# 一种基于聚类的文本迁移学习算法①

杜俊卫　李爱军　(山西财经大学 信息管理学院 山西 太原 030006)

**摘　要：** 当现有训练数据过期，而新数据又非常少时，运用迁移学习能够有效提高分类器性能。本文提出一种基于聚类的文本迁移学习算法，给出了算法的主要思想及实现步骤。然后，在中文文本语料库上进行了实验，并与非迁移学习算法进行了比较。实验证明该方法能有效提高分类器性能。

**关键词：** 训练数据过期；新数据非常少；迁移学习；聚类；文本

## Transfer Learning Algorithm for Text Classification Based on Clustering

DU Jun-Wei, LI Ai-Jun (Department of Information Management, Shanxi University of Finance and Economics, Taiyuan 030006, China)

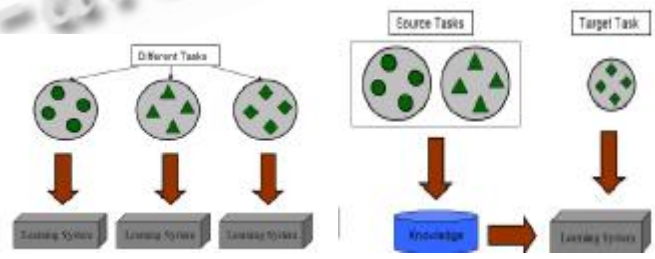
**Abstract:** Transfer learning can improve the performance of classifier effectively, when the training data are out of date, but the new data are very few. In this paper, we propose a transfer learning algorithm for text classification based on clustering. We describe the main idea and the step of the algorithm. Then have experiment on text corpus of Chinese, and compare the algorithm with transfer-unaware algorithm. The experiments demonstrate that this algorithm significantly outperforms the others.

**Keywords:** training data are out of date; new data are very few; transfer learning; clustering; text

## 1　引言

传统文本分类技术主要有：贝叶斯、支持向量机、决策树、K 最近邻和神经网络等等[1]。这些分类技术都需要有大量的训练数据。但是，在很多情况下，目标任务可能没有足够的训练数据。例如，在 Web 应用领域，大量新的领域不断涌现，从传统的新闻，到网页，到图片，再到博客、播客等等。这就常常导致现有训练数据过期，而新数据的收集又非常困难。此时传统的机器学习方法就显得力不从心了。解决这种问题的一种有效方法就是迁移学习。

从心理学的角度看，迁移学习是人类的基本技能。例如，学习认识苹果有助于认识梨；或者学习弹奏电风琴有助于学习钢琴等等。迁移学习的目标是将从一个环境中学到的知识迁移到新的环境中，帮助新环境下的学习。当现有训练数据过期，而新数据又非常少，或者标注新数据代价非常大时，利用现有的不同分布下的训练数据来帮助新数据的学习，这就是迁移学习的任务[2]。图 1[2]中显示了传统机器学习与迁移学习的区别。从图中可以看出：传统的机器学习绝大多数都是从零开始，并不借鉴以前学到的知识，而迁移学习尝试将以前任务中学到的知识迁移到新任务的学习中去，这样机器学习的能力就会大大增强。




(a)　传统机器学习　　(b)　迁移学习

图 1　传统机器学习与迁移学习的不同学习过程

目前，迁移学习的技术主要分为基于实例的迁移

① 基金项目:国家自然科学基金(60873100)
收稿时间:2010-04-13;收到修改稿时间:2010-05-23

学习[3,4]和基于特征的迁移学习[5,6]。在文本分类中，这两种迁移
技术的研究已经有了初步的成果[7,8]，但这些成果都有一定的局限性，还需要进一步的研究。本文借鉴基于实例迁移技术的思想[2]，提出一种基于聚类的文本迁移学习算法。采用聚类技术对现有数据进行过滤，找出与目标数据非常相似的数据，来帮助目标任务的学习。

## 2 基于聚类的文本迁移

虽然现有的辅助数据已经过期。但是在这些现有数据中，应该还会存在一部分数据与测试数据非常相似，能够用来帮助目标任务的学习[4]。因此考虑利用聚类技术从现有数据中找出与测试数据非常相似的数据。

### 2.1 聚类简介

聚类是一种重要的数据挖掘形式[9]。文本聚类的目的是为了将大规模的文本数据集分组成为多个类，并使同一类中的文本之间具有较高的相似度，而不同类之间的文本差别较大。作为数据挖掘的一项功能，聚类可以作为独立的工具，获得数据分布的情况，观察每个簇的特征，集中对某些特定的簇做进一步的分析。同时聚类技术也可以作为其他算法的预处理步骤，有效提高其他算法的分类性能。

### 2.2 本文使用的文本表示法和文本相似度公式

按照传统的向量空间模型 (VSM, vector space model)表示法，文本内容可表示成一个加权的特征向量。设 D 为文本集合，$d_i$ 表示集合中的一篇文本，$t_i$ 表示第 i 个特征词，wi 表示第 i 个特征词的权重，则一篇文档就表示成了向量的形式：$d_i=(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$。其中，权值 wi 可以用各个特征的 tf-idf 权重[10]来表示。tf-idf 公式如下：

$$tf-idf=\sum_{d\in D} tf(d,t)\cdot log\frac{|D|}{df(t)} \quad (1)$$

其中，tf(d,t)是词 t 在文本 d 中的词频，df(t)是文本集合 D 中包含词 t 的文本数目，|D|表示文本集合 D 所包含的文本数。

两个文本之间的相似度[11]可以通过两个向量之间的夹角 $\alpha$ 的余弦来计算，假设两篇文本分别为 $d_1=(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$和 $d_2=(t_1,\sigma_1;t_2,\sigma_2;\cdots;t_n,\sigma_n)$，则 $d_1$ 和 $d_2$ 之间的相似度表示为：

$$sim(d_1,d_2)=cos\alpha=\frac{\sum_{i=1}^{n}\omega^i\times\sigma^i}{\left(\sum_{i=1}^{n}\omega_i^{\,2}\times\sum_{i=1}^{n}\sigma_i^{\,2}\right)^{\frac{1}{2}}} \quad (2)$$

$sim(d_1,d_2)$的值越大，两个文本就越相似。

### 2.3 算法思想

首先，将辅助训练数据与目标训练数据一起进行聚类。聚类的结果是使得簇内数据间相似性较高，而簇间数据相异。因此，经过聚类后，没有和目标训练数据聚在同一簇的辅助数据就被过滤掉。剩下的就是和目标数据相似性较高的数据，将它们和目标数据一起进行训练，将会大大提高分类器的性能。

下面给出文中会用到的一些基本符号的定义。

定义 2.1.

* 设 Xb 为目标样例空间，Xa 为辅助样例空间。
* 设 Y={0，1}为类空间。

定义 2.2(测试数据集).

S={($x_i^t$)}，其中 $x_i^t\in$ Xb，i=1,2,⋯,k，k 是集合 S 的元素个数。

定义 2.3 (训练数据集).

训练数据集包括两部分：

Tb={($x_j^b$, c($x_j^b$))},其中 $x_j^b\in$ Xb, j=1,2,⋯,m;

Ta={($x_i^a$, c($x_i^a$))},其中 $x_i^a\in$ Xa， i=1,2,⋯,n;

其中，c(x)是实例的真实类标，c(x) Y。Tb 是目标训练数据集，Ta 是辅助训练数据集。m 和 n 分别是目标训练数据集和辅助训练数据集的大小。

### 2.4 算法步骤

输入：两个训练数据集 $T_a$ 和 $T_b$，一个测试数据集 S。

输出：分类结果 $h_t(x_t)$。

① 读入训练数据 $T_a$ 和 $T_b$；

② 将训练数据按照类标分为 N 类：$T_i$(i=1,⋯,N)，其中 Ti 表示类标为 i 的实例集；

③ For i←1 to ℕ⑥

1) 调用一个基本聚类算法，对 $T_i$ 进行聚类，并返回聚类结果；

2) 扫描 $T_i$，将辅助数据中没有和目标数据聚在一簇的实例删除；

④ end for；

⑤ 调用一个基本分类算法，根据过滤后的训练数据和测试数据 S，得到一个分类模型：

$h_t$：X→Y。

⑥ 在 S 上测试该分类模型的性能并输出。

## 3 实验结果及分析

### 3.1 数据集

采用有层次结构的数据集 TanCorp-12[12]，该数据集包含财经、科技和艺术等 12 个大类，每个大类下面又包含若干子类，共有 14150 篇文档。实验选取其中六个大类(财经、科技、艺术、教育、人才、电脑)，又在每个大类下选取了 4 个子类。将大类作为分类的目标，目标数据和辅助数据分别由不同的子类构成。这样，相对于测试数据，辅助数据是过期的。其中，财经/科技 表示选择财经和科技作为大类，即分类的目标，财经(财富、消费) 表示财经类的两个子类财富和消费，以此类推。具体的数据分布见表 1。

表 1　TanCorp-12 数据分布

| 数据集 | 目标源训练数据 | 辅助训练数据 |
|---|---|---|
| 财经/科技 | 财经(财富、消费)<br>科技(自然、天文) | 财经(金融、证券)<br>科技(考古、生命) |
| 财经/艺术 | 财经(证券、财富)<br>艺术(音乐、文学) | 财经(消费、金融)<br>艺术(舞台、美学) |
| 艺术/科技 | 艺术(文学、舞台)<br>科技(天文、生命) | 艺术(美学、音乐)<br>科技(自然、考古) |
| 教育/人次 | 教育(培训、就业)<br>人才(创业、管理) | 教育(招生、留学)<br>人才(猎取、应试) |
| 电脑/教育 | 电脑(软件、网络)<br>教育(培训、留学) | 电脑(病毒、游戏)<br>教育(就业、招生) |
| 人才/电脑 | 人才(管理、应试)<br>电脑(病毒、软件) | 人才(创业、猎取)<br>电脑(游戏、网络) |

### 3.2 评价指标

实验采取准确率(P)、召回率(R)以及 F 值作为评价指标。准确率是所有判断的文本中与分类结果吻合的文本所占的比率。召回率是分类结果应有的文本中与分类系统吻合的文本所占的比率。它们的定义分别为：

$$准确率 = \frac{分类正确的文本数}{实际分类的文本数} \tag{3}$$

$$召回率 = \frac{分类正确的文本数}{应有的文本数} \tag{4}$$

准确率和召回率反映了分类质量的两个不同方面，两者必须综合考虑，不可偏废，F 值就是两者综合考虑的评估指标，其定义如下：

$$F-meaasure=\frac{准确率*召回率*2}{准确率+召回率} \tag{5}$$

### 3.3 实验结果

进行实验前，首先对数据集进行了一系列预处理，将文本表示成传统的向量空间模型。实验中分别采用 EM 和 Na?ve Bayes 作为算法中的基本聚类和分类算法。表 2 给出了进行聚类前后，辅助数据数量的变化情况。由于过滤了和测试数据分布不同的数据，辅助数据的数量都有不同程度的减少。表 3 中给出了目标数据比例为 1%时，采用文中算法前后各项指标值的变化情况。表中的第二列 NB(Tb)表示采用 Na?ve Bayes 算法结合目标训练数据训练的结果。

表 2　聚类前后辅助数据数量的变化情况

| 数据集 | 聚类前 | 聚类后 |
|---|---|---|
| 财经/科技 | 495 | 361 |
| 财经/艺术 | 495 | 368 |
| 艺术/科技 | 495 | 415 |
| 教育/人才 | 473 | 269 |
| 电脑/教育 | 495 | 251 |
| 人才/电脑 | 480 | 347 |

表 3　目标数据为 1%时的效果

| 数据集 | NB(Tb) | | | 本文中的算法 | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| 财经/科技 | 0.702 | 0.67 | 0.656 | 0.881 | 0.875 | 0.874 |
| 财经/艺术 | 0.75 | 0.695 | 0.677 | 0.935 | 0.935 | 0.93 |
| 艺术/科技 | 0.898 | 0.89 | 0.889 | 0.975 | 0.975 | 0.975 |
| 教育/人才 | 0.636 | 0.62 | 0.62 | 0.788 | 0.74 | 0.718 |
| 电脑/教育 | 0.878 | 0.875 | 0.874 | 0.96 | 0.96 | 0.96 |
| 人才/电脑 | 0.795 | 0.745 | 0.734 | 0.965 | 0.965 | 0.965 |

从表 3 中，我们可以看出：在目标训练数据不足的情况下，使用本文中提出的算法，各项性能指标均得到了较大的改善。

图 2 和图 3 中分别显示了三种算法：(1)NB 结合

目标训练数据；(2)NB 结合全部训练数据；(3)本文中的算法，在数据集(财经/科技)和(教育/人才)上的性能。其中，横坐标表示目标数据所占比例，纵坐标表示分类准确度。从图中可以看出，随着目标数据比例的不断增大，三种算法的性能都成递增趋势，但是本文中的算法表现得更突出。

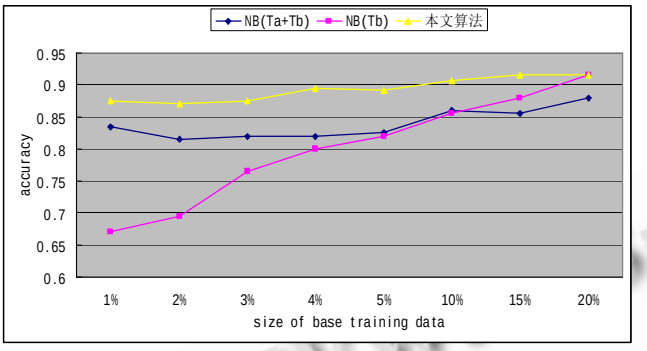



图 2 (财经/科技)数据集上的效果

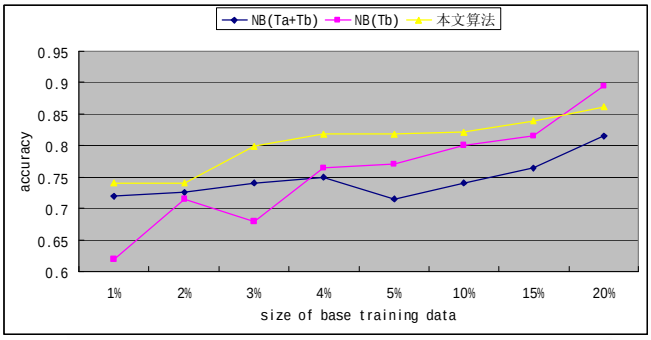



图 3 (教育/人才)数据集上的效果

从以上两幅图中还可以看出：当目标训练数据达到一定数量后，本文中的算法反而稍显逊色，说明此时目标训练数据的数量已经足够用来训练一个可靠的分类模型了，加入辅助数据后反而会影响分类器的性能。

## 4 结论

当目标任务训练数据不足时，基于聚类的文本迁移学习算法能够有效的利用现有数据，提高分类模型的准确率。但迁移学习不是总能提高学习效果。有的时候，使用迁移学习反而会降低学习的效果，这就是负迁移。像在实验中，当目标训练数据达到一定数量后，迁移的效果反而不如传统机器学习方法。这就启发我们去寻求一种机制来合理避免负迁移。

## 参考文献

1 牛延莉，张化. 文本自动分类研究进展. 教育技术导刊, 2008,7(4):24－26.

2 Sinno Jialin Pan, Yang Q. A Survey on Transfer Learning. IEEE TKDE, 2009.

3 Caruana R. Multitask learning. Machine Learning, 1997,28(1):41－75.

4 Dai WY, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. Proceedings of the Twenty-Fourth International Conference on Machine Learning, 2007:193－200.

5 Dai WY, Chen YQ, Xue GR, Yang Q, Yu Y. Translated learning:Transfer learning across different feature spaces. Advances in Neural Information Processing Systems 21, 2009.

6 Ling X, Xue GR, Dai WY, Jiang Y, Yang Q, Yu Y. Can Chinese Web Pages be Classified with English. Proceedings the Seventeenth International World Wide Web Conference (WWW 2008), Beijing, China, 2008:969－978.

7 Dai WY, Xue GR, Yang Q, Yu Y. Transfer naive bayes classifiers for text classification. Proceedings of the Twenty-Second National Conference on Artificial Intelligence, 540－545.

8 Do C, Ng A. Transfer learning for text classification. Advances in Neural Information Processing System 18, 2006:299－306.

9 吴启明,易云飞.文本聚类综述.河池学院学报, 2008, 28(2):86－91.

10 Salton G, Buckley C. Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 1998,24(5):513－523.

11 江涛,陈小莉,张玉芳,熊忠阳.基于聚类算法的 KNN 文本分类算法研究. 计算机工程与应用, 2009,45(7):153－155.

12 谭松波.王月粉中文文本分类语料库-TanCorpV1.0 2010-1-16:http://www.searchforum.org.cn/tansongbo/corpus.htm.