一种基于数据相关性的半监督模糊聚类集成方法

冯晨菲 杨 燕 王红军 徐英歌 王 韬

(西南交通大学信息科学与技术学院 成都 610031)

摘 要 现有的半监督聚类集成方法能利用先验信息,使集成的准确性、鲁棒性和稳定性得到提高,但在集成阶段加入成对约束信息时,只考虑了给定的约束信息而忽视了约束点与被约束点的邻域点之间的关系。针对此问题,提出了一种基于数据相关性的半监督模糊聚类集成方法。该方法首先利用半监督模糊聚类算法建立集成信息矩阵,并将其转换为相似性矩阵;然后,利用已知的约束信息及约束点与被约束点的邻域点之间的关系来修改相似性矩阵;最后,利用图划分算法得到最终的聚类结果。真实数据上的实验结果表明,提出的方法可以有效提高聚类质量。

关键词 半监督聚类集成,模糊聚类,成对约束,邻域点

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.009

Semi-supervised Fuzzy Clustering Ensemble Approach with Data Correlation

FENG Chen-fei YANG Yan WANG Hong-jun XU Ying-ge WANG Tao

(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract Semi-supervised clustering ensemble has emerged as a powerful machine learning paradigm that provides improved precision, robustness and stability by taking advantage of prior information, while most of them only consider the given pairwise constraints and do not consider the neighbors around the data points constrained in the ensemble step. In this paper, a semi-supervised fuzzy clustering ensemble with data correlation (SFCEDC) was proposed to overcome this defect. Firstly, an ensemble information matrix is built by primarily exploiting the results of semi-supervised fuzzy clustering and a similarity matrix is constructed by aggregating much information of the ensemble information matrix. And then this matrix is modified by using the given constraints and the neighbors around the data points constrained. Finally, a graph partitioning algorithm is employed to get the final clustering results. Experimental results on UCI datasets demonstrate that the proposed approach can improve clustering performance effectively.

Keywords Semi-supervised clustering ensemble, Fuzzy clustering, Pairwise constraints, Neighbors points

1 引言

聚类[1]是数据挖掘领域的一个重要研究方向,它根据某种相似度准则将数据对象划分为若干个类或簇,以发现数据内在的分布情况。尽管目前学者提出了许多聚类算法及其改进算法,但机器学习中仍存在一个普适定理——没有免费的午餐[2](No Free Lunch Theorem, NFL),NFL 定理表明:任何单一、超强的聚类算法都无法准确揭示各种数据集所呈现出的簇结构;面对可能具有各种形状或结构的高维数据集,寻找适合于该数据集的单一聚类算法变得愈发困难。

聚类集成[3]将不同算法或者同一算法下运用不同初始参量所获得的结果进行融合,以得到优于单个聚类算法的结果。Strehl 等引进图分割的思想,提出了 CSPA、HGPA、MCLA 这 3 种基于超图的算法[4]。罗会兰等提出了一种基于数学形态学的聚类集成方法[5]。周志华等通过建立投票机制来解决聚类集成问题,根据投票结果得到聚类集成结果[6]。Iam-on 等在考虑簇与簇相关性的基础上提出了基于链接的聚类集成方法[7]。Naldi 等提出了一种基于相对评价指标的选择性聚类集成方法[8]。

然而,多数情况下的聚类集成算法是建立在非监督方式之上的,但由于缺乏对先验知识的利用,致使聚类集成的准确性、鲁棒性和稳定性降低。为此许多学者提出在聚类集成中使用半监督学习技术来克服这些缺点,进而提高学习性能。Abdala 等提出一种基于成对约束的半监督聚类集成方法 Cop-EAC-SL,该方法较好地保证了集成结果不违反约束关系[9]。王红军等把半监督学习和聚类集成结合起来,设计出基于贝叶斯网络的半监督聚类集成模型,并对模型进行了变分法推理求解[10]。Iqbal 等提出基于投票法的半监督聚类集成方法,该算法在集成阶段使用已知标签信息来修改基聚类结果,在此基础上使用权重投票法获得最终结果[11]。Yang

到稿日期:2014-04-11 返修日期:2014-05-22 本文受国家自然科学基金(61170111,61134002),西南交通大学牵引动力国家重点实验室自主研究课题(2012TPL_T15)资助。

冯晨菲(1989—),男,硕士生,CCF 学生会员,主要研究方向为数据挖掘、集成学习;杨 燕(1964—),女,教授,博士生导师,主要研究方向为数据挖掘、计算智能、集成学习等,E-mail:yyang@swjtu.edu.cn(通信作者);王红军(1977—),男,副研究员,硕士生导师,主要研究方向为机器学习、数据挖掘、集成学习;徐英歌(1988—),女,硕士生,主要研究方向为数据挖掘、集成学习;王 韬(1985—),男,硕士生,主要研究方向为数据挖掘、云计算。

等提出基于多蚁群的半监督聚类集成模型，该算法先将成对约束信息加入到蚁群聚类算法中，然后在多蚁群结果合并的相似性矩阵中加入半监督信息，获得了较好的实验结果[12]。

本文提出一种基于数据相关性的半监督模糊聚类集成方法（Semi-supervised Fuzzy Clustering Ensemble Approach with Data Correlation，SFCEDC），该算法在集成阶段加入成对约束信息时不仅考虑约束点与被约束点之间的关系，同时考虑了约束点与被约束点邻域点之间的关系。这一设计将数据本身的相关信息与半监督知识相结合，在半监督信息的基础上对先验知识进行扩充，最终获得了较为理想的实验结果。

2 成对约束与半监督聚类集成

2.1 成对约束

在半监督学习中，先验知识具有不同的表现形式，其中成对约束是比较常用的一种。成对约束所描述的主体是两个数据实例。正关联约束关系表示两个数据实例属于相同的类，用 Must-Link 表示；而负关联约束关系表示两个数据实例属于不同的类，用 Cannot-Link 表示[13]。

例如，对于给定的同一数据集上的两个不同的数据对象 x_i 和 x_j，如果它们同属于一个类，表示 $(x_i,x_j)\in$ Must-Link；如果属于不同类，表示 $(x_i,x_j)\in$ Cannot-Link。Must-Link 约束和 Cannot-link 约束具有对称性和传递性[14]。

(a)对称性

$$(x_i,x_j)\in\text{Must-Link}\Rightarrow(x_j,x_i)\in\text{Must-Link} \tag{1}$$

$$(x_i,x_j)\in\text{Cannot-Link}\Rightarrow(x_j,x_i)\in\text{Cannot-Link} \tag{2}$$

(b)传递性

$$(x_i,x_j)\in\text{Must-Link}\&\&(x_j,x_k)\in\text{Must-Link}\Rightarrow(x_i,x_k)\in\text{Must-Link} \tag{3}$$

$$(x_i,x_j)\in\text{Cannot-Link}\&\&(x_j,x_k)\in\text{Cannot-Link}\Rightarrow(x_i,x_k)\in\text{Cannot-Link} \tag{4}$$

2.2 半监督模糊 C 均值聚类算法

由 Pedrycz 在文献[15]中提出的半监督 FCM 聚类(S-FCM)算法是用部分已标记的样本引导迭代优化过程。S-FCM的目标函数为：

$$J=\sum_{i=1}^{c}\sum_{j=1}^{N}U_{ij}^{p}d_{ij}^{2}+\alpha\sum_{i=1}^{c}\sum_{j=1}^{N}(U_{ij}-f_{ij}b_j)^2d_{ij}^{2} \tag{5}$$

其中，$d_{ij}=\|v_i-x_j\|$，v_i 表示第 i 个簇的中心，α 为已标记样本数 L 与样本总数 N 的比值，标记样本的隶属度矩阵用 f_{ij} 表示，b_j 是一个二值函数，表示是否为标记样本，p 为加权指数，c 为簇个数。由于目标函数的求解较为复杂，这里给出最小化式(5)的两个必要条件：

$$v_i=\frac{\sum_{j=1}^{N}U_{ij}^{2}x_j}{\sum_{j=1}^{N}U_{ij}^{2}} \tag{6}$$

$$U_{ij}=\frac{1}{1+\alpha}\left\{\frac{1+\alpha(1-b_j\sum_{i=1}^{c}f_{ij})}{\sum_{k=1}^{c}(d_{ij}/d_{kj})^2}+\alpha f_{ij}b_j\right\} \tag{7}$$

由上述两个条件可知，S-FCM 算法通过不断迭代式(6)与式(7)来更新聚类中心 v_i 和隶属度 U_{ij}，再根据式(5)计算目标函数。如果目标函数相对于上次目标函数的改变量小于某个阈值，则算法停止。

2.3 半监督聚类集成

作为数据挖掘的新技术，半监督聚类集成越来越受到研究人员的关注。半监督聚类集成将半监督学习用于聚类集成的方法，充分利用了半监督信息来指导聚类过程，同时融入集成学习的思想，将多个基聚类结果进行融合以得到最优的划分结果。半监督聚类集成的示意图如图 1 所示。

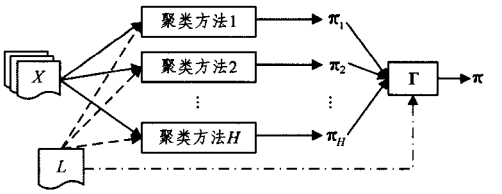

图 1 半监督聚类集成示意图

这里 X 表示数据集，L 表示半监督信息，$\pi_i(i=1,\cdots,H)$ 表示基聚类结果，Γ 表示共识函数，π 表示最终结果。在半监督聚类集成中，半监督信息可以先用于产生基聚类结果的过程，然后用于且必须用于指导集成过程。在上述两个阶段均加入半监督信息来提高聚类性能，这也是半监督聚类集成方法中半监督信息的主要作用。

3 考虑数据相关性的半监督模糊聚类集成方法

3.1 SFCEDC 算法

SFCEDC 算法首先构建集成信息矩阵，该矩阵实质上是一个表示数据对象与簇关系的图。假设对数据集 $X=\{x_1,x_2,\cdots,x_N\}$ 用初始化不同的 S-FCM 算法运行 H 次得到 H 个有差异性的结果，该结果的集合表示为 $\Pi=\{\pi_1,\pi_2,\cdots,\pi_H\}$。这里使用 S-FCM 作为基聚类器的原因是：S-FCM 算法不仅可以添加半监督信息用于指导聚类过程，而且其结果可以直接表示为数据与簇之间的相关信息。利用结果集将集成信息矩阵表示为 $\psi(H)=[\pi_1,\pi_2,\cdots,\pi_H]$，集成信息矩阵所表示的含义如图 2 所示。

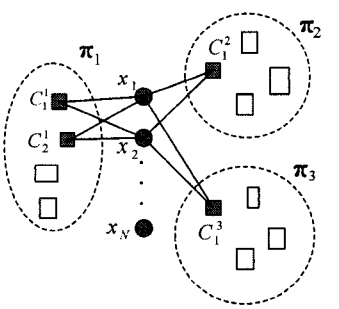

图 2 集成信息矩阵含义的示意图

假设数据集 X 在第 i 个基聚类器所产生的簇的个数为 k_i，则数据在聚类阶段所产生的簇的总个数 $m=\sum_{i=1}^{H}k_i$，那么集成信息矩阵 ψ 可以表示为一个 $N\times m$ 的矩阵，其中矩阵 ψ 中第 i 行第 l 列的元素 γ_{il} 表示第 i 个数据对第 l 个簇的隶属程度，也即图 2 中边所代表的含义。

为求得数据对象之间的相似性矩阵，首先需定义两个数据对同一个簇的隶属程度：

$$\omega_{ij}^{l}=\min(\gamma_{il},\gamma_{jl}) \tag{8}$$

ω_{ij}^{l} 的实质为图 2 中两个数据对象对同一个簇的两个边的最小权重值。同时，由于多个簇的存在，定义两个数据对象之间的相似度为：

$$W_{ij}=\sum_{l=1}^{m}\omega_{ij}^{l} \tag{9}$$

通过式(9)得到数据对象之间的相似度矩阵 W，对其做

归一化处理。处理规则如下：

$$W(i,j)=\begin{cases}\dfrac{W_{ij}}{W_{\max}}, & i\neq j\\ 1, & i=j\end{cases} \tag{10}$$

由于在聚类阶段获得了标记信息，这些标记信息可以很自然地转化为成对约束信息 Must-Link 和 Cannot-Link。数据点之间的约束关系可以通过图 3 很好地体现出来。

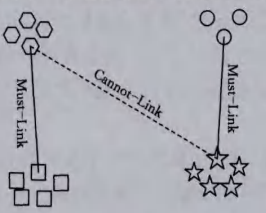

图 3　数据点之间的约束关系示意图

在集成阶段，单纯地加入给定的约束信息并不能很好地体现数据之间的关联信息，为此本文考虑约束点与被约束点的邻域点之间的关系，这样可以充分利用数据对象之间本身存在的相关信息。

接下来，首先求得数据对象之间的欧氏距离，并将结果表示为距离矩阵 D，然后将成对约束信息与距离矩阵 D 相结合。根据成对约束信息来遍历距离矩阵，若一个数据点与另外一个数据点为 Must-Link，那么就认为这个数据点与另外一个数据点相邻的 λ 个值为 Must-Link；若一个数据点与另外一个数据点为 Cannot-Link，那么就将与另外一个数据点相邻近的 λ 个数据点与这个数据点的约束关系定为 Cannot-Link。λ 的具体取值会在 3.2 节中详细讨论。

在扩充了上面的约束关系之后，下一步的任务便是根据以上所添加的成对约束信息修改数据对象之间的相似度矩阵 W。为方便起见，这里首先给出表示数据对象之间的关系为 Must-Link 的集合 M 与表示数据对象之间的关系为 Cannot-Link 的集合 C。假设存在数据对象 x_i 与数据对象 x_j，如果 x_i 与 x_j 满足上面所认为的 Must-Link 的条件，那么将成对约束(x_i,x_j)放入到集合 M 中；如果 x_i 与 x_j 满足上面所认为的 Cannot-Link 的条件，那么将成对约束(x_i,x_j)放入到集合 C 中。接着使用半监督信息的集合 M 与 C 对相似度矩阵进行修改，修改公式如下：

$$W(i,j)=\begin{cases}1, & (x_i,x_j)\in M\\ 0, & (x_i,x_j)\in C\\ W(i,j), & \text{其他}\end{cases} \tag{11}$$

经过式(11)的处理之后，数据点之间的约束关系可以用图 4 形象地表示出来。

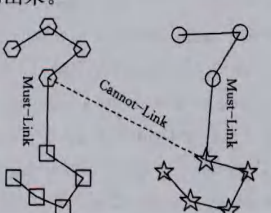

图 4　修改后的数据点之间的约束关系图

相似度矩阵 W 是体现数据相关性的权重图。在获得新的数据对象之间的相似度矩阵之后，接下来便是对这个权重图进行划分。图划分技术是聚类技术中的关键技术，Metis[16] 和 SPEC(Spectral graph partitioning)[17] 是两种著名的图划分聚类算法。本文采取 Metis 算法对权重图进行聚类来获得最终的聚类集成结果 π。

3.2　数据邻近点个数 λ 取值的讨论

λ 的取值是 SFCEDC 算法的关键。通常情况下盲目地确定邻域点个数并不具有科学性。过多的取值会引入更多的噪声，相反，过少的取值会造成信息的损失，都会使得相似度矩阵 W 不能很好地体现数据之间的相关信息，进而导致最终结果不理想。

假设距离矩阵 D 的每一行数据均服从高斯分布，那么第 i 行数据满足均值 μ 为 x_i、标准差为 σ_i 的正态分布。正态曲线下横轴上一定区间的面积反映该区间的例数占总例数的百分比，这种比例关系可以通过图 5 得到很好的体现。

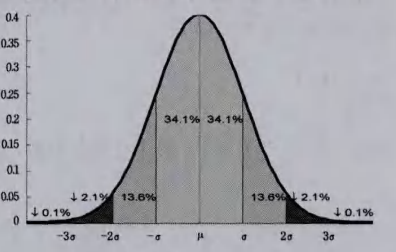

图 5　正态曲线的面积分布

由图 5 可知正态分布曲线下，横轴区间$[\mu-\sigma,\mu+\sigma]$内所占面积约为 68%，表明与数据点 x_i 相邻近的数据点占到数据总数的 68%，为此将横轴区间范围缩小为$[\mu-\sigma/5,\mu+\sigma/5]$。当横轴区间为$[\mu-\sigma/5,\mu+\sigma/5]$时，其面积的计算如下：

$$\int_{\mu-5/\sigma}^{\mu+5/\sigma}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x\approx 16\% \tag{12}$$

实验数据表明，在此横轴范围内取值能很好地获取与数据点 x_i 最邻近的点，意味着与数据点最邻近的点为落在横轴区间$[\mu-\sigma/5,\mu+\sigma/5]$内的点。因此，$\lambda$ 的取值与数据集所拥有的数据总个数 N 有关，而并不是单纯的一个固定值。除了此方法，还可以以数据点为圆心、一定的距离 R 为半径画圆来选取数据的邻近点，但是由于不同的数据点间所选取的半径 R 不同，而且不知道 R 到底取值多少为好，因此只能设定不同的半径来尝试获得最好值。

3.3　算法描述

综合以上分析，SFCEDC 算法描述如下。

输入：数据集 $X=\{x_1,x_2,\cdots,x_N\}$，标记数据集 L，未标记数据集 E，簇个数 c

输出：数据对象标签集 π

SFCEDC(X,L,E,c)

1. 对数据集 X 运行 H 个不同初始条件的半监督 FCM 聚类算法，得到 H 个有差异性的结果集$\Pi=\{\pi_1,\pi_2,\cdots,\pi_H\}$。
2. 将结果集Π转化为表示数据与簇相关性的集成信息矩阵 $\psi(H)=[\pi_1,\pi_2,\cdots,\pi_H]$。
3. 将集成信息矩阵 $\psi(H)$转换为体现数据之间相关性的相似性矩阵 W。
4. 根据约束点与被约束点的邻近点之间的关系扩充半监督信息，并利用半监督信息来修改相似度矩阵 W。
5. 使用图划分算法 Metis 对相似度矩阵 W 进行划分，得到最终结果 π。

4　实验

4.1　实验数据与评价标准

本文选用在 UCI(University of California, Irvine)的机器学习库[18] 上下载的部分数据集作为实验数据集。表 1 给出

了这些数据的相关信息描述。

表 1　实验数据集的相关信息描述

数据集	样本个数	属性	分类数
Glass	214	9	6
Iris	150	4	3
Bupa	345	6	2
Wdbc	569	30	2
Ionosphere	351	34	2
Magic04	19020	10	2
Pima	768	8	2
Wine	178	13	3
Abalone	4177	8	29
Segmentation	2100	19	7

实验采用 Micro-precision[6,10,19] 标准来对各算法的聚类效果进行评价，其计算公式如下：

$$MP=\frac{\sum_{i=1}^{c}(m_i)}{N} \tag{13}$$

其中，c 为簇个数，m_i 为在第 i 个簇中分对的个数，N 为数据总个数。为避免集成结果的偶然性，进行多次重复实验，将集成结果的平均准确率作为最终聚类结果。该过程的计算公式为：

$$AMP=\frac{1}{H\times N}\sum_{h=1}^{H}\sum_{i=1}^{c}m_i \tag{14}$$

其中，H 为重复进行实验的次数，本文中 H 在基聚类中的取值为 10，在集成过程中的取值为 20。

4.2　实验步骤与实验结果

实验首先选取 S-FCM 作为基聚类器，运行 H 次，由于初始值选取的随机性，将会得到 H 个有差异性的基聚类结果；然后根据 SFCEDC 算法对 H 个基聚类结果进行融合，得到集成后的结果 π。

实验中采取的对比算法有本文基聚类算法 S-FCM[15]（半监督信息为 10%）和文献[4]中提出的 3 种经典的聚类集成算法 CSPA、HGPA、MCLA。在运行 SFCEDC 算法的同时将半监督信息分别设置为 5%、8%、10%。表 2 给出了运行各种算法 20 次后每个数据集上的 AMP 值。最好的结果加粗标记，未获取的值用“N/A”表示。

表 2　各种算法的 AMP 值

数据集	S-FCM	CSPA	HGPA	MCLA	SFCEDC (5%)	SFCEDC (8%)	SFCEDC (10%)
Glass	0.6075	0.5607	0.4673	0.5748	0.5935	0.6028	**0.6355**
Iris	0.9000	0.8933	0.6067	0.8933	0.9267	0.9333	**0.9400**
Bupa	0.5797	0.5797	0.5797	0.5797	0.5797	0.5826	**0.5971**
Wdbc	0.8541	0.6696	0.6274	0.8541	0.8524	0.8330	**0.8576**
Ionosphere	0.7037	0.6667	0.6410	0.7094	0.7066	0.7293	**0.7464**
Magic04	0.6484	N/A	0.6484	0.6484	0.6484	0.6489	**0.6510**
Pima	0.6602	0.6510	0.6510	0.6589	0.6510	0.6641	**0.6745**
Wine	0.6910	0.6629	0.6461	0.6854	0.6966	0.7079	**0.7191**
Abalone	0.2696	0.2703	0.2449	0.2657	0.2818	0.3002	**0.3142**
Segmentation	**0.5848**	0.5342	0.4442	0.5693	0.4580	0.4766	0.4584

由表 2 可以看出：当加入半监督信息为 10%时，SFCEDC 算法在总共的 10 个数据集上有 9 次获得了最好的平均准确率；剩余 1 个数据集上，S-FCM 取得了最好的平均准确率。通过分析表 2 可以得出以下两个结论：

(1)与单一的半监督算法或者聚类集成算法相比较，半监督聚类集成都取得了较好的结果，说明了将半监督学习与聚类集成二者相结合的优越性；

(2)通过分析半监督聚类集成算法实验可以得出：使用较多半监督信息，可以提升 SFCEDC 算法的性能。因为从总体上来说，SFCEDC (10%) > SFCEDC (8%) > SFCEDC (5%)。

结束语　本文提出了一种基于数据相关性的半监督模糊聚类集成方法 SFCEDC，其利用邻近数据之间的相关信息来扩充半监督信息，并在集成阶段加入半监督信息以修改相似度矩阵。实验结果表明了该算法的优越性。然而在选取邻域点个数 λ 时未考虑聚类的簇个数 c，造成了局部数据集上的效果不理想；但考虑聚类簇个数 c 会增加选取邻域点个数 λ 的复杂度。在今后的工作中将进一步寻找有效选取邻域点个数 λ 的方法。

参 考 文 献

[1] Han J, Kamber M, Pei J. Data Mining Concepts and Techniques [M]. Morgan Kaufmann Press, 2012

[2] Wolpert D H, Macready W G. No free lunch theorems for search [R]. Technical Report SFI-TR-9502010. Santa Fe Institute, 1995

[3] Topchy A, Jain A K, Puneh W. Clustering ensembles: models of consensus and weak partition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866-1881

[4] Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions [J]. Journal of Machine Learning Research, 2003, 3(3): 583-617

[5] 罗会兰，危辉. 基于数学形态学的聚类集成算法[J]. 计算机科学，2010，37(8)：214-218
Luo Hui-lan, Wei Hui. Clustering Enseble Algorithm Based on Mathematical Morphology[J]. Computer Science, 2010, 37(8): 214-218

[6] Zhou Zhi-hua. Ensemble Methods: Foundations and Algorithms [M]. CRC Press, 2012

[7] Iam-on N, Boongone T, Garrett S, et al. Link-based cluster ensemble approach for categorical data clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24: 413-425

[8] Naldi M C, Carvalho A C P L F, Campello R J G B. Cluster ensemble selection based on relative validity indexes [J]. Data Mining and Knowledge Discovery, 2013, 27(2): 259-289

[9] Abdala D D, Jiang X. An evidence accumulation approach to constrained clustering combination[C]//Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition. Leipzig, Germany, 2009: 361-371

[10] 王红军，李志蜀，戚建淮，等. 基于贝叶斯网络的半监督聚类集成模型[J]. 软件学报，2010，21(11)：2814-2825
Wang Hong-jun, Li Zhi-shu, Qi Jian-huai, et al. Semi-supervised Cluster Ensemble Model Based on Bayesian Network[J]. Journal of Software, 2010, 21(11): 2814-2825

[11] Iqbal A M, Moh'd A, Khan Z A. Semi-supervised clustering en-

semble by voting [C]// Proceeding of the International Conference on Information and Communication System. Amman, Jordan, 2009: 1-5

[12] Yang Yan, Wang Hong-jun, Lin Chao, et al. Semi-supervised Clustering Ensemble Based on Multi-ant Colonies Algorithm [C]// Rough Sets and Knowledge Technology 7th International Conference. Chengdu, China, 2012: 302-309

[13] Wagstaff K, Cardie C, Rogers S. Constrained k-means clustering with background knowledge [C]// Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA, USA, 2001: 577-584

[14] Klein D, Kamvar S D, Manning C. From instance-level constraints to space-level constraints: marking the most of prior knowledge in data clustering [C]// Proceedings of the 19th International Conference on Machine Learning. San Francisco, CA, USA, 2002: 307-314

[15] Pedrycz W, Waletzky J. Fuzzy Clustering with Partial Supervision [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1997, 27(5): 787-795

[16] Karypis G, Kumar V. Multilevel K-Way partitioning scheme for irregular graphs [J]. Journal of Parallel Distributed Computing, 1998, 41(2): 278-300

[17] Ng A, Jordan M, Weiss Y. On Spectral Clustering: Analysis and an Algorithm[C]// Advances in Neural Information Processing Systems. 2001, 14: 849-856

[18] Blake C L, Merz C J. UCI repository of machine learning databases [EB/OL]. 2012-05-01[2012-12-01]. http://archive.ics.uci.edu/ml

[19] Modha D, Spangler W S. Feature weighting in k-means clustering [J]. Machine Learning, 2003, 52(3): 217-237

(上接第 31 页)

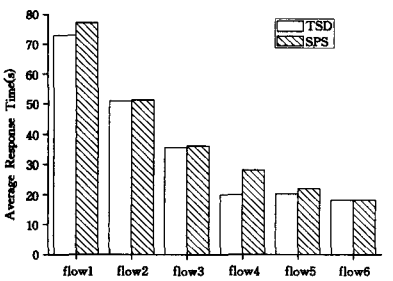

图 6 负载二作业平均响应时间

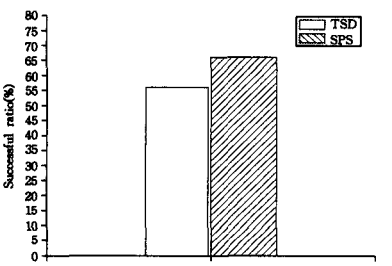

图 7 负载三作业成功率

结束语 针对不同作业类别，本文从提高集群资源利用率的角度出发，设计了一种基于作业类别和截止时间的调度算法。该调度算法包括两部分：1)将作业划分为 CPU 密集型作业和 I/O 密集型作业，并将其置于两种不同的队列中；2)根据作业执行截止时间，为作业设置不同的优先级，截止期限越近的作业优先级越高。该算法充分利用了集群的 CPU 和 I/O 资源，优于 Hadoop 默认调度算法。后续研究的重点在于对类型动态变化的作业调度的研究，如某些类型的作业可能早期使用 CPU 较多但后期则使用 I/O 较多，此种类型作业的调度尚有待进一步深入研究。

参 考 文 献

[1] White T. Hadoop 权威指南[M]. 周敏，译. 北京：清华大学出版社，2011：23-55
White T. Hadoop: The Definitive Guide[M]. Zhou Min. Beijing: Tsinghua University Press, 2011: 23-55

[2] Schwarakopf M, Konwinski A. Omega: flexible, scalable schedulers for large compute clusters[J]. EuroSys'13 Proceeding of the 8th ACM European Conference on Computer Systems. 2013: 351-364

[3] 范帆. Hadoop 中基于优先级的调度算法研究[D]. 上海：复旦大学，2012，8
Fan Fan. A Priority-basd Scheduling Algorithm for Hadoop[D]. Shanghai: Fudan Unversity, 2012, 8

[4] Tian Chao, Zhou Hao-jie, He Yong-qiang, et al. A Dynamic MapReduce Scheduler for Heterogeneous Workloads[C]// Eigth International Conference on Grid and Cooperative Computing (GCC '09). 2009: 218-224

[5] Teng Fei, Yang Hao, Li Tian-rui, et al. Scheduling real-time workflow on MapReduce-based cloud[C]// 2013 Third International Conference on Innovative Computing Technology. 2013: 117-122

[6] Kc K, Anyanwu K. Scheduling Hadoop Jobs to Meet Deadlines [C]// 2010 IEEE Second International Coference on Cloud Computing Technology and Science. 2010: 388-392

[7] Zhang Xiao-hong, Ju Shuai, Jiao Zhi-bin. A Scheduling Method Based on Deadlines in MapReduce[J]. Electrical, Information Engineering and Mechatronics 2011 Lecture Notes in Electrical Engineering, 2012, 138: 1585-1592

[8] Tang Zhuo, Zhou Jun-qing, Li Ken-li, et al. MTSD: A task scheduling algorithm for MapReduce base on deadline constraints[C]// 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum. 2012: 2012-2018

[9] Ning Wen-yu, Wu Qing-bo, Tan Yu-song. MapReduce oriented self-adaptive delay scheduling algorithm [J]. Computer Engineering & Science, 2013(3): 52-57

[10] 杨浩，滕飞，李天瑞，等. Hadoop 平台中空闲时间调度器的设计与实现[J]. 计算机工程与科学，2013(10)：125-131
Yang Hao, Teng Fei, Li Tian-rui, et al. Design and implementation of a least spare time scheduler for Hadoop[J]. Computer Engineering & Science, 2013(10): 125-131

[11] 陈国营. 基于 MapReduce 模型文本分类算法的研究[D]. 辽宁：辽宁大学，2013：1-10
Chen Guo-ying. Design and Implementation of Text Classification Algorithm Based on Hadoop[D]. Liaoning: Liaoning University, 2013: 1-10

[12] 韩定一. 云推荐—大数据时代的个性化互联网服务解决之道[J]. 程序员，2013(3)：16-17
Han Ding-yi. Cloud recommend—The road to solve personalized service on the era of big data[J]. Programmer, 2013(3): 16-17