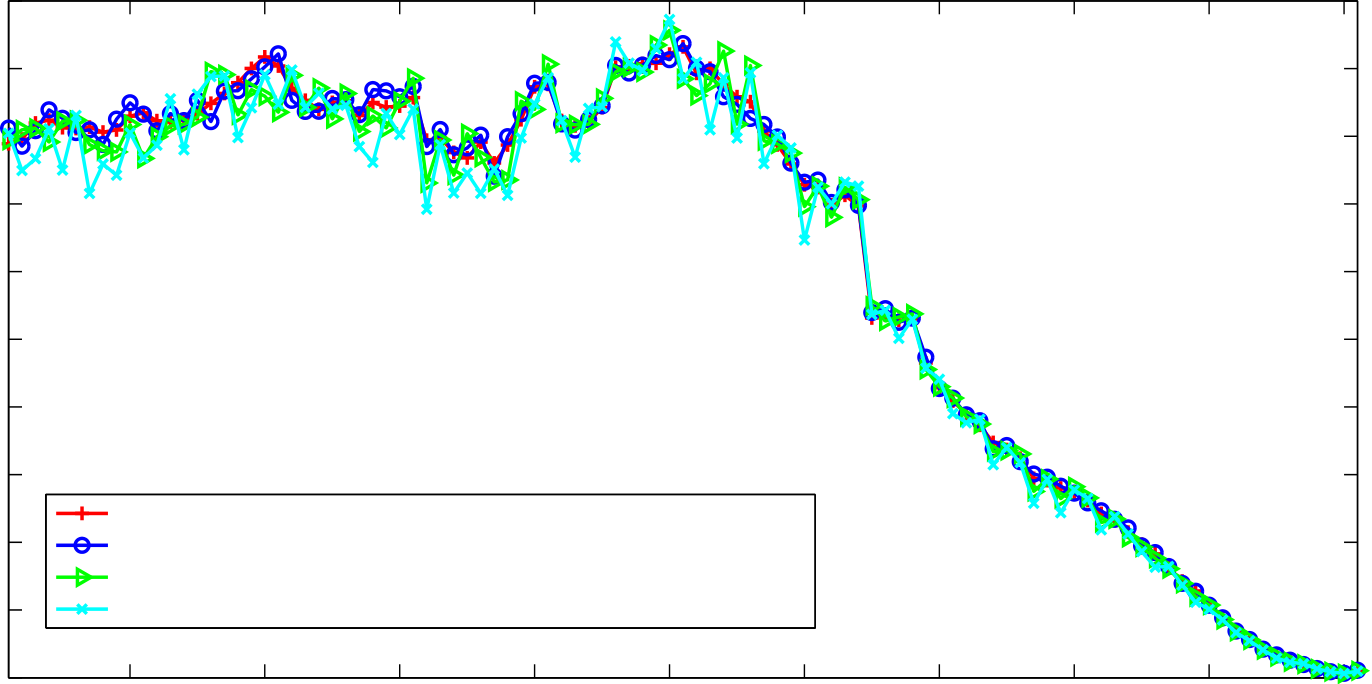
and showed the experimental results with 12 figures, while since the limitation of paper length, we won’t show all the results. For convenience, we will only show the average difference between the estimation and the real values of age in year 2010. SMVMED and the proposed method in [43] (we call it EDT) are used for com-parison. Fig. 3 shows the related experiments. From this figure, it is found that compared with SMVMED and EDT, the predicted age distribution of our SSMVMED can accord with the real age distribu-tion to a large extent.

4.5. Application to regression problem

Here, we apply SSMVMED to regression problem. The regres-sion problem discussed here aims to estimate full joint distribu-tions from incomplete information which has also been discussed in [44]. In [44], authors proposed generalized cross entropy model (GCEM) and estimated the distribution of Singapore household profile (<http://www.singstat.gov.sg>) with the joint probability dis-tribution of the household dwelling type (HD), household size (HS) and home ownership (HO) measures. In those experiments, authors conducted the experiments on three different cases: (1) pure entropy with constraints, (2) minimum discrimination infor-mation without constraints, and (3) minimum discrimination information with constraints. Moreover, authors adopted (a) accu-

racy heat maps to compare the accuracy of the three cases, (b) KullbackCLeibler (KL) distance to compare the estimated joint dis-tribution and the observed one, and (c) Linfoots measures to com-pare a wide range of spatiotemporal signals from brain waves to human dynamics. For accuracy heat maps, if the estimated heat map is close to the true heat map, we say the estimation is better. For KL distance, the smaller the KL distance between any two dis-tributions is, the closer are their profiles. For Linfoots measures, it has three indexes, C measures the relative structural content, F looks at the fidelity or peak alignment, and Q reflects the correla-tion quality [44]. If these three indexes are more closer to 1, then we say the estimated distribution is more closer to the true distri-bution. From those experiments, it was found that case 1 and case 2 had comparable performance and case 3 had a best performance.

In our experiments, for the limitation of paper length, we won’t conduct all experiments given in [44]. We will only adopt GCEM and SSMVMED for comparison on the Singapore household profile when case 3 is considered. In order to show the results clearly, we combine the results of accuracy heat maps (here, the differences between estimated heat maps and the true ones are given), KL dis-tance, and Linfoots measures together and only show that whether SSMVMED brings a better performance or not. Table 6 shows the results in a simple manner. From this table, it is found that with SSMVMED, the error of estimation is decreased and the difference



|  |  |  |
| --- | --- | --- |
| 180 | 6  5x 10 | C. Zhu, Z. Wang / Applied Computing and Informatics 15 (2019) 172–181 |
| difference between the estimation and the real values of age in year 2010 |

4.5

4

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| number of people | 3.5 | real population distribution in the year 2010 | | | | | | 70 | 80 | 90 | 100 |
| 3 |
| 2.5 |
| 2 |
| 1.5 |
| 1 | estimated population distribution in the year 2010 by SSMVMED | | | | | |
| 0.5 | estimated population distribution in the year 2010 by SMVMED | | | | | |
| estimated population distribution in the year 2010 by EDT | | | | | |
| 0 | 10 | 20 | 30 | 40 | 50 | 60 |
| age | | | | | |

Fig. 3. The predicted age distribution of US population for the year 2010 based on the population data in the years 2000 and 2006 with SSMVMED, SMVMED, and EDT used.

Table 6   
Comparison between GCEM and SSMVMED on Singapore household profile with minimum discrimination information with constraints in terms of the results of heat maps, KL distance, and Linfoots measures.

|  |  |  |
| --- | --- | --- |
| Measures | SSMVMED | SMVMED |

SMVMED, the average test accuracy of SSMVMED has a 2% enhancement; (2) SSMVMED costs more training time than SMVMED and the extra time is not more than 10% which is accept-able for us; (3) in terms of the generation of additional unlabeled instances, ‘mid’ strategy has a better test accuracy than ‘self’ and

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy heat maps | C | 0.017 | 0.012 | taking all instances to get the center brings a better test accuracy |
| KL distance | 0.0039 | 0.0028 | as well; (4) with SSMVMED, the applications to estimation prob- |
| Linfoots measures | 1.016 | 1.003 |
| lem and regression problem will be more feasible. |
| Q | 0.9991 | 0.9994 |
| F | 1.007 | 1.004 |

Acknowledgment

between the estimated distribution and the true distribution is more smaller. In other words, SSMVMED brings a better regression performance.

This work is supported by (1) Natural Science Foundation of Shanghai – China under Grant No. 16ZR1414500 (2) National Nat-ural Science Foundation of China – China under Grant Nos. 61602296, 41701523, 51575336 and the authors would like to

|  |  |
| --- | --- |
| 5. Conclusions   Traditional multi-view learning machines do not consider the | thank their supports.  References |

uncertainties over model parameters. Thus maximum entropy dis-crimination (MED) and its extended versions multi-view maxi-mum entropy discrimination (MVMED) and alternative MVMED (AMVMED) are developed for this issue. While for processing multi-view data sets, they only use the hard margin consistency principle that the decision of margin parameter c is related to clas-sifier parameter H directly. As we know, the decision always be indirectly in practice. So soft margin consistency based multi-view maximum entropy discrimination (SMVMED) has been pro-posed. Although related experiments have validated the effective-ness of SMVMED, it is only adaptive to supervised problems. Indeed, in real-world, most data sets are semi-supervised, namely, the data sets consist of labeled instances and unlabeled instances. So this paper extends the model of SMVMED to the semi-supervised problems and develop a semi-supervised SMVMED (SSMVMED). Furthermore, in order to get more useful discriminant information, we propose some schemes to generate more addi-tional unlabeled instances. Moreover, these generated additional unlabeled instances will also be used in the model of SSMVMED along with the original labeled and unlabeled instances so that the performance of a learning machine can be boosted. Related experiments on multi-view data sets from different aspects have validated the effectiveness of SSMVMED theoretically and empiri-cally. From the experiments, it is found that (1) compared with

[1] R. Bach, G.R. Lanckriet, M.I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm, in: Proceedings of the 21st International Conference on Machine Learning, 2004, pp. 6–13.

[2] C. Cortes, M. Mohri, A. Rostamizadeh, Two-stage learning kernel algorithms, in: Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 239–246.

[3] [A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, SimpleMKL, J. Mach. Learn.](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0015)  [Res. 9 (2008) 2491–2521](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0015).

[4] M. Kloft, U. Brefeld, S. Sonnenburg, A. Zien, Non-sparse Regularization and Efficient Training with Multiple Kernels, 2010. Available from: arxiv preprint <arXiv:1003.0079>.

[5] M. Gönen, E. Alpaydin, Localized multiple kernel learning, in: Proceeding of the 25th International Conference on Machine Learning, 2008, pp. 352–359.

[6] G. Ye, D. Liu, I.H. Jhuo, S.F. Chang, Robust late fusion with rank minimization, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012, pp. 3021–3028.

[7] [A. Iosifidis, A. Tefas, N. Nikolaidis, I. Pitas, Multi-view human movemen](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0035)t [recognition based on fuzzy distances and linear discriminant analysis, Comput. Vis. Image Underst. 116 (3) (2012) 347–360](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0035).

[8] J. Rupnik, J. Shawe-Taylor, Multi-view canonical correlation analysis, in: Proceeding of Slovenian KDD Conference on Data Mining Data Warehouses, 2010, pp. 1–4.

[9] [X. Yin, Q. Huang, X. Chen, Multiple view locality preserving projections with](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0045)  [pairwise constraints, Commun. Syst. Inform. Technol. 100 (2011) 859–866](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0045). [10] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Eleventh Conference on Computational Learning Theory, 1998, pp. 92–100. [11] [M.L. Zhang, Z.H. Zhou, Cotrade: confident co-training with data editing, IEE](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0055)E [Trans. Syst., Man, Cybernet., Part B: Cybernet. 41 (6) (2011) 1612–1626](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0055).

[12] V. Sindhwani, P. Niyogi, M. Belkin, A co-regularization approach to semi-supervised learning with multiple views, in: Proceeding of ICML workshop on Learning With Multiple Views, 2005, pp. 74–79.

C. Zhu, Z. Wang / Applied Computing and Informatics 15 (2019) 172–181 181

[13] [T. Jaakkola, M. Meila, T. Jebara, Maximum entropy discrimination, Adv. Neural](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0065)  [Inform. Process. Syst. 12 (2000) 470–476](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0065).

[14] [T. Jebara, Machine Learning: Discriminative and Generative, Kluwer Academic,](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0070)

[29] [V. Cherkassky, W.Y. Dai, Empirical study of the Universum SVM learning for](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0145)  [high-dimensional data, Lect. Notes Comput. Sci. 5768 (2009) 932–941](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0145).

[30] D. Zhang, J. Wang, L. Si, Document clustering with Universum, in: International

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [2004](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0070). | | | | | Multi-view | | maximum | | entropy | | discrimination, | | | in: | Conference on Research and Development in Information Retrieval, 2011, pp. |
| [15] S.L. | Sun, | G.Q. Chao, | | | 873–882. |
| Proceedings | | | of | the | 23rd | International | | Joint | | Conference | | on | Artificial | | [31] [B. Peng, G. Qian, Y.Q. Ma, View-invariant pose recognition using multilinear](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0155) |

Intelligence, 2013, pp. 1706–1712.

[16] [G.Q. Chao, S.L. Sun, Alternative multiview maximum entropy discrimination,](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0080)  [IEEE Trans. Neural Networks Learn. Syst. 99 (2015) 1–12](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0080).

[17] L. Mao, S.L. Sun, Soft margin consistency based scalable multi-view maximum entropy discrimination, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 1839–1845.

[18] G. Tzortzis, A. Likas, Kernel-based weighted multi-view clustering, in: 2012 IEEE 12th International Conference on Data Mining, 2012, pp. 675–684.

[19] [S.L. Sun, Q.J. Zhang, Multiple-view multiple-learner semi-supervised learning,](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0095)

[analysis and the Universum, Adv. Visual Comput. 5359 (2008) 581–591](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0155).

[32] [C. Shen, P. Wang, F. Shen, H. Wang, Uboost: boosting with the Universum, IEEE](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0160)  [Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 825–832](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0160).

[33] [X.H. Chen, S.C. Chen, H. Xue, Universum linear discriminant analysis, Electron.](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0165)  [Lett. 48 (22) (2012) 1407–1409](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0165).

[34] [Z. Wang, Y.J. Zhu, W.W. Liu, Z.H. Chen, D.Q. Gao, Multi-view learning with](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0170)  [Universum, Knowl.-Based Syst. 70 (2014) 376–391](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0170).

[35] J. Weston, R. Collobert, F. Sinz, L. Bottou, V. Vapnik, Inference with the Universum, in: The 23rd International Conference on Machine Learning, 2006,

[Neural Process. Lett. 34 (2011) 229–240](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0095). pp. 1009–1016.

[20] [M.Q. Deng, C. Wang, Q.F. Chen, Human gait recognition based on deterministic](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0100)  [36] [D.L. Liu, Y.J. Tian, R.F. Bie, Y. Shi, Self-Universum support vector machine, Pers.](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0180)

[learning through multiple views fusion, Pattern Recogn. Lett. 78 (C) (2016) 56–](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0100) [Ubiquit. Comput. 18 (2014) 1813–1819](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0180).

[63](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0100). [37] [Y.M. Xu, C.D. Wang, J.H. Lai, Weighted multi-view clustering with feature](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0185)

[21] [F. Wu, X.Y. Jing, X.G. You, D. Yue, R.M. Hu, J.Y. Yang, Multi-view low-rank](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0105)  [selection, Pattern Recogn. 53 (2016) 25–35](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0185).

[dictionary learning for image classification, Pattern Recogn. 50 (2016) 143](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0105)– [38] [P. Sen, G.M. Namata, M. Bilgic, L. Getoor, B. Gallagher, T. Eliassi-Rad, Collective](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0190)

[154](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0105). [classification in network data, AI Mag. 29 (3) (2008) 93–10](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0190)6.

[22] [S.H. Zhu, X. Sun, D.L. jin, Multi-view semi-supervised learning for image](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0110)  [classification, Neurocomputing 208 (2016) 136–14](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0110)2.

[23] [H.Y. Wang, X. Wang, J. Zheng, J.R. Deller, H.Y. Peng, L.Q. Zhu, W.G. Chen, X.L. Li, R.J. Liu, H.J. Bao, Video object matching across multiple non-overlapping camera views based on multi-feature fusion and incremental learning, Pattern](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0115)

[39] H.J. Ye, D.C. Zhan, Y. Miao, Y. Jiang, Z.H. Zhou, Rank consistency based multi-view learning: a privacy-preserving approach, in: ACM International on Conference on Information and Knowledge Management, 2015, pp. 991-1000. [40] G. Bisson, C. Grimal, Co-clustering of multi-view datasets: a parallelizable approach, in: Proceedings of the IEEE 12th International Conference on Data

[Recogn. 47 (12) (2014) 3841–3851](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0115). Mining, 2012, pp. 828-833.

[24] [R. Sheikhpour, M.A. Sarram, S. Gharaghani, M.A.Z. Chahooki, A survey on semi-](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0120) [41] M.R. Amini, N. Usunier, C.Goutte, Learning from multiple partially observed

[supervised feature selection methods, Pattern Recogn. 64 (2017) 141–158](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0120). viewsan application to multilingual text categorization, in: Neural Information

[25] T.P. Xie, N.M. Nasrabadi, I.A.O. Hero, Semi-supervised multi-sensor Processing Systems (NIPS), 2009, pp. 28–36.

classification via consensus-based multi-view maximum entropy [42] <http://multilingreuters.iit.nrc.ca/ReutersMultiLingualMultiView.htm>.

discrimination, in: IEEE International Conference on Acoustics, Speech and [43] [G.Q. Li, D.X. Zhao, Y. Xu, S.H. Kuo, H.Y. Xu, N. Hu, G.S. Zhao, C. Monterola,](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0215)

Signal Processing, 2015, pp. 1936–1940. [Entropy based modelling for estimating demographic trends, Plos One 10.9](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0215)

[26] A.N. Erkan, Y. Altun, Y.W. Teh, M. Titterington, Semi-supervised learning via [(2015) e0137324](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0215).

generalized maximum entropy, in: International Conference on Artificial Intelligence and Statistics, 2010, pp. 209–216.

[27] G.Q. Chao, S.L. Sun, Semi-supervised multitask learning via self-training and maximum entropy discrimination, in: Neural Information Processing, Springer

[44] [H.Y. Xu, S.H. Kuo, G.Q. Li, E.F.T. Legara, D.X. Zhao, C.P. Monterola, Generalized Cross Entropy Method for estimating joint distribution from incomplete information, Physica A 453 (2016) 162–172](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0220).

[45] [V.N. Vapnik, Statistical Learning Theory, Wiley Interscience Press, New York,](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0225)

Berlin/Heidelberg, 2012. [1998](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0225).

[28] [V. Vapnik, S. Kotz, Estimation of Dependences based on Empirical Data,](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0140)  [46] [J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0230).

[Springer, 1982](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0140). [Learn. Res. 7 (2006) 1–30](http://refhub.elsevier.com/S2210-8327(17)30241-7/h0230).