

# 基于双向选择调整策略的半监督聚类算法

刘 明, 宣照国, 吴江宁

(大连理工大学 系统工程研究所, 辽宁 大连 116024)

**摘 要:**半监督聚类算法通常利用标注数据优化类别描述参数(如类的中心),然后通过类别描述参数划分无标注数据的类别,但是没有考虑标注数据对其周围无标注数据的类别划分的直接作用。文中提出一种双向选择调整策略,在根据类别描述参数对数据进行类别划分之后,利用标注数据调整其周围未标注数据的类别标签,从而提高类别划分的准确度。该方法根据标注数据周围的数据密度来动态确定数据调整范围,并采用新的相似度计算方法提高被调整的数据准确度。文中利用双向选择调整策略改进了基于多项式模型的半监督聚类算法和半监督模糊聚类算法,并使用多个标准数据集进行实验。实验结果表明改进的算法有效提高了半监督聚类的准确性。

**关键词:**半监督聚类;未标注数据;标注数据;相似度;多项式模型;模糊聚类

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2013)02-0001-06

doi:10.3969/j.issn.1673-629X.2013.02.001

## Semi-supervised Clustering Algorithm Based on Double Adjustable Strategy

LIU Ming, XUAN Zhao-guo, WU Jiang-ning

(Institute of Systems Engineering, Dalian University of Technology, Dalian 116024, China)

**Abstract:** Usually, semi-supervised clustering algorithms utilize a small amount of labeled data to improve cluster parameters which guide the clustering of unlabeled data. However, the existing semi-supervised clustering algorithms (such as cluster centroid) ignore the labeled data could directly affect the clustering of unlabeled data. It proposes a double adjustment strategy which adjusts unlabeled data clustering with the labeled information, after the data is clustered according to the cluster parameters. Thus, the proposed method improves the clustering accuracy. The adjustment extension is changed dynamically by the local density around the labeled data. And a novel similarity measure is proposed to improve the accuracy of the adjusted unlabeled data. It modifies two algorithms, based on multinomial model semi-supervised clustering algorithm and semi-supervised fuzzy clustering algorithm, with the double adjustment method. Experimental results show that the method could improve the accuracy of semi-supervised clustering.

**Key words:** semi-supervised clustering; unlabeled data; labeled data; similarity; multinomial model; fuzzy clustering

## 0 引言

在数据挖掘实际应用中会遇到这样的数据,仅有一少部分数据具有标注的类别信息,大部分数据没有标注信息。当数据量很大时,如果完全采用人工方式对数据进行类别标注,将会花费大量的人力和时间。针对这种类型的数据,常采用半监督聚类算法对数据进行类别标注。

近年来,半监督聚类方面的研究得到了广泛的关注。Kunlun Li 等人利用标注的数据对模糊 c 均值的目标函数进行修改,使其更能适应半监督聚类的模式,从而得到更好的聚类效果<sup>[1]</sup>。Guan 等人利用标注数据改进相似度度量方法,来提高聚类的效果<sup>[2]</sup>。Dang 等人利用标注的数据优化聚类的初始化阶段,优化运算时间,提高聚类的精度<sup>[3]</sup>。而较多的半监督聚类算法在聚类过程中,利用标注信息对中间结果进行调整。如 Basu 提出的基于种子的半监督聚类算法(SSCS)<sup>[4,5]</sup>,Shi 提出的基于确定退火多项式模型的半监督聚类(DAMNSC)<sup>[6]</sup>和基于标签的半监督模糊聚类(SFCM)<sup>[7]</sup>等,在聚类过程中,根据类别描述参数计算所有的数据所属类别,然后纠正划分错误的标注数据。Shi 改进的半监督反馈算法<sup>[8]</sup>在聚类过程中,标注数据的类别标签随其同类标注数据分布最多的类

收稿日期:2012-05-11;修回日期:2012-08-17

基金项目:国家自然科学基金重点项目(71031002);国家自然科学基金资助项目(70871016)

作者简介:刘 明(1985-),男,山东日照人,硕士研究生,研究方向为文本聚类;宣照国,博士,讲师,研究方向为中文信息处理、知识管理、个性化推荐等;吴江宁,博士,教授,研究方向为知识发现与获取、文本挖掘、信息检索、知识可视化。

标号而改变。上述半监督聚类算法关注标注信息对聚类准则的优化,优化类别描述参数(如类中心)的搜索,但却忽视了标注数据对周围数据类别划分的直接影响,可能会降低算法的准确度和运算的效率。文中提出的改进方法,不仅通过类别描述参数得到无标注数据的类别划分,还利用标注数据调整其周围无标注数据的类别划分,并且通过标注数据和被调整数据的相互选择来确定被调整数据的类别划分,从而提高聚类的效果。

很明显,相似的数据倾向属于同一个类别。传统的方法一般采用相似度(或距离)来判断数据是否相近。但是当不同类别的数据密度差别较大时,传统方法的有效性不高<sup>[9]</sup>。文中提出一种双向选择策略来判断数据是否相近。首先选定一个标注数据,得到此标注数据与所有未标注数据的相似度,选择相似度大的未标注数据作为候选者。然后从候选者的角度观察该标注数据是否也是与该候选者最相似的数据。双向选择策略能够更好地适应类别密度分布差别较大的数据集。在同一个类别中,数据的密度分布也不是均匀的,一个类中的数据分布往往是在内部分布较密集,外围分布稀疏。根据标注数据周围的数据密度,来确定可调整的未标注数据的数量。如果标注数据周围的数据密度大,则增加其可调整数据的数量,从而加强该标准数据的作用,提高数据类别划分的准确度。

文中提出有效使用标注数据的双向选择调整策略,利用标注数据调整其周围数据的类别标签,并在聚类解搜索过程中,动态改变调整范围。文中应用双向选择调整策略,对两种典型的半监督聚类算法进行了改进,并选取了标准文本数据集进行了实验,验证算法有效性。

## 1 两种典型的半监督聚类算法

本节介绍两种典型的半监督聚类算法,即基于多项式模型的半监督聚类算法(MNSC)<sup>[6,8]</sup>和半监督模糊聚类算法(SFCM)<sup>[7,9]</sup>。首先定义 $N$ 个数据 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ 和 $Y = \{y_1, y_2, \dots, y_N\}$ ,其中 $\mathbf{x}$ 是 $V$ 维的向量数据,第 $w$ 维数据用 $\mathbf{x}(w)$ 表示, $y$ 表示数据 $\mathbf{x}$ 所属的类别, $y \in \{1, 2, \dots, K\}$ , $K$ 为划分的类别数目。数据 $\mathbf{X}$ 由两部分组成 $\mathbf{X} = \{\mathbf{X}^{(l)}, \mathbf{X}^{(u)}\}$ ,包括 $N_l$ 个标注数据 $\mathbf{X}^{(l)} = \{\mathbf{x}_1, \dots, \mathbf{x}_{N_l}\}$ 以及对应的类别标签 $Y^{(l)} = \{y_1^{(l)}, \dots, y_{N_l}^{(l)}\}$ , $y^{(l)} \in \{1, \dots, k\}$ , $k \leq K$ 和 $N_u$ 个未标注的数据向量 $\mathbf{X}^{(u)} = \{\mathbf{x}_{N_l+1}, \dots, \mathbf{x}_N\}$ 。半监督聚类的目标是对无标注数据进行类别划分,确定 $\mathbf{X}^{(u)}$ 对应的 $Y^{(u)}$ 的值。

基于多项式模型的半监督聚类算法是基于EM算法<sup>[10]</sup>发展起来的,它采用多项式模型对划分的每个类

别进行类的描述。聚类的过程中,在E步骤根据类别的描述参数,算法重新计算每个数据的类别所属情况;在M步骤根据数据的类别划分情况,计算每个类的描述参数。通过EM步骤的不断迭代循环直到算法收敛,得到最终的聚类结果。

半监督模糊聚类算法采用模糊数学<sup>[11]</sup>对数据隶属类别的划分情况进行计算描述,并根据最优化目标函数推导出数据对类别的隶属度和类中心的计算方法。通过迭代计算数据对类别的隶属度和类的中心直到收敛,得到最终的聚类结果。

上述两种半监督聚类算法能够利用已有的标注数据得到较好的初始类别划分,通过迭代优化获得较好的结果。但是,上述算法没有利用已标注数据对周围数据类别划分的直接影响,而且由于周围数据密度的不同,标注数据的影响范围也会不同<sup>[5,7]</sup>。文中提出双向选择调整策略,根据标注数据周围的数据密度选择被标注数据,并且应用双向选择的思想,通过标注数据和被标注数据的相互选择,确定对哪些未标注数据进行类别调整。

## 2 基于双向调整策略改进半监督聚类算法

当标注信息非常少的情况下,可以利用标注数据对其相近的数据进行类别标注。但是,很难保证被标注数据的准确性,错误的标注很可能会降低聚类的效果。文中借鉴文献[12]和[13]中提出的类别与数据双向选择的思想,给出选择未标注数据的方法。根据标注数据周围的数据密度选择被标注数据,并且应用双向选择的思想,通过标注数据和被标注数据的相互选择,确定对哪些未标注数据进行类别调整。由此,文中提出应用双向选择调整策略的半监督聚类算法,通过类中心与标注数据的共同作用来对数据进行类别划分。这样通过双向选择调整能够提高数据类别划分的准确度。

### 2.1 双向选择调整策略

传统半监督聚类算法在聚类过程中,通过类的参数(如类的中心)与未标注数据的相似度(或距离),对数据进行类别划分。文中提出的双向选择调整策略,在传统类别划分之后,利用标注数据对周围数据重新划分类别。在聚类的初始阶段,因为类的参数描述不够准确,通过标注数据对周围数据进行重新调整类别划分,能够提高类别划分的准确度;随着调整的反复进行,慢慢减弱标注数据对局部数据的调整范围。这样,采用双向选择调整的方法能够提高聚类的效果,加快算法的收敛。

双向选择调整策略首先是利用标注数据选择合适的候选数据,然后确定标注数据与候选数据是否相似。

传统的确定数据是否相似的方法是通过比较相似度与给定阈值来判断。

文中给出新的方法确定数据的相似性。如果从标注数据  $\mathbf{x}_i^{(y)}$  发现与其最近的未标注的数据  $\mathbf{x}_j$  作为候选者,从  $\mathbf{x}_j$  的角度观察  $\mathbf{x}_i^{(y)}$  也是与  $\mathbf{x}_j$  最近的数据,那么可以基本认定  $\mathbf{x}_i^{(y)}, \mathbf{x}_j$  属于同一个类别。这个具有相对性的相似度能够更好地适应类别密度分布差别较大的数据集。在数据集的同一个类中,数据的密度分布也不是均匀的。当标注的数据非常少的时候,这种方法虽然保证了增加信息的有效性,但是调整作用相对较小,需要适当放宽范围,增加候选者的数量。对于标注数据  $\mathbf{x}_i^{(y)}$ ,新方法发现与其最近的  $q_i$  个未标注的数据  $\mathbf{x}$  作为候选者。新的相似度测量方法采用这两个距离的比值见式(1)。

$$\text{sim}(\mathbf{x}_i^{(y)}, \mathbf{x}_j) = \frac{s_{\mathbf{x}_j \in X^{(u)}}(\mathbf{x}_i^{(y)}, \mathbf{x}_j)}{\max_{p \neq j}(s(\mathbf{x}_j, \mathbf{x}_p))} \quad (1)$$

如果  $\text{sim}(\mathbf{x}_i^{(y)}, \mathbf{x}_j) > \gamma$ , 那么这两个数据属于同一类。 $\gamma$  是判断数据是否属于同一类的阈值参量。其中,  $\mathbf{x}_p \in \mathbf{X}$ , 为数据集中的任意一条数据,  $s(\mathbf{x}_i, \mathbf{x}_j)$  是  $\mathbf{x}_i$  与  $\mathbf{x}_j$  之间的余弦相似度。

文中根据数据密度来确定数据的标注范围。当标注数据  $\mathbf{x}_i^{(y)}$  周围的数据密度较大,则对  $\mathbf{x}_i$  周围较多数据进行标注。首先,文中计算标注数据  $\mathbf{x}_i^{(y)}$  和与其最相似的  $Q$  个数据的平均相似度  $\bar{s}_i$  来估算它周围的数据密度。然后利用公式(2)来计算选取候选数据的数量。

$$q_i = Q \cdot \exp\left(-\left(\frac{\sigma}{\bar{s}_i}\right)^2\right) \quad (2)$$

其中,  $Q$  为最大候选数据数量的允许值,参数  $\sigma$  为调控候选数据数量,通过调控参量  $\sigma$  慢慢增大,减少标注数量。

双向选择调整策略在整个半监督聚类过程中,外层全局范围内采用概率计算每个数据对于各个类的隶属度,内层在标注数据的局部范围内对未标注数据进行类别标注。局部范围不断缩小,最后达到算法收敛,得到聚类结果。

## 2.2 改进的半监督聚类算法

将双向选择调整策略引进第1节两种半监督聚类算法中,形成两种改进算法,即基于多项式模型的半监督双向选择调整聚类算法(DMNSC)和半监督模糊双向选择调整聚类算法(SDFCM)。改进后的两种算法根据类别的描述参数重新计算每个数据的类别所属,然后由标注数据对其周围数据的类别进行调整。这样可以使划分的类别更加准确,从而提高聚类的准确度。

改进后的两种半监督聚类算法流程如下。

●基于多项式模型的半监督双向选择调整聚类算法(DMNSC):

输入:  $N$  个向量数据 ( $N_l$  个已知标注的数据  $\mathbf{X}^{(l)} = \{\mathbf{x}_1, \dots, \mathbf{x}_{N_l}\}$  与对应的类别标签  $Y^{(l)} = \{y_1^{(l)}, \dots, y_{N_l}^{(l)}\}$ ,  $y^{(l)} \in \{1, \dots, k\}$ ,  $k \leq K$  和  $N_u$  个未标注的数据向量  $\mathbf{X}^{(u)} = \{\mathbf{x}_{N_l+1}, \dots, \mathbf{x}_N\}$ ), 聚类模型  $\Lambda = \{\lambda_1, \dots, \lambda_K\}$ 。参数  $\sigma$  的初始值,参数  $Q$  的值。

输出: 所有数据的类别划分。

第一步,初始化:

a) 根据公式(1)和(2)将部分未标注数据进行类别标注。

b) 利用修改的 KKZ 方法得到初始的类别划分和模型  $\Lambda$  的参数。

第二步,迭代优化:

a) 更新参数  $\sigma$  的值;

b) 对于  $n \in \{N_{l+1}, \dots, N\}$ , 根据公式(1)和(2)将部分无标签数据进行标注;

c) 对于每个标注数据  $\mathbf{x}_n$ , 当  $y = y_n^{(i)}$ , 设定  $p(y | \mathbf{x}_n) = 1$ , 当  $y \neq y_n^{(i)}$ , 设定  $p(y | \mathbf{x}_n) = 0$ ;

d) M 步骤,对所有未标注数据  $\mathbf{x}_n$  更新  $p(y | \mathbf{x}_n)$ ;

e) E 步骤,更新类模型  $\Lambda$ ;

f) 判断更新后的数据类别与更新前是否发生变化,如无变化,则转到第三步,否则返回第二步 a)。

第三步,结果输出,对于每一个数据  $\mathbf{x}_n$ , 得到其类别标签  $y_n = \arg \max_j \log p(\mathbf{x}_n | \lambda_j)$ 。

●半监督模糊双向选择调整聚类算法(SDFCM):

输入:  $N$  个向量数据,  $N_l$  个已知标注的数据  $\mathbf{X}^{(l)} = \{\mathbf{x}_1, \dots, \mathbf{x}_{N_l}\}$  与对应的类别标签  $Y^{(l)} = \{y_1^{(l)}, \dots, y_{N_l}^{(l)}\}$ ,  $y^{(l)} \in \{1, \dots, K\}$ ,  $N_u$  个未标注的数据向量  $\mathbf{X}^{(u)} = \{\mathbf{x}_{N_l+1}, \dots, \mathbf{x}_N\}$ , 聚类中心  $C = \{c_1, \dots, c_K\}$ 。参数  $\sigma$  的初始值,参数  $Q$  的值。

输出: 所有数据的类别划分。

第一步,初始化:

a) 根据公式(1)和(2)将部分未标注数据进行类别标注。

b) 利用修改的 KKZ 方法得到初始的类别划分和聚类中心  $C$ 。

第二步,迭代优化:

a) 更新参数  $\sigma$  的值;

b) 对于  $n \in \{N_{l+1}, \dots, N\}$ , 根据公式(1)和(2)将部分无标签数据进行标注;

c) 对于每个标注数据  $\mathbf{x}_n$ , 当  $y = y_n^{(i)}$ , 设定  $p(y | \mathbf{x}_n) = 1$ , 当  $y \neq y_n^{(i)}$ , 设定  $p(y | \mathbf{x}_n) = 0$ ;

d) 对所有未标注数据  $\mathbf{x}_n$  更新  $\mu_{y_n}$ ;

e) 更新类中心  $C$ ;



f) 如果隶属矩阵  $U$  中元素最大的变动值小于阈值  $\varepsilon$ , 则转到第三步, 否则返回第二步 a)。

第三步, 结果输出, 对于每一个数据  $x_n$ , 得到其类别标签  $y_n = \arg \max_j \mu_{jn}$ 。

从上面算法的实现步骤可以看到, 在半监督聚类算法的前期阶段, 聚类划分可能偏差较大, 算法强化已标注数据的作用, 通过内层的局部调整加快聚类的收敛。在后期, 聚类基本收敛, 算法减小局部调整范围, 提高反馈调整的有效性, 来加快聚类的收敛。

3 实验与结果分析

3.1 实验数据

文中采用 Karypis Lab 中的部分文本数据集<sup>[14]</sup>进行实验来验证文中方法的有效性, 实验数据用 CLUTO 中相关方法进行预处理 (<http://www.cs.umn.edu/~cluto>)。

表 1 列出了文中用到的实验数据集及其相关的属性数据, 其中, 数据集的数据数量用数据量表示, 数据集的类别数量用类别数表示, 数据集中最大的一个类别的数据数量用最大类数据量表示, 数据集中最小的一个类别的数据数量用最小类数据量表示, 数据集中的特征词维度用特征维度表示, 数据集中最小类别的数据量与最大类别的数据量的比值用平衡度表示。

表 1 实验数据统计

数据集名称	数据量	类别数	最大类数据量	最小类数据量	特征维度	平衡度
C400	7049	4	3203	1033	41681	0.3225
tr11	414	8	132	11	6424	0.0833
NG5	1943	5	404	371	43586	0.918

3.2 实验方案

实验中, 提出的双向选择调整算法中的最大候选数据量  $Q$  为 5, 阈值  $\gamma$  为 0.75。针对文本数据, 文中采用夹角余弦公式计算数据的相似性<sup>[15]</sup>, 设计调控参数  $\sigma$  的范围为  $[0.25, 0.7]$ , 初始  $\sigma_0$  为 0.25,  $\sigma$  在聚类过程中的变化为  $\sigma = 0.25 + 0.45(1 - \exp(-0.1 * t))$ , 其中  $t$  为迭代次数。

实验分为两组, 分别为完全标签和不完全标签两种方案。在完全标签方案中, 对所有类别的部分数据做类别标注, 而在不完全标签方案中, 仅对部分类别的部分数据做类别标注。

●完全标签: 在数据集的所有类别中, 按照一定的比例 (包括 0.5%、1%、3%、5%、10%、20%、30%、40%), 随机选取部分数据以及其所属的类别作为标注数据。

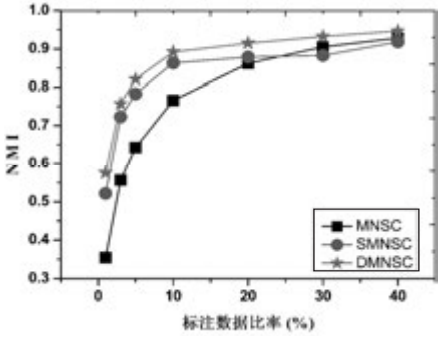
●不完全标签: 在数据集的所有类别中, 首先随机选取一半的类别 (NG5 中随机选择 3 个类别), 然后对

选取的类别中随机选取一定的比例 (包括 0.5%、1%、3%、5%、10%、20%、30%、40%) 的数据以及其所属的类别作为标注数据。

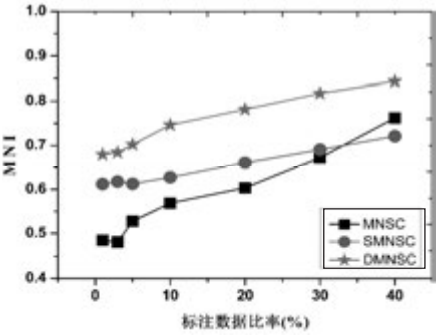
在实验中, 不仅对原半监督聚类算法与改进后的半监督聚类算法进行比较, 还在半监督双向选择调整聚类算法中分别采用余弦相似度和双向选择策略两种计算方法进行结果对比, 来验证双向选择策略的有效性。在后面的实验结果中, 用 MNSC、DFCM 分别表示基于多项式模型的半监督聚类算法和半监督模糊聚类算法, 用 SMNSC、DMNSC 分别表示采用余弦相似度和双向选择策略的基于多项式模型的半监督双向选择调整聚类算法, 用 SSFCM、SDFCM 分别表示采用余弦相似度和双向选择策略的半监督模糊双向选择调整聚类算法。实验用 NMI 指标<sup>[8]</sup>来评价聚类结果。

3.3 实验结果及分析

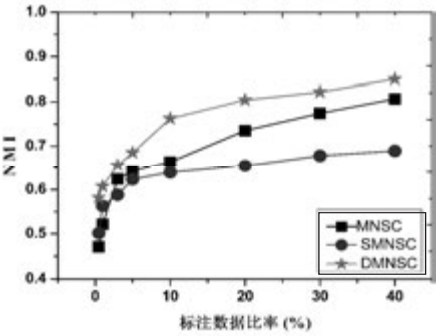
完全标签情况下的实验结果如图 1 和 2 所示。其



(a)C400 实验结果



(b)tr11 实验结果



(c)NG5 实验结果

图 1 在完全标签数据条件下基于多项式模型的半监督聚类算法的比较

中,图1是基于多项式模型半监督聚类算法的实验结果;图2是基于半监督模糊聚类算法的实验结果。

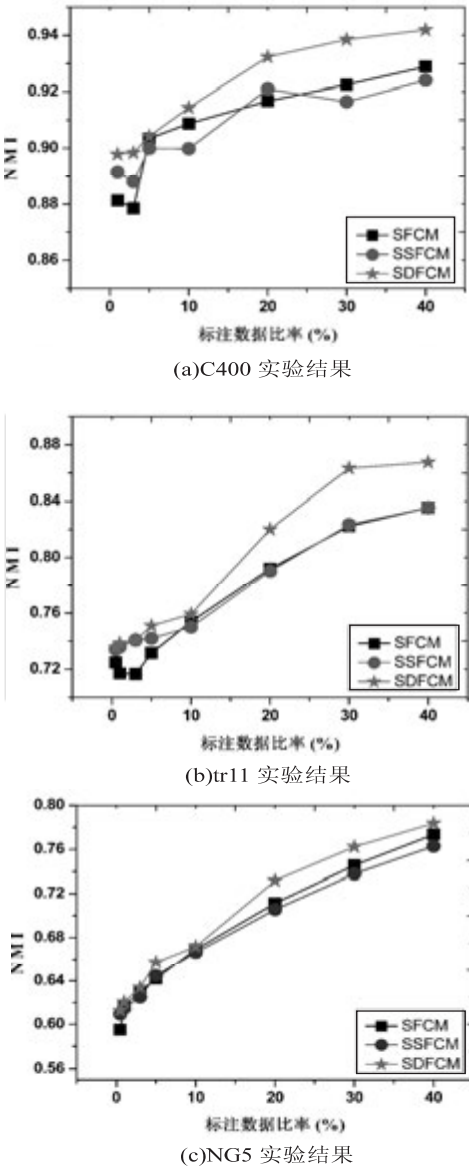


图2 在完全标签数据条件下半监督模糊聚类算法的比较

通过图1和图2可以发现基于双向选择调整策略的改进算法能够较好地提高聚类的效果。特别是标注数据在10%~30%时,基于双向选择调整策略的半监督聚类算法对聚类准确度的提高较大。这是因为在算法初始阶段,根据局部数据密度控制被调整的无标注数据数量,使生成的类别参数(如类中心)更接近实际值。当已标注数据在10%以下时,基于双向选择调整策略的半监督算法对聚类准确度的提高较小。这是因为双向选择调整策略虽然保证了调整数据有效性,却限制了被调整数据的数量,使得聚类准确度的提高较小。而当已标注数据较多时,其保证准确度的作用得到了较好的体现,图1、2中,在标注数据比率大于10%后,NMI值有了较大的提高。但是,采用余弦相似

度的调整策略,因为无法保证被调整数据的准确度,所以在标注数据较多时可能会限制聚类准确度的提高。如图1、2,已标注数据在20%~40%时,采用余弦相似度的SMNSC和SDFCM中的调整策略反而降低了聚类的准确度。

不完全标签情况下的实验结果如图3和4所示。其中,图3是基于多项式模型半监督聚类算法的实验结果;图4是半监督模糊聚类算法的实验结果。

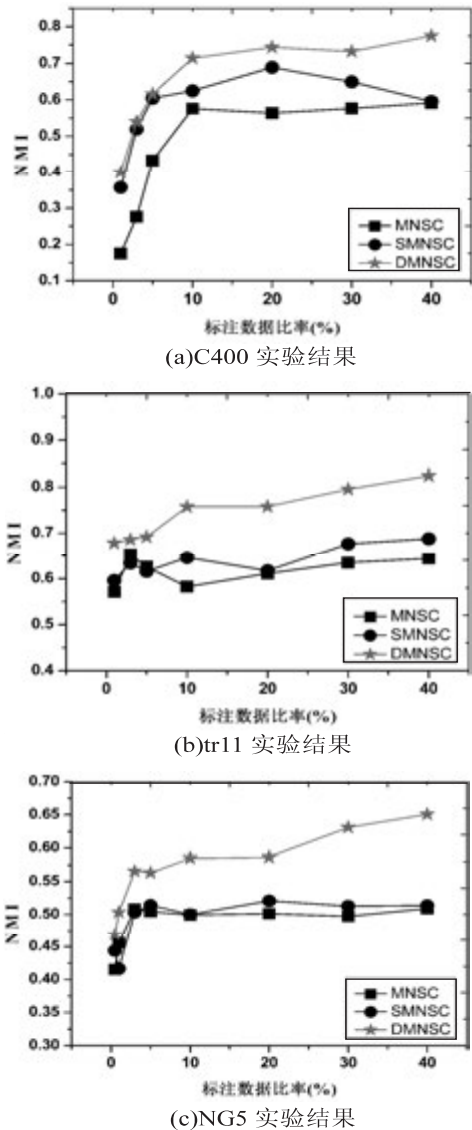
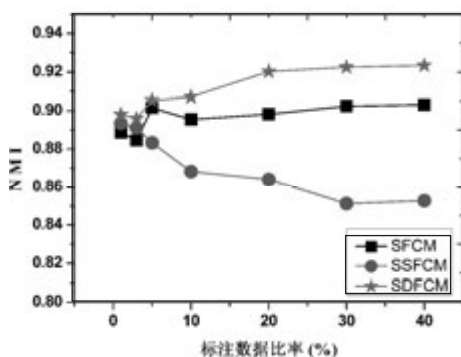


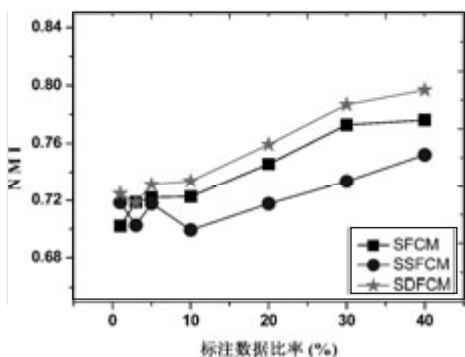
图3 在不完全标签数据条件下基于多项式模型的半监督聚类算法的比较

图3和图4显示基于双向选择调整策略的改进算法在不完全标签情况下的半监督聚类效果明显。在不完全标签情况下,初始的类别划分对聚类结果的精确度有重要的影响。只有能够更好地描述已知的类别,才能准确地得到未知的类别。基于双向选择调整策略的改进算法充分利用已标注的数据以及可调整的数据计算已知的类别描述参数(如类的中心),然后利用较准确的已知类别来确定未知的类别。此外,通过图3

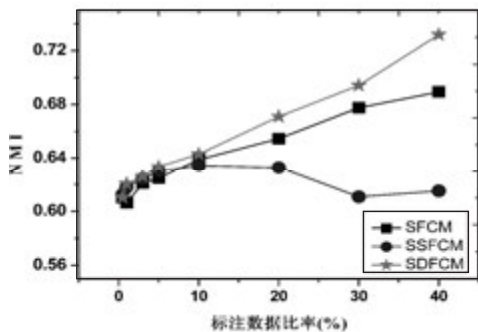
和图 4 还能看出当已标注数据数量较少时,基于多项式模型的半监督双向选择调整聚类算法和半监督模糊双向选择调整聚类算法聚类结果的 NMI 值,随着已标注数据数量的增加较快增加。这是因为,当已标注数据非常少时,由于调控参数  $\sigma$  和判断阈值  $\gamma$  的限制,可调整的数据数量很少,甚至为 0,聚类的效果与原半监督聚类算法的效果接近。但是,当可利用的标注数据增加时,可被调整的数据也有所增加,聚类结果的精度有较大提高。图 3 和图 4 的结果还显示出采用双向选择调整的有效性明显高于采用余弦相似度。而通过图 4 还发现当标注数据较多时,原半监督聚类算法的聚类结果优于采用余弦相似度的基于双向选择调整策略的改进算法的聚类结果。这充分说明采用双向选择调整策略选择数据进行调整,保证了被调整数据的有效性,提高了聚类的效果。



(a)C400 实验结果



(b)tr11 实验结果



(c)NG5 实验结果

图 4 在不完全标签数据条件下半监督模糊聚类算法的比较

通过完全标签和不完全标签两种情况下对双向选择调整策略的有效性进行验证,实验结果显示双向选择调整策略在多种半监督聚类算法中都能取得较好的效果。

## 4 结束语

文中提出了一种应用于半监督聚类问题的双向选择调整策略,该方法在聚类过程中能够有效地利用标注类别标签的数据对其周围局部数据进行类别标注,提高无标注数据类别划分的准确度。双向选择策略保证了被调整的数据的有效性。文中应用双向选择调整策略对基于多项式模型的半监督聚类算法和半监督模糊聚类算法进行了改进,实验结果验证了双重调整方法的有效性。

除了提供数据的类别标签,半监督聚类中还存在其他多种形式的已知信息,比如标注一对数据是同一类或者不属于同一类的约束信息。在后续工作中研究在其他形式的已知信息条件下,如何应用文中提出的方法对半监督聚类算法进行改进。

## 参考文献:

- [1] Li Kunlun, Cao Zheng, Cao Liping, et al. A Novel Semi-supervised Fuzzy C-Means Clustering Method [C]//22nd International Conference on Control and Decision. Hebei: IEEE Press, 2009: 3761-3765.
- [2] Guan Renchu, Shi Xiaohu, Marchese M. Text Clustering with Seeds Affinity Propagation [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 23(4): 627-637.
- [3] Dang Yanzhong, Xuan Zhaoguo, Rong Lili, et al. A Novel Initialization Method for Semi-supervised Clustering [C]//4th International Conference on Knowledge Science Engineering and Management. Belfast: Springer Publisher, 2010: 317-328.
- [4] Basu S, Banerjee A, Mooney R. Semi-supervised Clustering by Seeding [C]//19th International Conference on Machine Learning. Sydney: ACM Inc, 2002: 19-26.
- [5] Basu S. Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments [D]. Knoxville: University of Texas at Austin, 2005.
- [6] Shi Zhong, Ghosh J. A Unified Framework for Model-based Clustering [J]. Journal of Machine Learning Research (JMLR), 2003, 4(11): 1001-1037.
- [7] Zhang Huaxiang, Lu Jing. Semi-supervised Fuzzy Clustering: A Kernel-based Approach [J]. International Journal of Knowledge Based Systems, 2009, 22(6): 477-481.
- [8] Shi Zhong. Semi-supervised Model-based Document Clustering: A Comparative Study [J]. Machine Learning, 2006, 65(1): 3-29.
- [9] Basu S, Bilenko M, Mooney R J. A Probabilistic Framework for



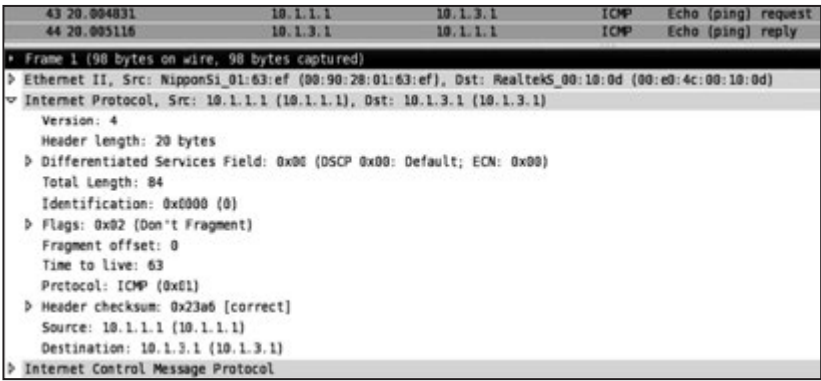


图 5 ASR1 的 eth2 上抓到的包

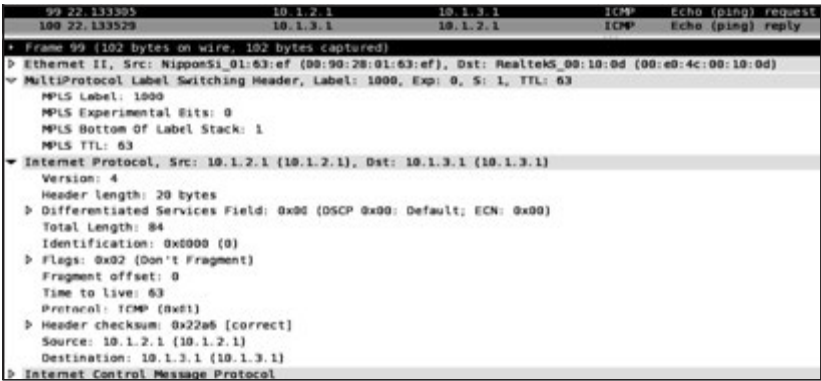


图 6 在 ASR1 的 eth3 上抓到的包

网络虚拟化基本功能测试通过,假定的不同性质的数据流在 ASR 处被分离开,服务质量要求较高或传送文件较大的数据流通过 MPLS 模拟电路交换被发送出去,而其余的数据流则通过 IP 分组交换被发送出去。

4 结束语

文中设计并实现了一种基于身份与位置分离映射的混合交换路由系统。在该系统中,接入路由器对数据流的某些特性(如流所传送文件的大小或流所要求的服务质量等)进行判定,并根据判定结果使不同的数据流通过不同的分组/电路平面传输。例如,对于服

务质量要求高或者传送大文件的流,使其通过电路交换传输,直接走底层的光纤链路,而其他的数据流则通过分组交换传输。文中设计了该系统的所有功能模块,并且进行了功能测试。

参考文献:

[1] Guichard J, Faucheur F L, Vasseur J P. Definitive MPLS network designs [M]. Indianapolis: Cisco Press, 2005.

[2] 张宏科, 苏伟. 新网络体系基础研究——一体化网络与普适服务[J]. 电子学报, 2007, 35(4): 593-598.

[3] 杨冬, 周华春, 张宏科. 基于一体化网络的普适服务研究[J]. 电子学报, 2007, 35(4): 607-613.

[4] 董平, 秦亚娟, 张宏科. 支持普适服务的一体化网络研究[J]. 电子学报, 2007, 35(4): 599-606.

[5] Farinacci D, Fuller V, Oran D. Locator/ID separation protocol (LISP) [M]. [s.l.]: IETF, 2007.

[6] Rosen E. Multiprotocol Label Switching Architecture [S]. RFC3031, 2001.

[7] Bovet D, Cesati M. Understanding The Linux Kernel [M]. [s.l.]: [s.n.], 2005.

[8] 陈启美, 吴政, 刘海. MPLS 组件与框架—MPLS 体系结构解析[J]. 电力自动化设备, 2002(2): 87-90.

[9] 肖宇峰, 李昕, 时岩. Linux 网络内核分析与开发 [M]. 北京: 电子工业出版社, 2010.

[10] DeGhein L. MPLS 技术架构 [M]. 陈麒帆译. 北京: 人民邮电出版社, 2008.

[11] 赵强, 鲁昆生. 多协议标记交换 (MPLS) 技术研究及应用 [J]. 武汉理工大学学报, 2004(3): 94-96.

+++++ (上接第 6 页)

Semi-supervised Clustering [C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle: ACM Inc, 2004: 59-68.

[10] Nigam K, McCallum A K, Thrun S, et al. Text Classification from Labeled and Unlabeled Documents Using EM [J]. Machine Learning, 2000, 39(1): 103-134.

[11] 张敏, 于剑. 基于划分的模糊聚类算法 [J]. 软件学报, 2004, 15(6): 858-869.

[12] Frey B J, Dueck D. Clustering by Passing Messages between Data Points [J]. Science, 2007, 315(5814): 972-976.

[13] 肖宇, 于剑. 基于紧邻传播算法的半监督聚类 [J]. 软件学报, 2008, 19(11): 2803-2813.

[14] CLUTO Document Datasets Toolkit [EB/OL]. [2010-06-01]. <http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/datasets.tar.gz>.

[15] Luo Congnan, Li Yanjun, Chung S M. Text Document Clustering Based on Neighbors [J]. Data and Knowledge Engineering, 2009, 68(11): 1271-1288.

## 基于双向选择调整策略的半监督聚类算法

作者: [刘明](#), [宣照国](#), [吴江宁](#)  
作者单位: [大连理工大学 系统工程研究所, 辽宁 大连 116024](#)  
刊名: [计算机技术与发展](#)  
英文刊名: [Computer Technology and Development](#)  
年, 卷(期): 2013 (2)

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_wjtz201302003.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjtz201302003.aspx)