ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.11, November 2010, pp.2814−2825
doi: 10.3724/SP.J.1001.2010.03683

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于贝叶斯网络的半监督聚类集成模型*

王红军[1,2+], 李志蜀[2], 戚建淮[1], 成 飏[2], 周 鹏[2], 周 维[2]

[1](西南交通大学 信息化研究院,四川 成都 610031)

[2](四川大学 计算机学院,四川 成都 610054)

Semi-Supervised Cluster Ensemble Model Based on Bayesian Network

WANG Hong-Jun[1,2+], LI Zhi-Shu[2], QI Jian-Huai[1], CHENG Yang[2], ZHOU Peng[2], ZHOU Wei[2]

[1](Information Research Institute, Southwest Jiaotong University, Chengdu 610031, China)

[2](School of Computer Science, Sichuan University, Chengdu 610054, China)

+ Corresponding author: E-mail: wanghongjun@cs.scu.edu.cn

Wang HJ, Li ZS, Qi JH, Cheng Y, Zhou P, Zhou W. Semi-Supervised cluster ensemble model based on Bayesian network. *Journal of Software*, 2010,21(11):2814−2825. http://www.jos.org.cn/1000-9825/3683.htm

Abstract: The existing algorithms are mostly unsupervised algorithms of a cluster ensemble, which cannot take advantages of known information of datasets. As a result, the precision, robustness, and stability of a cluster ensemble are degraded. To conquer these disadvantages, a semi-supervised cluster ensemble (SCE) model, based on both semi-supervised learning and ensemble learning technologies, is designed in this paper. There are three main works in this paper. The first is that SCE is proposed, and the variational inference oriented SCE is illustrated in detail. The second is based on the above work: an EM (expectation maximization) algorithm of SCE is illustrated step by step. The third is that some datasets are drawn from the UCI (University of California, Irvine) machine learning database for experiments which show that both SCE and its EM algorithm are good for semi-supervised cluster ensemble and outperforms NMFS (algorithm of nonnegative-matrix-factorization based semi-supervised), semi-supervised SVM (support vector machine), and LVCE (latent variable model for cluster ensemble). The Semi-Supervised Cluster Ensemble model is first stated in this paper, and this paper includes the advantages of both the semi-supervise learning and the cluster ensemble. Therefore, its result is better than the results of semi-learning clustering and cluster ensemble.

Key words: semi-supervised cluster ensemble; variational inference; must link; can not link

摘 要: 已有的聚类集算法基本上都是非监督聚类集成算法,这样不能利用已知信息,使得聚类集成的准确性、鲁棒性和稳定性降低.把半监督学习和聚类集成结合起来,设计半监督聚类集成模型来克服这些缺点.主要工作包括:第一,设计了基于贝叶斯网络的半监督聚类集成(semi-supervised cluster ensemble,简称SCE)模型,并对模型用变分法进行了推理求解;第二,在此基础上,给出了 EM(expectation maximization)框架下的具体算法;第三,从 UCI(University of California, Irvine)机器学习库中选取部分数据来做实验.实验结果表明,SCE 模型本身及其变分推理后所设计的

* Supported by the National Natural Science Foundation of China under Grant No.61003142 (国家自然科学基金); the Program of the Ministry of Railways of China under Grant Nos.2009X010-A, 2009X010-B (国家铁道部资助项目)

Received 2009-02-11; Revised 2009-04-27; Accepted 2009-07-09

EM 算法都能进行半监督聚类集成,总的来说,效果比 NMFS(algorithm of nonnegative-matrix-factorization based semi-supervised)、半监督 SVM(support vector machine)、LVCE(latent variable model for cluster ensemble)等算法要好.该半监督聚类集成模型聚集了半监督学习和聚类集成两者的优点,最后的聚类结果比单纯的半监督聚类或聚类集成的效果都要好.

关键词: 半监督聚类集成;变分推理;必连;不连

中图法分类号: TP181 **文献标识码**: A

聚类集成[1]的基本思想是,用若干独立的基聚类器分别对原始数据进行聚类,然后对这些基聚类器的结果进行组合,最终获得对原始数据的聚类结果.聚类集成使用了多个基聚类结果,可以分布式处理数据;同时,噪声和孤立点对结果的影响较小,增强了聚类结果的稳定性和鲁棒性.半监督学习是利用已知的一些数据信息,使建立的模型或者模型参数更准确的一种学习方式,能够增加模型对数据处理的准确性、稳定性和鲁棒性.其中有基于图正则化框架的半监督学习算法,其代表包括文献[2,3]等;还有以生成模型为分类器[4,5],采用 EM (expectation maximization)算法来进行模型参数估计的半监督学习.文献[6,7]使用两个或多个学习器,在学习过程中,挑选若干个置信度高的未标记示例进行相互标记,从而使模型得以更新.对于半监督聚类或者半监督聚类集成,有基于成对约束集的聚类,如必连(must-link)、不连(can-not-link)[8];也有使用训练集来训练得到概率模型的先验概率[4,5].

已有的聚类集成算法基本上都是非监督聚类集成算法,这样不能利用已知信息,使得聚类集成的准确性、鲁棒性和稳定性降低.如果在聚类集成中使用半监督学习技术,就可以克服这些缺点.虽然文献[9]中体现了半监督聚类集成的思想,文献[10]用集成学习的技术来选择半监督学习的一些参数,提高聚类效果,但它们都没有把半监督和聚类集成真正结合起来.本文设计半监督聚类集成模型来克服非监督聚类集成算法的缺点.半监督聚类集成模型是聚类集成和半监督学习两者优点的集中体现,使最终的聚类结果比单纯的半监督聚类或者聚类集成的效果都要好,以便更好地解决实际问题.

本文针对这些问题,设计了半监督聚类集成(semi-supervised cluster ensemble,简称 SCE)模型,使模型本身能够进行半监督聚类集成.并且采用变分法推理,在 EM 框架下设计新算法,也使算法可以进行半监督聚类集成.实验中既可以采用小部分标记数据训练来得到模型的初始化参数,又可以对数据点进行专家知识和先验知识矫正,使模型更具稳定性、鲁棒性和准确性.

1 聚类集成概述

1.1 聚类集成问题

聚类集成主要分为两个阶段.第 1 个阶段是基聚类器对原始数据进行聚类,得到基聚类结果.在这个阶段主要是用一些成熟的算法对原数据聚类,重复 m 次,并且每次使用不同的初始化得到对原数据的 m 个有差别的聚类结果,也可以采用几种算法得到这 m 个结果.假设有 n 个原数据对象 $OX=\{ox_1,ox_2,\ldots,ox_n\}$和 m 个基聚类器 $BC=\{bc_1,bc_2,\ldots,bc_m\}$,那么基聚类结果的数据表示为

$$x_i=\{x_{i,j}\}=\{bc_j(ox_i)|i\in(1,\ldots,n);j\in(1,\ldots,m)\} \quad (1)$$

第 2 个阶段是基聚类结果集成,根据聚类集成算法对前一个阶段采集的基聚类结果进行处理,使之能够最大限度地分享这些结果,从而得到一个对原始数据最好的聚类集成结果.m 个基聚类结果中不可能完全聚类正确,原因在于每一个基聚类结果的数据中都包含噪声;如果基聚类结果允许丢失数据出现,那么这些基聚类结果数据中还存在丢失的数据标签.这两者都加大了聚类集成的难度.然后,通过一个聚类集成函数或者算法模型对这 m 个基聚类结果处理,从而得到最终的聚类集成结果.如果聚类集成函数为Γ,那么第 2 个阶段可以表示为

$$C^*=\Gamma(x_i) \quad (2)$$

其中,C^*是聚类集成结果.一般来说,最终的聚类结果是在m个基聚类结果中最好的.聚类集成的关键也是寻找聚

类集成函数.

1.2 相关工作

美国 Texas 大学计算机学院的 Strehl[11]最早明确提出了聚类集成问题以及解决方法.他认为,要最大限度地分享 m 个基聚类结果,可以通过计算 m 个基聚类结果的相关信息和它们之间的信息熵来度量其信息熵和相关信息之比,而获得最好的结果[11].周志华等学者提出,解决聚类集成问题可以通过投票机制,根据投票的结果来得到一个聚类集成结果[12−14].也有主要基于图分割[11,15,16]的方法.这类方法的一般过程是,先把基聚类结果转换成图的顶点和边,或者超图的顶点和超边,然后再基于最小切或者最小权重的方法开始切割图,最后切割成顶点和边不交叉的几个子图,而每一个子图表示一个类别.也有主要基于矩阵运算的算法[9].这类方法是先把基聚类结果转化为矩阵,然后再进行矩阵变换得到数据点的相似性,最后按其相似性聚类.文献[9]提到了半监督聚类集成,但其主要目的不是研究半监督聚类集成.还有基于概率模型的[17−19]的方法.这类方法主要是先求基聚类结果数据在统计上的特征,基聚类的权重与其置信度成正比[18].文献[17]就是每一个基聚类结果服从一个多项式分布,而聚类集成结果就是这些多项式分布的集成结果.

这些算法的共同点是,都是非监督聚类集成算法,因此这些算法不能利用已知信息,使得聚类集成的准确性、鲁棒性和稳定性降低.本文设计半监督聚类集成模型来克服这些缺点.并且,此模型可以直接和间接地进行半监督聚类集成,有效地利用各数据的先验知识和专家知识,从而提高集成的准确性和鲁棒性.

2 半监督聚类集成模型(SCE)

2.1 SCE介绍

Semi-supervised Cluster Ensemble(SCE)建立在概率理论之上,如图 1 所示.它实际上是一个贝叶斯网,主要阐述基聚类结果与聚类集成结果变量之间的依赖关系.本模型假定每一个基聚类的结果服从一个有先验概率的多项式分布,模型中的符号定义:$x_{i,j}$ 为第 j 个基聚类器对第 i 个原数据对象的聚类的类别索引,是聚类集成中的原子元素,其中,i,j 范围为 $\{i_1^n, j_1^m\}$,在 SCE 模型中是观察数据,为图 1 中的 x.x_i 表示一个向量,其中元素为 $x_{i,(\cdot)}$,也就是 m 个基聚类器对第 i 个原数据对象各自的聚类结果;$c_{i,j}$ 表示对 $x_{i,j}$ 的聚类集成结果,$c_{i,j}$ 可以重复取值,在 SCE 模型中 C 是隐含变量,为图 1 中的 c;θ_i 表示对 x_i 的聚类集成结果,当 i 取不同值时,θ_l 可能等于 θ_k,这说明第 I 个和第 k 个原数据对象属于同一个类别,为图 1 中的 θ;α和β都是预先设定模型参数,其中,α是狄利克雷(Dirichlet)分布的参数,β是多项式分布参数;由于β参数与聚类集成结果的类别数量成线性关系,为了解决这个问题,设ϕ是对β参数的分布,因此在实际计算中,β用ϕ的分布代替.L 是先验知识和监督学习条件参数.此模型有两个假设:一个是假使α的维数已知并且固定不变,也就是聚类集成的类别的数量已知和类别数量在特定数据集是固定的;另一个是假设 $x_{i,j}$ 是由 $c_{i,j}$ 和一个参数β决定的.也就是说,$x_{i,j}$ 依赖 $c_{i,j}$ 和β.其现实意义就是基聚类数据可以通过某个参数转化得到聚类集成结果.

SCE 的模型主要有 3 个作用:其一,模拟产生基聚类器结果;其二,如果知道基聚类器的结果,则可以推导出隐含变量 C 和θ的值,也就可以知道聚类集成的结果;其三,如果知道一些先验知识和专家知识,就可以对模型的参数进行较好的设置,并且可以直接输入模型中进行半监督学习,使聚类集成的结果非常好.

SCE 模型与 Latent Variable Model for Cluster Ensemble(LVCE)[20]和 Latent Dirichlet Allocation(LDA)[21]两个模型都有一些区别和联系,接下来我们重点分析它们的区别.

SCE 模型与 LVCE 属于同一系列模型(LVCE 模型的图表示如图 2 所示),但两者有 3 个方面的差别:(1) SCE 和 LVCE 的图表示中,SCE 比 LVCE 多一个变量并且多一条依赖边.这样,SCE 模型本身可直接进行半监督学习,而 LVCE 不能;(2) LVCE 使用的是 MCMC 推理算法,而 SCE 使用的是变分法推理并使用 EM 框架的算法.也就是说,从推理后所设计的算法上讲,SCE 的 EM 算法也可以进行半监督学习,而 LVCE 的 MCMC 推理算法则不能.由于 SCE 可直接进行半监督学习,使用变分法推理并使用 EM 框架的算法也可以进行半监督学习,这样就容易求出全局最优解.即使 LVCE 使用变分推理并设计算法,由于只有算法能进行半监督学习,这样就容易陷入局部

最优解;(3) SCE 在使用变分法推理后应用 EM 算法框架,使其算法复杂度比 LVCE 的 MCMC 算法复杂度要低.使用变分法推理应用 EM 算法框架所设计的算法的复杂度为 $o((t+1)n)$,其中,t 是指算法迭代次数,n 是数据对象的数量.本算法中 t 的值一般介于 10~100 之间;而 LVCE 使用 MCMC 算法的复杂度为 $o((tk+1)n)$,其中,t 和 n 与前者意义相同,k 是数据对象类别的数量.本算法中 t 的值一般介于 200~400 之间.

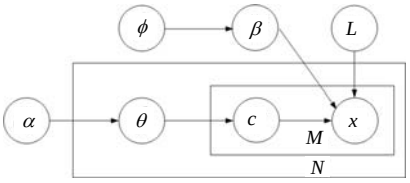

Fig.1 Graphic model of semi-supervised cluster ensemble

图 1 半监督聚类集成的图模型

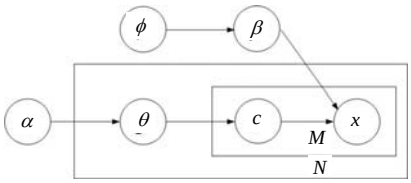

Fig.2 Graphic model of latent variable cluster ensemble

图 2 LVCE 的图模型

SCE 模型与 LDA[21]的区别可以先从 LDA 的缺陷来阐述,LDA 模型在理论上存在着 3 个缺陷:一是只对离散数据起作用,不能用于实数;二是假使每一个数据列是同一分布;三是在算法中,所有列数据(列数据这里指数据集的每一个属性的所有取值)的后验概率也是统一更新.SCE 和 LDA 的区别为:首先是 SCE 与 LDA 有 3 个本质的区别:(1) LDA 只能应用于离散数据,其数据分布假设为多项式分布,而 SCE 模型本身可以把这个假设扩展到指数族分布(如高斯分布、泊松分布、多项式分布等),因此,SCE 模型本身就可以对实数类的数据进行处理.在本文中,由于主要讨论半监督聚类集成,故未讨论 SCE 模型对实数类数据的处理;(2) LDA 模型假设所有的列数据都服从同一参数的多项式分布,而 SCE 模型本身可以假设不同列的数据服从不同指数族的分布,主要由 SCE 模型中∅决定.即使同属于多项式分布,也假设这个分布的参数是不同的,SCE 就更灵活;(3) LDA 对于所有的列数据都假设服从同一参数的多项式分布,并且在计算各数据分布的后验概率时,也不区分列之间的差异性.而 SCE 在更新各列数据的后验概率时,是各自根据自己列的数据进行更新,这样使计算的结果更准确.其次,除了 SCE 和 LDA 的本质区别之外,SCE 比 LDA 多两个变量和两条依赖边,并且各变量的意义和 LDA 所表示的意义完全不同.LDA 模型也有一些改进的半监督学习模型[22,23],这些模型都是基于 LDA 改进的且应用于文本的分类,也继承了 LDA 的特点.这些模型与 SCE 都存在着前 3 个本质的区别.在聚类集成中,所有的基聚类的结果不是服从同一个先验概率分布,并且其后验概率也是分别更新,这样就决定了 LDA 及其相关半监督算法如果用于聚类集成算法,其假使条件就是错误的,因为每个基聚类的结果的分布是不同的.在 SCE 模型中,即使在很特殊的情况下,假使所有的基聚类的结果服从同一个先验概率分布,也是和 LDA 完全不同的.因为即使假使先验概率相同,但其分别更新,后验概率也是不同的.

2.2 模型模拟基聚类结果生成

图 1 是 SCE 的各变量之间的依赖关系,SCE 是一个生成模型,是对 $x_{i,j}$ 如何生成过程的模拟.如果要产生一个 $x_{i,j}$,则首先根据 α 的分布,抽样一个分布 θ_i;根据 θ_i 对真正类别 $c_{i,j}$ 的先验概率和参数 β 抽样出一个 $x_{i,j}$.假设对于每一个观测数据 $x_{i,j}$,对于聚类集成来说,就是对每一个基聚类的结果.

(1) 首先根据 $Dirichlet(\alpha)$ 分布参数选择一个聚类集成结果 θ_i;

(2) 根据 θ_i 对真正类别 $c_{i,j}$ 的先验概率和多项式分布参数 β;

(3) 根据专家知识和先验知识 L;

(4) 抽样出 $x_{i,j}$.

如果假设 α,β,L 已知,那么 x,θ,c 的联合分布密度函数为

$$p(X,C,\theta\mid\alpha,\beta,L)=p(\theta\mid\alpha)\prod_{i=1,j=1}^{i=n,j=m}\sum_{c_{i,j}}p(c_{i,j}\mid\theta)p(x_{i,j}\mid c_{i,j},\beta,L) \tag{3}$$

于是,对于每一个数据点,其边缘概率是

$$p(x\mid\alpha,\beta,C,L)=\int_{\theta}P(\theta\mid\alpha)\prod_{j=1}^{M}\sum_{z_j}p(z_j\mid\theta)p(x_j\mid\beta,z_j,c_j,L)\mathrm{d}\theta \tag{4}$$

必连(must-link)[8]是指几个数据点属于同一类,假如 H 个数据点$\{x_1,x_2,\ldots,x_H\}$属于同一类,那么对于这些数据点的边缘概率,模型为

$$p(x_i\mid\alpha,\beta,C,L)=\frac{1}{H}\sum_{1}^{H}\left(\int_{\theta}P(\theta\mid\alpha)\prod_{j=1}^{M}\sum_{z_j}p(z_j\mid\theta)p(x_{i,j}\mid\beta,z_j,c_j,L)\mathrm{d}\theta\right)\{i_1^H\} \tag{5}$$

不连(can-not-link)[8]是指几个数据点不属于同一类,一般情况是指两个数据点.假如 H 个数据点$\{x_1,x_2,\ldots,x_H\}$不属于同一类,对 H 的最好取值是 2,并且可以最好地体现模型的意义.那么对这两个数据点的边缘概率,模型为

$$p(x_i\mid\alpha,\beta,C,L)=\frac{\left(\int_{\theta}P(\theta\mid\alpha)\prod_{j=1}^{M}\sum_{z_j}p(z_j\mid\theta)p(x_{i_j}\mid\beta,z_j,c_j,L)\mathrm{d}\theta\right)^2}{\sum_{i=1}^{H}\left(\int_{\theta}P(\theta\mid\alpha)\prod_{j=1}^{M}\sum_{z_j}p(z_j\mid\theta)p(x_{i_j}\mid\beta,z_j,c_j,L)\mathrm{d}\theta\right)^2}\{i_1^H\} \tag{6}$$

在聚类集成中,基聚类结果是已知的,所以主要是对 $c_{i,j}$ 感兴趣,而其又很难直接求出,但可以通过近似算法得到.这些近似算法有变分法、Laplace 近似法、MCMC 等.

3 SCE 的变分法推理及 EM 算法

3.1 SCE的变分法推理过程

对于公式(4)~公式(6),最主要的是推理求出公式(4).而公式(5)、公式(6)则是公式(4)的变种,只是算法不同.对公式(4)的求解,可以根据公式(3)进行.在聚类集成中,对于给定的基聚类结果,最主要的是计算两个潜在变量θ和 c 的后验概率.按照 Bayes 规则,

$$p(c,\theta\mid x,\alpha,\beta)=\frac{p(\theta,c,x\mid\alpha,\beta)}{p(x\mid\alpha,\beta)} \tag{7}$$

公式(7)很难求解,所以我们假设用另外一个分布来替代 $p(X,C,\theta|\alpha,\beta)$.假设这个分布为

$$q(C,\theta\mid\gamma,\eta)=q(\theta\mid\alpha)\prod_{i=1,j=1}^{i=n,j=m}q(c_{i,j}\mid\eta_{i,j}) \tag{8}$$

这个分布与原来的分布的接近程度越高,求解的精度就越高.因此,通过优化来选择一个这样的分布:

$$\begin{aligned}\log p(x\mid\alpha,\beta)&=\log\int\sum_{C}p(\theta,c,x\mid\alpha,\beta)\mathrm{d}\theta\\&=\log\int\sum_{C}\frac{p(\theta,c,x\mid\alpha,\beta)q(\theta,c)}{q(\theta,c)}\mathrm{d}\theta\\&\geqslant\int\sum_{C}q(\theta,c)\log p(\theta,c,x\mid\alpha,\beta)\mathrm{d}\theta-\int\sum_{C}q(\theta,c)\log q(\theta,c)\mathrm{d}\theta\\&=E_q[\log p(\theta,c,x\mid\alpha,\beta)]-E_q[\log q(\theta,c)]\end{aligned} \tag{9}$$

这样,就得到了这个分布的一个下界(lower bound),于是,

$$\log p(x|\alpha,\beta)=L(\alpha,\beta:\gamma,\eta)+D(p(X,C,\theta|\alpha,\beta)\|q(C,\theta|\gamma,\eta)) \tag{10}$$

如果要使 $L(\alpha,\beta:\gamma,\eta)$最大,那么只需要优化:

$$(\gamma^*,\eta^*)=\mathrm{argmin}_{\gamma,\eta}(D(p(X,C,\theta|\alpha,\beta)\|q(C,\theta|\gamma,\eta))) \tag{11}$$

这两个最优的变分参数可以通过求变分分布和真实分布之间的 Kullback-Leibler(KL)差得到.只需对 $D(p(X,C,\theta|\alpha,\beta)\|q(C,\theta|\gamma,\eta))$求导并且使其为 0,就可以得到:

$$\eta_{i,j} \propto \beta_{i,j} \exp\left\{\Psi(\gamma_{i,j}) - \Psi\left(\sum_{i=1,j=1}^{i=m,j=n} \gamma_{i,j}\right)\right\} \tag{12}$$

$$\gamma_{i,j} = \alpha_{i,j} + \sum_{i=1,j=1}^{i=m,j=n} \eta_{i,j} \tag{13}$$

对于必连(must-link)的数据点,假如 H 个数据点$\{x_1,x_2,\ldots,x_H\}$是属于同一类,那么可以用下式进行计算:

$$\eta_{i,j} \propto \frac{1}{H}\sum_{i=1}^{i=H}\left\{\beta_{i,j} \exp\left\{\Psi(\gamma_{i,j}) - \Psi\left(\sum_{i=1,j=1}^{i=m,j=n} \gamma_{i,j}\right)\right\}\right\} \tag{14}$$

对于不连(can-not-link)的数据点,假如 H 个数据点$\{x_1,x_2,\ldots,x_H\}$不属于同一类,那么其变化为

$$\eta_{i,j} \propto \frac{\left\{\beta_{i,j} \exp\left\{\Psi(\gamma_{i,j}) - \Psi\left(\sum_{i=1,j=1}^{i=m,j=n} \gamma_{i,j}\right)\right\}\right\}^2}{\sum_{i=1}^{i=H}\left\{\beta_{i,j} \exp\left\{\Psi(\gamma_{i,j}) - \Psi\left(\sum_{i=1,j=1}^{i=m,j=n} \gamma_{i,j}\right)\right\}\right\}^2} \tag{15}$$

如果给定基聚类的结果,则希望找到最好的参数α和β,使这些数据的 H 对数似然 H 估计最大:

$$L(\alpha,\beta) = \sum_{i=1}^{i=m} \log p(x_{(i,*)}\alpha,\beta) \tag{16}$$

根据公式(12)、公式(13),就可以设计 EM 算法来估计这些参数:

$$\beta_{i,j} \propto \sum_{i=1,j=1}^{i=m,j=n} \eta_{i,j} x_{i,j} \tag{17}$$

如果已知$\gamma_{i,j}$,则$\alpha_{i,j}$需要通过牛顿-拉斐逊迭代算法得到.

3.2 SCE的EM算法

SCE 的 EM 算法:

循环开始(条件为α,β,γ,η维数固定,并且在算法中维数不再改变):

E 步骤:输入参数α,β和基聚类结果 x.根据公式(12)~公式(15)找到每一个$\gamma_{i,j},\eta_{i,j}$,这样实际上是找到了一个变分分布的下界值.对于第 1 次循环,α,β的值为初始值.

M 步骤:输入$\gamma_{i,j},\eta_{i,j}$和 x,根据公式(17)可以更新β,根据牛顿-拉斐逊迭代算法得到$\alpha_{i,j}$.更新是为了找到对 x 最大 H对数似然H 估计.

循环结束.

Algorithm 1.

Input: $\{x,\alpha,\beta,constrain\{set\}\}$.

Output: $\{\eta,\gamma\}$.

```
Initialize α, β randomly;
Begin loop until η, γ is stable;
E-Step:
  If x_i is must-link
    (a) Calculate η, γ according to the formulae (13) and formulae (14);
    Else if x_i is can-not-link
    (b) Calculate η, γ according to the formulae (13) and formulae (15);
    Else
    (c) Calculate η, γ according to the formulae (12) and formulae (13);
    (d) Submit the η, γ to M-step;
M-Step:
  (a) Update α, β according to the formula (17) and Newton-Raphson methods;
  (b) Submit the α, β to E-step;
```

14. End Loop.

此算法中,*constrain*{*set*}(数据点不连和必连的成对集合)可以为空.在数据集较小的情况下,本算法的结果与先验概率也即参数α,β有关,会陷入局部最优值的情况.如果参数较好,或者 *constrain*{*set*}较大,一般会达到全局最优值.针对这种情况,本算法很适合作半监督学习中的算法.先采用一部分数据作为训练数据来确定整个数据集的初始化参数.对于较大的数据集,参数α,β对结果的影响不大.

本算法可以得到 x_i 属于各类别的概率.算法 1 的复杂度为 $o((t+1)n)$,其中,n 是数据对象的数量,t 为收敛的迭代次数.本算法中,t 的值一般介于 10~100 之间.

根据文献[24,25],如果数据是指数族的分布形式,那么 EM 算法收敛.本算法首先属于 EM 算法框架,其次,所有的数据属性在模型中被假设为指数族分布,并且基聚类结果实际上是多项分布,是指数分布的一种形式.所以,本算法是收敛的.

4 实 验

4.1 数据集和评价标准

本文使用 UCI(University of California, Irvine)的机器学习库中的部分数据集作为实验数据集.表 1 中列出了这些数据的样本、属性和类别数量.

Table 1 Number of instances, features and classes of datasets

表 1 实验数据集的样本、属性和类别数量

Dataset	Instances	Features	Classes
Pima	768	8	2
Glass	214	9	6
Iris	150	4	3
Ionosphere	351	34	2
Wdbc	569	30	2
Bupa	345	6	2
Wine	178	13	3
Magic04	19 020	10	2
Balance	625	4	3
Segmentation	2 100	19	7

聚类中有很多评价标准,本文选用 Micro-precision[14,26]标准.这个标准实际上是聚类正确率的一种表示方法.这个标准的计算公式如下:

$$MP=\frac{1}{N}\sum_{h=1}^{h=K}a_h.$$

其中,a_h 表示对数据某一类分类正确的数量,N 表示数据集中数据对象的数量,K 表示此数据集合中的类别的数量.为了更好地衡量聚类的正确率,我们要进行重复的实验,采用平均正确率来衡量聚类将更准确.所以,使用下面的公式来计算:

$$AMP=\frac{1}{T\times N}\sum_{t=1}^{T}\sum_{h=1}^{h=K}a_h.$$

其中,T 为重复实验的次数,本文 T 在基聚类中选择为 20,在聚类集成中选择为 50.

4.2 实验步骤和实验结果

本节首先介绍综合实验,选取了标准 *K*-means,LVCE[20],NMFS(algorithm of nonnegative-matrix-factorization based semi-supervised)[9],半监督 SVM(support vector machine)[27]和 SCE 等算法来做实验.*K*-means 只用于对原始数据聚类,也就是作为基聚类器算法;半监督 SVM 用于对原始数据的半监督聚类;LVCE 主要是针对聚类集成实验,NMFS 和 SCE 作为半监督聚类集成算法.如果是半监督聚类或者聚类集成实验,我们选取原数据的 10%作为训练集,或者选择 10%的数据作为成队约束集来进行实验.我们把这些实验的结果记录在表 2 中,如

“0.575±0.041”表示平均正确率是 0.575,标准差是 0.041.为了更好地观察这些结果,我们把这些结果也做了配对 T 检验,在做 T 检验时,我们使用传统的假设,如果这两组值的 P 值小于 0.05,那么这两组值是存在着差异的.这里,也把 P 值也记录在表 2 中,如“S-LVCE p-values”表示是 SCE 和 LVCE 两种算法的结果来做配对 T 检验.

Table 2 Results of the algorithms and pair-wised test

表 2 各算法在这些数据集上的运行结果以及配对 T 检验结果

Dataset	Algorithm							
	K-means	LVCE	NMFS	SVM	SCE	S-LVCE *p*-value	S-NMFS *p*-value	S-SVM *p*-value
Pima	0.575±0.041	0.661±0.103	0.684±0.041	0.665±0.098	**0.691±0.004**	0.008	0.041	0.023
Glass	0.514±0.020	0.553±0.018	0.600±0.002	0.550±0.086	**0.608±0.013**	0.021	0.030	0.041
Iris	0.864±0.058	0.900±0.187	0.920±0.010	0.890±0.100	**0.940±0.064**	0.010	0.000	0.008
Ionosphere	0.624±0.023	0.702±0.044	0.743±0.007	0.695±0.106	**0.765±0.042**	0.022	0.031	0.042
Wdbc	0.761±0.011	0.884±0.051	0.884±0.002	0.884±0.075	**0.889±0.025**	0.038	0.032	**0.612**
Bupa	0.462±0.029	0.571±0.030	0.578±0.050	0.565±0.082	**0.583±0.018**	0.008	0.004	0.041
Wine	0.590±0.034	0.721±0.071	**0.741±0.000**	0.722±0.058	0.724±0.008	0.000	0.040	0.050
Magic04	0.625±0.028	0.649±0.018	0.649±0.002	0.649±0.023	**0.651±0.000**	**0.242**	**0.462**	**0.142**
Balance	0.511±0.031	0.529±0.089	**0.601±0.011**	0.518±0.011	0.596±0.009	**0.052**	0.025	0.041
Segmentation	0.557±0.022	0.585±0.004	0.610±0.014	0.579±0.029	**0.630±0.017**	0.006	0.001	0.012

在表 2 中,我们把关键值都突出地表示出来.在这 10 个数据集上,SCE 有 8 次获得了最好的平均正确值;另外 2 个数据集上,NMFS 取得了最大的平均正确率.在 30 个 P 值中,只有 5 个值大于 0.05,说明各算法结果的差异性的确是存在的.在这 5 个值中,有 3 个是针对同一个数据集 Magic04 的,说明这个数据集的类别分界线比较明显,用各算法对它都取得很接近的结果.所以,表 2 总体上体现了以下两个结论:

(1) 半监督聚类集成比单一的半监督聚类或者聚类集成的结果都要好,说明了把半监督学习和集成学习结合起来的优越性;

(2) 总的来说,SCE 比 NMFS 的正确率要高,更适宜各类数据.这是由于 SCE 模型本身就可以进行半监督聚类集成,可以通过训练集确定模型的先验概率,同时也可以把专家知识集输入模型中进行计算.

本节还将介绍另外两类实验:第 1 类是采用每组数据集的部分数据作为训练集(训练集不断增加),根据成对约束集 *constrain*{*set*}(数据点不连和必连的成对集合)为空的情况来确定算法 1 的初始化参数 α 和 β,进行半监督聚类集成;第 2 类是随机初始化参数 α 和 β,而成对约束集(*constrain*{*set*})的数量逐步增加,进行半监督聚类集成.关于聚类集成中基聚类的算法,本文选择标准的 *K*-means.聚类集成算法本文选用 NMFS[9]和本文的 SCE 模型算法.在本实验中,假设实验数据的聚类的类别数量为表 1 中数据的类别数量,是已知和固定不变的.在 SCE 模型算法中,这个假设用算法中的 *Dirichlet*(α)分布的参数 α 的维数来控制.

第 1 类实验的第 1 步是使用 *K*-means 算法来对原始数据集聚类,得到基聚类结果.在这步实验中,基聚类的数量选择为 20,如果数据集合中的对象的数量为 N,那么当 *K*-means 在不同初始化条件下,就可以得到一个 $N\times20$ 的基聚类结果矩阵.实验第 2 步是使用 NMFS 和 SCE 作为聚类集成算法,用第 1 步得到的基聚类结果矩阵作为这两种算法的输入数据集,并且随机选用基聚类结果矩阵的部分数据作为训练集来确定算法的参数.其中,NMFS 和 SCE 在使用基聚类结果矩阵的不同比例的数据作为训练集,最后的结果如图 3 所示.图 3 的横轴是选取基聚类结果矩阵作为训练集的百分比,而纵轴是在相应训练集下的 AMP 值.从这 10 个标准数据来看,SCE 在 8 个数据集上的正确率明显比 NMFS 要高;但在另外 2 个数据集上,NMFS 算法的正确率比 SCE 要高.总的来说,作为半监督聚类集成算法,SCE 要优于 NMFS 算法.

第 2 类实验的第 1 步若和第 1 类实验的第 1 步完全一样,就可以得到一个 $N\times20$ 的基聚类结果矩阵.实验的第 2 步也是使用 NMFS 和 SCE 作为聚类集成算法,在这一步中,使用专家知识来初始化算法 1 中的成对约束集(*constrain*{*set*}).其中,NMFS 和 SCE 在不同数量的 *constrain*{*set*}集合下得到的最后结果如图 4 所示.图 4 的横轴是 *constrain*{*set*}集合所包含的成对约束集的数量,纵轴是在相应训练集下的 AMP 值.从图中可知,在 8 个数据集上,SCE 要明显优于 NMFS 算法;在另外的 2 个数据集上,SCE 在成队约束集合较小的情况下,NMFS 的效果较好.但随着成队约束集的增加,SCE 的效果明显增加,最后在成队约束集的数量大于 50 以后,SCE 的效果就明

显地优于 NMFS 算法.总的来说,SCE 的效果比 NMFS 的效果要好.

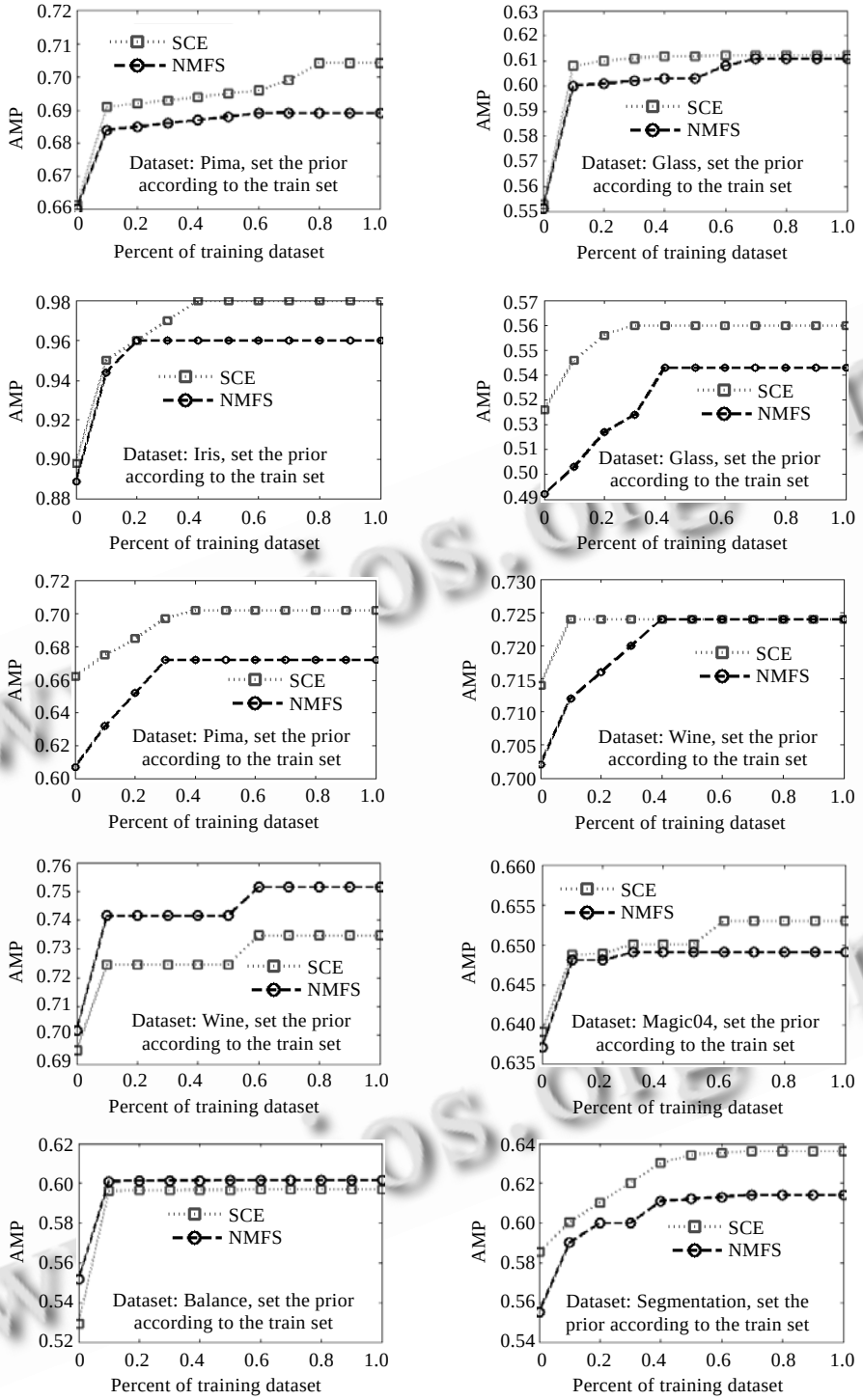

Fig.3 Results of algorithms with increase the training dataset

图 3 训练集按百分比增加时,各聚类集成算法的实验结果

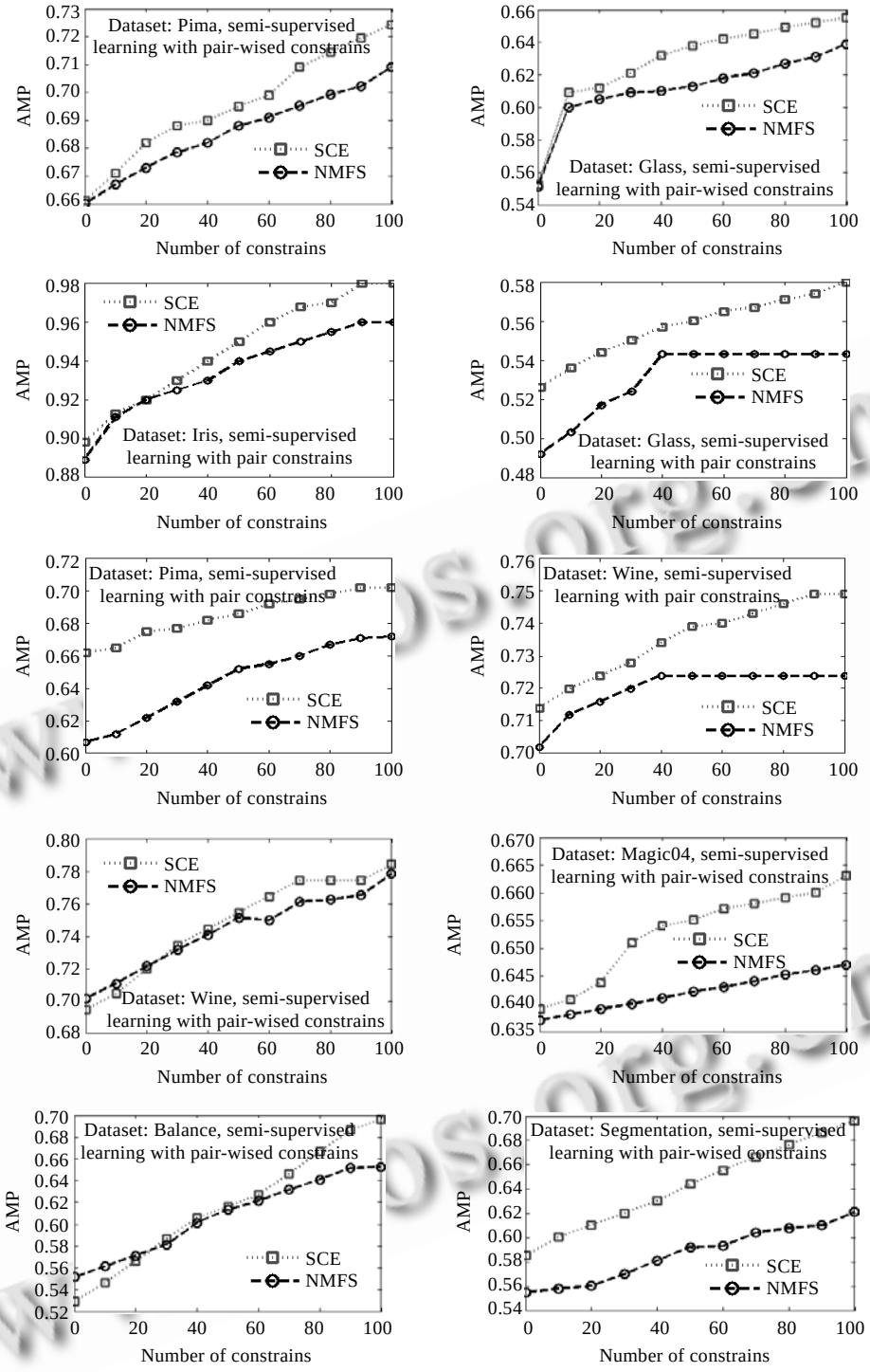

Fig.4 Results of algorithms with increase pair constrains

图 4 成队约束集增加时,各聚类集成算法的实验结果

5 结束语

本文针对目前的非监督聚类集成的缺点,即不能利用已知信息,使得聚类集成的准确性、鲁棒性和稳定性降低,提出半监督聚类集成概念,把半监督学习和集成学习结合起来,设计半监督聚类集成模型 SCE 来克服这些缺点.并且对 SCE 模型用变分法进行了推理求解,设计算法.然后用 UCI 的标准数据对算法进行检验,实验结果也说明了半监督聚类集成体现半监督聚类和聚类集成两者的优点.对于实际的应用,还涉及到半监督聚类集成的分布式和并行式的问题.在今后的工作中,我们将对其进行探讨.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是美国明尼苏达州立大学(University of Minnesota)计算机科学技术系的 Banerjee 教授和 Shan 博士研究生表示感谢.

References:

[1] Tang W, Zhou ZH. Bagging-Based selective clusterer ensemble. Journal of Software, 2005,16(4):496−502 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/496.htm

[2] Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. In: Thrun S, Saul L, Schölkopf B, eds. Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2004. 321−328.

[3] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. In: Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). 2001. 19−26.

[4] Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Machine Learning, 2000,39(2-3):103−134. [doi: 10.1023/A:1007692713085]

[5] Miller DJ, Uyar HS. A mixture of experts classifier with learning based on both labelled and unlabelled data. In: Mozer M, Jordan MI, Petsche T, eds. Advances in Neural Information Processing Systems 9. Cambridge: MIT Press, 1997. 571−577.

[6] Zhou ZH, Li M. Semi-Supervised regression with co-training. In: Proc. of the 19th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2005). 2005. 908−913.

[7] Zhou ZH, Li M. Semi-Supervised regression with co-training style algorithms. IEEE Trans. on Knowledge and Data Engineering, 2007,19(11):1479−1493. [doi: 10.1109/TKDE.2007.190644]

[8] Wagstaff K, Cardie C, Rogers S, Schroedl S. Constrained *K*-means clustering with background knowledge. In: Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). 2001. 577−584.

[9] Li T, Ding C, Jordan MI. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In: Proc. of the 7th IEEE Int'l Conf. on Data Minning. IEEE Computer Society, 2007. 577−582.

[10] Choi I, Shin H. Semi-Supervised learning with ensemble learning and graph sharpening. In: Proc. of the Intelligent Data Engineering and Automated Learning—IDEAL 2008. 2008. 172−179.

[11] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 2002,3:583−617. [doi: 10.1162/153244303321897735]

[12] Nguyen N, Caruana R. Consensus clusterings. In: Proc. of the 7th IEEE Int'l Conf. on Data Mining. IEEE Computer Society, 2007. 1−34. http://www.ist.unomaha.edu/icdm2007/papers/papers.php

[13] Windeatt T. Vote counting measures for ensemble classifiers. Pattern Recognition, 2003,12(36):2743−2756.

[14] Zhou ZH, Tang W. Clusterer ensemble. Knowledge-Based Systems, 2006,19(1):77−83. [doi: 10.1016/j.knosys.2005.11.003]

[15] Asur S, Parthasarathy S, Ucar D. An ensemble approach for clustering scale-free graphs. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2006. 1−8. http://kt.ijs.si/Dunja/LinkKDD2006/Papers/asur.pdf

[16] Ludmila IK, Stefan TH. Solving cluster ensemble problems by bipartite graph partitioning. In: Proc. of the 21st Int'l Conf. on Machine Learning. 2004. 281−288.

[17] Topchy A, Jain AK, Punch W. A mixture model for clustering ensembles. In: Proc. of the 4th SIAM Int'l Conf. on Data Mining. 2004. 22−24.

[18] Muna AR, Domeniconi C. Weighted cluster ensemble. In: Proc. of the Society for Industry and Applied Mathematics Conf. on Data Mining. 2006. 258−269.

[19] Topchy A, Behrouz MB, Anil KJ, William FP. Adaptive clustering ensembles. In: Proc. of the 17th Int'l Conf. on Pattern Recognition (ICPR 2004). 2004. 272−275.

[20] Wang HJ, Li ZS, Cheng Y, Zhou P, Zhou W. A latent variable model for cluster ensemble. Journal of Software, 2009,20(4): 825−833 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3431.htm [doi: 10.3724/SP.J.1001.2009.03431]

[21] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022. [doi: 10.1162/jmlr.2003.3.4-5.993]

[22] Kristina T, Mark J. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In: Proc. of the Neural Information Processing Systems (NIPS 2007). 2007. Http://books.nips.cc/papers/files/nips20/NIPS2007_0964.pdf

[23] Li WB, Sun L, Zhang DK. Text classification based on labeled-LDA model. Chinese Journal of Computers, 2008,31(4):620−627 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2008.00620]

[24] Boyles RA. On the convergence of the EM algorithm. Journal of the Royal Statistical Society, Series B—Methodological, 1983, 45(1):47−50.

[25] Xu L, Jordan MI. On convergence properties of the EM algorithm for Gaussian mixtures. Neural Computer, 1996,8:129−151. [doi: 10.1162/neco.1996.8.1.129]

[26] Modha D, Spangler WS. Feature weighting in *k*-means clustering. Machine Learning, 2003,52(3):217−237. [doi: 10.1023/A:1024016609528]

[27] Bennett KP, Demiriz A. Semi-Supervised support vector machines. In: Proc. of the Neural Information Processing Systems. 1998.

附中文参考文献:

[1] 唐伟,周志华.基于 Bagging 的选择性聚类集成.软件学报,2005,16(4):496−502. http://www.jos.org.cn/1000-9825/16/496.htm

[20] 王红军,李志蜀,成扬,周鹏,周维.基于隐含变量的聚类集成模型.软件学报,2009,20(4):825−833. http://www.jos.org.cn/1000-9825/3431.htm [doi: 10.3724/SP.J.1001.2009.03431]

[23] 李文波,孙乐,张大鲲.基于 Labeled-LDA 模型的文本分类新算法.计算机学报,2008,31(4):620−627.

王红军(1977－),男,四川广安人,博士,主要研究领域为计算机网络信息处理,机器学习,人工智能.

成飚(1979－),男,博士生,主要研究领域为网络与信息系统,智能计算.

李志蜀(1946－),男,教授,博士生导师,主要研究领域为计算机网络信息处理,机器学习,人工智能,Web 数据集成.

周鹏(1975－),男,博士生,主要研究领域为网络与信息系统,智能计算.

戚建淮(1965－),男,博士,高级工程师,主要研究领域为信息安全,人工智能.

周维(1975－),男,博士生,主要研究领域为网络与信息系统,智能计算.