Transductive Multi-view Modeling with Interpretable Rules, Matrix Factorization and Cooperative Learning

Wei Zhang, Zhaohong Deng, *Senior Member, IEEE*, Jun Wang, Kup-Sze Choi, Te Zhang, Xiaoqing Luo, Hongbin Shen, Shitong Wang

***Abstract*--Multi-view fuzzy systems aim to deal with fuzzy modeling in multi-view scenarios effectively and to obtain interpretable model through multi-view learning. However, current studies of multi-view fuzzy systems still face several challenges, one of which is how to achieve efficient collaboration between multiple views when there are few labeled data. To address this challenge, this paper explores a novel transductive multi-view fuzzy modeling method. The dependency on labeled data is reduced by integrating transductive learning into the fuzzy model to simultaneously learn both the model and the labels using a novel learning criterion. Matrix factorization is incorporated to further improve the performance of the fuzzy model. In addition, collaborative learning between multiple views is used to enhance the robustness of the model. The experimental results indicate that the proposed method is highly competitive with other multi-view learning methods.**

***Index Terms* - transductive multi-view learning, fuzzy system, collaboratively learning, matrix factorization.**

I. Introduction

THE complexity of data is ever increasing with rapid development of data acquisition technology, which brings great challenges to traditional machine learning methods. Data can be complex due to various factors, e.g. the number of subjects, distribution of data attributes, timeliness and scalability. Complex data can be described from multiple views, i.e., multiple representations of a subject. For example, in content-based retrieval applications, a target can be described by both visual and text features to represent different characteristics of the subjects.

Multi-view learning has received increasing attention in recent years. It has been applied in various fields [1, 2], especially for clustering and classification tasks.

(1) Multi-view learning for clustering: By integrating multi-kernel technology with traditional clustering, Zhao *et al*. proposed multiple kernel clustering (MKC) [3], which can simultaneously find the maximum margin hyperplane, the optimal clustering and the optimal kernel. Based on the concept of MKC, Du *et al*. proposed a new multiple kernel clustering method [4] by integrating multi-kernel K-means clustering with $L_{2,1}$ regularization. By fusing the graph structure from different views, a new multi-view clustering method was proposed [5]. Further, Zhan *et al*. proposed a new graph-based multi-view clustering method [6] by adding a rank constraint to the Laplacian matrix and a new disagreement cost function into the process of learning the consensus graph. A new method was proposed by considering not only the information from different views, but also mining the correlation between each individual feature [7]. In order to generate more discriminative representation, a method integrating deep feature extraction technique with semi-nonnegative matrix factorization (semi-NMF) was also proposed [8].

(2) Multi-view learning for classification: A multi-view support vector machine (SVM) was proposed by adding both manifold regularization and collaborative regularization between views into the SVM framework [9]. Zhang *et al*. proposed a novel image classification method by considering visual, semantic and view consistency between classifiers [10]. By using the strategy of marginal consistency under the maximum entropy discrimination (MED), Sun and Chao proposed the Multi-view MED (MVMED) algorithm [11]. Further, based on MVMED, a multi-view classification method was proposed in [12] by assigning relative entropy to each view. Low-rank discriminant embedding was also used for multi-view learning based classification [13], where the complementarity between the maximization of empirical

This work was supported in part by the NSFC under Grants 61772239 and 61772237, Jiangnan University State Key Laboratory of Food Science and Technology Free Exploration Project under Grant SKLF-ZZB-201901, the National First-Class Discipline Program of Light Industry Technology and Engineering under Grants LITE2018-02 and LITE2018-03, the Six Talent Peaks Project in Jiangsu Province under Grants XYDXX-056 and XYDXX-030, the General Project of the National Social Science Fund of China under Grant 19BTQ030, the Jiangsu Province Natural Science Fund under Grant BK20181339, the Hong Kong Innovation and Technology Commission under Grant MRF/015/18, and the Hong Kong Research Grants Council under Grant PolyU 152006/19 (Corresponding author: Zhaohong Deng).

W. Zhang, Z. H. Deng, J. Wang, X. Q. Luo and S. T. Wang are with the School of Artificial Intelligence and Computer Science, Jiangnan University and Jiangsu Key Laboratory of Media Design and Software Technology, Wuxi 214122, China (e-mail: 6181610010 @stu.jiangnan.edu.cn; dengzhaohong@jiangnan.edu.cn; wangjun_sytu@hotmail.com; xqluo@jiangnan.edu.cn, wxwangst@aliyun.com).

K.S. Choi is with The Centre for Smart Health, the Hong Kong Polytechnic University, Hong Kong (e-mail: thomasks.choi@polyu.edu.hk).

T. Zhang is with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China (e-mail: zhangt7@mail.sustech.edu.cn)

H. B. Shen is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China (e-mail: hbshen@sjtu.edu.cn).

likelihood and the preservation of geometric structure is considered.

Although various multi-view intelligent models have been developed, most of them only focus on enhancing the generalizability of the models, ignoring the importance of interpretability which is also essential in real-world applications. Therefore, development of interpretable multi-view learning method is of great significance. Fuzzy systems are well-known for possessing strong interpretability and generalizability [14-16]. Significant progress has also been made to create multi-view fuzzy systems with good interpretability. For example, a two-view Takagi-Sugeno-Kang (TSK) fuzzy system is developed by combining collaborative learning with the maximum margin learning criterions [17]. Furthermore, multi-view TSK fuzzy system is built by fully incorporating both visible and hidden views [18]. A view-reduction based multi-view TSK fuzzy system is also developed by filtering the weak views with an adaptive threshold [19].

While these methods succeed in interpretable multi-view fuzzy modeling, their practicability is still limited with challenges to be resolved. For example, the modeling methods are developed based on inductive learning framework that always requires sufficient training data. However, it is usually difficult and expensive to obtain labeled training data in many applications. The performance of existing modeling methods for multi-view fuzzy system is thus greatly reduced when labeled data are inadequate for training.

To deal with the inadequacy of labeled training data, many transductive learning methods have been proposed. The existing work includes two main categories: the *classifier generation* based transductive algorithms [20-24] and the *label propagation* based transductive algorithms [25-29]. While these algorithms achieve better performance than the traditional inductive algorithms, there are still issues that require in-depth investigations. First, the robustness of classifier generation based multi-view transductive algorithms [22-24] needs to be enhanced with more effective mechanisms. Second, the consistency information between views is often ignored in the label propagation based multi-view transductive algorithms [27-29]. When two classes of a dataset overlap significantly, the performance of this category of algorithms may degrade severely [30]. Third, the interpretability of most existing algorithms is usually poor, which calls for more interpretable methods.

To this end, a novel transductive multi-view TSK fuzzy system modeling method is proposed in this study for multi-view modeling scenes with inadequate labeled data. First, pseudo labels are assigned to the unlabeled data, and then the labeled and unlabeled data are both considered for the multi-view TSK fuzzy modeling based on transductive learning framework. Since the adopted base model, *i.e.*, TSK fuzzy system, is premised on fuzzy rules and fuzzy logic inference, the proposed transductive multi-view learning method is more interpretable. Second, the pseudo label matrix is decomposed into two matrices through optimization. The product of the two matrices can yield soft labels to effectively promote transductive learning abilities. Third, the view weighting mechanism and the collaborative learning mechanism are used to further mine the information among different views to enhance the robustness of the proposed transductive multi-view fuzzy system. The main contributions of this paper are as follows:

1) Transductive learning mechanism is introduced into the modeling of multi-view fuzzy systems, which alleviates the challenge due to inadequate labeled training samples.

2) The performance of transductive multi-view learning is enhanced by introducing the decomposition of the pseudo label matrix.

3) The inter-view collaborative learning mechanism and the view weighting mechanism are integrated into transductive multi-view fuzzy modeling.

4) The effectiveness of the proposed method is verified using real-world multi-view datasets. Outstanding performance over existing methods is demonstrated with extensive experiments.

The rest of this paper is organized as follows. In Section II, classical single-view TSK-FS model, transductive learning mechanism and traditional multi-view learning frameworks are briefly reviewed. In Section III, the details of the transductive multi-view fuzzy system for cooperative learning of labeled and unlabeled data are presented. In Section IV, the proposed method is extensively evaluated on real-world datasets. Conclusions are given in Section V.

II. BACKGROUND

A. *TSK Fuzzy System*

As a classic intelligent model, fuzzy system has good interpretability and powerful data-driven learning ability [31]. The popular fuzzy models include TSK model [32] and Mamdani model [33], with the former being more widely studied in the data driven modeling scenes. In this paper, we construct fuzzy systems for multi-view data scenes based on the TSK fuzzy system.

For a TSK fuzzy system, it contains a fuzzy rule base with the fuzzy inference rules defined as follows:

$$\begin{aligned}
&R^i : IF\ x_1\ is\ A_1^i \wedge x_2\ is\ A_2^i \wedge ... \wedge x_d\ is\ A_d^i \\
&THEN\ f_i^1(\boldsymbol{x}) = p_0^{i,1} + p_1^{i,1}x_1 + ... + p_d^{i,1}x_d, \\
&\qquad \vdots \\
&\qquad f_i^c(\boldsymbol{x}) = p_0^{i,c} + p_1^{i,c}x_1 + ... + p_d^{i,c}x_d, \\
&\qquad \vdots \\
&\qquad f_i^C(\boldsymbol{x}) = p_0^{i,C} + p_1^{i,C}x_1 + ... + p_d^{i,C}x_d \\
&\qquad\quad i = 1, 2...I, c = 1, 2 \dots C
\end{aligned} \tag{1}$$

where $\boldsymbol{x} = [x_1, x_2...x_d]^{\mathrm{T}}$ is an input vector, A_d^i is a fuzzy subset of the input space, $\wedge$ is a fuzzy conjunction operator, *I* is the number of fuzzy rules and *C* is the number of outputs. When multiplicative conjunction, multiplicative implication and additive disjunction are adopted, the output of a classical TSK fuzzy system is given by:

$$\boldsymbol{y}^o = \sum_{i=1}^{I} \tilde{\varphi}^i\left(\boldsymbol{x}\right) \boldsymbol{f}_i\left(\boldsymbol{x}\right) \tag{2a}$$

where $\boldsymbol{f}_i = [\, f_i^1(\boldsymbol{x}), f_i^2(\boldsymbol{x}),\ldots, f_i^C(\boldsymbol{x})]$, $\tilde{\varphi}^i\left(\boldsymbol{x}\right)$ is the firing strength associated with the corresponding fuzzy rule and it can be calculated as follows:

$$\tilde{\varphi}^i\left(\boldsymbol{x}\right) = \frac{\varphi^i\left(\boldsymbol{x}\right)}{\sum_{i'=1}^{I} \varphi^{i'}\left(\boldsymbol{x}\right)} \tag{2b}$$

$$\varphi^i\left(\boldsymbol{x}\right) = \prod_{j=1}^{d} \varphi_{A_j^i}\left(x_j\right) \tag{2c}$$

where $\varphi_{A_j^i}\left(x_j\right)$ is the fuzzy membership of x_j to the fuzzy subset A_j^i. By using the Gaussian function as the membership function, $\varphi_{A_j^i}\left(x_j\right)$ in (2c) can be expressed as:

$$\varphi_{A_j^i}\left(x_j\right) = \exp\left(-\left(x_j - e_j^i\right)^2 \big/ 2\delta_j^i\right) \tag{2d}$$

where parameter e_j^i and δ_j^i are the center and variance of the Gaussian function, respectively. Classical TSK fuzzy systems typically use Fuzzy C-Means Clustering (FCM) [34] to estimate these two parameters. Since FCM contains a random initialization process that does not always produce stable results, a more robust clustering algorithm, *i.e.*, Var-Part [35], is used to partition the input data and estimate the parameters in (2d). See [17, 18] for the details of the parameter estimation.

When the antecedent parameters of the rules are determined, the output of the TSK fuzzy system can be formulated as a linear output, i.e.,

$$\boldsymbol{y}^o = \boldsymbol{x}_g^{\mathrm{T}} \boldsymbol{P}_g \in R^{1\times C} \tag{3a}$$

where $\boldsymbol{x}_g$ represents the feature vector obtained through fuzzy mapping of the antecedent and $\boldsymbol{P}_g$ represents the combination of the consequent parameters of the TSK fuzzy system. These two parts can be further defined as follows:

$$\boldsymbol{x}_g = \left(\left(\tilde{\boldsymbol{x}}^1\right)^{\mathrm{T}}, \left(\tilde{\boldsymbol{x}}^2\right)^{\mathrm{T}}, \ldots, \left(\tilde{\boldsymbol{x}}^I\right)^{\mathrm{T}}\right)^{\mathrm{T}} \in R^{I(1+d)\times 1} \tag{3b}$$

$$\tilde{\boldsymbol{x}}^i = \tilde{\varphi}^i\left(\boldsymbol{x}\right) \boldsymbol{x}_e \in R^{(1+d)\times 1} \tag{3c}$$

$$\boldsymbol{x}_e = \left(1, \boldsymbol{x}^T\right)^T \in R^{(1+d)\times 1} \tag{3d}$$

$$\boldsymbol{P}_g = \left(\boldsymbol{p}^1, \boldsymbol{p}^2, \ldots, \boldsymbol{p}^C\right) \in R^{I(1+d)\times C} \tag{3e}$$

$$\boldsymbol{p}^c = \left(\left(\boldsymbol{p}_1^c\right)^{\mathrm{T}}, \left(\boldsymbol{p}_2^c\right)^{\mathrm{T}}, \ldots, \left(\boldsymbol{p}_I^c\right)^{\mathrm{T}}\right)^{\mathrm{T}} \in R^{I(1+d)\times 1} \tag{3f}$$

$$\boldsymbol{p}_i^c = \left(p_0^{i,c}, p_1^{i,c}, \ldots, p_d^{i,c}\right)^{\mathrm{T}} \in R^{(1+d)\times 1} \tag{3g}$$

B. *Transductive Learning*

Transductive learning can handle prediction problem effectively when the test data in the training procedure are known. It is an important branch of semi-supervised learning and has attracted much attention in recent years. Transductive learning methods are mainly used in classifier generation and label propagation.

1) Transductive learning based on classifier generation: This category of methods regards the label of unlabeled samples as optimization variables and iteratively updates them during the training process. For example, a transductive support vector machine (TSVM) based on low-density separation is proposed to integrate unlabeled data information into the optimization of support vector machine (SVM) [20]. To alleviate the dependence of deep neural network on a large labeled data, Shi *et al*. introduced transductive learning into deep neural network and proposed transductive semi-supervised deep learning (TSSDL) [21]. By imposing a global constraint in two classifiers, Li *et al*. proposed a two view TSVM [22]. Li *et al*. proposed a robust transductive multi-view SVM [23] by introducing margin distribution into TSVM. In [24], Zhuge *et al*. proposed a new multi-view transductive method and a novel classification loss named probabilistic square hinge loss.

2) Transductive learning based on label propagation: This category of methods employs graphical representation in which the labeled and unlabeled data are both involved in the same graph, where data samples are represented with nodes. The weight of an edge is used to represent the similarity between the two connecting nodes. Then, the labels assigned to the unlabeled samples are inferred by measuring the similarity between the training samples, whereas the labels of the labeled samples are spread to the nearby unlabeled samples. Iscen *et al*. proposed a novel method by using label propagation to refine the pseudo-labels [25]. Yi *et al*. proposed a new feature extraction method by introducing label propagation into semi-NMF. With this method, the problem that the distribution relationships between the labeled and unlabeled data cannot be used in semi-NMF [26] is solved. Karasuyama and Mamitsuka proposed Sparse Multiple Graph Integration (SMGI) [27] by integrating multiple adjacency graphs for label propagation and using the graph weights to eliminate irrelevant graphs. Nie *et al*. proposed Auto-weighted Multiple Graph Learning (AMGL) [28] by optimizing the adjacency graph and the labels at the same time. Based on AMGL, Nie *et al*. proposed Multi-view Learning with Adaptive Neighbours (MLAN) [29] and introduced an adaptive weightings mechanism to improve the performance of the model.

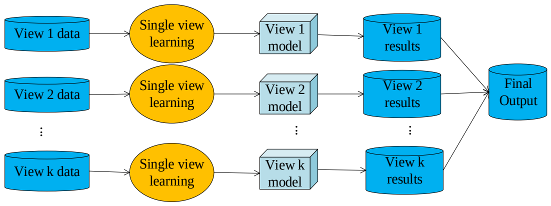


Fig. 1 Single-view learning algorithm in multi-view scenarios

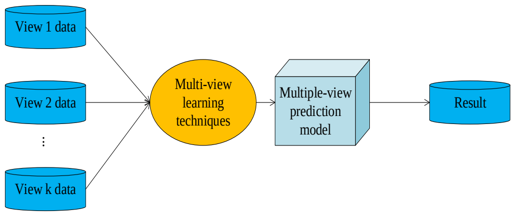

View 1 data
View 2 data
⋮
View k data
Multi-view learning techniques
Multiple-view prediction model
Result

Fig. 2 The framework of traditional multi-view learning

C. Traditional Multi-view Learning Frameworks

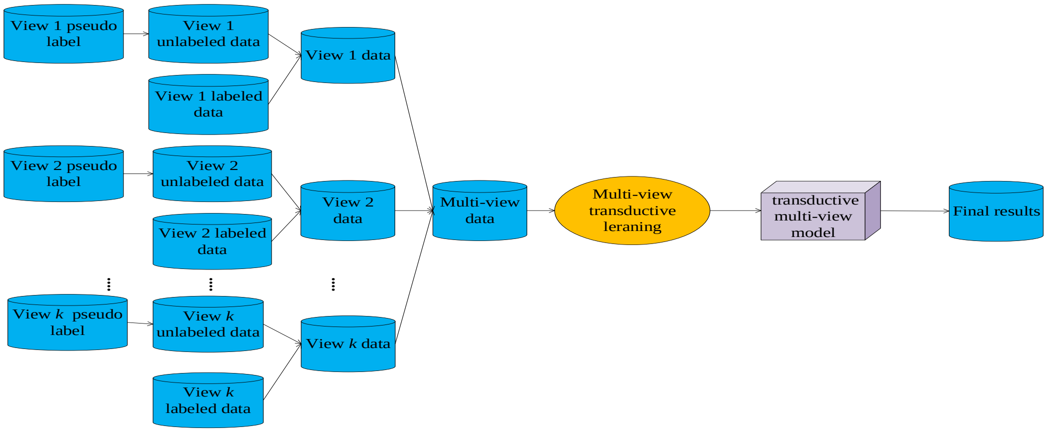


Fig. 3 The framework of the proposed transductive multi-view learning

Conventional single-view learning methods usually construct a separate model for each view and then fuse the outputs of the multiple single-view models to obtain the final results. Fig. 1 shows a conventional single-view learning methods that is applied to multi-view scenarios. However, this learning framework destroys the correlation between the views and the final performance is usually poor. Multi-view learning can make good use of the information between multiple views and thus obtain better results. Fig. 2 shows the framework of classical multi-view learning.

When sufficient labeled data are available from multi-view scenarios, the performance of existing multi-view learning methods can readily surpass the conventional single-view learning methods. However, multi-view algorithms remain unsatisfactory if the labeled data are inadequate. It is therefore necessary to explore multi-view learning methods that can work effectively even with few labeled data.

III. TRANSDUCTIVE MULTI-VIEW FUZZY SYSTEM

A. The Framework of Transductive Multi-view Learning

The novel transductive multi-view learning framework proposed in this study is illustrated with Fig. 3. First, pseudo labels are assigned to the unlabeled data. Then, transductive multi-view learning is performed by using both the labeled and unlabeled data to learn the final labels for the unlabeled data. Compared with traditional multi-view modeling, the proposed framework not only makes full use of the labeled data, but also mines the information of the unlabeled data.

B. Transductive Single-view Fuzzy System

By considering the unlabeled data in the training procedure, the learning criterion of transductive single-view fuzzy system can be formulated as:

$$\min_{\boldsymbol{P}_g,\boldsymbol{Y}^U} \frac{1}{2}\left(\left\|\boldsymbol{X}_g^L\boldsymbol{P}_g-\boldsymbol{Y}^L\right\|^2+\alpha\left\|\boldsymbol{X}_g^U\boldsymbol{P}_g-\boldsymbol{Y}^U\right\|^2\right) \tag{4a}$$

where $\boldsymbol{X}_g^L=[\boldsymbol{x}_{g,1}^L,\boldsymbol{x}_{g,2}^L,...,\boldsymbol{x}_{g,M}^L]^{\mathrm{T}}\in R^{M\times d_g}$ is the labeled data, $\boldsymbol{X}_g^U=[\boldsymbol{x}_{g,1}^U,\boldsymbol{x}_{g,2}^U,\ldots,\boldsymbol{x}_{g,N}^U]^{\mathrm{T}}\in R^{N\times d_g}$ is the unlabeled data. $\boldsymbol{Y}^L\in R^{M\times C}$ is the label matrix of the labeled data, $\boldsymbol{Y}^U\in R^{N\times C}$ is the pseudo label matrix of the unlabeled data. $\boldsymbol{P}_g\in R^{d_g\times C}$ is the matrix of the consequent parameters. *M* and *N* are the number of the labeled and unlabeled data respectively, d_g is the number of the features generated by fuzzy rules, *C* is the number of the output spaces, and α is a parameter used to control the influence of the unlabeled data.

Since $\boldsymbol{Y}^U$ is a hard sparse matrix, with only one unity element in each row and all the remaining elements being zero, it is difficult to solve for $\boldsymbol{Y}^U$ directly [36, 37]. Recently, the matrix decomposition technique is introduced to improve the performance of the algorithms [36-40]. It has been shown experimentally that performance enhancement can be achieved by replacing the target matrix with the product of the basis matrix and the coefficient matrix [38]. Besides, by decomposing the pseudo label matrix into basis matrix and coefficient matrix, the product of the two matrices can generate continuous values (soft labels) to provide useful information for modeling [36]. Inspired by the work in [36, 38], the matrix decomposition technique is adopted in this study to alleviate the problems due to the direct optimization of $\boldsymbol{Y}^U$. Here, we decompose the pseudo label matrix into the product of $\boldsymbol{H}\in R^{K\times C}$ and $\boldsymbol{B}\in R^{K\times N}$, where $\boldsymbol{H}$ is the coefficient matrix and $\boldsymbol{B}$ is the basis matrix, *K* is the intrinsic dimension. Then, Eq. (4.a) can be updated as follows:

$$\min_{\boldsymbol{P}_g,\boldsymbol{B},\boldsymbol{H}} \frac{1}{2}\left(\left\|\boldsymbol{X}_g^L\boldsymbol{P}_g-\boldsymbol{Y}^L\right\|^2+\alpha\left\|\boldsymbol{X}_g^U\boldsymbol{P}_g-\boldsymbol{B}^{\mathrm{T}}\boldsymbol{H}\right\|^2\right)$$

(4b)

It should be noted that although $\boldsymbol{Y}^L$ and $\boldsymbol{Y}^U$ are both hard sparse matrices, we only decompose $\boldsymbol{Y}^U$. The reasons are as follows. First, although $\boldsymbol{Y}^L$ can also be decomposed, we need to add the constraint $\boldsymbol{Y}^L=\boldsymbol{AZ}$ to (4b), which makes the optimization more difficult and increases the computation burden. Second, while $\boldsymbol{Y}^U$ is the variable to be solved, $\boldsymbol{Y}^L$ is the known label matrix of the labeled samples. If we decompose it into soft labels, it is easy to introduce noise into the true labels, which reduces the performance of the trained model. Based on the above analyses, we only decompose $\boldsymbol{Y}^U$ in the proposed method.

C. The Proposed Method

Based on the mechanism of transductive single-view fuzzy system described above and the introduction of multi-view collaborative learning, we propose a novel learning criterion for transductive multi-view fuzzy system as follows:

$$\min_{\boldsymbol{P}_g^v, w^v, \boldsymbol{B}^v, \boldsymbol{H}^v} \quad J = \Gamma + \Lambda + \Omega$$
$$\text{s.t.} \sum_{v=1}^{V} w^v = 1,\ w^v \geq 0 \tag{5a}$$

$$\Gamma = \frac{1}{2}\sum_{v=1}^{V} w^v \left(\left\| \left(\boldsymbol{X}_g^L\right)^v \boldsymbol{P}_g^v - \boldsymbol{Y}^L \right\|^2 + \alpha \left\| \left(\boldsymbol{X}_g^U\right)^v \boldsymbol{P}_g^v - \boldsymbol{B}^{v\mathrm{T}} \boldsymbol{H}^v \right\|^2 \right) + \lambda_{\boldsymbol{P}_g} \sum_{v=1}^{V} \left\| \boldsymbol{P}_g^v \right\|^2 \tag{5b}$$

$$\Lambda = \frac{\lambda_{co}}{2}\sum_{v=1}^{V} \left(\left\| \left(\boldsymbol{X}_g^L\right)^v \boldsymbol{P}_g^v - \frac{1}{V-1}\sum_{l=1,l\neq v}^{V} \left(\boldsymbol{X}_g^L\right)^l \boldsymbol{P}_g^l \right\|^2 \right) + \frac{\lambda_{co}}{2}\sum_{v=1}^{V} \left(\left\| \left(\boldsymbol{X}_g^U\right)^v \boldsymbol{P}_g^v - \frac{1}{V-1}\sum_{l=1,l\neq v}^{V} \left(\boldsymbol{X}_g^U\right)^l \boldsymbol{P}_g^l \right\|^2 \right) + \frac{\lambda_{\boldsymbol{BH}}}{2}\sum_{v=1}^{V} \left\| \boldsymbol{B}^{v\mathrm{T}} \boldsymbol{H}^v - \frac{1}{V-1}\sum_{l=1,l\neq v}^{V} \boldsymbol{B}^{l\mathrm{T}} \boldsymbol{H}^l \right\|^2 \tag{5c}$$

$$\Omega = \lambda_w \sum_{v=1}^{V} w^v \ln w^v \tag{5d}$$

where $(\boldsymbol{X}_g^L)^v \in R^{M\times d_g}$, $(\boldsymbol{X}_g^U)^v \in R^{N\times d_g}$ are the labeled and unlabeled data of the *v*-th view, respectively, $\boldsymbol{B}^{v\mathrm{T}}\boldsymbol{H}^v$ is the soft label matrix of the *v*-th view, $\boldsymbol{Y}^L$ is the label matrix of the labeled data, $\boldsymbol{P}_g^v \in R^{d_g \times C}$ is the matrix of the consequent parameters of the *v*-th view, and w^v is the weight of the *v*-th view. Further details about the learning criterion are given below.

1) Eq. (5b) considers both the training error and the prediction error across different views, where α is used to control the influence of the unlabeled data.

2) Since the labels learned from the multiple views may be inconsistent, multi-view collaborative learning mechanism [38] is introduced to deal with the issue, where the mechanism is tailored for transductive learning scenario. The first two terms in Eq. (5c) correspond to collaborative learning, which are used to make the output of multiple views consistent and to mine useful information between the views, thereby improving the generalizability of the trained model.

3) In Eq. (5d), $\sum_{v=1}^{V} w^v \ln w^v$ is the negative Shannon entropy which has been used to adaptively balance the importance of different variables, such as features, views or models, in the learning procedure [41-44]. Inspired by these works, we use the negative Shannon entropy to mine the different information among different views and adaptively obtain the weights of the views. According to the maximum entropy principle, the probability of different views, *i.e.*, the importance of the views, can be balanced adaptively by minimizing the negative entropy to avoid the domination of a certain view in the final output.

4) The regularization parameter $\lambda_{\boldsymbol{P}_g}$ and the view consistency parameters λ_{co}, $\lambda_{\boldsymbol{BH}}$ are used to control the influence of the consequent parameters, collaborative learning and the soft labels respectively. These parameters can be determined manually, or by using parameter optimization strategies such as cross-validation.

D. Optimization

The optimization problem in Eq. (5) is non-convex, which can be solved using alternate iterative algorithm. The Lagrange function of Eq. (5a) is given by:

$$\begin{aligned} L\left(\boldsymbol{P}_g^v, w^v, \boldsymbol{B}^v, \boldsymbol{H}^v\right) = {} & \frac{1}{2}\sum_{v=1}^{V} w^v \\ & \left(\left\| \left(\boldsymbol{X}_g^L\right)^v \boldsymbol{P}_g^v - \boldsymbol{Y}^L \right\|^2 + \alpha \left\| \left(\boldsymbol{X}_g^U\right)^v \boldsymbol{P}_g^v - \boldsymbol{B}^{v\mathrm{T}} \boldsymbol{H}^v \right\|^2 \right) \\ & + \frac{\lambda_{co}}{2}\sum_{v=1}^{V} \left(\left\| \left(\boldsymbol{X}_g^L\right)^v \boldsymbol{P}_g^v - \frac{1}{V-1}\sum_{l=1,l\neq v}^{V} \left(\boldsymbol{X}_g^L\right)^l \boldsymbol{P}_g^l \right\|^2 \right) \\ & + \frac{\lambda_{co}}{2}\sum_{v=1}^{V} \left(\left\| \left(\boldsymbol{X}_g^U\right)^v \boldsymbol{P}_g^v - \frac{1}{V-1}\sum_{l=1,l\neq v}^{V} \left(\boldsymbol{X}_g^U\right)^l \boldsymbol{P}_g^l \right\|^2 \right) \\ & + \frac{\lambda_{\boldsymbol{BH}}}{2}\sum_{v=1}^{V} \left\| \boldsymbol{B}^{v\mathrm{T}} \boldsymbol{H}^v - \frac{1}{V-1}\sum_{l=1,l\neq v}^{V} \boldsymbol{B}^{l\mathrm{T}} \boldsymbol{H}^l \right\|^2 \\ & + \lambda_w \sum_{v=1}^{V} w^v \ln w^v + \lambda_{\boldsymbol{P}_g} \sum_{v=1}^{V} \left\| \boldsymbol{P}_g^v \right\|^2 \end{aligned} \tag{6}$$

which can be optimized by alternately solving $\boldsymbol{P}_g^v$, w^v, $\boldsymbol{B}^v$ and $\boldsymbol{H}^v$ using the scheme below:

1) Update $\boldsymbol{P}_g^v$ with w^v, $\boldsymbol{B}^v$ and $\boldsymbol{H}^v$ fixed: By taking the derivative of Eq. (6) with respect to $\boldsymbol{P}_g^v$ and setting it to zero, we can obtain the update rule for $\boldsymbol{P}_g^v$ as follows:

$$\boldsymbol{P}_g^v = \begin{pmatrix} w^v \left(\boldsymbol{X}_g^L\right)^{v\mathrm{T}} \left(\boldsymbol{X}_g^L\right)^v + \alpha w^v \left(\boldsymbol{X}_g^U\right)^{v\mathrm{T}} \left(\boldsymbol{X}_g^U\right)^v + \\ \lambda_{\boldsymbol{P}_g} \boldsymbol{I} + \lambda_{co} \left(\left(\boldsymbol{X}_g^L\right)^{v\mathrm{T}} \left(\boldsymbol{X}_g^L\right)^v + \left(\boldsymbol{X}_g^U\right)^{v\mathrm{T}} \left(\boldsymbol{X}_g^U\right)^v \right) \end{pmatrix}^{-1} \begin{pmatrix} w^v \left(\boldsymbol{X}_g^L\right)^{v\mathrm{T}} \boldsymbol{Y}^L + w^v \left(\boldsymbol{X}_g^U\right)^{v\mathrm{T}} \boldsymbol{B}^{v\mathrm{T}} \boldsymbol{H}^v + \\ \frac{\lambda_{co}}{V-1} \left(\left(\boldsymbol{X}_g^L\right)^{v\mathrm{T}} \sum_{l=1,l\neq v}^{V} \left(\boldsymbol{X}_g^L\right)^l \boldsymbol{P}_g^l + \left(\boldsymbol{X}_g^U\right)^{v\mathrm{T}} \sum_{l=1,l\neq v}^{V} \left(\boldsymbol{X}_g^U\right)^l \boldsymbol{P}_g^l \right) \end{pmatrix} \tag{7}$$

2) Update w^v with $\boldsymbol{P}_g^v$, $\boldsymbol{B}^v$ and $\boldsymbol{H}^v$ fixed: By taking the derivative of Eq. (6) with respect to w^v and setting it to zero, we can obtain the update rule for w^v as follows:

$$w^v = \frac{\exp\left(-\left(\left\| \left(\boldsymbol{X}_g^L\right)^v \boldsymbol{P}_g^v - \boldsymbol{Y}^L \right\|^2 - \alpha \left\| \left(\boldsymbol{X}_g^U\right)^v \boldsymbol{P}_g^v - \boldsymbol{B}^{v\mathrm{T}} \boldsymbol{H}^v \right\|^2 \right) \Big/ 2\lambda_w \right)}{\sum_{l=1}^{V} \exp\left(-\left(\left\| \left(\boldsymbol{X}_g^L\right)^j \boldsymbol{P}_g^j - \boldsymbol{Y}^L \right\|^2 - \alpha \left\| \left(\boldsymbol{X}_g^U\right)^j \boldsymbol{P}_g^j - \boldsymbol{B}^{j\mathrm{T}} \boldsymbol{H}^j \right\|^2 \right) \Big/ 2\lambda_w \right)} \tag{8}$$

3) Update $\boldsymbol{B}^v$ with w^v, $\boldsymbol{P}_g^v$ and $\boldsymbol{H}^v$ fixed: By taking the derivative of Eq. (6) with respect to $\boldsymbol{B}^v$ and setting it to zero, we can obtain the update rule for $\boldsymbol{B}^v$ as follows:

$$\begin{aligned} \boldsymbol{B}^v = {} & \left(\left(\alpha w^v + 2\lambda_{\boldsymbol{BH}} \right) \boldsymbol{H}^v \boldsymbol{H}^{v\mathrm{T}} \right)^{-1} \\ & \left(\alpha w^v \boldsymbol{H}^v \left(\left(\boldsymbol{X}_g^U\right)^v \boldsymbol{P}_g^v \right)^{\mathrm{T}} + \frac{2\lambda_{\boldsymbol{BH}}}{V-1} \boldsymbol{H}^v \left(\sum_{l=1,l\neq v}^{V} \boldsymbol{B}^{l\mathrm{T}} \boldsymbol{H}^l \right)^{\mathrm{T}} \right) \end{aligned} \tag{9}$$

4) Update $\boldsymbol{H}^v$ with w^v, $\boldsymbol{P}_g^v$ and $\boldsymbol{B}^v$ fixed: By taking the derivative of Eq. (6) with respect to $\boldsymbol{H}^v$ and setting it to zero, we can obtain the update rule for $\boldsymbol{H}^v$ as follows:

$$\boldsymbol{H}^v = \left(\left(\alpha w^v + 2\lambda_{\boldsymbol{BH}}\right)\boldsymbol{B}^v\boldsymbol{B}^{v\mathrm{T}}\right)^{-1} \left(\alpha w^v \boldsymbol{B}^v\left(\left(\boldsymbol{X}_g^U\right)^v \boldsymbol{P}_g^v\right) + \frac{2\lambda_{\boldsymbol{BH}}}{V-1}\boldsymbol{B}^v\left(\sum_{l=1,l\neq v}^{V}\boldsymbol{B}^{l\mathrm{T}}\boldsymbol{H}^l\right)\right) \tag{10}$$

By iteratively optimizing Eqs. (7), (8), (9) and (10), Eq. (5) converges to the local minimum and the local optimal solution can be obtained.

Finally, the labels for the unlabeled data are computed by:

$$\boldsymbol{Y}^U = \sum_{v=1}^{V} w^v \boldsymbol{X}_g^U \boldsymbol{P}_g^v \tag{11}$$

Remark 1: The parameters $\boldsymbol{P}_g^v$, w^v, $\boldsymbol{B}^v$ and $\boldsymbol{H}^v$ should be initialized before the optimization process. Here, we use TSK-FS in [45] to initialize $\boldsymbol{P}_g^v$, whose initial value is given by

$$\boldsymbol{P}_g^v = \left(\boldsymbol{I} + \lambda_1 \boldsymbol{X}_g^{L\mathrm{T}}\boldsymbol{X}_g^L\right)^{-1}\left(\lambda_{\boldsymbol{P}_g}\boldsymbol{X}_g^{L\mathrm{T}}\boldsymbol{Y}^L\right) \tag{12}$$

We further assume that the views have an equal weight and w^v is initialized with $w^v = 1/V, v = 1, 2, \ldots, V$.

Remark 2: For $\boldsymbol{B}^v$ and $\boldsymbol{H}^v$, they are initialized through the decomposition of $(\boldsymbol{Y}^U)^v$. Here, $(\boldsymbol{Y}^U)^v$ is initialized with $(\boldsymbol{Y}^U)^v = \boldsymbol{X}^v\boldsymbol{P}_g^v$, which can be achieved using various matrix decomposition methods. In this study, the non-negative matrix factorization techniques (NMF) is adopted for its simplicity and good interpretability.

```
Algorithm 1 TMV-TSK
Input: Set the number of fuzzy rules I; set the parameters $\lambda_{\boldsymbol{P}_g}$, $\lambda_{co}$, $\lambda_W$, $\lambda_{\boldsymbol{BH}}$, set the maximum iteration numbers T, set view weight $w^v = 1/V$, set the threshold $\varepsilon = 1e-6$.
Output: $\boldsymbol{P}_g^v$ and $w^v$.
1: Use the Var-Part clustering algorithm to calculate the antecedent parameters of TSK fuzzy system.
2: Use Eqs. (3b) – (3d) to construct a new labeled dataset $D_L^v = \{(\boldsymbol{X}_g^L)^v, \boldsymbol{Y}^L\}$ and an unlabeled dataset $D_U^v = \{(\boldsymbol{X}_g^U)^v\}$ in the new feature space generated by the fuzzy rules for each view.
3: Initialize $\boldsymbol{P}_g^v$ and $(\boldsymbol{Y}^U)^v$.
4: Decompose $(\boldsymbol{Y}^U)^v$ into $\boldsymbol{B}^v$, $\boldsymbol{H}^v$.
5: Compute $J(0)$ using Eq. (5a);
6: For $t \leftarrow 1, 2, \ldots, T$ do
7:   For $v \leftarrow 1, 2, \ldots, V$ do
8:       Update $\boldsymbol{P}_g^v$ base on Eq. (7);
9:       Update $w^v$ base on Eq. (8);
10:      Update $\boldsymbol{B}^v$ base on Eq. (9);
11:      Update $\boldsymbol{H}^v$ base on Eq. (10);
12:   End for
13:   Compute $J(t)$ using Eq. (5a);
12:   If $\left|J(t)\text{-}J(t\text{-}1)\right| < \varepsilon$
13:       return;
14:   End
15: End for
```

E. *Algorithm*

Based on the above analyses and derivations, the details of the proposed TMV-TSK algorithm are given in Algorithm 1. The source code of this algorithm is available at https://github.com/BBKing49/TMV-TSK.

F. *Complexity Analyses*

The computational complexity of the proposed TMV-TSK algorithm is discussed in this subsection. Refer to Algorithm 1, the time complexity of step 1 is $O(2MdI)$, where *M*, *d* and *I* are the number of labeled data, the dimension of original feature space and the number of fuzzy rules, respectively. The time complexity of step 2 is $O(2(M+N)I(d+1))$, where *N* is the number of unlabeled data. $O(d_g M d_g^2 MC + M d_g C)$ is the time complexity of step 3, where d_g is the feature space dimension generated by the fuzzy rules and *C* is the number of the output spaces. The time complexity of the decomposition process in step 4 is $O((M+C)K)$, where *K* is the number of rows of the $\boldsymbol{B}^v$, $\boldsymbol{H}^v$. The time complexity of step 7 is $O((M+N)^2 CTd_g^3(d_g V+1))$, where *V* is the number of views, and *T* is the number of iterations. Finally, the time complexity of steps 8, 9 and 10 are given by. $O((Md_g + N(d_g+K)^2)C^2VT)$, $O((KCK)^2 NVT + (KC)^2 Kd_g NT)$, $O((KCK)^2 CVT + (KN)^2 Kd_g CT)$. According to [46], and because step 7 is the rate determining step, the overall computational cost is dominated by the complexity of step 7, i.e., $O((M+N)^2 CTd_g^3(d_g V+1))$.

G. *Interpretability*

While considerable research is being conducted to enhance the interpretability of intelligent models, maintaining interpretability and accuracy at the same time remains a difficulty. The issue for some representative models is discussed in detail in [47, 48]. As shown in Fig. S1 of the *Supplementary Materials* section, we can see that rule-based models have better interpretability but poor accuracy, whereas deep models have high accuracy but poor interpretability. Nevertheless, it has been shown that interpretability and accuracy can both be achieved by introducing data-driven learning techniques into the training of the classical rule-based TSK FS model [14-16]. In this study, TSK FS is adopted as the basis model of the proposed TMV-TSK for multi-view data modeling, so that it can inherit the superior interpretability of TSK FS.

Besides, the interpretability of the proposed TSK FS based TMV-TSK can be appreciated from another viewpoint. Given that linear models are usually regarded the most interpretable models [49], it is beneficial that TSK FS is indeed the form of dynamically weighted linear models. As shown in Fig. S2 of the *Supplementary Materials* section, the decision function of TKS FS is a combination of multiple linear models with dynamic weighting. Each linear model corresponds to a rule which covers a fuzzy space. Since a linear model that is associated with one rule is readily interpretable, we can interpret the decision result of TSK FS through each of the linear models. In this sense, we can infer that the proposed TMV-TSK has good interpretability.

Table I Statistics of datasets

Dataset	**Size**	**Views/(Dimensions)**	**Classes**
Toy	600	2(4-4)	2
Forest Type	523	2(9-18)	4
Multiple Features	2000	2(64-47)	10
Corel Images	1000	2(300-256)	10
NUS-WIDE	12451	2(64-73)	15
Image Segmentation	2310	2(10-9)	7
MSRCv1	240	3(24-256-254)	7

IV. Experimental Studies

A. Datasets

1) Toy multi-view dataset

To intuitively illustrate the effectiveness of the proposed algorithm, we have constructed a toy multi-view dataset for experimental studies. The brief descriptions of the toy datasets are given in Table I. The toy dataset, visualized based on t-SNE [50], is shown in Fig. S3 of the *Supplementary Materials* section. Two views of the toy dataset are shown in Fig. S3 (a) and (b), where the plum dots represent the labeled positive samples, and the pink dots represent the labeled negative samples. In Fig. S3 (c) and (d), we show the data of the two views when the proportion of labeled data is 20%, where the red dots represent the unlabeled samples.

2) Real world dataset

In the experiments, we used six benchmarking real world multi-view datasets for performance evaluation. Four datasets were obtained from the UCI repository, and the remaining two from the NUS-WIDE dataset [51] and MSRCv1 dataset [52], respectively. A brief descriptions of the datasets are given in Table I. Further detailed are provided as follows:

1) Forest Type: It is a remote sensing image dataset of forest satellite, which can be divided into two views, the image band view and the spectral view.
2) Multiple Features: It is a multi-view handwritten numerals dataset containing five views. The Karhunen-love value view and Zernike moment view were selected for the experiments [18].
3) Corel Images: It is an image classification dataset, from which 10 classes of images with obvious foreground objects were selected and preprocessed to produce a two-view dataset. Specifically, the SIFT [53] and LBP [54] features were extracted as the two views for the experiments.
4) NUS-WIDE(NUS): It is an image classification dataset, from which 15 classes of images with obvious foreground objects were selected to produce a two-view dataset, with color histogram and edge orientation histogram as the two views.
5) Image Segmentation: It is a dataset consisting of 2,310 instances randomly selected from a database of 7 classes of outdoor images. The feature dimension of this dataset is 19, and can be naturally divided into two views, *i.e.*, shape view and RBG view.
6) MSRCv1: It is a scene recognition dataset containing 7 classes, 220 images in total. The CMT [55], LBP [54] and GENT [56] features were extracted as the three views for the experiments.

7) Table II Experimental settings

Algorithm	Hyperparameters
AMGL	Parameter free
TwoV-TSK-FS	Number of fuzzy rules $I \in \{2,\ 4,\ 6,\ 8,\ 10\}$, regularization parameter $\tau_A, \tau_B \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.
MV-F-ELM	Number of fuzzy rules $I \in \{2,\ 4,\ 6,\ 8,\ 10\}$.
MLAN	Regularization parameter $\lambda \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.
AMMSS	Regularization parameter $\lambda \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$, weight coefficient $\gamma \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.
SMGL	Weight coefficient $\gamma \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$, regularization parameter $\lambda \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.
JCD	adaptive parameter $\beta \in \{1.1,\ 1.3,\ \ldots\ ,\ 3.3\}$, regularization parameter $\lambda \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.
FMSSL	regularization parameter $\lambda \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.
TMV-TSK	Number of fuzzy rules $I \in \{2,\ 4,\ 6,\ 8,\ 10\}$, regularization parameter $\lambda_{\boldsymbol{P}_g}$, λ_{W} are set with values in $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$, unlabeled data balance parameters $\alpha \in \{0.1,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1\}$, multi-view collaborative parameter λ_{co} and $\lambda_{\boldsymbol{BH}}$ are set to 0.01.

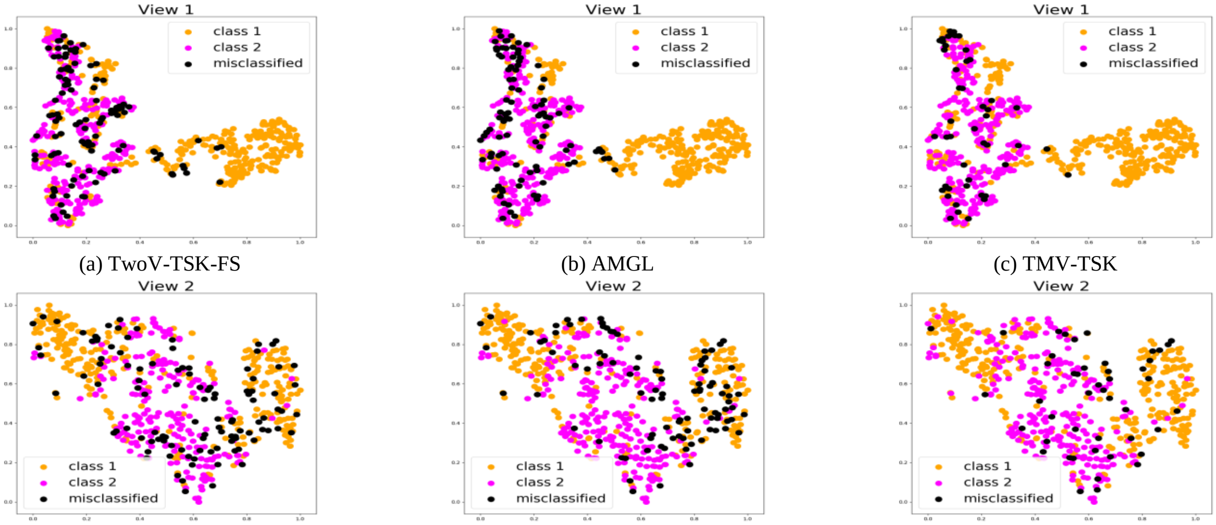
B. Experimental Settings

1) Algorithms under comparison

We compared the proposed TMV-TSK with eight algorithms: two inductive multi-view fuzzy classifiers, *i.e.*, TwoV-TSK-FS [17] and MV-F-ELM (an improved versions of F-ELM [57]), and six transductive multi-view learning methods, i.e. MLAN [29], AMGL [28], AMMSS [58], SMGL [27], JCD [24] and FMSSL[59]. For the training of MV-F-ELM, we first trained several single-view F-ELM classifiers separately with an individual view of the data for each classifier. Then, we averaged the outputs of the multiple F-ELM classifiers to give the final output of the MV-F-ELM. Since TwoV-TSK-FS can only handle two-view dataset, but the MSRCv1 dataset has three views, we implemented TwoV-TSK-FS on any two of the views and the best results were reported for comparison. For each multi-view dataset, we randomly choose 10%-50% labeled instances. To prevent potential bias that may be caused by random selection of labeled instances in the generation of multi-view datasets with few labels, the above process was repeated 10 times and the average accuracy was compared.

2) Parameter Settings

The grid searching strategy was used to find the optimal hyper-parameters of all the algorithms. Table II shows the details on the experimental settings.



(a) TwoV-TSK-FS (b) AMGL (c) TMV-TSK

(d) TwoV-TSK-FS (e) AMGL (f) TMV-TSK

Fig. 4 The results of different algorithms on the toy multi-view dataset at view 1 and view 2, visualized by t-SNE when the proportion of labeled data is 20%: (a) and (d) show the results of inductive multi-view fuzzy classifier TwoV-TSK-FS; (b) and (e) show the results of label propagation based transductive algorithm AMGL; (c) and (f) show the results of the proposed TMV-TSK.

Table III Accuracy (Mean ± SD) of the eight algorithms on the toy dataset with 20% labeled data

Dataset	TwoV-TSK-FS	MV-F-ELM	MLAN	AMMSS	SMGL	AMGL	JCD	FMSSL	TMV-TSK
Toy	0.7896 ±0.0059	0.7635 ±0.0120	0.8842 ±0.0042	0.8479 ±0.0059	0.8271 ±0.0030	0.8375 ±0.0147	0.8625 ±0.0035	0.8521 ±0.0074	**0.9177** **±0.0105**

Table IV Accuracy (Mean ± SD) of the eight algorithms on real world datasets with 20% labeled data

Dataset	Algorithms								
	TwoV -TSK-FS	MV-F-ELM	MLAN	AMGL	AMMSS	SMGL	JCD	FMSSL	TMV-TSK
Forest Type	0.8175 ±0.0209	0.8708 ±0.0083	0.8772 ±0.0097	0.8644 ±0.0215	0.8636 ±0.0205	0.8772 ±0.0018	0.8355 ±0.0203	0.8617 ±0.0111	**0.8852** **±0.0127**
Multiple Features	0.7008 ±0.0316	0.6965 ±0.0223	**0.9669** **±0.0051**	0.9592 ±0.0058	0.9598 ±0.0035	0.9502 ±0.0025	0.9477 ±0.0042	0.9469 ±0.0052	0.9579 ±0.0064
Corel Images	0.4459 ±0.0411	0.3563 ±0.0239	0.5180 ±0.0161	0.3583 ±0.0138	0.4974 ±0.0052	0.4967 ±0.0026	0.5750 ±0.0065	0.3725 ±0.0115	**0.5846** **±0.0080**
Image Segmentation	0.6918 ±0.0252	0.8393 ±0.0092	0.8956 ±0.0130	0.6059 ±0.0182	0.5931 ±0.0057	0.8636 ±0.0156	0.8636 ±0.0057	0.6878 ±0.0153	**0.8987** **±0.0153**
MSRCv1	0.5595 ±0.0663	0.5391 ±0.0091	0.7520 ±0.0269	0.7639 ±0.0035	0.7540 ±0.0035	0.7599 ±0.030	0.8036 ±0.0315	0.7016 ±0.0374	**0.8075** **±0.0149**
NUS	0.3634 ±0.0098	0.3777 ±0.0016	0.4173 ±0.0026	0.3619 ±0.0031	0.3755 ±0.0091	0.3911 ±0.0031	0.4360 ±0.0004	0.4044 ±0.0026	**0.4411** **±0.0006**
Average	0.5965 ±0.0325	0.6133 ±0.0101	0.7378 ±0.0122	0.6523 ±0.0110	0.6739 ±0.0079	0.7231 ±0.0093	0.7436 ±0.0114	0.6625 ±0.0091	**0.7625** **±0.0097**

C. *Experimental Results*

1) *Results on Toy Dataset*

The experimental results on the toy multi-view dataset are given in Fig. 4, where the black dots represent misclassification. Table III shows the detailed results of all the algorithms on the toy dataset when the proportion of labeled data is 20%. Fig. 4 (a) and (d) show the results of the inductive multi-view fuzzy classifier, *i.e.*, TwoV-TSK-FS. The result demonstrates that inductive multi-view algorithm is not reliable when labeled data is inadequate. Fig. 4 (b) and (e) show the results of label propagation based transductive algorithm, i.e., AMGL, whose performance is deteriorated when two classes overlap significantly. Fig. 4 (c) and (f) show the results of the proposed TMV-TSK, showing that it is more robust against inadequacy of labeled data and has more competitive performance when compared with the label propagation based transductive multi-view learning algorithm.

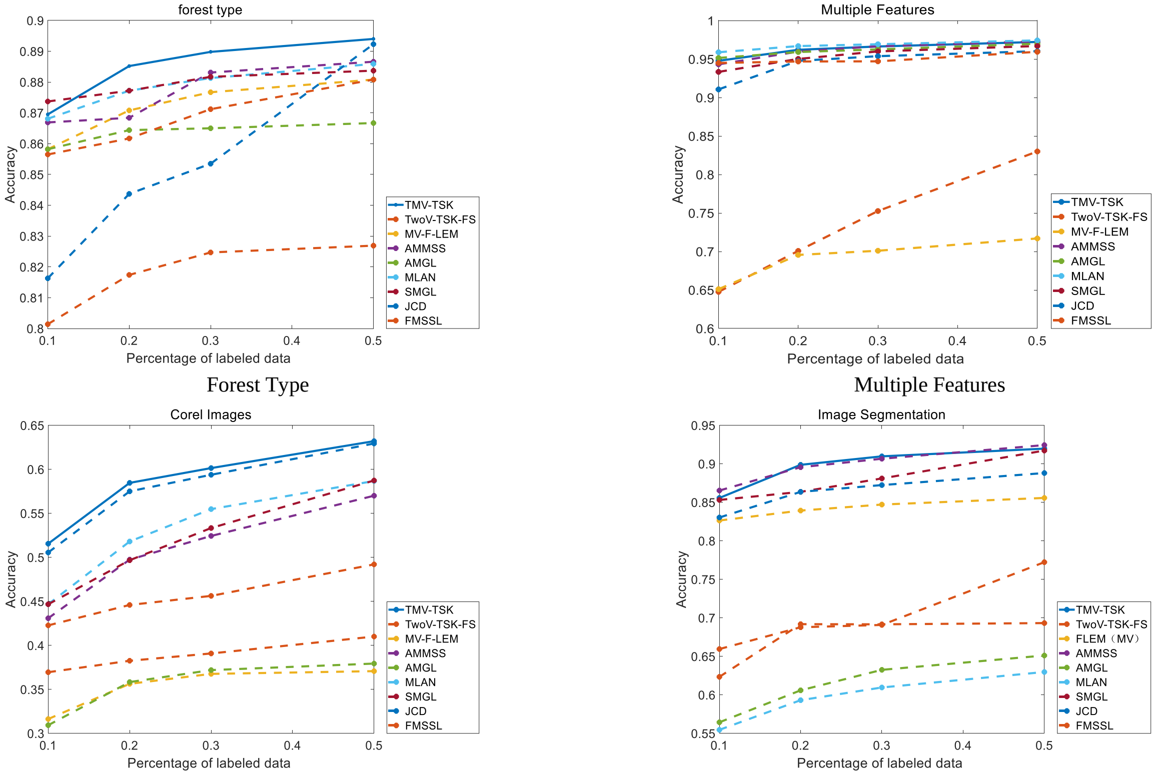
2) *Results on Real World Dataset*

Next, we compare the performance of all the algorithms on the six real world datasets. The representative results are shown in Fig. 5. Tables IV shows the mean and standard deviation (SD) of the classification accuracy of the algorithms when the proportion of labeled data is 20%. The results with other proportions of labeled data are detailed in Tables S1-S3 and Fig. S4 of the *Supplementary Materials* section. The following observations can be drawn from the results.

First, in most cases, the performance of the inductive multi-view algorithms is poor. Taking 20% labeled data as an example, even the mediocre transductive multi-view algorithm AMGL can outperform the inductive multi-view algorithms MV-F-ELM and TwoV-TSK-FS by about 4% and 5% on average. This indicates that the inductive multi-view algorithm is not reliable when labeled data are inadequate.

Second, compared with other transductive algorithms, JCD and TMV-TSK perform better in most cases. Taking the case of 20% labeled data as an example. For all the datasets, the accuracy of JCD and TMV-TSK are about 1% and 3% higher on average when compared to MLAN, where MLAN is the best among the four transductive algorithms (MLAN, AMGL, AMMSS, SMGL, FMSSL). Meanwhile, the accuracy of JCD and TMV-TSK are 9% and 11% higher when compared to AMGL, and that AMGL is relatively less effective among the four transductive algorithms. This is because most transductive algorithms only consider the difference of the information among different views, whereas JCD and TMV-TSK also mine the consistency information among multiple views.

Third, in most case, the performance of TMV-TSK is better than that of the other algorithms. Taking the case of 20% labeled data as an example, TMV-TSK has the highest average accuracy of 0.7625, which is 2% higher than the second best, i.e., 0.7436 obtained by JCD, and 17% higher than the worst, i.e., 0.5965 obtained by TwoV-TSK-FS. These results show the reliability of the proposed TMV-TSK in handing cases with few labeled data and that the performance can be greatly improved through multi-view collaborative learning and the decomposition of the pseudo label matrix. While the performance of TMV-TSK is only slightly better than that of JDC, TMV-TSK is still a favorable choice considering further that it has better interpretability with fuzzy rules and fuzzy logic inference.

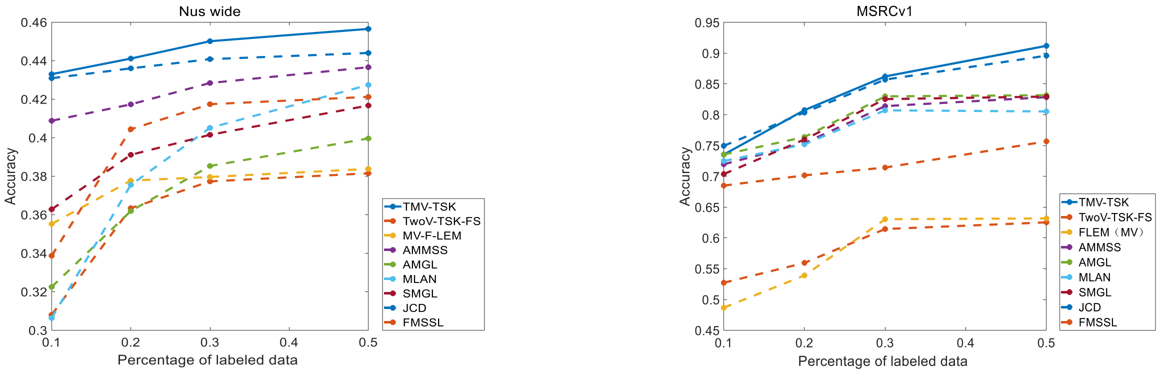


Forest Type

Multiple Features

Corel Images

Image Segmentation



NUS

MSRCv1

Fig. 5 Accuracy of the algorithms on different datasets

D. *Statistical Analysis*

The eight algorithms were further compared in terms of their performance under different proportions of labeled data. Experiments were conducted with 10%, 20%, 30% and 50%, labeled data respectively. The results were analyzed using the Friedman test and the post-hoc Holm test.

Table V Results of Friedman tests at different proportions of labeled data

Proportion of label data	10%	20%	30%	50%
p-value	0.001219	0.002282	0.000299	0.00012

Table VI Ranking of the algorithms

Algorithm	Proportion of labeled data			
	Ranking (10%)	Ranking (20%)	Ranking (30%)	Ranking (50%)
TMV-TSK	**1.9167**	**1.5**	**1.333**	**1.5**
TwoV-TSK-FS	7.8333	7.5	7.8333	8
MV-F-ELM	6.9167	6.5833	7.3333	7.4167
MLAN	2.5833	3	3.1667	3
AMGL	5.6667	6.1667	6.3333	6.3333
AMMSS	6.1667	5.3333	4.6667	5.3333
SMGL	3.9167	4.3333	4.1667	4.1667
JCD	4	4.0833	3.75	3
FMSSL	6	6.5	6.4167	6.25

First, Friedman test [60] is used to evaluate whether the difference in classification performance of the eight algorithms on different datasets at four different proportions of labeled data is statistically significant. The null hypothesis is that the classification performances of all the algorithms are the same. In the test, assuming that the null hypothesis is correct, the *p*-value is the probability of obtaining test results that are at least as extreme as the results observed [61]. In general, when *p*-value<0.05, the null hypothesis is rejected. As shown in Table V, the p-values are less than 0.05, indicating that the difference in performance is statistically significant. Table VI shows the ranking of the algorithms at the four different proportions of labeled data.

Next, the post-hoc Holm test was used to further verify whether there was a significant difference in performance between the best algorithm (*i.e.*, TMV-TSK) and the other seven algorithms. The test results for the case when the proportion of labeled data is 20% are shown in Table VII; the results for 10%, 30% and 50% labeled data are shown Tables S4-S6 of the *Supplementary Materials* section. In the test, R_i , R_o are the mean rank of the *i*-th algorithm and TMV-TSK respectively, *SE* is the standard error, z is the statistic (the smaller the value of z, the smaller the difference in performance). The value $\text{Holm} = \alpha / i$ is used for hypothesis, where α is the pre-specified level of signifi-

Table VII Results of post-hoc Holm test with 20% labeled data (null hypothesis if *p*-value <0.0125)

i	Algorithm	$z = \left(R_o - R_i\right) / SE$	*p*-value	$\text{Holm} = \alpha / i, \alpha = 0.05$	Null hypothesis
8	TwoV-TSK-FS /TMV-TSK	3.794733	0.000148	0.00625	Reject
7	MV-F-ELM / TMV-TSK	3.214982	0.001305	0.007143	Reject
6	FMSSL /TMV-TSK	3.162278	0.001565	0.008333	Reject
5	AMGL /TMV-TSK	2.951459	0.003163	0.01	Reject
4	AMMSS /TMV-TSK	2.424413	0.015333	0.0125	Not Reject
3	SMGL /TMV-TSK	1.791957	0.07314	0.016667	Not Reject
2	JCD/TMV-TSK	1.633843	0.102292	0.25	Not Reject
1	MLAN/TMV-TSK	0.948683	0.342782	0.05	Not Reject

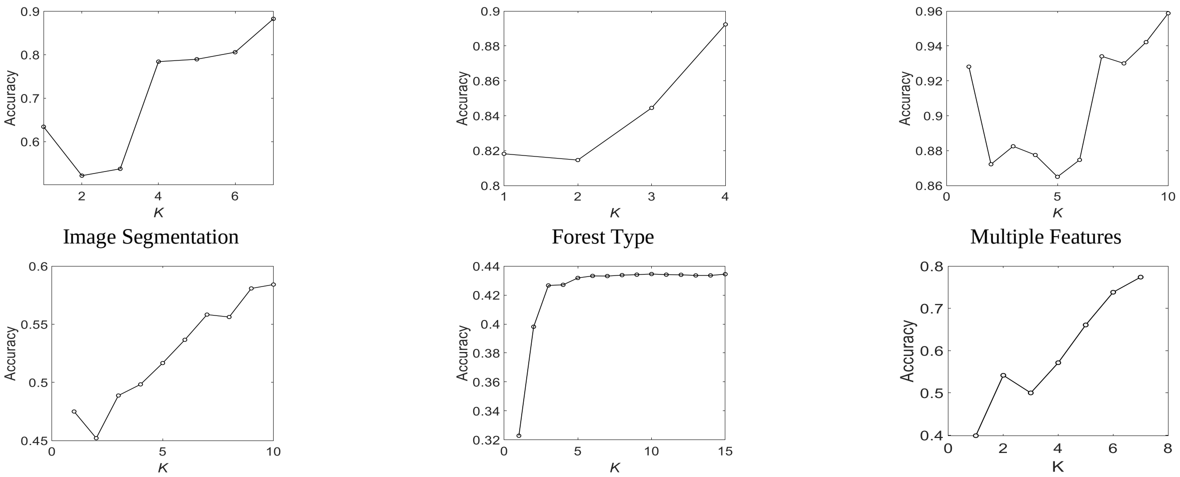


Image Segmentation Forest Type Multiple Features

Corel Images NUS MSRCv1

Fig.6 Influence of the matrix decomposition parameter K on the classification accuracy.

-cance (set to 0.05). When the *p-value* is less than Holm, the null hypothesis is rejected. The results of the post-hoc Holm test show that, for all the four proportions of labeled data, the difference in performance between TMV-TSK and TwoV-TSK-FS, MV-F-ELM, FMSSL and AMGL are statistically significant, but not so for AMMSS, SMGL, JCD and MLAN. Nevertheless, based on the results in Section IV-C, TMV-TSK is clearly advantageous over AMMSS, SMGL, JCD and MLAN.

Table VIII Classification accuracy of TMV-TSK with and without pseudo label matrix decomposition

Dataset	Without decomposition	With decomposition
Forest Type	0.8740±0.0091	**0.8852±0.0127**
Image Segmentation	0.7589±0.0159	**0.8987±0.0153**
Multiple Features	0.9525±0.0022	**0.9579±0.0064**
Corel Images	0.5388±0.0025	**0.5796±0.0181**
NUS	0.42±0.0014	**0.4411±0.0006**
MSRCv1	0.7381±0.0158	**0.7837±0.0338**
Average	0.7137±0.0087	**0.7590±0.0148**

E. Effectiveness Analysis

The effectiveness of the decomposition of the pseudo label matrix was analyzed when the proportion of labeled data was 20%. The classification accuracy achieved without and with matrix decomposition is shown in Table VIII. It is clear that with matrix decomposition, TMV-TSK always shows better performance, indicating that the decomposition of pseudo label matrix is helpful to improve the transductive learning of the classifier.

Next, we conducted experiments to study the influence of the intrinsic dimension *K* in $\boldsymbol{H} \in R^{K \times C}$ and $\boldsymbol{B} \in R^{K \times N}$ on the classification accuracy of the proposed method. In the experiments, NMF was used to initialize matrix decomposition under the condition that $K \le C$. The results are shown in Fig. 6. It can be seen that the classification accuracy generally increases with *K* for the datasets Dermatology, Forest Type, NUS, Corel Images and MSRCv1. For the Multiple Features dataset, the performance of TMV-TSK fluctuates with *K*, but higher accuracy can usually be obtained when *K* takes a larger value.

F. Sensitivity Analysis

The sensitivity of the proposed TMV-TSK algorithm to the three parameters α , λ_{P_g} and λ_w are analyzed in the subsection. In the analysis, one parameter is varied while the other two are fixed at the respective optimal values. Like the above experiments, all the datasets contain 20% labeled data. The results are show in Fig. S5 of the *Supplementary Materials* section, from which the following conclusions can be drawn.

1) It can be seen from Fig. S5 (a) that the performance of TMV-TSK increases with λ_{P_g} , which demonstrates the usefulness of multi-view regularization. Obviously, TMV-TSK achieves satisfactory results when the parameter takes a value from {0.001, 0.01, 0.1}.

2) The results in Fig. S5 (c) show that TMV-TSK is not sensitive to λ_w. Satisfactory performance is attained when λ_w takes a value from {0.001, 0.01}.

3) It can be seen from Fig. S5 (b) that the performance of TMV-TSK is poor when α is large. This is because α controls the influence of unlabeled data on the learning criterion. As indicated by Eq. (5), increasing α will put more weight on the unlabeled data, and introduce more noise to the classification. The performance of TMV-TSK is reasonable when α takes a value from {0.1, 0.2, 0.4}.

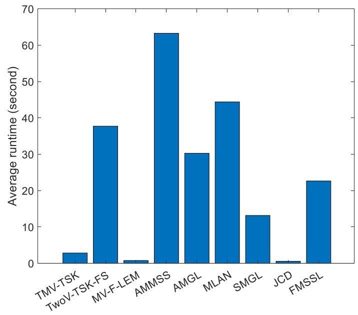
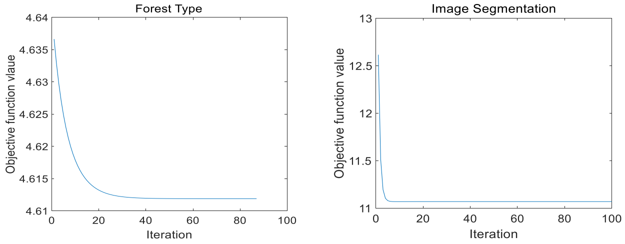


Fig. 7 The average runtime of the eight algorithms



(a) (b)

Fig. 8 Convergence of the TMV-TSK on the Forest Type and Image Segmentation datasets

G. *Runtime and Convergence Analysis*

In this subsection, first, we evaluate the actual machine runtime of TMV-TSK and the other seven algorithms. Fig. 7 shows the runtime of the eight algorithms, averaged over all the datasets with 20% labeled data. It can be seen that TMV-TSK has a smaller runtime than most algorithms. Although the runtime of TMV-TSK is larger than that of MV-F-ELM and JCD, it is considered acceptable given the superior classification accuracy.

Next, we illustrate the convergence of TMV-TSK empirically on the Forest Type and Image Segmentation datasets. Fig. 8 shows the convergence curves on these datasets with 20% labeled data. It can be seen that TMV-TSK converges after 40 iterations on the Forest Type dataset, and 10 iterations on the Image Segmentation dataset. In fact, in the experiments, we find that TMV-TSK usually converges within 60 iterations. Thus, we set the maximum number of iterations *T* to 100 for all the datasets.

H. *Interpretability Analysis*

In this subsection, the interpretability of TMV-TSK is analyzed by using the Forest Type dataset with 20% labeled data as an illustration. In the analysis, the rule number is set to four, *i.e.*, *I*=4. Table S7 and Fig. S6 of the *Supplementary Materials* section shows the four rules for the image band view of the trained TMV-TSK and the corresponding membership functions, respectively. According to the order of membership center, these four membership functions can be represented with the linguistic terms "High", "Medium", "Little Low" and "Low". Similarly, other features can be divided into four fuzzy subsets.

With the linguistic expressions of the IF-part and the corresponding linear function of the THEN-part, four fuzzy rules can be defined. For example, the first fuzzy rule can be expressed as follows:

The first fuzzy rule:

IF the Band1 of ASTER image is Medium, and the Band2 of ASTER image is High, and the Band3 of ASTER image is High, and the Band4 of ASTER image is Medium, and the Band5 of ASTER image is High, and the Band6 of ASTER image is High, and the Band7 of ASTER image is Medium, and the Band8 of ASTER image is High, and the Band9 of ASTER image is High, THEN the decision values of the four outputs are given as follows:

$$f^1(\boldsymbol{x}) = \begin{bmatrix} 7.333 + 2.16x_1 + 2.31x_2 + 4.31x_3 - 1.22x_4 \\ -2.53x_5 + 5.76x_6 - 1.65x_7 - 2.62x_8 + 6.89x_9; \end{bmatrix}$$

$$f^2(x) = \begin{bmatrix} -2.82 - 0.29x_2 + 0.98x_2 - 3.90x_3 + 2.31x_4 \\ 3.56x_5 - 3.11x_6 + 3.19x_7 + 2.04x_8 + 0.054x_9; \end{bmatrix}$$

$$f^3(x) = \begin{bmatrix} -0.43 - 3.03x_1 - 3.41x_2 - 1.92x_3 - 3.30x_4 \\ -1.20x_5 - 1.52x_6 + 0.57x_7 - 0.10x_8 + 0.74x_9; \end{bmatrix}$$

$$f^4(x) = \begin{bmatrix} -2.57 - 3.89x_1 - 4.32x_2 - 4.29x_3 - 3.79x_4 \\ -0.71x_5 - 4.39x_6 + 1.16x_7 + 0.77x_8 + 0.38x_9 \end{bmatrix}$$

In the *Supplementary Materials* section, Fig. S7 further explains the usage and the importance of the rules generated by the proposed method. The fuzzy system generated for the spectral view can be explained in the same way.

V. CONCLUSION

In this paper, a transductive multi-view TSK fuzzy system is proposed by integrating transductive learning into the framework of multi-view fuzzy systems, where the unlabeled data is used to enhance the performance of the multi-view fuzzy model. Furthermore, collaborative learning between multiple views, view weighting mechanism, and the strategy of pseudo label matrix decomposition are adopted to improve the robustness of the model. The results of the extensive experiments indicate that the proposed TMV-TSK is superior to both traditional multi-view fuzzy systems and the existing transductive multi-view algorithms.

The TMV-TSK is currently limited to single-label applications. Future work will focus on the development of efficient algorithms for multi-label scenarios which are very common in real-world applications. Another future work is to improve the proposed method by leveraging transfer learning mechanisms [62-66]. For example, we can regard different views as heterogeneous domains in transfer learning scene. Thus, the consistency of the views can be extracted based on the distribution divergence of the different domains using transfer learning strategies, such as the maximum density divergence based domain adaptation [62].

REFERENCES

[1] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion,* vol. 38, pp. 43-54, 2017.

[2] Y. Yang and H. Wang, "Multi-view clustering: A survey," *Big Data Mining and Analytics,* vol. 1, no. 2, pp. 83-107, 2018.

[3] B. Zhao, J. T. Kwok, and C. Zhang, "Multiple kernel clustering," in *Proceedings of the 2009 SIAM International Conference on Data Mining*, 2009, pp. 638-649: SIAM.

[4] L. Du, P. Zhou, L. Shi and H. wang *et al.*, "Robust multiple kernel k-means using l21-norm," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[5] K. Zhan, C. Niu, C. Chen and F. Nie *et al*, "Graph structure fusion for multiview clustering," *IEEE Transactions on Knowledge and Data Engineering,* vol. 31, no. 10, pp. 1984-1993, 2018.

[6] K. Zhan, F. Nie, J. Wang, and Y. Yang, "Multiview consensus graph clustering," *IEEE Transactions on Image Processing,* vol. 28, no. 3, pp. 1261-1270, 2018.

[7] K. Zhan, X. Chang, J. Guan and L. Chen *et al*, "Adaptive structure discovery for multimedia analysis using multiple features," *IEEE transactions on cybernetics,* vol. 49, no. 5, pp. 1826-1834, 2018.

[8] H. Zhao and Z. Ding, "Multi-view clustering via deep matrix factorization," in *AAAI*, 2017.

[9] S. Sun, "Multi-view Laplacian support vector machines," in *International Conference on Advanced Data Mining and Applications*, 2011, pp. 209-222: Springer.

[10] C. Zhang, J. Cheng, and Q. Tian, "Multi-View Image Classification With Visual, Semantic and View Consistency," *IEEE Transactions on Image Processing,* vol. 29, pp. 617-627, 2019.

[11] S. Sun and G. Chao, "Multi-view maximum entropy discrimination," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[12] G. Chao and S. Sun, "Alternative multiview maximum entropy discrimination," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 7, pp. 1445-1456, 2015.

[13] J. Li, Y. Wu, J. Zhao, and K. Lu, "Low-rank discriminant embedding for multiview learning," *IEEE transactions on cybernetics*, vol. 47, no. 11, pp. 3516-3529, 2016.

[14] W. W. Tan and T. W. Chua, "Uncertain rule-based fuzzy logic systems: introduction and new directions (Mendel, JM; 2001)[book review]," *IEEE Computational Intelligence Magazine,* vol. 2, no. 1, pp. 72-73, 2007.

[15] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]," *IEEE Transactions on automatic control,* vol. 42, no. 10, pp. 1482-1484, 1997.

[16] J. Wang, D. Lin, Z. Deng and Y. Jiang *et al.*, "Multitask TSK Fuzzy System Modeling by Jointly Reducing Rules and Consequent Parameters," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019.

[17] Y. Jiang, Z. Deng, F.-L. Chung, and S. Wang, "Realizing two-view TSK fuzzy classification system by using collaborative learning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 1, pp. 145-160, 2016.

[18] T. Zhang, Z. Deng, D. Wu, and S. Wang, "Multiview Fuzzy Logic System With the Cooperation Between Visible and Hidden Views," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 6, pp. 1162-1173, 2018.

[19] Y. Zhang, J. Li, X. Zhou and M. Zhang *et al.*, "A view-reduction based multi-view TSK fuzzy system and its application for textile color classification," *Journal of Ambient Intelligence and Humanized Computing,* pp. 1-11, 2019.

[20] T. Joachims, "Transductive inference for text classification using support vector machines," in *Icml*, 1999, vol. 99, pp. 200-209.

[21] W. Shi, Y. Gong, C. Ding and Z. Ma *et al.*, "Transductive semi-supervised deep learning using min-max features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 299-315.

[22] G. Li, S. C. Hoi, and K. Chang, "Two-view transductive support vector machines," in *Proceedings of the 2010 SIAM International Conference on Data Mining*, 2010, pp. 235-244: SIAM.

[23] Y. Li, Y. Wang, J. Zhou, and X. Jiang, "Robust transductive support vector machine for multi-view classification," *Journal of Circuits, Systems and Computers*, vol. 27, no. 12, p. 1850185, 2018.

[24] W. Zhuge, C. Hou, S. Peng, and D. Yi, "Joint consensus and diversity for multi-view semi-supervised classification," *Machine Learning,* pp. 1-21, 2019.

[25] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5070-5079.

[26] Y. Yi, Y. Shi, H. Zhang and J. Wang *et al.*, "Label propagation based semi-supervised non-negative matrix factorization for feature extraction," *Neurocomputing*, vol. 149, pp. 1021-1037, 2015.

[27] M. Karasuyama and H. Mamitsuka, "Multiple graph label propagation by sparse integration," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 12, pp. 1999-2012, 2013.

[28] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *IJCAI*, 2016, pp. 1881-1887.

[29] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[30] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning,* vol. 3, no. 1, pp. 1-130, 2009.

[31] M. F. Azeem, M. Hanmandlu, and N. Ahmad, "Generalization of adaptive neuro-fuzzy inference systems," *IEEE Transactions on Neural Networks,* vol. 11, no. 6, pp. 1332-1346, 2000.

[32] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE transactions on systems, man, and cybernetics,* no. 1, pp. 116-132, 1985.

[33] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," in *Proceedings of the sixth international symposium on Multiple-valued logic*, 1976, pp. 196-202: IEEE Computer Society Press.

[34] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences,* vol. 10, no. 2-3, pp. 191-203, 1984.

[35] T. Su and J. G. Dy, "In search of deterministic methods for initializing K-means and Gaussian mixture clustering," *Intelligent Data Analysis,* vol. 11, no. 4, pp. 319-338, 2007.

[36] S. Du, Y. Ma, S. Li, and Y. Ma, "Robust unsupervised feature selection via matrix factorization," *Neurocomputing,* vol. 241, pp. 115-127, 2017.

[37] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[38] D. Han and J. Kim, "Unsupervised simultaneous orthogonal basis clustering feature selection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5016-5023.

[39] W. Xu and Y. Gong, "Document clustering by concept factorization," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 202-209: ACM.

[40] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," *Pattern Recognition,* vol. 48, no. 1, pp. 10-19, 2015.

[41] L. Jing, M. K. Ng, and J. Z. Huang, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data," *IEEE Transactions on knowledge and data engineering,* vol. 19, no. 8, pp. 1026-1041, 2007.

[42] S. Chakraborty, D. Paul, S. Das, and J. Xu, "Entropy weighted power k-means clustering," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 691-701: PMLR.

[43] Z. Deng, S. Wang, and X. Wang, "Robust maximum entropy clustering algorithm RMEC and its outlier labeling," *Engineering Science,* vol. 6, no. 9, pp. 38-45, 2004.

[44] X. Wang and S. An, "Research on learning weights of fuzzy production rules based on maximum fuzzy entropy," *Journal of Computer Research Development,* vol. 43, no. 4, pp. 673-678, 2006.

[45] Z. Deng, K.-S. Choi, Y. Jiang, and S. Wang, "Generalized hidden-mapping ridge regression, knowledge-leveraged inductive transfer learning for neural networks, fuzzy systems and kernel methods," *IEEE transactions on cybernetics,* vol. 44, no. 12, pp. 2585-2599, 2014.

[46] S. Arora and B. Barak, *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[47] A. Fernandez, F. Herrera, O. Cordon and M. J. del Jesus *et al.*, "Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to?," *IEEE Computational Intelligence Magazine,* vol. 14, no. 1, pp. 69-81, 2019.

[48] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web,* vol. 2, no. 2, 2017.

[49] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 150-158.

[50] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* vol. 9, no. Nov, pp. 2579-2605, 2008.

[51] T.-S. Chua, J. Tang, R. Hong and H. Li *et al*., "NUS-WIDE: a real-world web image database from National University of Singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, p. 48: ACM.

[52] J. Winn and N. Jojic, "Locus: Learning object classes with unsupervised segmentation," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005, vol. 1, pp. 756-763: IEEE.

[53] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision,* vol. 60, no. 2, pp. 91-110, 2004.

[54] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proceedings of 12th International Conference on Pattern Recognition*, 1994, vol. 1, pp. 582-585: IEEE.

[55] J. Wu and J. M. Rehg, "Where am I: Place instance and category recognition using spatial PACT," in *2008 Ieee Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1-8: IEEE.

[56] H. Yu, M. Li, H.-J. Zhang, and J. Feng, "Color texture moments for content-based image retrieval," in *Proceedings. International Conference on Image Processing,* 2002, vol. 3, pp. 929-932: IEEE.

[57] S. Y. Wong, K. S. Yap, H. J. Yap, S. C. Tan, and S. W. Chang, "On equivalence of FIS and ELM for interpretable rule-based knowledge representation,"

IEEE transactions on neural networks and learning systems, vol. 26, no. 7, pp. 1417-1430, 2014.

[58] X. Cai, F. Nie, W. Cai, and H. Huang, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1737-1744.

[59] B. Zhang, Q. Qiang, F. Wang, and F. Nie, "Fast Multi-view Semi-supervised Learning with Learned Graph," *IEEE Transactions on Knowledge and Data Engineering*, 2020, DOI: 10.1109/TKDE.2020.2978844.

[60] J. H. Friedman, "On bias, variance, 0/1—loss, and the curse-of-dimensionality," *Data mining and knowledge discovery*, vol. 1, no. 1, pp. 55-77, 1997.

[61] C. Aschwanden, "Not even scientists can easily explain p-values," *FiveThirtyEight. com, Nov*, vol. 24, p. 2015, 2015.

[62] J. Li, E. Chen, Z. Ding and L. Zhu *et al.*, "Maximum Density Divergence for Domain Adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, DOI: 10.1109/TPAMI.2020.2991050.

[63] J. Li, M. Jing, K. Lu and L. Zhu *et al.*, "Locality preserving joint transfer for domain adaptation," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6103-6115, 2019.

[64] J. Li, K. Lu, Z. Huang and L. Zhu *et al.*, "Heterogeneous domain adaptation through progressive alignment," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 5, pp. 1381-1391, 2018.

[65] P. Xu, Z. Deng, J. Wang, Q. Zhang, K.-S. Choi, and S. Wang, "Transfer Representation Learning with TSK Fuzzy System," *IEEE Transactions on Fuzzy Systems*, 2019, DOI: 10.1109/TFUZZ.2019.2958299.

[66] C. Yang, Z. Deng, K.-S. Choi, and S. Wang, "Takagi–Sugeno–Kang transfer learning fuzzy logic system for the adaptive recognition of epileptic electroencephalogram signals," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 5, pp. 1079-1094, 2015.