ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.4, April 2009, pp.825−833
doi: 10.3724/SP.J.1001.2009.03431

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

基于隐含变量的聚类集成模型*

王红军+, 李志蜀, 成 飏, 周 鹏, 周 维

(四川大学 计算机学院,四川 成都 610054)

A Latent Variable Model for Cluster Ensemble

WANG Hong-Jun+, LI Zhi-Shu, CHENG Yang, ZHOU Peng, ZHOU Wei

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

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

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. http://www.jos.org.cn/1000-9825/3431.htm

Abstract: Cluster ensemble becomes a research focus due to its success in privacy protection, distributing computing and reusing knowledge. Furthermore, the noise and isolation have little effect on the final result. There are two contributions in this paper. First, by regarding every base clustering as one attribute of the original data, it has found that the algorithm based on that is more extendable and flexible. Second, it designs a latent variable cluster ensemble (LVCE) model in this way and infers the algorithm of the model with Markov chain Monte Carlo (MCMC) approximation. At the end of the paper, the experimental results show that the MCMC algorithm of LVCE has a better result and can show the compactedness of data points clustering.

Key words: cluster ensemble; latent variable; LVCE (latent variable cluster ensemble); MCMC (Markov chain Monte Carlo)

摘 要: 聚类集成能成为机器学习活跃的研究热点,是因为聚类集成能够保护私有信息、分布式处理数据和对知识进行重用,此外,噪声和孤立点对结果的影响较小.主要工作包括:第一,分析了把每一个基聚类器看成是原数据的一个属性这种处理方式的优越性,发现按此方法建立起来的聚类集成算法就具有良好的扩展性和灵活性;第二,在此基础之上,建立了 latent variable cluster ensemble(LVCE)概率模型进行聚类集成,并且给出了 LVCE 模型的 Markov chain Monte Carlo(MCMC)算法.实验结果表明,LVCE 模型的 MCMC 算法能够进行聚类集成并且达到良好的效果,同时可以体现数据聚类的紧密程度.

关键词: 聚类集成;隐含变量;聚类集成模型;MCMC(Markov chain Monte Carlo)
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,然后对这些结果加以组合,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,可以分布式地处理数据.同时,噪声和孤立点对结果的影响较小.聚类集成主要有 3 个方面的作用:

* Supported by the China Scholarship Council Foundation under Grant No.2007U24068 (国家留学基金委员会资助项目)
Received 2008-03-13; Accepted 2008-08-11

第一是保护私有信息,在对原数据进行数据挖掘的时候,有时需要对敏感的数据进行保护.可以通过对原数据加密或者数据转换[2]的方法,更为简单的方法就是将该数据标为丢失数据.这样处理的结果使得最后的挖掘结果不够准确,精度降低.而聚类集成不仅不访问原数据,而且在大部分情况下,其集成结果还比单个的聚类器要准确;

第二是知识重用,对于部分公司和政府,堆积了很多的数据信息.而这些数据信息的格式可能多种多样,有很多的数据由于维护不当,也导致丢失的情况,而只剩下对这些数据的一些结果信息,如类别、分布等.那么聚类集成就可以重用这些数据和新数据标签一起聚类集成,达到知识重用的目的;

第三是分布式计算,对于庞大的分布式数据,不可能把它们先集中起来进行处理.现在,一般的处理方式是进行分布式计算,最后集成的方法.聚类集成就是对基聚类器的结果进行集成,这些基聚类器可以完全是分布的,也可以是异步的,只要提交基聚类的结果就可以进行聚类集成.这样处理速度更快,传输代价更小.

1 聚类集成工作原理

聚类集成工作原理如图1所示,图1中的C_m是基聚类结果,C^*是聚类集成结果.聚类集成主要分为两个阶段:第1个阶段是采集基聚类器的结果数据,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;第2个阶段是数据集成,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结果.m个基聚类结果里面不可能完全聚类正确,原因在于,每一个基聚类结果的数据里都包含噪声.如果基聚类结果允许丢失数据出现,那么这些基础数据里还存在丢失的数据标签.然后,通过一个聚类集成函数或者算法模型对这m个基聚类结果进行处理,从而得到最终的聚类集成结果.一般来说,最终的聚类结果是在m个基聚类结果中最好的.

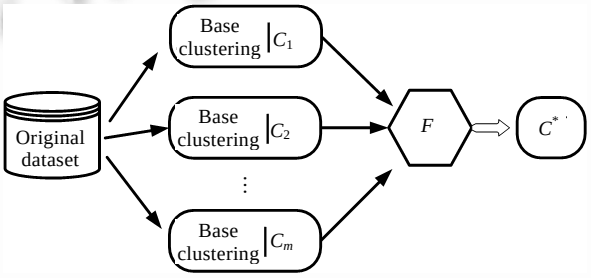

Fig.1 Procedure of cluster ensemble

图1 聚类集成的工作原理

Texas大学计算机学院的Alexander Strehl最早在机器学习研究中明确提出聚类集成问题以及解决方法.他认为,要最大程度地分享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计算它们之间的信息熵来实现.这样来度量其信息熵和相关信息之比,从而获得最好的结果[3].周志华等学者提出,解决聚类集成问题可以通过投票机制,从而根据投票的结果来得到一个聚类集成结果[4−6].对于聚类集成算法,目前主要有3类:第1类是基于图形分割[3,7,8]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,然后在基于最小切或者最小权重的方法开始切割图形,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;第2类是基于矩阵相似算法[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;第3类是基于概率统计[10−12]的方法,这类方法主要是先求基聚类结果数据的统计上的特征,基聚类的权重与其置信度成正比[11].文献[10]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.

基于图形分割和矩阵相似的聚类集成算法粗糙地把数据点归属这个类别或者不归属这个类别,这样的分类结果根本无法看出数据点之间的距离关系,不知道数据点聚类的紧密程度.如CSPA,MCLA和HGPA在把基

聚类结果转换成超图的时候,简单地把数据点属于某类的用 1 表示,不属于某类的用 0 表示,基于图和矩阵相似的分类方法基本上采用的都是这一方法.这样忽略了聚类中的两个关系:一是同类数据之间的关系,因为此方法中同类数据都用 1 表示,没有区别;二是不同类之间的数据关系,在聚类问题中,很多算法都忽略不同类数据之间的关系.实际上,即使是不同类的数据,它们之间也是有关系的.聚类最重要的就是能够表示各数据点之间的关系,这种聚集关系可以直观地感觉这些数据点之间是结合得紧密还是稀疏,并不是简单地将其分类.这两类算法适应于低噪声的数据,如果数据噪声高于 50%,那么这些算法的正确率将下降比较快,这是由其逻辑处理方式(1 或者 0)所决定的.这是前两类算法的主要缺陷.基于概率统计的算法克服了前两类算法的缺陷,但目前所提出算法的初始化都依赖于具体的数据,如基于权重的算法,对每一个具体的数据集来说这个权重是不同的,这就为初始化带来了困难;并且目前所提出的这一类算法不具有扩展性.

上述许多方法的一个共同点是把 m 个基聚类结果看成 m 个向量,然后对其进行处理,如图 2(a)所示.现在的实际数据每天都会增加(如销售、顾客等数据),按此方式设计的聚类集成算法处理此类数据,将每次都会对上一次的数据进行重复处理.如集成算法对 n 个数据进行聚类集成,已经得到了聚类集成结果.如数据从 n 增加到 $n+h$ 个数据,是否可以利用聚类集成算法只对 h 个数据进行聚类集成?从后文公式(1)可以知道,这些算法都要重复前面的工作;还需要对 $n+h$ 数据进行聚类集成,这样就对 m 个数据聚类重复了一次,增加了开销和代价.本文分析了把基聚类的结果看成是原始数据属性[13]的聚类集成的处理方式,发现这种方式更具灵活性和扩展性.并且根据此方式提出一个新的聚类集成模型,此模型是属于概率模型,不仅可以表示数据点的聚类结果,而且可以表示聚类紧密程度,发现真实的数据结构,表示各数据点的稀疏关系.可以有效地避免这两类问题的出现.

	C_1	C_2	C_3	C_4		C_m
x_1	2	3	1	1	…	2
x_2	1	4	1	1	…	6
x_3	1	5	2	4	…	1
x_4	2	2	3	2	…	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_n	3	2	2	4	…	6

(a)

	C_1	C_2	C_3	C_4		C_m
x_1	2	3	1	1	…	2
x_2	1	4	1	1	…	6
x_3	1	5	2	4	…	1
x_4	2	2	3	2	…	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_n	3	2	2	4	…	6

(b)

Fig.2 Different ways of processing base clusterings' results

图 2 处理基聚类结果的不同方式

2 聚类集成问题

设原数据集有 n 个对象,有 m 个基聚类器对原数据进行聚类,那么我们可以得到 m 个长度为 n 的基聚类结果向量.按照上面所述的处理方式,则聚类集成问题为[3]

$$F_1:\{c_i \mid c_1,c_2,...,c_m\}\rightarrow C^* \tag{1}$$

其中,C_i 和 C^*是维数为 n 的向量,C_i 是基聚类结果,C^*为聚类集成结果.本文把 m 个基聚类的结果换个角度来分析,如图 2(b)所示.把这 m 个基聚类结果不再看成是 m 个向量,而看成是 n 个数据对象,把每一个基聚类看成是数据对象的特征属性.那么,每一个基聚类器的结果则是属性值.

那么,聚类集成问题可以表示为

$$F_2:\{x_j \mid x_1,x_2,...,x_n\}\rightarrow C' \tag{2}$$

其中,x_j 是编号为 j,维数为 m 的向量对象,这里,x_j 里不带有原数据的属性和属性值,其每个属性都是基聚类器,属性值都是基聚类结果.所以,数据对象的维数为 m,C'为聚类集成结果.对比图 2(a)和图 2(b)可以发现数据都相同,只是对向量对象的划分标准上行列进行了变换.对比公式(1)和公式(2),这两者有着重要的区别和意义.

第一:按公式(2)建立起来的聚类集成模型具有很好的扩展性,当原数据对象增加到 $n+h$ 时,这个聚类集成模型只需对后面的 h 个数据对象的基聚类结果进行处理;如果按公式(1)建立起来的聚类集成模型,在面对这种情况时,需要对整个基聚类结果数据重新处理.

第二:公式(2)实际上是一个聚类问题,任何一种聚类算法理论上都可以作为聚类集成算法,那么实际上就把聚类集成问题简单化了.在实验中,可以用 K-Means,SVM,NB 等系列的算法作为聚类集成算法.

3 隐含变量聚类集成模型(LVCE)

3.1 隐含变量聚类集成模型LVCE及符号定义

LVCE(latent variable cluster ensemble)建立在概率理论之上,如图 3 所示,实际上是一个贝叶斯网,主要阐述基聚类结果和聚类集成结果之间的依赖关系.模型中的符号定义:$x_{i,j}$ 为图 2 矩阵中第 i 行第 j 列的元素,它是第 j 个基础分类器对原数据对象 i 的聚类类别索引,也是聚类集成中的原子元素,其中,i,j 的范围为 $\{i_1^n, j_1^m\}$,在 LVCE 模型中是观察数据,为图 3 中的 x.x_i 表示一个向量,其中的元素为 $x_{i,(.)}$,也就是 m 个基础分类器对原数据对象 i 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 LVCE 模型中,C 是隐含变量,为图 3 中的 c;θ_i 表示对 x_i 的聚类集成结果,θ_i 在 i 取不同值时,θ_i 可能等于 θ_k,则说明原数据对象 i 和 k 属于同一个类别,为图 3 中的 θ.α和β都是预先设定的模型参数,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呈线性关系,因此,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.

此模型有两个假设:第一是假使α的维数已知并且固定不变,也就是聚类集成的类别数量已知和类别数量在特定数据集是固定的;第二是假设 $x_{i,j}$ 是由 $c_{i,j}$ 和一个参数β决定的,也就是说,$x_{i,j}$ 依赖于 $x_{i,j}$ 和β,其现实意义就是基聚类数据可以通过某个参数转化而得到聚类集成结果.

LVCE 的模型主要有两个作用:第一是模拟产生基聚类器结果;第二是如果知道基聚类器的结果,可以推导出隐含变量 C 和 Z 的值,也就可以知道聚类集成的结果.模型借鉴了 Latent Dirichlet Allocation(LDA)[14]的各变量之间依赖关系的表示.LDA 模型也称为主题模型,它是一个文档基于主题的生成模型,最主要地用于文档的聚类分类和模拟文档的产生.从其产生以来,有很多的文本聚类分类都用到这个模型.对比 LDA 模型,LVCE 主要与其有两个方面的不同:首先是 LDA 模型中β参数会呈线性增长的趋势,LVCE 模型中,ϕ是β的分布,这样会使模型更加稳定,并且在实际应用中,β是一个中间结果,也克服了参数线性增长的弊端;其次,LVCE 的各变量和 LDA 的各变量所表示的意义完全不同.在聚类集成问题中,聚类集成结果就是一个隐含在观察数据中的变量,观察数据受隐含变量的制约,依赖于隐含变量.图 3 详细地表示了各变量及观察数据之间的依赖关系.LVCE 模型也是基于此思路,把隐含变量表示在模型中,然后根据观察数据得到隐含变量.

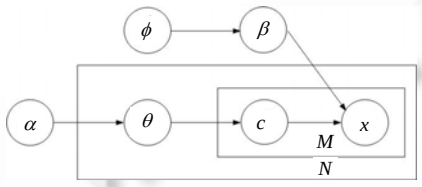

Fig.3 Graphic model of latent variable cluster ensemble

图 3 LVCE 的图形表示

3.2 模型生成抽样机制

图 3 是 LVCE 各变量之间的依赖关系,LVCE 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 $x_{i,j}$,首先需要根据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 $x_{i,j}$.生成的抽样算法如下:

Algorithm 1.

Input: α and ϕ;

Output: x.

1. For i=1:N
2. Sample θ_i; Calculate the β;
3. For j=1:M
4. Sample $c_{i,j}$ according to θ_i;
5. Sample $x_{i,j}$ according to $c_{i,j}$ and β;
6. End for Step 3

7. End for Step 1

当这个算法执行完毕,就可以模拟生成出图 2(b)中的数据.在聚类集成中主要是对 $c_{i,j}$ 感兴趣,而其又很难直接求出,但可以通过近似算法得到,这些近似算法如变分法、Laplace 近似法、MCMC 等.

3.3 LVCE的MCMC推理及算法

根据前面的符号定义和变量之间的依赖关系,LVCE 整个概率模型表示为

$$\left.\begin{aligned}\theta &\sim Dirchlet(\alpha)\\ c_{i,j}\mid\theta_i &\sim Discrete(\theta_i)\\ \beta &\sim Dirchlet(\phi)\\ x_{i,j}\mid c_{i,j},\beta_{c_{i,j}} &\sim Discrete(\beta_{c_{i,j}})\end{aligned}\right\} \tag{3}$$

关于公式(3)狄利克雷(Dirichlet)的分布θ~*Dirichlet*(α),其分布函数为

$$p(\theta\mid\alpha)=\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(a_i)}\theta_1^{\alpha_1-1}...\theta_k^{\alpha_k-1} \tag{4}$$

其中,α为 k 维的向量参数并且$\alpha_t>0$,$\Gamma(x)$是 Gamma 函数,这里选择 Dirichlet 分布的主要原因是其与多项式分布共轭,并且是属于指数族分布,这样可以使模型计算简单和方便.公式(3)中的 Discrete 分布是多项式分布.

对于此模型,本文是对 $p(c)$感兴趣.根据马尔可夫链蒙特卡尔方法(Markov chain Monte Carlo,简称 MCMC)[15],从复杂的概率分布里抽样,并且每一个马尔可夫链上状态的改变只依赖前一个状态,那么 $c_{i,j}$ 的条件后验概率[16]可以表示为

$$p(c_{i,j}=k\mid\overline{c}_{i,j},x)\propto p(x_{i,j}\mid c_{i,j}=k,\overline{c}_{i,j},\overline{x}_{i,j})\,p(c_{i,j}=k\mid\overline{c}_{i,j}) \tag{5}$$

其中,$\overline{c}_{i,j}$ 是指除当前 $c_{i,j}$之外的所有其他 $c_{i,j}$, $\overline{x}_{i,j}$ 是指除当前 $x_{i,j}$之外的所有其他 $x_{i,j}$,k 为类别编号.公式(5)只通过 $\overline{c}_{i,j}$ 和 x 得到 $c_{i,j}$ 的条件概率,这是因为 MCMC 方法.公式(5)右边两项实际上是对θ和β求边缘分布而得到的.这样,θ和β就可以不出现在公式(5)中.

$$p(x_{i,j}\mid c_{i,j}=k,\overline{c}_{i,j},\overline{x}_{i,j})=\int p(x_{i,j}\mid c_{i,j}=k,\beta_{c_{i,j}})\,p(\beta_{c_{i,j}}\mid\overline{c}_{i,j},\overline{x}_{i,j})d_{(\beta_{c_{i,j}})} \tag{6}$$

其中, $\beta_{c_{i,j}}$ 是指数据 $x_{i,j}$ 对聚类集成类别 $c_{i,j}$ 的多项式分布,即 $x_{i,j}$ 属于 $c_{i,j}$ 的概率.根据 Bayes 规则,公式(6)右边第 2 项为

$$p(\beta_{c_{i,j}}\mid\overline{c}_{i,j},\overline{x}_{i,j})\propto p(\overline{x}_{i,j}\mid\beta_{c_{i,j}},\overline{c}_{i,j})\,p(\beta_{c_{i,j}}) \tag{7}$$

因为 $p(\beta_{c_{i,j}})$ 是 *Dirichlet*(*f*)分布并且与 $p(\overline{x}_{i,j}\mid\beta_{c_{i,j}},\overline{c}_{i,j})$ 共轭,那么 $p(\beta_{c_{i,j}}\mid\overline{c}_{i,j},\overline{x}_{i,j})$ 的后验概率为

$$Dirichlet(f+n_{i,j}^{-(x_{i,j})}),$$

其中, $n_{i,j}^{-(x_{i,j})}$ 是指除当前 $x_{i,j}$ 外其他 $x_{i,j}$ 的数据值等于当前 $x_{i,j}$ 属于类别 $c_{i,j}$ 的数量.所以,只有当 $x_{i,j}$ 属于类别 $c_{i,j}$ 的数量时能够影响 $\beta_{c_{i,j}}$ 的分布.因此,

$$p(x_{i,j}\mid c_{i,j}=k,\overline{c}_{i,j},\overline{x}_{i,j})=\frac{n_{i,j}^{-(x_{i,j})}+\phi}{n_{i,j}^{-(\cdot)}+MN\phi} \tag{8}$$

其中, $n_{i,j}^{-(\cdot)}$ 是除当前 $x_{i,j}$ 外其他所有 $x_{i,j}$ 属于类别 $c_{i,j}$ 的总数量.刚才完成了对公式(5)右边第 1 项的推理,以同样的方式对其第 2 项进行推理.

$$p(c_{i,j}=k\mid\overline{c}_{i,j})=\int p(c_{i,j}=k\mid\theta_i)\,p(\theta_i\mid\overline{c}_{i,j})d_{\theta_i}=\frac{n_{i,k}^{-(i)}+\alpha}{n_{i,(\cdot)}^{-(i)}+K\alpha} \tag{9}$$

其中, $n_{i,k}^{-(i)}$ 是指编号为 i 的向量中有多少个 $x_{i,j}$ 属于类别 k(不包括当前的 $x_{i,j}$); $n_{i,(\cdot)}^{-(i)}$+1 是指编号为 i 的向量共有多少个有效元素 $x_{i,j}$.如果某个基聚类结果中有丢失数据,那么这个数量就不相等;如果完全没有丢失数据,那么 $n_{i,(\cdot)}^{-(i)}$ 的值为基聚类器数量减 1.因此,从公式(5)、公式(8)和公式(9)可以得到:

$$p(c_{i,j}=k\mid \bar{c}_{i,j},x)\propto \frac{n_{i,j}^{-(x_{i,j})}+\phi}{n_{i,j}^{-(\cdot)}+MN\phi}\frac{n_{i,k}^{-(i)}+\alpha}{n_{i,(\cdot)}^{-(i)}+K\alpha} \tag{10}$$

根据推导和 MCMC,初始化变量状态为任意状态,并且 MCMC 与初始化无关,设计算法如下:

Algorithm 2.

Input: $\{x,\alpha,\phi\}$;

Output: $\{p(C|X)\}$.

1. Step one: Initialize:
2. (a) Instantiate $p_{i,j}$ to one of its possible labels $c_{i,j}$ of $x_{i,j}$, $1\le c_{i,j}\le K; i_1^n; j_1^m$;
3. (b) Let $p^0=\{p_{1,1},p_{1,2},\ldots,p_{n,m}\}$;
4. Step two: For t=1:*IteratorTimes*; *IteratorTimes* is the iterative number of the Gibbs sampling
5. Pick index $x_{i,j}$ at random one by one;
6. For k=1:K; K is the number of classes
7. Calculate $p(c_{i,j}=k\mid \bar{c}_{i,j},x)$ one by one;
8. End for Step 6
9. Choose MAX $p(c_{i,j}=k\mid \bar{c}_{i,j},x)$;
10. Update p^t;
11. End for Step 4

本算法可以得到 $x_{i,j}$ 属于各类别的概率,要求 x_i 对象所属类别,可以用贝叶斯规则计算得到.算法 2 的复杂度为 $o((tk+1)n)$,其中,n 是数据对象的数量,k 为数据对象的类别数,t 为收敛的循环次数.本算法中,t 的值一般介于 200~400 之间.

4 实验和结果

我们使用两组数据集作为实验:第 1 组是按照模型生成的数据,也可以叫做人工模拟的数据;第 2 组选用 UCI 的数据.见表 1(a),这些数据把样本、属性和类别数目都列出来.对于聚类集成实验,一般会做 3 种类型的实验:首先是检查聚类集成结果是否比基聚类的结果要好;第二是在没有丢失数据的情况下,与其他聚类集成算法相比较;第三是在基聚类结果中随机丢失数据,测试几种算法的稳定性.

Table 1 Datasets and accuracy of base clusterings and LVCE

表 1 实验数据集及基聚类器和 LVCE 的正确率

(a) Number of instances, features and classes of datasets

(a) 实验数据集的样本、属性和类别数目

Dataset	Instances	Features	Classes
Artificial data 1	240	6	4
Artificial data 2	500	15	2
UCI iris	150	4	3
UCI ionosphere	351	34	2
UCI wdbc	569	30	2
UCI bupa	345	6	2

(b) Accuracy of base clusterings and LVCE

(b) 基聚类器和 LVCE 的正确率

Algorithm / Dataset	Base clustering K-means		LVCE
	Max	Average	
Artificial data 1	.872	.850±0.027	.904±0.003
Artificial data 2	.862	.822±0.328	.898±0.018
UCI iris	.887	.864±0.113	.900±0.187
UCI ionosphere	.68.2	.624±0.510	.702±0.059
UCI wdbc	.88.0	.805±0.601	.884±0.037
UCI bupa	.529	.502±0.082	.571±0.029

关于聚类和聚类集成算法有很多标准来衡量,本文以这些数据的真实类别标签为标准.以聚类的平均正确率来衡量算法,为了更好地检验这个平均正确率,对算法的分类结果作配对 t 检验(paired t-test).因为聚类集成结果依赖于基聚类器的结果,所以采用平均正确率较好.

首先,对这些数据集进行基聚类,基聚类器的数量为 10~50,也就是 m 取值为[10,50],其聚类的平均和最好结果见表 1(b).其中,.850±0.027 表示此算法的平均正确率是 0.850,其标准差为 0.027.而 LVCE 的结果比基聚类中任何一个聚类的结果都要好,所以,这一点说明 LVCE 用来作为聚类集成是有效的.

其次,对基聚类的结果进行聚类集成,5 种算法的结果见表 2,其中,L-HGPA 表示把 LVCE 和 HGPA 的分类结果作配对 t 检验,其 p 值的结果见表 2.可以看出,不论是人工模拟数据还是 UCI 的数据集 LVCE 的结果,都比 *K*-means,HGPA,CSPA 和 MCLA 效果好或者相同,而配对检验中的标准差和 P 值也证明了这一点,只有两组的 P 值超过了 0.05 的临界值.关于部分 UCI 数据的聚类集成的正确率可以参考文献[9,10].

Table 2 Cluster ensemble results of the algorithms

表 2 聚类集成算法在这些数据集上的运行结果

Algorithm / Dataset	*K*-means	HGPA	CSPA	MCLA	LVCE	L-HGPA *p*-value	L-CSPA *p*-value	L-MCLA *p*-value
Artificial data 1	.850±0.032	.767±0.591	.896±0.161	.892±0.098	.904±0.003	.009	.012	.023
Artificial data 2	.822±0.041	.764±0.467	.880±0.152	.858±0.086	.898±0.018	.007	.025	.041
UCI iris	.864±0.058	.633±0.817	.887±0.190	.887±0.100	.900±0.187	.003	.016	.008
UCI ionosphere	.624±0.023	.606±0.300	.675±0.157	.695±0.106	.702±0.059	.011	.019	.502
UCI wdbc	.805±0.011	.552±0.670	.834±0.432	.884±0.075	.884±0.037	.017	.041	.612
UCI bupa	.502±0.029	.526±0.423	.568±0.260	.565±0.082	.571±0.029	.006	.022	.041

本文使用 5 种聚类集成算法:第 1 类是任何的聚类算法,如 *K*-means,SVM,NB 等;本文选择标准 *K*-means,是为了说明任何一种聚类算法按照本文的数据处理都可以作为聚类集成算法.另外,本文的基聚类器也全部选择标准 *K*-means;第 2 类是已有的聚类集成算法,如 MCLA,CSPA,HGPA 代码可以从 Alexander Strehl 主页上下载;第 3 类是本文的聚类集成算法 LVCE.在本实验中,各数据集上聚类的类别数就是表 1(a)中类别数,并且在实验中固定不变.在 LVCE 模型中,这个假设也就是模型中第 1 个假设 *Dirichlet*(α)分布参数α的维数已知并且固定不变.

第三,在基聚类结果中随机丢失数据,测试几种算法的稳定性.在本实验中,基聚类器的数目为 30,丢失数据为随机丢失,但是必须保证这些数据的每一行至少有一个数据存在,丢失比例从 8%开始,每次增加 8%,直到 80%截止.图 4 分别是对 UCI 数据的实验结果,图 4 的横坐标是基聚类结果的丢失比例,最多丢失数据为 80%;纵坐标是这些丢失数据的平均正确率.

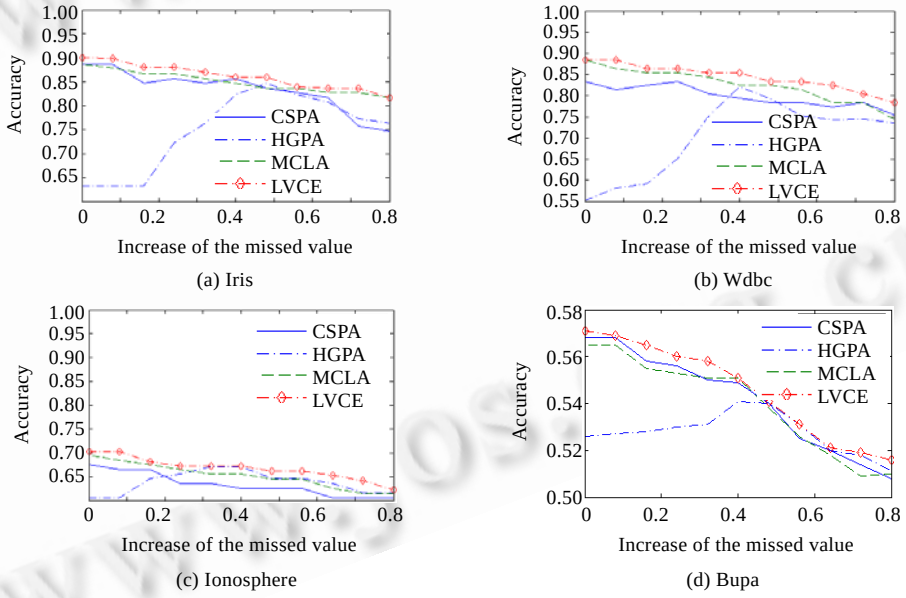

Fig.4 Changing results of algorithms with increase the missed value

图 4 丢失数据增加时,聚类集成算法的结果变化情况

从图 4 中可以看出,在处理丢失数据上,CSPA,MCLA 和 LVCE 都比较稳定,出现这种结果主要是因为基聚类器比较多,如果基聚类器数为 5 以下,那么在丢失数据很高的情况下,算法稳定性要差一些.从图中还可以看出,LVCE 的平均正确率略高于其他算法,这主要是因为在没有丢失数据的情况下,LVCE 的正确率高于其他算

法.这里可以看出,LVCE 在丢失数据的情况下具有较高的稳定性.在实验中发现,HGPA 出现一些随丢失数据增加而正确率增高的现象,这在文献[3]中有简单解释,那就是 HGPA 处理数据噪声比较强.

LVCE 不仅可以表示数据的类别,而且还可以表示数据间的联系和整个数据的稀疏关系.而 HGPA,CSPA 和 MCLA 只是能够简单地对数据分类,而不能表示这些数据间的关系.本文采用 ISOMAP[17]来将 LVCE 的输出结果映射到二维空间,如图 5 所示,这样可以直观地看出数据的空间距离关系,也可以看出其聚类的紧密度,图中不同形状的图形表示不同的类别,其中数字表示该类的数量.

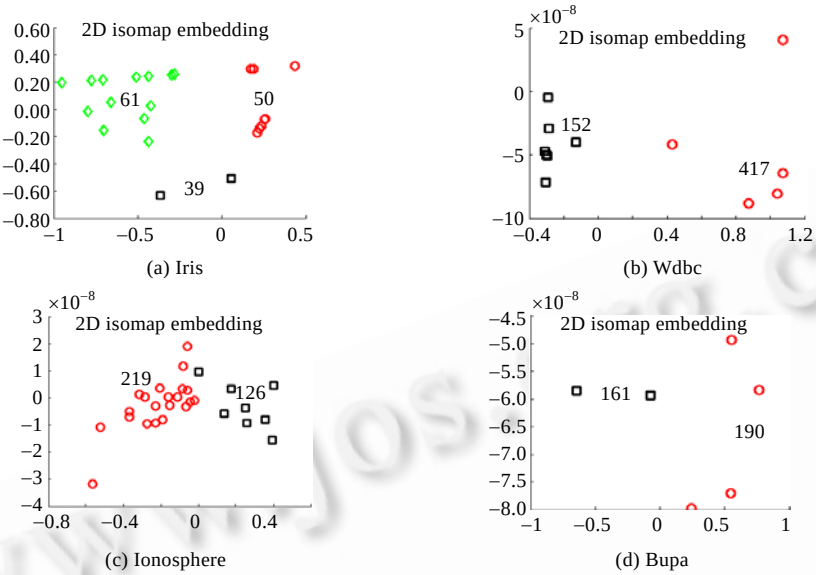

Fig.5 Visualization of data points clustering of LVCE outputs

图 5 LVCE 算法输出的数据点聚类二维图表示

由于聚类集成的结果与基聚类的结果很有关系,一般来说,基聚类结果好,那么聚类集成结果也较好;还有基聚类的多样性[18],也可以得到比较好的聚类集成结果.有可能有更好的聚类算法在这些数据集的绝对结果比 LVCE 好,但本文主要研究聚类集成算法的有效性.

5 结束语

本文分析了处理聚类集成的方式,发现把每一个基聚类器看成是原数据的一个属性这种方式有很大的优越性,通过这一方法建立起来的聚类集成算法就具有良好的扩展性和灵活性.本文在此基础上建立了 LVCE 生成概率模型,使模型能够处理聚类集成,处理高噪声的数据效果比较好.而且克服了其他聚类算法的一些缺陷,这一算法不仅可以表示数据的类别,也可以体现聚类的紧密度.本文给出了 LVCE 的 MCMC 算法,当然,本模型还可以使用变分法、Laplace 近似法等来进行推理.我们进一步的工作主要集中在 LVCE 的其他近似推导算法上,并且从理论上研究 LVCE 的分布式和并行聚类集成.

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 Minnesota 大学计算机科学技术系 Arindam Banerjee 教授和 Shan Hanhuai 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] Oliveira SRM, Zaane OR. Privacy preserving clustering by data transformation. In: Proc. of the 18th Brazilian Symp. on Databases. Manaus, 2003. 304−318. http://citeseer.ist.psu.edu/article/oliveira03privacy.html

[3] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 2002,3:583−617. http://jmlr.csail.mit.edu/papers/volume3/strehl02a/strehl02a.pdf

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

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

[6] Zhou ZH, Tang W. Clusterer ensemble. Knowledge-Based Systems, 2006,19(1):77−83.

[7] Asur S, Parthasarathy S, Ucar D. An ensemble approach for clustering scale-free graphs. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Philadelphia, 2006. http://kt.ijs.si/Dunja/LinkKDD2006/Papers/asur.pdf

[8] Kuncheva LI, Hadjitodorov ST. Solving cluster ensemble problems by bipartite graph partitioning. In: Proc. of the 21st Int'l Conf. on Machine Learning. Banff, 2004. 281−288. http://portal.acm.org/citation.cfm?id=1015414

[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. Omaha, 2007. http://www.cs.berkeley.edu/~jordan/papers/li-ding-jordan-icdm07.pdf

[10] Topchy A, Jain AK, Punch W. A mixture model for clustering ensembles. In: Proc. of the 4th SIAM Int'l Conf. on Data Mining. Lake Buena Vista, 2004. 22−24. http://www.siam.org/proceedings/datamining/2004/dm04_035topchya.pdf

[11] Al-Razgan M, Domeniconi C. Weighted cluster ensemble. In: Proc. of the Society for Industry and Applied Mathematics Conf. on Data Mining. 2006. 258−269. http://www.siam.org/meetings/sdm06/proceedings/024alrazganm.pdf

[12] Topchy A, Minaei-Bidgoli B, Jain AK, Punch WF. Adaptive clustering ensembles. In: Proc. of the 17th Int'l Conf. on Pattern Recognition (ICPR 2004), Vol.1. 2004. 272−275. http://www.lon-capa.org/papers/adaptive.pdf

[13] Zhou ZH, Zhang ML. Solving multi-instance problems with classifier ensemble based on constructive clustering. Knowledge and Information Systems, 2007,11(2):155−170.

[14] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[15] Robert CP, Casella G. Monte Carlo Statistical Methods. 2nd ed., New York: Springer-Verlag, 2004.

[16] Casella G, George EI. Explaining the Gibbs sampler. The American Statistician, 1992,46:167−174.

[17] Tenenbaum J, Silva V, Langford J. A global geometric framework for nonlinear dimensionality reduction. Science, 2000,290: 2319−2323.

[18] Kuncheva LI, Hadjitodorov ST. Using diversity in cluster ensembles. In: Proc. of the IEEE Int'l Conf. on Systems, Man and Cybernetics. 2004. 1214−1219. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1399790

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

王红军(1977－),男,四川广安人,博士生,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.

李志蜀(1946－),男,教授,博士生导师,CCF 高级会员,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.

成飏(1979－),男,博士生,主要研究领域为网络与信息系统,智能计算.

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.

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.