文章编号：1003-0077（2011）00-0000-00

利用源域结构的粒迁移学习及词性标注应用

孙世昶[1,2]，林鸿飞[1]，孟佳娜[2]，刘洪波[3]

（1. 大连理工大学计算机学院，辽宁 大连，116023；

2. 大连民族大学计算机学院，辽宁 大连，116600；

3. 大连海事大学信息科学技术学院，辽宁 大连，116026）

摘要：迁移学习在一定程度上减轻了目标域的数据稀疏问题对泛化能力的影响，然而泛化能力的提高仍然受到负迁移等问题的影响。为了解决负迁移问题，本文提出使用源域结构的文本语料的信息粒化方法，用区间信息粒表示出源域数据集的结构对数据集中统计量的影响。然后提出区间二型模糊隐马尔可夫模型(Interval type-2 fuzzy Hidden Markov Model, IHMM) 以处理区间信息粒。给出了 IHMM 的构建方法和去模糊化方法。在文本的词性标注任务中进行了多个实验，可以证实利用源域结构信息的粒迁移学习方法避免了负迁移，提高了模型的泛化能力。

关键词：迁移学习；粒计算；区间信息粒；词性标注

中图分类号：TP391　　**文献标识码：**A

Exploiting Source Domain Structure in Granular Transfer Learning with Part-of-speech Tagging Appllication

Sun Shichang[1,2], Lin Hongfei[1], Meng Jiana[2], Liu Hongbo[3]

(1. Dalian University Of Technology, Liaoning Dalian, 116023

2. Dalian Nationality University, Liaoning Dalian, 116600

3. Information Science and Technology College, Dalian Maritime University, Liaoning Dalian, 116626)

Abstract: Transfer learning solves the data sparseness problem to some extent, but the generalization capacity is still hindered by negative-transfer problem. For this, we propose an information granulation method for text corpora based on source domain structure. Interval granules are employed to express the influence of source domain structure on statistics of the dataset. Then we propose Interval type-2 fuzzy Hidden Markov Model (IHMM) to deal with the interval granules. The construction and defuzzification methods are given. Many experiments in text sequence labeling tasks verify that source domain structure based granular transfer learning avoids negative-transfer and improves generalization capacity.

Key words: transfer learning; granular computing; interval granules; part-of-speech tagging

1 引言

词性标注被认为是自然语言处理中的一个基础部分，并且是信息抽取与检索的重要的预处理工具。词性标注是通过计算的方式在文本上下文中确定词性标签。随着微博等网络应用的发展，词性标注任务常常需要被“迁移”到新的文本域。近年来，迁移学习[1]成为一个快速发展的研究领域；迁移学习的目的是把已有领域（称为源域）中的模型和信息移植到新领域（称为目标域）。虽然迁移学习在一定程度上减轻了目标域的数据稀疏问题对泛化能力的影响，然而泛化能力的提高仍然受到负迁移等问题的影响。

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金(61472058, 61572102)；中央高校自主基金(DC201502030202)

作者简介：孙世昶，男，1979 年生，博士生，讲师，研究方向为机器学习与文本挖掘；林鸿飞，男， 1962 年生，博士，教授，研究方向为文本挖掘和信息检索；孟佳娜，女，1972 年生，博士，教授，研究方向为文本挖掘；刘洪波，男，1971 年生，博士，教授，研究方向为智能信息处理.

在迁移学习中，负迁移是指源域的知识和任务对目标域的任务造成负面影响。某个粒度的迁移对象可能成为“负迁移”的条件。因此作者将粒计算的思想引入到迁移学习中，提出“粒迁移学习”方法。粒迁移学习将信息粒化方法、粒模型和粒度思想用在迁移学习中，以解决负迁移等问题，达到提高迁移学习模型泛化能力的目标。将粒计算思想用于迁移学习的研究刚刚开始，缺少详细的实施方法。为了将粒计算的思想和基础方法引入迁移学习，首先需要找到具体领域中适合知识迁移的信息粒化方法。本文探索适合文本应用的信息粒度的表示和处理方法，首先用区间信息粒表示带有结构信息的源域中的粒度，然后将机器学习模型扩展为粒模型以处理区间信息粒并得到适合迁移学习的粒度。

下面对粒计算的有关概念进行介绍。信息的粒度(granularity)泛指对信息进行观察和表示的不同抽象程度，例如信息粒的相对大小或粗糙程度。信息粒(granules)是由于相似性而聚集起来的复杂信息实体，包括区间、模糊集等多种形式[2]。粒化(granulation)是以粒的方式表示信息的过程，即将研究对象根据某种相似性而形成的聚集表示为可以处理的形式。粒模型(granular model)是指通过信息粒完成机器学习任务的模型。Pedrycz 认为粒模型可以作为对原模型的一种抽象，适用于向目标环境进行知识迁移[3]；并把信息粒度看成是知识迁移和复用中重要的设计资产，在诸如数据覆盖率等标准下进行信息粒度的最优分配。

我们提出使用源域结构的文本语料信息粒化方法。文本数据集带有自然的结构，例如 Brown 语料取自 500 多种文章来源并被分类为多种文体风格，包括“新闻”，“小说”等。语料中的句子聚集为不同的文章，而文章属于不同的类型；这些都属于源域的结构性信息，反映源域数据集在采样过程中的组织结构。传统机器学习方法通常假定数据集是论域中数据的具有代表性的采样，即代表了论域的分布，因而对于这些源域结构并不通过其他方式加以利用。然而在迁移学习中，我们要建立的不是完全拟合源域的分布，而是以源域分布为基础的抽象的、有利于知识迁移的模型。所以可以利用源域结构来得到关于源域的粒度的信息。对于文章标签这样的源域结构，我们采用 Pedrycz 提出的区间粒化方法来建立文本域特征的区间信息粒[2]。

本文提出利用源域结构的粒迁移学习方法，包括区间信息粒化和称为区间二型模糊隐马尔可夫模型的粒模型。通过 Brown 语料的大量词性标注实验，表明利用源域结构的粒迁移学习方法避免了负迁移，提高了模型的泛化能力。

2 相关工作

2．1 词性标注

词性标注是自然语言处理领域的基础性研究课题，其作用是通过上下文等信息计算决定符号的词性标签。目前的国内外研究大多采取基于概率统计的技术路线。词性标注是一种序列标注问题，可以通过很多基于统计学习的模型来处理，包括隐马科夫模型（HMM）[4]，最大熵模型（MEMM）[5]，条件随机域模型（CRF）[6, 7]等。许多模型通过特征选择和调整可以取得较好的性能，而 HMM 的优点在于计算量小和模型简单。在文献[8]的比较中经过平滑和未登录词处理的 HMM 的性能超过其他模型。

2．2 迁移学习

根据 Pan 的综述[1]，迁移学习方法主要可以分为基于表示、基于实例和基于参数的方法。在基于参数的迁移学习中，最大后验(Maximum a posteriori, MAP)[9]和最大似然线性回归(Maximum likelihood linear regression, MLLR)[10]可以用于以 HMM 为基础的序列识别方法中。但是这些工作需要观察值分布的连续性假设。对于离散型 HMM，可以把通过源域估计出的参数作为先验并通过对目标域的学习得出目标模型，这种方法在本文中称为 DT-HMM。

2．3 粒计算

信息粒和粒计算是人工智能领域的一个研究热点，其特点是将整个问题抽象为更容易计算的子问题。但是基于粒化机理的数据建模理论和方法还不能满足复杂任务的要求，仍需要

针对不同问题用粒计算的思想展开建模方法的研究。粒计算提出逐步精确化的特征表示和处理思想，即信息粒化表示。山西大学的郭虎升和王文剑中图像处理中通过数据粒化来提高支持向量机的学习效率[11]. 对于文本的粒化方法研究较少。邱桃荣[12]根据词的多属性来进行词的粒化，采用本体以获取领域概念以及概念之间的关系，这种将词作为对象的方法需要使用多个方面的属性信息，对于数量庞大而结构简单的文本语料并不适用。Pedrycz 认为粒模型可以作为原模型的一种抽象，适用于向目标环境进行知识迁移[3]，Song 和 Pedrycz[13]在神经网络中使用区间连接并输出区间结果。

3 利用源域结构的区间信息粒化

计算对象的粒化是粒计算中具有挑战性的问题，粒化是指以粒的方式表示信息的过程，即将研究对象根据某种相似性而形成的聚集表示为可以处理的形式。这在一定程度上是对人的认知方法的一种模拟，人在面对大量复杂信息时往往会将其简化为不同的聚集，每个聚集便是一个粒。目前在很多粒计算研究中，将连续特征采用区间的形式进行粒化，从而得到有效的处理；例如时间特征[14]就是一种可以区间粒化的连续特征。文本是一种由离散的词特征组成的复杂数据，而词特征本身难以用区间的形式表示出来。为了解决负迁移问题，本文找到一种文本的区间粒化方法，即通过对数据集结构信息进行粒化以利用语料库中篇章、类别等组织结构方面的信息。

数据集结构信息是一种反映数据聚集情况和采样方式的信息，对迁移学习有一定的影响。样本分布的不均衡性可能体现于各种粒度，有时一篇文章里某些特征非常显著，有时特征的显著性从某种文体中表现出来。可以推断，不同粒度上的源域数据对迁移学习的适宜程度是不同的。使用信息粒作为对数值特征的一种抽象，可以增加模型对负迁移问题的处理能力。因此，有必要在迁移学习中引入基于数据集结构信息的粒化表示方法。Pedrycz[2]认为区间通过引入二分法实现了对数据的抽象，即数据元素属于或不属于一个信息粒，依赖于从数据中提取信息的粒度；并将区间作为信息粒的一种实现形式。区间信息粒的构造方法如下。

为了使粒度具有合理性，从实验事实中建立信息粒的基本原则是满足两个相互竞争的要求：

① 实验事实的充分性。在区间粒的边界内有更多的数值型数据作为支撑。

② 语义的具体性。区间长度越短则对语义的表述越具体。

假定实验事实来自一维的数值型数据向量D，通常认为其中数$med(D)$具有代表性和估计的鲁棒性。可以通过考量中数左侧或右侧的数值，分别求解下界J和上界$\overline{J}$. 对于$\overline{J}$，实验事实由中数右侧数值的基数来考量，即$card\{med(D) \leqslant x_k \leqslant \overline{J}\}$。在构造优化目标时，使用增函数$f_1(\cdot)$来表示希望获得更多的实验事实。同时使用减函数$f_2(\cdot)$表示希望获得更具体的表示，即较小的区间长度。所以$\overline{J}$可以通过如下优化问题来求解。

$$\arg\max_{\overline{J}} \ V(\overline{J}) \tag{1}$$

$$V(\overline{J}) = f_1(card\{x_k \in D | med(D) \leqslant x_k \leqslant \overline{J}\}) * f_2(|med(D) - \overline{J}|) \tag{2}$$

同理，J的实验事实由中数左侧数值的基数来考量，即$card\{\underline{J} \leqslant x_k \leqslant med(D)\}$，并通过如下优化问题来求解：

$$\arg\max_{J} V(\underline{J}) \tag{3}$$

$$V(\underline{J}) = f_1(card\{x_k \in D | \underline{J} \leqslant x_k \leqslant med(D)\}) * f_2(|med(D) - \underline{J}|) \tag{4}$$

函数$f_1(\cdot)$和$f_2(\cdot)$可以选择为公式(5)和(6)的形式，其中α为粒化提供了一定的灵活性。

$$f_1(u) = u \tag{5}$$

$$f_2(u) = exp(-\alpha u) \tag{6}$$

较大的α起到强调语义具体性的效果，较小的α起到强调事实充分性的效果。

在文本应用中，语料结构信息构成对数据集的划分，例如文件这种结构信息对语料中的句子构成划分。这样，各子集对应的统计量组成数值型数据向量D. 这样我们可以求解公式(1)和(3)中的优化问题，得到区间信息粒$J=[\underline{J},\overline{J}]$, 其中$J$和$\overline{J}$是区间信息粒的下限和上限。$J$的含义是对实验事实的充分性和语义的具体性的平衡。在实验中采用文件作为划分的单位，因此文本中的信息粒度是指语料中对统计量起作用的文件子集的大小。然后可以将区间信息粒用于粒模型的构造。

4 区间二型模糊隐马尔可夫模型

对应于区间粒形式的输入，需要建立粒模型进行处理。考虑隐马尔可夫模型对于序列标注任务的高效性，本节建立区间二型模糊隐马尔可夫模型(Interval type-2 fuzzy Hidden Markov Model, IHMM)以完成序列迁移学习任务。

区间二型模糊隐马尔可夫模型 IHMM 通过其参数$\widetilde{\lambda}=(\widetilde{A},\widetilde{B},\widetilde{\pi})$来刻画。

① 二型模糊状态迁移向量：

$\widetilde{A}=\{\widetilde{V_i}|1\le i\le N\}$代表二型模糊状态迁移向量，其中$\widetilde{V_i}=\sum_{j\in\mathbb{N}}\sum_{u\in J_j}1/(j,u)$，$J_j\subseteq[0,1],\mathbb{N}=[1,..,N]$. 这里$\widetilde{V_i}$是一个区间二型模糊集(IT2 FS)，意义是状态$S_j$对状态$S_i$的隶属程度。$J_j$为二级隶属度函数的域(domain), 称为主隶属(primary membership).

② 二型模糊符号发射向量

$\widetilde{B}=\{\widetilde{U_j}|1\le j\le N\}$代表二型模糊符号发射向量，其中二型模糊集$\widetilde{U_j}=\sum_{k\in\mathbb{M}}\sum_{u\in J_k}1/(k,u)$, 主隶属$J_k\subseteq[0,1],\mathbb{M}=[1,..,M]$. 这里$\widetilde{U_j}$是一个区间二型模糊集，意义是符号$V_k$对状态$S_j$的隶属程度。

③ 二型模糊初始状态

二型模糊集$\tilde{\pi}=\sum_{i\in\mathbb{N}}\sum_{u\in J_i}1/(i,u)$ 代表二型模糊初始状态，其中主隶属$J_i\subseteq[0,1],\mathbb{N}=[1,..,N]$代表二型模糊初始状态，其中$\pi_i=P[q_1=S_i]$, q_1是初始时刻的状态，$0\le\pi_i\le1$, $\sum_{i=1}^{N}(\pi_i)=1$.

由于区间二型模糊集的二级度为常数 1， 因此 IT2 FS 可以由其不确定覆盖域(Footprint of uncertainty, FOU)来描述。FOU 由主隶属的并集构成。例如对于$\widetilde{U_j}$, 其 FOU 由公式(7)给出。如图 1, 区间二型模糊集$\widetilde{U_j}$可以由二维平面上的 FOU 描述。

$$FOU(\widetilde{U_j})=\bigcup_{k\in[1,..,N]}J_k \tag{7}$$

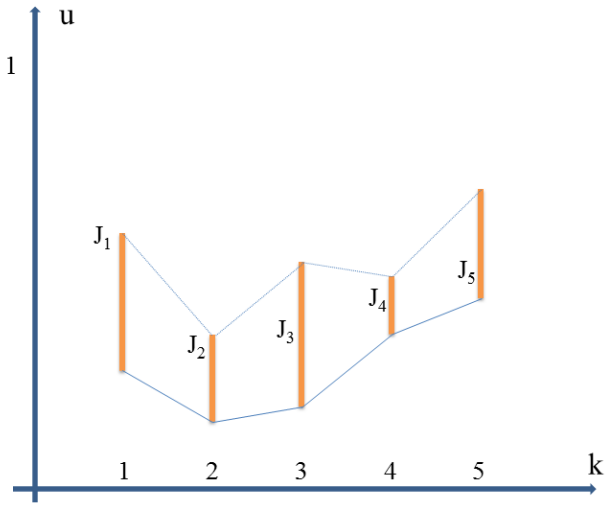


图 1 $\widetilde{U_j}$的不确定覆盖域

本文通过粒度控制负迁移产生的条件。用区间信息粒表示带有结构信息的源域数据对应的模型参数，这样建立的粒模型中包含区间信息粒中的粒度。在迁移学习设置下，可以通过验证集选取合适的粒度，类似于基于实例的迁移学习方法中的对实例赋权，不同在于本文是从粒度层面控制影响迁移学习模型的源域数据子集，从而控制负迁移产生的条件。这样，通过选择适合目标域的粒度，粒计算的思想方法被引入到迁移学习中。

本文将 IHMM 用于两种迁移学习设置。其一是只在源域中有已标注文本，称为转导学习设置；其二是目标域也有少量已标注文本，称为归纳学习设置。

IHMM 的构建可以分成以下三个步骤：

① 如果目标域没有已标注数据，在以语料结构为单位的子集上进行 HMM 监督学习，从而构建参数的统计量；这样对应于每个参数得到一组数值型数据，作为 IHMM 的输入。如果目标域也有少量已标注数据，以语料结构为单位的子集，以及目标域的验证集可以采用诸如 DT-HMM 等归纳迁移学习方法建立模型，然后通过监督学习得到对应于每个参数的一组数值型数据。

② 将对应于 HMM 每个参数的一组数值型数据，根据公式(1)至公式(6)建立N^2+NM+N个区间信息粒，作为$\widetilde{V_i}$, $\widetilde{U_j}$, $\widetilde{\pi}$的主隶属J_j, J_k, J_i，并表示为$\widetilde{A}, \widetilde{B}, \widetilde{\pi}$。

③ 通过$\widetilde{V_i}$, $\widetilde{U_j}$, $\widetilde{\pi}$来建立 IHMM：$\widetilde{\lambda} = (\widetilde{A}, \widetilde{B}, \widetilde{\pi})$.

为了在序列标注中使用高效的 Viterbi 算法，需要对 IHMM 进行去模糊化。由于 IHMM 使用了模糊集的概念，其参数的含义是隶属度而不是概率。去模糊化后得到的脆性(crisp, 与 fuzzy 相对)参数值的 HMM 模型中不再满足原有的诸如$\sum_{k=1}^{M} b_j(k) = 1$的概率约束。

对于不同的迁移学习设置，可以采用不同的去模糊化方法。对于转导学习，采用区间中点得到脆性值；对于归纳学习，可以利用目标域的少量已标注文本，采用粒子群(PSO)等优

化方法选取适合目标域的脆性值参数的 HMM，记为λ^*.

去模糊化算法如下：

算法 1. IHMM 的去模糊化算法

输入：$\widetilde{\lambda} = (\widetilde{A}, \widetilde{B}, \widetilde{\Pi})$, 目标域验证集

输出：λ^*

```
if 目标域验证集==NULL：
    在Ṽ_i, Ũ_j, Π̃中取各主隶属的中点作为λ*的参数值
else：
    设定 PSO 的粒子的维度为 HMM 参数的个数
    将λ̃的主隶属J_i, J_j, J_k作为 PSO 粒子在对应维度上的取值范围
    使用 bestFitness 保持最高准确率，并初始化为 0
    λ* = NULL
    启动 PSO 迭代
    while 迭代的结束条件不满足:
        生成候选粒子 R
        根据候选粒子得到一个对应的 HMM, 记为λ(R)
        测试λ(R)在目标域验证集上的准确率，记为 Fitness
        if Fitness > bestFitness:
            bestFitness = Fitness
            λ* = λ(R)
        更新粒子 R 的位置
    end while
返回λ*
```

5 实验结果与分析

词性标注是评估序列学习方法的经典任务。为了把 IHMM 和其他方法相比较，并说明参数设置和正则项参数的选取，我们进行了 Brown 语料的实验。语料中的不同类型的文本分布被用作源域数据和目标域数据。

5．1 实验数据

Brown 语料被编辑为英语语言文本的通用语料，是第一个百万词集的英文电子语料，并且包含 500 多种文本的来源。这些文本来源通过语体被分类为形式小说、新闻、社论等多种类别。在每个类别中，根据语料的自然结构划分为 40 多个文件。因此，Brown 语料适合用来验证基于自然结构的区间粒化方法及模型的有效性。

我们使用 Brown 语料构建了 20 个迁移学习任务，选择 5 种有代表性的类型，使用一种类型作为源域，并用另一种类型作为目标域。对于每一个迁移学习任务“源域-目标域”，使用如下方式构造训练集、测试集，以及验证集的数据。在转导学习设置中，我们取源域中前 36 个文件，计约 3600 个已标注句子作为训练集；取目标域中的 800 个句子作为测试集。在归纳学习设置中，我们使用源域中的 36 个文件作为训练集；取目标域中的前 200 个句子作为验证集、接下来 800 个句子作为测试集。对于 IHMM，将源域的每个文件作为一次对 HMM

各参数的采样。

5．2 实现细节

本文在两种设置下进行 IHMM 和其他方法的性能比较。其一是只在源域中有已标注文本，称为转导学习设置，此时 IHMM(transductive)简记为 IHMM(t)；与 HMM 进行比较。其二是目标域也有少量已标注文本，称为归纳学习设置，此时 IHMM(inductive)简记为 IHMM(i)；与 DT-HMM 进行比较。

① HMM：采用传统的机器学习模型[4]，利用源域数据进行训练。

② DT-HMM：将源域数据作为先验，以最大后验(MAP)[8]方式估计最终模型的参数。

在 IHMM 中，参数α代表粒度，是粒化方法中的影响序列识别性能的因素。对于参数α的设置，使用点列 $[0.2,0.5,1,2,5,8]$ 来进行数值实验。我们使用了 4 个任务来展示参数 对准确率的影响，结果如图 2 所示。实验表明 时通常有较好的表现。在下面的实验中取$\alpha = 0.5$. PSO 算法参数的设置如表 1.

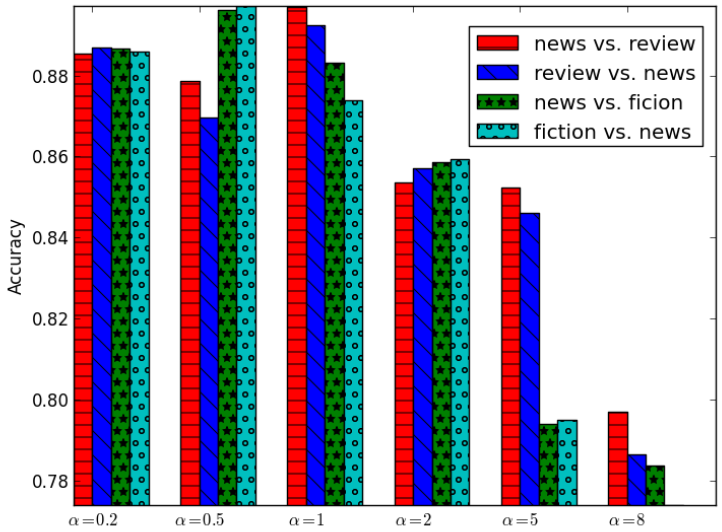


图 2 粒度的选取

表 1 PSO 算法参数的设置：

参数名称	参数值
迟滞度	0.5
认知比率	1.0
社会化比率	1.0
种群大小	5
最大迭代次数	15

5．3 实验结果与分析

首先进行转导迁移学习的实验，结果见图 3 和表 2.

然后在上面的 20 个任务中进行归纳迁移学习的实验，在实验中与 DT-HMM 进行比较。由于使用了随机优化方法，所以在每一个任务中进行了四次运行，以观察运行的稳定性。结果见图 4 和表 3.

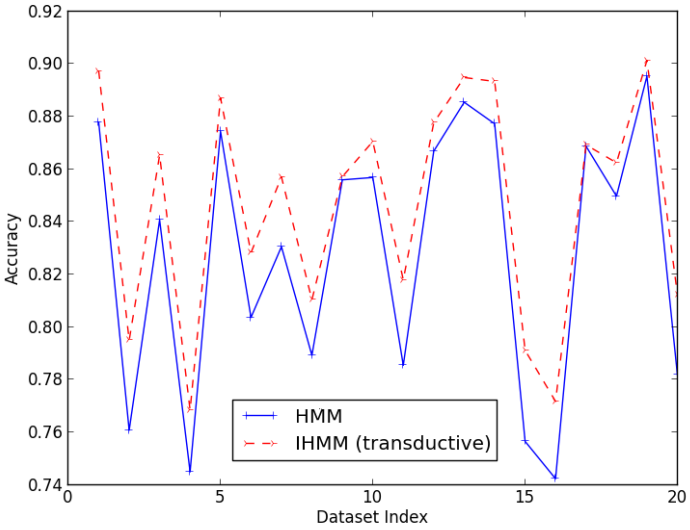


图 3 转导迁移学习方式下的词性标注结果

表 2 转导迁移学习的结果和 t 检验

源域	目标域	HMM	IHMM(transductive)
editorial	news	87.78	89.73
fiction		76.07	79.49
government		84.07	86.53
adventure		74.49	76.84
news	editorial	87.43	88.70
fiction		80.36	82.82
government		83.05	85.70
adventure		78.92	81.03
news	fiction	85.59	85.71
editorial		85.68	87.05
government		78.54	81.79
adventure		86.68	87.76
news	government	88.54	89.48
editorial		87.72	89.33
fiction		75.65	79.11
adventure		74.23	77.17
news	adventure	86.87	86.92
editorial		84.97	86.24
fiction		89.53	90.13
government		78.18	81.24
平均		82.72	84.60
成对 t 检验	t 值		8.08
	p 值		$1.43*10^{-7}$

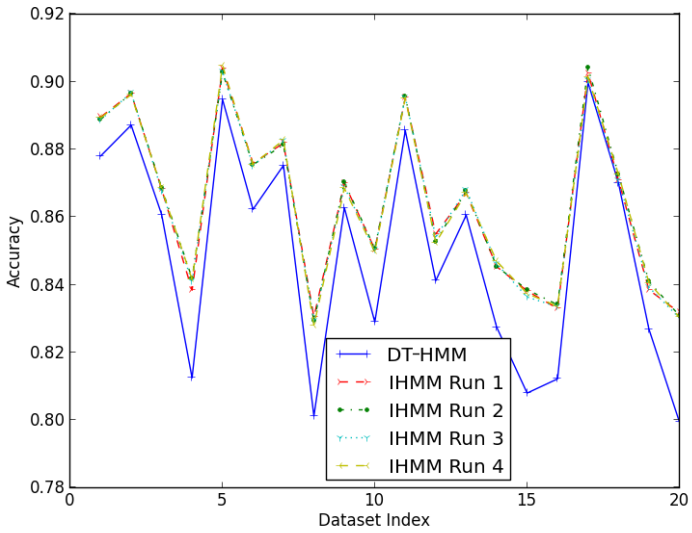


图 4 归纳迁移学习方式下的词性标注结果

表 3 归纳迁移学习的结果和 t 检验

源域	目标域	DT-HMM	IHMM (inductive)			
			运行 1	运行 2	运行 3	运行 4
editorial	news	87.81	88.96	88.89	88.88	88.94
fiction		88.72	89.66	89.66	89.68	89.62
government		86.08	86.83	86.87	86.79	86.85
adventure		81.25	83.84	84.19	84.09	84.16
news	editorial	89.48	90.39	90.28	90.31	90.48
fiction		86.23	87.59	87.55	87.53	87.56
government		87.53	88.19	88.14	88.25	88.28
adventure		80.12	83.05	82.93	82.95	82.80
news	fiction	86.28	87.00	87.03	86.89	86.85
editorial		82.90	85.05	85.07	85.03	84.98
government		88.56	89.53	89.57	89.55	89.52
adventure		84.12	85.50	85.28	85.41	85.25
news	government	86.06	86.71	86.77	86.82	86.71
editorial		82.75	84.54	84.54	84.62	84.70
fiction		80.80	83.82	83.86	83.64	83.74
adventure		81.22	83.30	83.41	83.35	83.37
news	adventure	89.98	90.26	90.42	90.06	90.14
editorial		87.01	87.10	87.26	87.10	87.27
fiction		82.69	83.85	84.10	83.93	84.06
government		79.95	83.18	83.09	83.02	83.04
平均		84.98	86.42	86.45	86.40	86.42

成对t检验	t 值	6.90	7.06	6.80	6.99
	p值	$1.42*10^{-6}$	$1.02*10^{-6}$	$1.72*10^{-6}$	$1.17*10^{-6}$

由图可见，IHMM 在每个任务中都取得了准确率的提升。由表中“平均”一行可知，在转导迁移学习中，IHMM(transductive)比 HMM 平均高出 1.88 个百分点；在归纳迁移学习中，IHMM (inductive)比 DT-HMM 平均高出 1.42 个百分点。并且，在每个任务的每次运行中，IHMM 的准确率都高于采用标准监督学习方法的 HMM 和 DT-HMM, 没有出现负迁移。

为了评估性能提高的统计显著性，我们进行了成对 t 检验。零假设为 IHMM 的正确率没有比原模型提高，备择假设为 IHMM 的正确率比原模型有了显著提高。显著性水平设为 $\alpha_H = 0.05$，从表 1 中可见 p 值远低于α_H，因此可以得出 IHMM 已经获得显著的正确率提升的结论。

以上实验结果证明，基于数据集结构信息的粒化方法避免了序列迁移学习模型的负迁移，提高了泛化能力。

6 结论

本章提出基于数据集结构信息的粒迁移学习方法。为了能够避免负迁移，首先提出基于数据集结构信息的粒化方法。用区间信息粒表示出源域数据集的结构对数据集中统计量的影响。然后提出区间二型模糊隐马尔可夫模型(IHMM)以处理区间信息粒。给出了 IHMM 的构建方法和去模糊化方法。在文本的词性标注任务中进行了多个实验，可以证实基于数据集结构信息的粒迁移学习方法避免了负迁移，提高了模型的泛化能力。

我们将进一步研究使用其他数据集结构信息的粒迁移学习方法，并将应用扩展到组块分析等文本的序列标注任务。

参考文献

[1] Pan S J, Yang Q. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10):1345–1359.

[2] Pedrycz W. Granular computing: analysis and design of intelligent systems[M]. CRC press, 2013.

[3] Pedrycz W, Russo B, Succi G. Knowledge transfer in system modeling and its realization through an optimal allocation of information granularity[J]. Applied Soft Computing, 2012, 12(8):1985–1995.

[4] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2):257–286.

[5] Walder C J, Kootsookos B C, Peter J. andLovell. Towards a Maximum Entropy Method for Estimating HMM Parameters[C]//INTERSPEECH. 2003:45–49.

[6] Liu J, Yu K, Zhang Y, et al. Training Conditional Random Fields Using Transfer Learning for Gesture Recognition[C]//Proceedings of IEEE International Conference on Data Mining. 2010:314–323.

[7] Sutton C, McCallum A. Composition of conditional random fields for transfer learning[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005:748–754.

[8] Brants T. TnT: a statistical part-of-speech tagger[C]//Proceedings of the Sixth Conference on Applied Natural Language Processing. 2000:224–231.

[9] Ait-Mohand K, Paquet T, Ragot N. Combining structure and parameter adaptation of HMMs for printed text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,

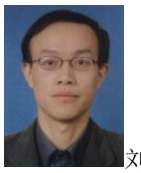
2014, 36(9):1716–1732.
[10] Kim N S, Sung J S, Hong D H. Factored MLLR Adaptation[J]. Signal Processing Letters, 2011, 18(2):99–102.
[11] 郭虎升，王文剑. 动态粒度支持向量回归机[J]. 软件学报, 2013, 24(11):2535－2547.
[12] 邱桃荣. 面向本体学习的粒计算方法研究[D]. 北京交通大学, 2009.
[13] Song M, Pedrycz W. Granular neural networks: concepts and development schemes[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(4):542–553.
[14] 孟军. 相容粒计算模型及其数据挖掘研究[D]. 大连理工大学, 2012.

作者简介：

 孙世昶，（1979 年--）,男,博士生,讲师,研究方向为机器学习与文本挖掘。Email: ssc@dlnu.edu.cn。

 林鸿飞，（1962 年--）,男,博士,教授,CCF 高级会员,研究方向为文本挖掘和信息检索。Email: lhf@dlut.edu.cn。

 孟佳娜，（1972 年--）,女,博士,教授,研究方向为文本挖掘。Email: mjn@dlnu.edu.cn。

 刘洪波，（1971 年--）,男,博士,教授,研究方向为智能信息处理。Email: lhb@dlut.edu.cn。