Neurocomputing 214 (2016) 126–133

Contents lists available at ScienceDirect

Neurocomputing

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

Granular transfer learning using type-2 fuzzy HMM for text sequence recognition

Shichang Sun [a,b], Jian Yun [b], Hongfei Lin [a], Nanxun Zhang [c], Ajith Abraham [d], Hongbo Liu [a,e,*]

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China
[b] School of Computer Science and Engineering, Dalian Nationality University, Dalian 116600, China
[c] School of Japanese Studies, Dalian University of Foreign Languages, Dalian 116044, China
[d] Machine Intelligence Research Labs, Scientific Network for Innovation and Research Excellence, Auburn, WA 98071, USA
[e] Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 22 December 2015
Received in revised form
8 May 2016
Accepted 20 May 2016
Available online 7 June 2016

Keywords:
Transfer learning
Information granules
Text sequence recognition
Type-2 fuzzy set
(Hidden Markov Model) HMM

ABSTRACT

Context information plays an important role in text sequence recognition, but it is difficult to harness the uncertainty caused by conflicting implications. In this paper, we propose a novel Granular Transfer (GT) learning with type-2 fuzzy Hidden Markov Model (HMM) called GT2HMM, in which interpretable granules' representation is introduced to describe the contextual uncertainty for its transfer learning. The correspondences among words are transformed into information granules using fuzzy c-means. To fulfill the utilization of granular information in sequence recognition, we construct a type-2 fuzzy HMM which fuses labeled data and unlabeled observations. With a tunable granularity, correspondence information is refined in a coarse-to-fine manner in GT2HMM. Experiments on transductive and inductive transfer learning in part-of-speech (POS) tagging tasks verify the effectiveness of our proposed GT2HMM.

1. Introduction

In the recent decade transfer learning has become an expanding research area and has found many real-world applications for various kinds of domains [1–3]. Many transfer learning methods seek to find a common feature representation that minimizes the divergence among domains [4–6]. Specific to language structure, Blitzer [7] used correspondence to estimate the correlation between word features, and the difference between text domains is thus reduced. The correspondences among features are identified by modeling feature correlations with pivot features. It can be seen that context information is of importance in transfer learning.

Context information such as correspondence often has complex structures that allow different levels of abstraction and various types of arrangement. Current methods represent the context information as extended features in the form of numeric vectors [7], which can be viewed as implications for classification. But such implications may have conflicting influence on classification models, and such uncertainty is not well handled yet. Besides, the numeric vector representation is not interpretable and is not easily used in models which require symbolic observations as input. Granular computing (GrC) provides new ideas for the feature representation and utilization. Recent research in GrC shows that granular models [8] can be used as an abstraction of the original model so that it is more in rapport with the target environment. Therefore, we introduce GrC into the processing of contextual uncertainty.

With the help of information granules (IGs), we address the problems in context information representation for transfer learning. To make further utilization of the context information, it is necessary to handle the uncertainty in conflicting implications. Besides, it is desirable that the influence of context information can be interpreted. Our key idea is to granulate the correspondence knowledge and then build a granular model for text sequence transfer learning. The correspondence information is granulated using fuzzy c-means. Words are represented as pivot vectors, and are then clustered in the pivot space. A word can have different distances with different clusters. Such arrangement of the correspondence information allows flexible processing. For an exemplar word *extent*, an IG is built as in Fig. 1. The IG exists in the pivot space of *extent* and consists of clusters, showed as tables in the figure. The rows of the tables contain in-cluster words and the weights for each pivot word. In this example, the pivots are *required*, *of* and *to*. The circles around *extent* show that there are various distances to the clusters, which can be used as a granularity on how many correspondences are to be used. It can be seen

* Corresponding author at: Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093, USA.
E-mail addresses: scsun@dlut.edu.cn (S. Sun), yj@dlnu.edu.cn (J. Yun), hflin@dlut.edu.cn (H. Lin), zhangnanxun@foxmail.com (N. Zhang), ajith.abraham@ieee.org (A. Abraham), hongbo@sccn.ucsd.edu (H. Liu).

http://dx.doi.org/10.1016/j.neucom.2016.05.077
0925-2312/

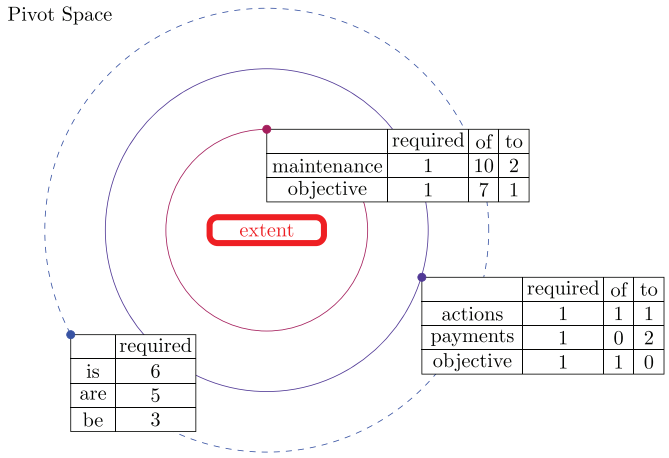


Fig. 1. An example of information granule in pivot space. The clusters are shown as tables with pivot features as columns. The top left corners of the tables indicate the positions of the cluster centers in the pivot space.

that the most similar cluster for *extent* is the one with words *maintenance* and *objective*. Given a certain granularity, part of the clusters can be selected to estimate the parameters for the word *extent*.

The implications of correspondences can be viewed as incomplete information with two types of uncertainties: randomness and fuzziness [9]. Randomness is processed by probabilistic models. For example, hidden Markov model (HMM) is an efficient model for sequence recognition problems. However, the contextual uncertainty, which lies in the conflicting implications from correspondence information, has not been processed accordingly. In this paper, such fuzziness is preserved in information granules and is then processed by a granular model, which makes a coarse-to-fine utilization of the correspondence information. Considering the discreteness of text applications, we use fuzzy set to represent the uncertainty in HMM parameters. As a result, HMM is extended to a type-2 fuzzy HMM. Overall, we present a novel granular transfer learning approach called GT2HMM which consists of information granulation of correspondence knowledge and a granular model using type-2 fuzzy HMM.

Compared with existing literature, our distinctive contribution is the introducing of GrC into the processing of contextual uncertainty for transfer learning. The major advantage of GT2HMM is the capacity to deal with the correspondence fuzziness using interpretable information granules' representation. Besides, the granules' representation is readily used in the HMM-like models which require symbolic observations.

The rest of the paper is organized as follows. We briefly review related works in Section 2. GT2HMM is proposed for transfer learning in text sequence recognition in Section 3. In Section 4 some experiments in POS tagging are conducted. Finally we conclude this paper in Section 5.

2. Related works

2.1. Sequence recognition

As we known, sequence recognition has been modeled with many machine learning techniques. Among them, HMM achieves a good tradeoff between expressive power and computational complexity, and is generally more computationally efficient than exponential family models [10,11]. So it has been applied in many challenging fields [12–14]. As shown in [15], HMM is one of the best sequence recognizers for tasks including Part-of-speech (POS) tagging.

POS tagging is a fundamental part of text mining and provides useful preprocessing tools for information extraction and retrieval [16,17]. For POS tagging tasks, context information can be used in an unsupervised manner [18,19], but these methods cannot work with predefined tag sets.

2.2. Transfer learning

Recently, transfer learning methods thrive in text mining applications [20–22]. In sequence recognition tasks, the distributional differences among domains are exacerbated. For example, a sentence has meaningful word orders and cannot be treated as bag-of-words.

Many existing works in transfer learning can be categorized into representation based methods according to Pan's survey [23]. Based on an assumption that classifiers learned using common feature representation will generalize better when used in a new domain, such works seek to learn a feature representation to encode the knowledge for transferring. Maximum mean discrepancy is used in [24] to reduce the difference between domains where the features can be represented in the reproducing kernel Hilbert space. The structural risk function is optimized in [5] together with joint and marginal distributions, so that adaptive classifiers can be learned under a regularization framework. However it is often difficult to find such intermediate representations for sequence observation spaces as in above methods. So some works seek to find contextual connections among the observations. For example, Blitzer [7] treats similarly the cross-domain observations that are correlated with many of the same pivot features. But the binary feature representation is not easily used to HMMs in text applications, where discrete observation space is used.

Parameter-based approaches, such as maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR), assume that the source model and the adapted model share some parameters or prior distributions. MAP and MLLR have been used for HMM-based transfer learning in applications including voice recognition [25–27] and text recognition [28]. However, the above works assume that the distributions of observations are continuous. For discrete HMM modeling with labeled data, the priors can be directly transferred from source domain using MAP, which is referred to as DT-HMM in this paper.

2.3. Granular models

Granular models are useful in transfer learning since it serves as a more abstract version of the original model. Pedrycz et al. [8,29,30] view information granularity as an important design asset for knowledge transfer and reusability and focus on the problem of distributing a certain level of granularity optimally among the parameters of the model on a certain criterion such as data coverage. By abstracting the original model through granulation, the granular model becomes more in rapport with the new domain. Song and Pedrycz [31] used interval connections in neural networks and output interval results. However, these models have not been investigated with the contextual information.

3. Granular transfer learning with type-2 fuzzy HMM

As portrayed in Fig. 2, our granular transfer learning with type-2 fuzzy HMM (GT2HMM) approach granulates the correspondence information from unlabeled data, and then processes the correspondence information granules with a type-2 fuzzy HMM. This type-2 fuzzy HMM is constructed on top of a HMM which is learned using labeled source domain data, and using labeled target domain data if provided. In order to keep the computing efficiency

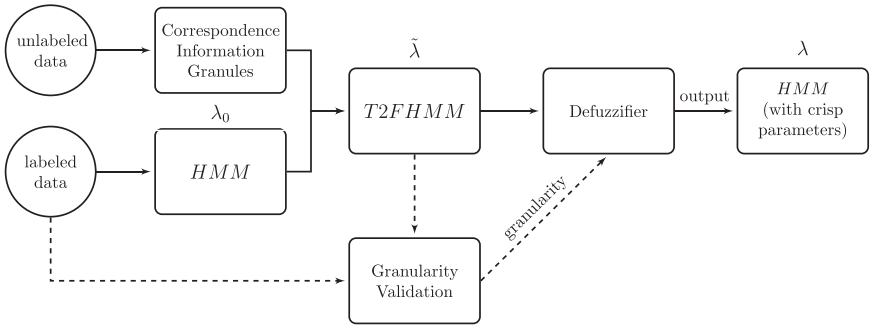


Fig. 2. Framework of granular transfer learning with type-2 fuzzy HMM.

of Viterbi algorithm, a defuzzifier is applied to GT2HMM before inference. If some labeled data in target domain is available, the architecture provides granularity validation as an option to choose domain specific granularity. As a result, a crisp HMM is output to evaluate the target domain sequences.

3.1. Sequence transfer learning

For clarity purposes, the notations on transfer learning of text sequence are presented as follows.

Let $V = \{V_1, \ldots, V_M\}$ denote the symbol set in texts, where M is the size of the symbol set, then V^t is the set of symbol sequences, where t is a variable that denotes the length of the sequence. Most methods for text sequence recognition assume a hidden layer which consists of a set of hidden states denoted by $S = \{S_1, \ldots, S_N\}$, where N is the size of the state set.

Inside V^t, sequence feature space $\mathcal{O}$ is a language dependent subset. Then an observation sequence of $\mathcal{O}$ can be denoted by $O = o_1 o_2 \ldots o_t$. Accordingly, inside S^t, sequence state space $\mathcal{Q}$ is a subset associated with $\mathcal{O}$. Then a labeled sequence of $\mathcal{Q}$ can be denoted by $Q = q_1 q_2 \ldots q_t$, where q_i is corresponding to o_i for $1 \le i \le t$.

Definition 1 (*Sequence domains*). By the condition whether labeled data exist, sequence domains $\mathcal{D}$ can be categorized into labeled domain $\mathcal{D}_l$ and unlabeled domains $\mathcal{D}_u$. $\mathcal{D}_u = \langle \mathcal{O}, P(\mathcal{O}) \rangle$ is a tuple of $\mathcal{O}$ and a marginal distribution $P(\mathcal{O})$, where $O \in \mathcal{O}$. $\mathcal{D}_l = \langle \mathcal{O}, \mathcal{Q}, P(\mathcal{O}), P(\mathcal{O}, \mathcal{Q}) \rangle$ is a tuple of $\mathcal{O}$, $\mathcal{Q}$, marginal distribution $P(\mathcal{O})$ and joint distribution $P(\mathcal{O}, \mathcal{Q})$, where $O \in \mathcal{O}$ and $Q \in \mathcal{Q}$.

In many problems, we can assume that observations o_i and tags q_i are drawn independently and identically from some joint distribution, but text domains $\mathcal{D}$ has significant sequential correlation. Therefore $P(\mathcal{O}, \mathcal{Q})$ cannot be estimated directly using $P(o_i, q_i)$. In text sequence recognition applications include POS tagging, text chunking and named entity recognition, the orders in sequences have significance and cannot be treated as bag-of-words.

Definition 2 (*Sequence recognition task*). A sequence recognition task $\mathcal{T}$ for a specific domain $\mathcal{D}$ is to find $Q \in \mathcal{Q}$ for $O \in \mathcal{O}$ that is optimal under certain criterion.

Such criterion can be the conditional probability:

$$\underset{Q}{\operatorname{argmax}} P(Q|O, \lambda) \tag{1}$$

where λ represents the model parameters learned from $\mathcal{D}$. Note that in sequence recognition we seek to find a $Q = q_1 q_2 \ldots q_t$ which is optimal for the whole sequence. HMM is one of the most efficient sequence recognizers which models the sequence generation with N hidden states and M symbols.

Definition 3 (*Hidden Markov Model*). Hidden Markov model [32] is characterized by its parameters $\lambda = \{\pi, A, B\}$, where $\pi = \{\pi_i\}$ is initial states distribution, $A = \{a_{ij}\}$ is transition probability matrix and $B = \{b_j(k)\}$ is emission probability matrix for $1 \le i, j \le N$ and $1 \le k \le M$.

Definition 4 (*Sequence transfer learning*). In order to assist the sequence recognition tasks $\mathcal{T}^T$ in target domain, sequence transfer learning enables the transferring of knowledge in $\{\mathcal{D}^S, \mathcal{D}^T\}$ and source domain task $\mathcal{T}^S$.

3.2. Correspondence information granules

Correspondence is important context information in unlabeled text and is usually abundant in both domains. We use the co-occurrences between pivot features and common features to build correspondences. In modeling the cross-domain correspondences among features, Blitzer [7] emphasizes *pivot* features which behave the same way among different domains. We put forward this method and use a ⟨*word on the right*⟩ type of pivot features to extract the correlations between words and their context. After automatically identifying pivots, vectors in pivot space become building blocks for correspondence information granules.

Information granules arrange data as complex information entity and serve as a basis of coarse-to-fine processing in granular models. Information granulation can be conducted by user-based approaches or algorithm-based approaches. As pointed out by Pedrycz in [33], user-based approaches are lack of problem specificity, whereas algorithm-based approaches are lack of semantics. Since correspondence information contains semantics knowledge of text domains, the usage of correspondences enables our approach to use algorithmic granulation without missing semantics.

Correspondence information is granulated using fuzzy c-means [34] in pivot space. As depicted in Fig. 1, a correspondence IG contains the correlation between words and a series of word clusters. The granularity is the extent to which the clusters of IG are included to represent the correspondence knowledge for words. In POS tagging task, the clusters are formed for each POS to reveal the contextual influences on the fact that some words share the same POS. Given a collection of M word vectors v_l, the cluster structure is formed by maximizing the objective function as follows:

$$Q = \sum_{l=1}^{M} \sum_{c=1}^{C} \omega_{lc}^{m} \parallel v_l - v_c \parallel^2 \tag{2}$$

where v_c denotes the prototypes of cluster c, C is the number of clusters in correspondence to IG and ω_{lc} stands for a partition matrix. The fuzzifier m is commonly set to 2.

One important clustering parameter is the initial centers vector. There are two methods to choose the initial centers. One is to randomly select p points, which are usually p objects to be clustered, as initial centers. The other is to choose p points by a fixed procedure. We prefer to use the fixed choosing procedure, since it produces more stable results for same datasets.

It can be seen that correspondence IG is a flexible way to utilize the complex context information, especially when the context implications are sometime conflicting. Correspondence IG helps to contain the conflicting implications into a fuzzy model.

3.3. Type-2 fuzzy HMM for correspondence IG

The correspondence IG will be input into a granular model, thus the fuzzy implications of correspondences can be exploited in sequence recognition. In order to utilize this context information as well as to preserve the simplicity and computing efficiency of HMM, we extend HMM to a type-2 fuzzy HMM (T2FHMM).

Type-2 fuzzy models [35–37] have emerged as a extension to type-1 fuzzy. The merits of type-2 fuzzy sets are that both random and fuzzy uncertainties can be modeled. For a discrete symbol set V indexed by k, a type-2 fuzzy set is defined as:

Definition 5 (*Type-2 fuzzy set*). A type-2 fuzzy set (T2 FS) $\tilde{U}$ is characterized by a type-2 membership function (T2 MF) $\mu_{\tilde{U}}(k, u)$, where u is the primary grade and its domain J_k is called primary membership [35].

$$\tilde{U} = \{((k, u), \mu_{\tilde{U}}(k, u)) | \forall\ k = 1, \ldots, N,\ \forall\ u \in J_k \subseteq [0, 1]\} \tag{3}$$

Definition 6 (*Type-2 fuzzy HMM*). T2FHMM extends HMM with all its parameters $\tilde{\lambda} = \{\pi, A, \mathbb{U}\}$:

- Initial states' distribution
  The initial states' distribution is denoted by π, which is a vector of initial states where $\pi_i = P(q_1 = S_i),\ 1 \le i \le N$. Here q_1 is the state at initial time that satisfies the constraints $0 \le \pi_i \le 1$, $1 \le i \le N$, and $\sum_{i=1}^{N}(\pi_i) = 1$.
- Transition probability matrix
  The state transition matrix is denoted by $A = \{a_{ij}\}$, where $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, $\sum_{j=1}^{N} a_{ij} = 1$ and $1 \le i, j \le N$.
- Emission T2 FS vector
  Let $\mathbb{U} = \{\tilde{U}_j | j = 1, \ldots, N\}$ denote the emission T2 FS vector, where $\tilde{U}_j$ is a T2 FS specific to state S_j:

$$\tilde{U}_j = \{((k, u), \mu_{\tilde{U}_j}(k, u)) | \forall\ \ k = 1, \ldots, N,\ \forall\ u \in J_k \subseteq [0, 1]\} \tag{4}$$

  where $\tilde{U}_j$ is characterized by a type-2 membership function $\mu_{\tilde{U}_j}(k, u)$.

An example of T2 FS $\tilde{U}_j$ is depicted in Fig. 3, which also shows the concepts of primary grade, secondary grade and secondary MF. The T2 MF $\mu_{\tilde{U}_j}(k, u)$ reflects the uncertainty of primary grades of emissions. For a specific $k = k_0$, the secondary MF $\mu_{\tilde{U}_j}(k_0, u)$ is a vertical slice of T2 MF $\mu_{\tilde{U}_j}(k, u)$. The $\mu_{\tilde{U}_j}(k_0, u)$ represents our belief to the primary membership J_{k_0}.

The design of T2FHMM is a tradeoff between expressive power and efficiency. To preserve the simplicity of HMM, we choose to keep the crisp design of state transition matrix A. To accommodate the correspondence information granules of symbols, we replace the emission matrix B of HMM with a vector of T2 fuzzy sets. In POS tagging problem, A depicts the randomness of the transitions between POS tags, while $\mathbb{U}$ contains both randomness and fuzziness of tag-symbol emissions.

By utilizing our correspondence information granules, T2FHMM fuses information in labeled data and information in unlabeled target domain data. T2FHMM provides the procedure of obtaining $\mathbb{U}$, which is stated in Algorithm 1. For an out-of-vocabulary word V_k, lines 3–8 calculate the primary memberships J_k and secondary MFs $\mu_{\tilde{U}_j}(k, u)$.

As in Eq. (5), J_k is related to the clusters of $IG(S_j)$, which is the correspondence IG specific to state j. Here $J_k{}^c$ denotes a primary grade that relates to cluster c, and C denotes the number of clusters in $IG(S_j)$.

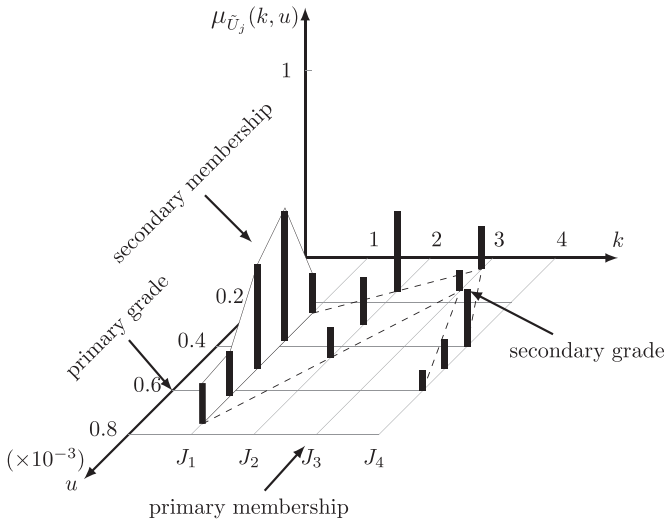


Fig. 3. Example of a T2 FS specific to state S_j.

$$J_k = \{J_k^c | c = 1, \ldots, C\} \tag{5}$$

To approximate $J_k{}^c$, we use the emission probabilities of within-cluster words l' in the weighted arithmetic mean of Eq. (7). Since the partition matrix ω of Eq. (2) gives membership grades of words to cluster c, $\omega_{l'c}$ can be viewed as membership functions for cluster-belonging fuzzy sets W^c for $1 \le c \le C$. For each cluster c, only words in the α-cut of fuzzy set W^c, as shown is Eq. (6), are used to approximate $J_k{}^c$.

$$W_\alpha^c = \{l' | \omega_{l'c} \ge \alpha\} \tag{6}$$

$$J_k^c = \frac{\sum_{l'} \omega_{l'c} b_j(l')}{\sum_{l'} \omega_{l'c}}, \tag{7}$$

After the cluster centers v_c are obtained in fuzzy c-means algorithm as in Eq. (2), the membership grades ω_{kc}^* of new data $\{V_k\}$ to the clusters can be predicted by fixing the centers and updating the partition matrix. These membership grades can be used as second grades as in the following equation:

$$\mu_{\tilde{U}_j}(k, u = J_k^c) = \omega_{kc}^* \tag{8}$$

For a labeled word V_k, J_k has only one element of which the belief is 1 as in lines 9–11 of Algorithm 1. Then a secondary MF $\mu_{\tilde{U}_j}(k, u)$ can be built using $\{(J_k^c, \mu_{\tilde{U}_j}(k, u = J_k^c))\}$ as in line 13. Finally the emission T2 FS vector $\mathbb{U} = \{\tilde{U}_j\}$ is returned from the algorithm.

Algorithm 1. Procedure of obtaining $\mathbb{U}$.

Input: Correspondence information granule set $\{IG(S_j)\}$, observation set $\{V_k\}$
Require: Symbol set V, hidden state set S, transition matrix $B = \{b_j(k)\}$
Output: $\mathbb{U} = \{\tilde{U}_j\}$
01. for each state $S_j \in S$:
02. for each target word $V_k \in V$:
03. if V_k is an out-of-vocabulary word:
04. for each cluster $c \in IG(S_j)$:
05. build $J_k{}^c$ as in Eq. (7)
06. calculate a secondary grade $\mu_{\tilde{U}_j}(k, u = J_k^c)$ as in Eq. (8)
07. save($J_k{}^c$, $\mu_{\tilde{U}_j}(k, u = J_k^c)$)
08. end for

09. else:
10. $J_k^1 = b_j(k)$
11. save ($J_k^1 = b_j(k), 1$)
12. build primary memberships $J_k = \{J_k^c\}$ as in Eq. (5)
13. build secondary MF $\mu_{\tilde{U}_j}(k, u)$ with $\{(J_k^c, \mu_{\tilde{U}_j}(k, u = J_k^c))\}$
14. end for
15. build $\tilde{U}_j$ as in Eq. (4)
16. end for
17. return $\mathbb{U} = \{\tilde{U}_j\}$

3.4. Using granularity

Through GT2HMM, the fuzziness of correspondence information can be controlled by granularity. Finding a right granularity is the procedure of searching for the right level of abstraction for target domain.

Given an allowable degree of granularity, T2FHMM can be defuzzified in order to work with the Viterbi algorithm. This degree can be controlled by α value when α-plane [38] of $\mu_{\tilde{U}_j}(k, u)$ is used as in Eq. (9).

$$\tilde{U}_j^{\alpha} = \{(k, u) | \mu_{\tilde{U}_j}(k, u) \geq \alpha\} \tag{9}$$

Then a crisp emission matrix $\{u_j(k)\}$ can be output using arithmetic means above the α-plane as in the following equation:

$$u_j(k) = \frac{\sum_{(k,u(\in \tilde{U}_j^{\alpha}} u \mu_{\tilde{U}_j}(k, u)}{\sum_{(k,u)\in \tilde{U}_j^{\alpha}} \mu_{\tilde{U}_j}(k, u)} \tag{10}$$

If some labeled data exist in target domain, GT2HMM can provide granularity validation as an option. The procedure is to use a small part of labeled target domain data to test a list of granularity degrees and to use the degree with best accuracy.

We can see that the uncertainty of correspondence information is exploited in a coarse-to-fine manner. First, the emission implications from correspondence information are contained in T2FHMM as primary memberships and T2 MFs. Second, an appropriate granularity is chosen to output a crisp emission matrix using T2 MFs above the alpha-plane.

4. Experiments

Recall that POS tagging is a classical candidate for testing sequence recognition methods. To demonstrate the characteristics of the proposed approach, we empirically evaluate our algorithms in POS tagging tasks, which are designed using different combinations of categories in Brown corpus.

4.1. Datasets

We experiment on cross-domain tasks using Brown corpus. The Brown corpus is the first million-word English corpus that compiled as a general style English-language text. It gathers 500 text samples, which have been categorized by genres, such as *news* and *editorial*.

We use 10 categories in Brown corpus to construct two groups of cross-domain tasks. The first group uses *news, editorial, fiction, government* and *adventure*, and the second group uses *mystery, hobbies, reviews, romance* and *learned*. In each experiment group we construct 20 cross-domain text tasks, each of which uses "source vs. target" pairs from different categories. The name and the Kullback–Leibler divergence (KLD) of the tasks are showed in the first two columns in Tables 1 and 2.

4.2. Experimental settings

We implement the experiment program based on natural language toolkit (NLTK) [39]. The performance metric used here is the accuracy on the predictions of token–tag pairs.

We compare the sequence recognition accuracy of GT2HMMs and other methods in two settings: transductive transfer learning where labeled data are not available in target domain, and inductive transfer learning where some labeled data are available in target domain. We compare GT2HMM (transductive) and GT2HMM (inductive) with HMM and DTHMM. As inductive transfer learning method, DTHMM can be compared with GT2HMM (inductive).

The procedures of choosing initial centers are optional in building granules. The straightforward choice is to initialize the centers randomly. But we can also use a fixed procedure in order to obtain steady clustering results. The fixed procedure chooses top p pivots which have most correlated words. In experiments, we control p by average cluster sizes which is set to 45. Besides, the αs in Eqs. (6) and (9) are set to 0.05.

Before experiments, 20 cross-domain tasks are constructed in each of the two cross-domain groups using Brown corpus. In transductive experimental setting, we use 500 labeled sentences in source domain and 500 unlabeled sentences in target domain. In inductive setting, we use 50 labeled sentences in target domain as validation sets.

4.3. Overall performance

For the first cross-domain group, the experimental results are shown in Table 1 and Fig. 4, and for the second cross-domain group in Table 2 and Fig. 5. Then we can make the following observations.

In transductive setting, GT2HMM always outperforms HMM. According to the "Average" rows in Tables 1 and 2, the average accuracy improvements are more than 3.01% and 3.66% for the two experimental groups. Sometimes GT2HMM can even outperform DTHMM, which is an inductive transfer learning method. These results verify the effectiveness of the utilization of correspondence information in GT2HMM. As the example illustrated in Fig. 1, the improvement of GT2HMM is interpretable by reading the word

Table 1
Performance comparison for tasks in the first group.

Tasks	KLD	HMM	GT2HMM (transductive)	DTHMM	GT2HMM (inductive)
new vs. edi	0.854	76.73	79.22	77.49	80.31
edi vs. new	0.789	74.99	78.33	77.15	79.98
new vs. fic	1.272	73.94	76.55	76.06	78.91
fic vs. new	1.118	65.27	69.61	69.12	72.90
new vs. gov	1.111	77.35	79.84	79.04	81.66
gov vs. new	1.055	67.39	72.44	71.89	75.47
new vs. adv	1.481	72.29	73.97	76.09	77.80
adv vs. new	1.359	60.98	63.31	66.31	67.97
edi vs. fic	1.062	76.17	78.88	78.58	80.99
fic vs. edi	1.009	68.56	72.10	71.83	75.27
edi vs. gov	0.931	76.77	78.92	79.08	81.41
gov vs. edi	1.005	71.50	76.01	72.92	77.43
edi vs. adv	1.239	73.59	75.55	76.47	78.76
adv vs. edi	1.229	63.74	67.56	68.36	72.32
fic vs. gov	1.301	66.20	69.61	71.94	75.20
gov vs. fic	1.479	67.74	71.24	73.18	75.86
fic vs. adv	0.752	80.63	83.56	81.76	84.31
adv vs. fic	0.751	77.44	79.70	78.99	81.32
gov vs. adv	1.676	64.75	66.56	72.53	72.71
adv vs. gov	1.548	61.80	64.97	67.82	72.18
Average		70.89	73.90	74.33	77.14
Paired t test	t value	–	14.41	–	12.49
	p value	–	1.11×10^{-11}	–	1.32×10^{-10}

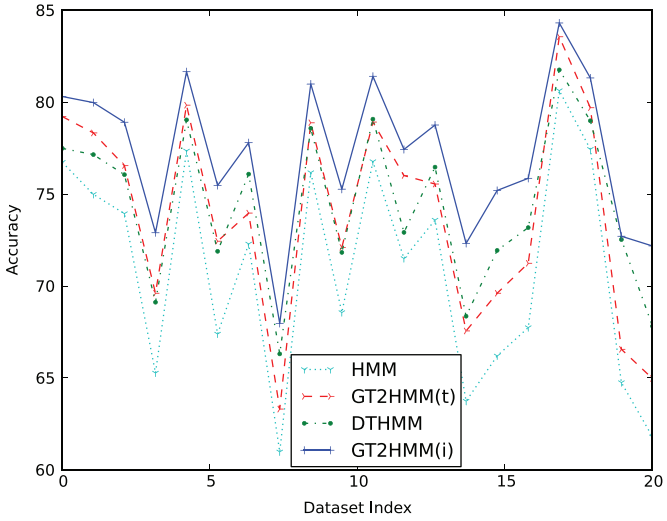


Fig. 4. Sequence recognition performance on the first group of 20 cross-domain tasks. In the legend, (t) stands for (transductive) and (i) stands for (inductive).

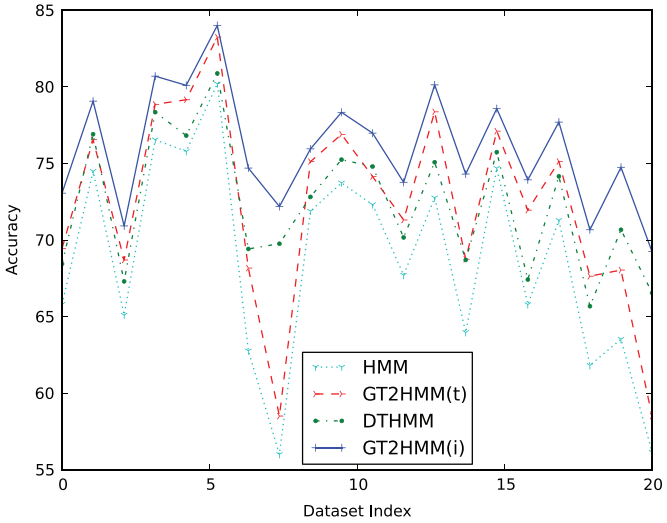


Fig. 5. Sequence recognition performance on the second group of 20 cross-domain tasks. In the legend, (t) stands for (transductive) and (i) stands for (inductive).

Table 2
Performance comparison for tasks in the second group.

Tasks	KLD	HMM	GT2HMM (transductive)	DTHMM	GT2HMM (inductive)
mys vs. hob	1.121	65.81	69.45	68.44	73.06
hob vs. mys	1.170	74.49	76.55	76.91	79.07
mys vs. rev	1.057	65.10	68.70	67.30	70.92
rev vs. mys	1.083	76.55	78.86	78.35	80.70
mys vs. rom	0.724	75.79	79.16	76.82	80.10
rom vs. mys	0.739	80.20	83.25	80.88	84.01
mys vs. lea	1.943	62.77	68.17	69.42	74.70
lea vs. mys	1.771	56.01	58.51	69.76	72.19
hob vs. rev	0.965	71.89	75.12	72.82	75.95
rev vs. hob	0.914	73.73	76.90	75.25	78.34
hob vs. rom	1.123	72.31	74.15	74.80	76.98
rom vs. hob	1.063	67.71	71.31	70.17	73.77
hob vs. lea	1.353	72.76	78.37	75.08	80.13
lea vs. hob	1.399	63.97	68.75	68.70	74.31
rev vs. rom	1.051	74.69	77.11	75.74	78.58
rom vs. rev	0.976	65.79	71.95	67.42	73.93
rev vs. lea	1.345	71.29	75.13	74.16	77.70
lea vs. rev	1.517	61.82	67.64	65.68	70.67
rom vs. lea	1.775	63.53	68.04	70.67	74.75
lea vs. rom	1.766	56.16	58.51	66.55	69.24
Average		68.62	72.28	72.25	75.96
Paired t test	t value	–	12.44	–	13.28
	p value	–	1.40×10^{-10}	–	4.58×10^{-11}

clusters used in approximation. In this POS tagging experiment, granularity has the meaning of a degree at which the clusters are allowed to approximate the target word.

In inductive setting, both DTHMM and GT2HMM have better results than HMM, but GT2HMM outperforms DTHMM with average improvements of 2.81% and 3.71% for the two experimental groups, according to the "Average" rows in Tables 1 and 2. Since Figs. 4 and 5 show the performance of the algorithms in every combination of source vs. target domain pairs $\{\mathcal{D}^S, \mathcal{D}^T\}$, the effectiveness and stableness of GT2HMM are testified.

Given the experimental results in Tables 1 and 2, the paired t test is conducted to assess the significance of performance improvements in "GT2HMM (transductive) vs. HMM" and "GT2HMM (inductive) vs. DTHMM". There are two hypotheses for each of the two paired t test:

(1) *The null hypotheses*: The performance of GT2HMM is no better than that of the other model.
(2) *The alternative hypotheses*: The performance of GT2HMM is significantly better than that of the other model.

The level of significance is set as $\alpha_H = 0.05$. As can be seen in the "paired t test" row of the tables, the values of p is far less than that of α_H. The null hypotheses is thus rejected in the paired t tests. Therefore, it is verified that GT2HMM models have gained significant performance improvements.

The sizes of datasets often have influence on the performance of machine learning algorithms. Here we conduct experiments to see how sequence recognition accuracies evolve when the sizes of source domain dataset vary. The sizes of both source and target domain data vary from 150 to 500. The typical experimental results are shown in Fig. 6. When corpus size increases, it can be seen that the accuracies of GT2HMM (inductive) and GT2HMM (transductive) increase more rapidly than those of HMM and DTHMM. When dataset sizes are over 350 sentences, GT2HMM (transductive) can outperform the inductive method DTHMM. Overall, GT2HMM can improve the performance in various sizes of datasets.

4.4. Parameter sensitivity

The number of clusters can influence the granules that built for granular transfer learning. We control the number of clusters by average cluster sizes. The following experiment shows sequence recognition accuracies on various cluster sizes. Both source and target domain use 500 sentences and target domain data are divided into 4 groups to test the performance. We use "new vs. edi" and "edi vs. fic" as two typical cases and show the experimental results in Figs. 7 and 8. From the figures we can find that the accuracies are not heavily influenced by the initial cluster numbers.

5. Conclusion

In this paper, we proposed GT2HMM to utilize complex correspondence information for text sequence recognition in a coarse-to-fine manner. GT2HMM consists of correspondence information granulation which constructs a useful abstraction of the correspondence information, and a type-2 fuzzy HMM which flexibly

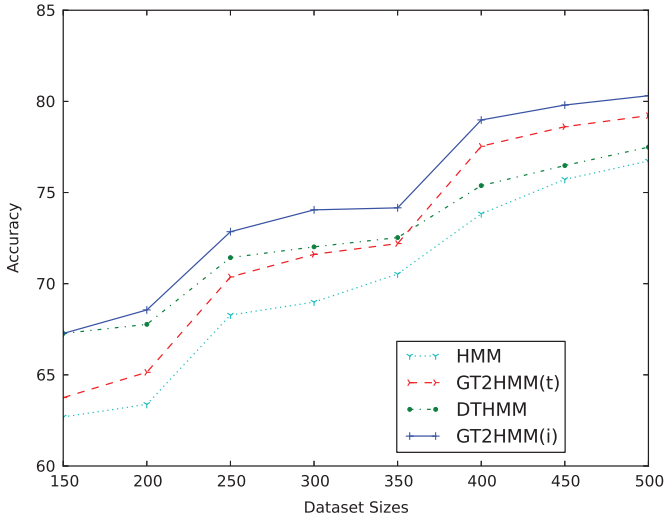


Fig. 6. Accuracies on various corpus sizes. In the legend, (t) stands for (transductive) and (i) stands for (inductive).

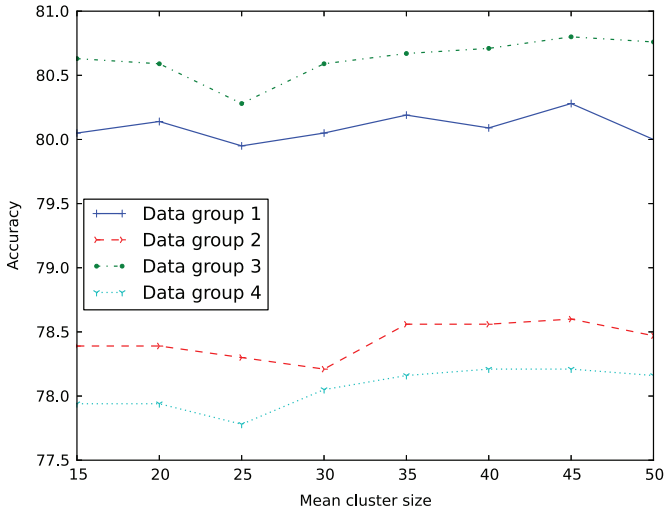


Fig. 7. Granular model performance on various cluster sizes in "new vs. edi" task.

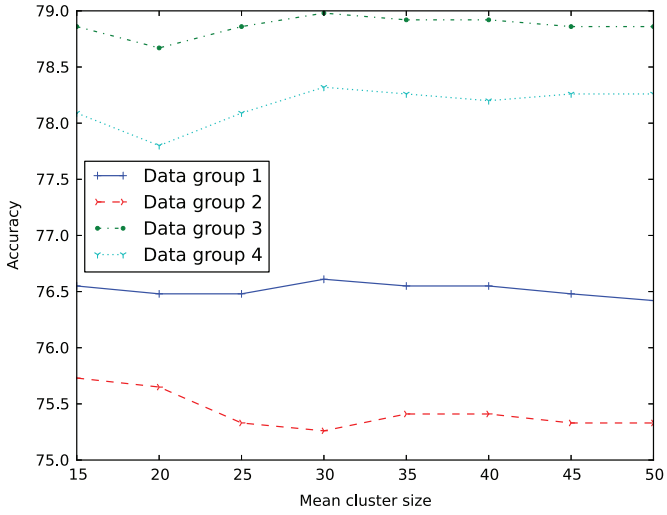


Fig. 8. Granular model performance on various cluster sizes in "edi vs. fic" task.

deals with the fuzziness of conflicting implications as well as the randomness of observations. The advantages of the context representation in this paper are the capacity to deal with contextual uncertainty, the applicability for the models using symbolic features, and the interpretability of the extended features. Experiments show performance improvements in various cross-domain combinations and with various corpus sizes for both transductive and inductive settings.

We will further research on the influence of clustering procedure on the granular model. On the one hand clustering method other than fuzzy c-means will be further investigated. On the other hand some criterion such as cluster validity indices [40] would be tested.

Acknowledgments

The authors would like to thank Liyong Zhang of Dalian University of Technology for research cooperation on information granulation. This work was supported partly by the National Natural Science Foundation of China (61173035 and 61472058), the Program for New Century Excellent Talents in University (NCET-11-0861) and the Humanity and Social Science Foundation of Ministry of Education of China (12YJCZH263).

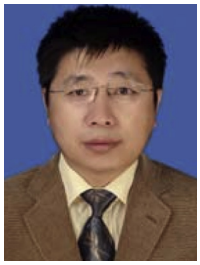
References

[1] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, F. Provost, Machine learning for targeted display advertising: transfer learning in action, Mach. Learn. 95 (1) (2014) 103–127.
[2] F. Zhuang, P. Luo, C. Du, Q. He, Z. Shi, H. Xiong, Triplex transfer learning: exploiting both shared and distinct concepts for text classification, IEEE Trans. Cybern. 44 (7) (2014) 1191–1203.
[3] W. Pan, A survey of transfer learning for collaborative recommendation with auxiliary data, Neurocomputing 177 (2016) 447–453.
[4] X. Zhou, P. Guo, C.L.P. Chen, Covariance matrix estimation with multi-regularization parameters based on MDL principle, Neural Process. Lett. 38 (2) (2013) 227–238.
[5] M. Long, J. Wang, G. Ding, S.J. Pan, P.S. Yu, Adaptation regularization: a general framework for transfer learning, IEEE Trans. Knowl. Data Eng. 26 (5) (2014) 1076–1089.
[6] L. Duan, D. Xu, I.W. Tsang, Domain adaptation from multiple sources: a domain-dependent regularization approach, IEEE Transactions on Neural Networks and Learning Systems 23 (3) (2012) 504–518.
[7] J. Blitzer, R. McDonald, F. Pereira, Domain adaptation with structural correspondence learning, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2006, pp. 120–128.
[8] W. Pedrycz, B. Russo, G. Succi, Knowledge transfer in system modeling and its realization through an optimal allocation of information granularity, Appl. Soft Comput. 12 (8) (2012) 1985–1995.
[9] J. Zeng, Z.-Q. Liu, Type-2 fuzzy hidden Markov models and their application to speech recognition, IEEE Trans. Fuzzy Syst. 14 (3) (2006) 454–467.
[10] C. Sutton, A. McCallum, Composition of conditional random fields for transfer learning, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 748–754.
[11] J. Liu, K. Yu, Y. Zhang, Y. Huang, Training conditional random fields using transfer learning for gesture recognition, in: Proceedings of IEEE International Conference on Data Mining, 2010, pp. 314–323.
[12] Y. Cao, Y. Li, S. Coleman, A. Belatreche, T. McGinnity, Adaptive hidden Markov model with anomaly states for price manipulation detection, IEEE Trans. Neural Netw. Learn. Syst. 26 (2) (2015) 318–330.
[13] O. Samanta, U. Bhattacharya, S. Parui, Smoothing of HMM parameters for efficient recognition of online handwriting, Pattern Recognit. 47 (11) (2014) 3614–3629.
[14] Z. Liu, S. Sarkar, Improved gait recognition by gait dynamics normalization, IEEE Trans. Pattern Anal. Mach. Intell. 28 (6) (2006) 863–876.
[15] T. Brants, TnT: a statistical part-of-speech tagger, in: Proceedings of the Sixth Conference on Applied Natural Language Processing, 2000, pp. 224–231.
[16] C. Li, A. Sun, J. Weng, Q. He, Tweet segmentation and its application to named entity recognition, IEEE Trans. Knowl. Data Eng. 27 (2) (2015) 558–570.
[17] T. Nakahara, H. Morita, Pattern mining in POS data using a historical tree, in: Sixth IEEE International Conference on Data Mining Workshops, IEEE, 2006, pp. 570–574.
[18] A. Clark, Inducing syntactic categories by context distribution clustering, in: Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning, vol. 7, 2000, pp. 91–94.
[19] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[20] K.D. Feuz, D.J. Cook, Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (FSR), ACM Trans. Intell. Syst. Technol. 6 (1) (2015) 1–27.
[21] M. Long, J. Wang, J. Sun, P.S. Yu, Domain invariant transfer kernel learning, IEEE Trans. Knowl. Data Eng. 27 (6) (2015) 1519–1532.

[22] K.-N. Tran, P. Christen, Cross-language learning from bots and users to detect vandalism on wikipedia, IEEE Trans. Knowl. Data Eng. 27 (3) (2015) 673–685.
[23] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[24] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Netw. 22 (2) (2011) 199–210.
[25] D.K. Kim, N.S. Kim, Maximum a posteriori adaptation of HMM parameters based on speaker space projection, Speech Commun. 42 (1) (2004) 59–73.
[26] N.S. Kim, J.S. Sung, D. Hong, Factored MLLR adaptation, Signal Process. Lett. 18 (2) (2011) 99–102.
[27] O. Siohan, C. Chesta, C.-H. Lee, Joint maximum a posteriori adaptation of transformation and HMM parameters, IEEE Trans. Speech Audio Process. 9 (2001) 417–428.
[28] K. Ait-Mohand, T. Paquet, N. Ragot, Combining structure and parameter adaptation of HMMs for printed text recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (9) (2014) 1716–1732.
[29] O.F. Reyes-Galaviz, W. Pedrycz, Granular fuzzy modeling with evolving hyperboxes in multi-dimensional space of numerical data, Neurocomputing 168 (2015) 240–253.
[30] W. Wang, W. Pedrycz, X. Liu, Time series long-term forecasting model based on information granules and fuzzy clustering, Eng. Appl. Artif. Intell. 41 (2015) 17–24.
[31] M. Song, W. Pedrycz, Granular neural networks: concepts and development schemes, IEEE Trans. Neural Netw. Learn. Syst. 24 (4) (2013) 542–553.
[32] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[33] W. Pedrycz, A. Gacek, Temporal granulation and its application to signal analysis, Inf. Sci. 143 (1) (2002) 47–71.
[34] H. Izakian, W. Pedrycz, Agreement-based fuzzy c-means for clustering data with blocks of features, Neurocomputing 127 (2014) 266–280.
[35] J.M. Mendel, R.I.B. John, Type-2 fuzzy sets made simple, IEEE Trans. Fuzzy Syst. 10 (2) (2002) 117–127.
[36] D. Bhattacharya, A. Konar, P. Das, Secondary factor induced stock index time-series prediction using self-adaptive interval type-2 fuzzy sets, Neurocomputing 171 (2016) 551–568.
[37] S.M. Chen, Y.C. Chang, J.S. Pan, Fuzzy rules interpolation for sparse fuzzy rule-based systems based on interval type-2 Gaussian fuzzy sets and genetic algorithms, IEEE Trans. Fuzzy Syst. 21 (3) (2013) 412–425.
[38] J.M. Mendel, F. Liu, D. Zhai, Alpha-plane representation for type-2 fuzzy sets: theory and applications, IEEE Trans. Fuzzy Syst. 17 (5) (2009) 1189–1207.
[39] E. Loper, S. Bird, NLTK: the natural language toolkit, in: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, 2002, pp. 62–69.
[40] A. Celikyilmaz, I.B. Trksen, Validation criteria for enhanced fuzzy clustering, Pattern Recognit. Lett. 29 (2) (2008) 97–108.

Shichang Sun is a Ph.D. candidate at Dalian University of Technology, and a lecturer at Dalian Nationality University. His main research interests are machine learning and data mining, especially transfer learning and sequence classification for imperfect and complex data.

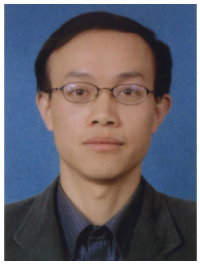
Jian Yun received his Ph.D. degree at Shanghai Normal University in 2010. He is currently an associate professor at Dalian Nationalities University. His current research interests include social computing.



Hongfei Lin received his Ph.D. at the Northeastern University, China. Currently he is a Professor at School of Computer Science and Technology, Dalian University of Technology. His research interests are in information retrieval, text mining, natural language processing and sentimental analysis.

Nanxun Zhang is a Ph.D. candidate at Dalian University of Foreign Languages and Josai International University. Her main research interests are natural language processing and machine translation.

Ajith Abraham is the Director of Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, WA, USA. He received the Ph.D. degree in Computer Science from Monash University, Melbourne, Australia. His research and development experience includes more than 25 years in the industry and academia. He works in a multidisciplinary environment involving machine intelligence, network security, various aspects of networks, e-commerce, Web intelligence, data mining and their applications to various real-world problems. He has authored/co-authored more than 900 publications, and some of the works have also won best paper awards at international conferences. He has given more than 60 plenary lectures and conference tutorials in these areas. He serves/has served the editorial board of over 50 International journals and has also guest edited over 40 special issues on various topics. Since 2008, he is the Chair of IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing and a Distinguished Lecturer of IEEE Computer Society representing Europe (since 2011). More information at: http://www.softcomputing.net.



Hongbo Liu received his three level educations (B.Sc., M.Sc., Ph.D.) at the Dalian University of Technology, China. Currently he is a Professor at Institute of Cognitive Information Technology (ICIT), with an affiliate appointment in the Institute for Neural Computation, University of California San Diego, USA. His research interests are in system modeling and optimization involving soft computing, probabilistic modeling, cognitive computing, machine learning, data mining, etc. He participates and organizes actively international conference and workshop and international journals/publications.