基于 Bagging 的选择性聚类集成*

唐 伟, 周志华+

(南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

Bagging-Based Selective Clusterer Ensemble

TANG Wei, ZHOU Zhi-Hua+

(National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

+ Corresponding author: Phn: +86-25-83686268, E-mail: zhouzh@nju.edu.cn, http://cs.nju.edu.cn/people/zhouzh/

Received 2003-11-03; Accepted 2004-07-27

Tang W, Zhou ZH. Bagging-Based selective clusterer ensemble. *Journal of Software*, 2005,16(4):496−502. DOI: 10.1360/jos160496

Abstract: This paper uses ensemble learning technique to improve clustering performance. Since the training data used in clustering lacks the expected output, the combination of component learner is more difficult than that under supervised learning. Through aligning different clustering results and selecting component learners with the help of mutual information weight, this paper proposes a Bagging-based selective clusterer ensemble algorithm. Experiments show that this algorithm could effectively improve the clustering results.

Key words: machine learning; ensemble learning; clustering; unsupervised learning; selective ensemble

摘 要: 使用集成学习技术来提高聚类性能.由于聚类使用的训练样本缺乏期望输出,与监督学习下的集成相比,在对个体学习器进行结合时更加困难.通过对不同的聚类结果进行配准,并基于互信息权进行个体学习器的选择,提出了基于 Bagging 的选择性聚类集成算法.实验表明,该算法能够有效地改善聚类结果.

关键词: 机器学习;集成学习;聚类;非监督学习;选择性集成

中图法分类号: TP181 **文献标识码**: A

聚类分析技术将未标记对象通过其相似度进行分组,使得组内对象的相似度最大而组间对象的相似度最小,从而发现对象中的内在特性.由于聚类分析技术在数据挖掘、模式识别等诸多领域有着广泛的应用前景,一直是机器学习领域的一个研究热点[1].

集成学习(ensemble learning)[2]技术利用基学习器的多个版本来解决同一个问题,可以显著地提高学习系统的泛化能力.最近几年,在机器学习、神经网络、统计学等领域的很多研究者都投入到集成学习的研究中,使得该领域成为了一个相当活跃的研究热点,并被认为是当前机器学习领域的 4 大研究方向之首[2].现在已经有很多集成学习算法,Bagging 算法[3]就是其中比较著名的一个.该算法在训练阶段,各学习器的训练集由原始训练集利用可重复取样(bootstrap sampling)技术获得,训练集的规模通常与原始训练集相当.这样,原始训练集中

* Supported by the National Outstanding Youth Foundation of China under Grant No.60325207 (国家杰出青年科学基金)

作者简介: 唐伟(1978－),男,湖南祁阳人,硕士,主要研究领域为机器学习,数据挖掘;周志华(1973－),男,博士,教授,博士生导师,主要研究领域为机器学习,数据挖掘,模式识别,信息检索,神经计算,进化计算.

某些示例可能在新的训练集中出现多次,而另外一些示例则可能一次也不出现.研究表明[3],Bagging 可以显著提高不稳定的基学习器的泛化能力.以往的集成学习算法在生成多个个体学习器之后,通常是对所有的个体都进行结合,因此很多研究者尝试使用大规模的集成来解决问题.Zhou 等人[4]提出了"选择性集成 (selective ensemble)"的概念,并证明通过选择部分个体学习器来构建集成可能要优于使用所有个体学习器构建的集成,这就意味着利用中小规模的选择性集成就可以获得很好的性能.

需要注意的是,Bagging 算法和其他大多数的集成学习算法都是为监督学习而设计的,对聚类这样的非监督学习来说,由于训练样本缺乏类别标记,聚类结果之间没有直接的对应关系,这将使得对个体学习器的结合难以直接进行.因此,用于非监督学习的集成学习算法比一般的用于监督学习的集成学习算法设计起来更加困难.实际上,利用集成学习技术来提高聚类分析的性能已经引起了一些研究者的关注.例如,在 Strehl 和 Ghosh[5]的工作中,他们利用互信息(mutual information)把聚类集成问题定义为一个基于互信息的优化问题,但同时又指出由于该优化问题的计算开销过于庞大,因此难以应用于实际领域.

本文对聚类集成进行研究,提出了基于 Bagging 的选择性聚类集成算法.第 1 节介绍本文所用到的标记和术语,并简单介绍作为聚类集成个体的 k 均值(k-means)算法.第 2 节具体介绍基于 Bagging 的选择性聚类集成算法.第 3 节通过实验对该算法进行性能测试并对实验结果进行分析.最后一节是对本文的工作进行总结并展望进一步的工作.

1 背景知识

1.1 标记和术语

为了下文讨论的方便,本节将对有关聚类集成中所涉及的标记和术语进行约定.假设 $\chi=\{x_1,x_2,\ldots,x_n\}\subset\Re^d$ 为 d 维特征空间中一组类别未知的向量集.该向量集中的第 i 个元素 x_i 为一个 d 维的特征向量$[x_{i1},x_{i2},\ldots,x_{id}]^{\mathbf{T}}$,$\mathbf{T}$ 表示矩阵的转置.不失一般性,假设特征向量中的每一个分量均为数值属性.对于某个聚类器(clusterer)可以将向量集 χ 中的元素划分为 k 个聚类,可用一个标记向量$\lambda^{(m)}=[\lambda_1,\lambda_2,\ldots,\lambda_n]^{\mathbf{T}}\in N^n$ 表示,其中$\lambda_i\in\{1,2,\ldots,k\}$为聚类标记.聚类集成算法首先通过 M 个聚类器对向量集χ进行聚类,然后将所有聚类器产生的标记向量$\{\lambda^{(1)},\lambda^{(2)},\ldots,\lambda^{(M)}\}$进行结合,最后得到结果标记向量$\lambda$.

1.2 k均值算法

MacQueen 提出的 k 均值算法[6]是一个著名的聚类学习算法.它根据相似度距离迭代地更新向量集的聚类中心,当聚类中心不再变化或者满足某些停止条件,则停止迭代过程得到最终的聚类结果.k 均值算法的具体步骤为:

(1) 随机选择 k 个数据项作为聚类中心;

(2) 根据相似度距离公式,将数据集中的每一项数据分配到离它最近的聚类中去;

(3) 计算新的聚类中心;

(4) 如果聚类中心没有发生改变,算法结束;否则跳转到第(2)步.

由于所选择的相似度距离公式的不同,k 均值算法所得到的聚类结果将存在较大的差异.为了简化讨论,本文采用欧氏距离作为 k 均值算法的相似度距离公式.但值得注意的是,本文工作对其他距离公式同样是适用的.

2 基于 Bagging 的选择性聚类集成

集成学习一般包含两个阶段,即个体生成阶段和个体学习器的结合阶段.在个体生成阶段,通过不同的个体生成方式产生不同的个体标记向量.在个体学习器的结合阶段,可以采用投票等方式将个体标记向量进行结合.

在个体生成阶段,考虑到需要聚类的对象数可能非常庞大,对所有对象进行聚类势必将增加聚类算法的运行开销.因此,可以考虑对整个空间中分布在某个局部的对象进行聚类从而得到结果,但是,由于聚类对象的空间分布可能很不均匀,所得到的聚类结果可能只代表某个局部信息而不是全局的聚类结果.为了解决这个问题,本文采用类似 Bagging 算法中产生个体训练集的方式产生用于聚类的训练集,即通过可重复取样技术从原向量

集χ中产生若干训练集$\{S_i\}$,i=1,…,T,对每个训练集S_i用k均值聚类器进行聚类.由于聚类对象受到某种程度的扰动,将进一步增大聚类结果中个体标记向量之间的差异,这有助于获得更好的集成.另一方面,由于k均值聚类器只对原向量空间中的某一个局部进行了聚类,从而降低了聚类器的算法开销.需要注意的是,通过可重复取样技术产生的训练集S_i并不代表原向量集χ,故其并不代表全局的聚类结果.本文首先记录下k均值聚类器对S_i进行聚类后达到稳定状态时的k个聚类中心,然后用这k个聚类中心对原向量集χ中的元素进行重新分配,最后得到针对向量集χ的标记向量.

然而,通过这种方式得到的标记向量由于缺乏先验的类别信息,并不能直接用于下一阶段的结论合成.例如,对于标记向量$[1,2,2,1,1,3,3]^{\mathrm{T}}$和$[2,3,3,2,2,1,1]^{\mathrm{T}}$,虽然它们的表达方式不同,但是却表示着同一个聚类结果.所以为了对聚类结果进行结合,个体标记向量必须经过匹配建立相互之间的对应关系.一般来说,有对应关系的聚类标记所覆盖的相同对象的个数应该是最大的.因此,可以根据这一启发式来对标记向量进行配准.假设存在着两个标记向量$\lambda^{(a)}$和$\lambda^{(b)}$,每个标记向量分别把原向量集划分为k个聚类,分别用聚类标记$\{C_1^{(a)},C_2^{(a)},\ldots,C_k^{(a)}\}$和$\{C_1^{(b)},C_2^{(b)},\ldots,C_k^{(b)}\}$表示.首先,将这两个标记向量中每一对聚类标记$C_i^{(a)}$和$C_j^{(b)}$所覆盖的相同对象的个数记录在$k\times k$的OVERLAP矩阵中,然后选择其中覆盖相同对象个数最大的聚类标记建立对应关系,并将其结果从OVERLAP矩阵中移除.重复以上过程,直到所有聚类标记都建立了对应关系为止.

当存在$M(M>2)$个聚类标记向量时,则随机选取某个标记向量作为匹配基准,将其他标记向量和基准标记向量进行匹配.匹配算法只需要对$M-1$个标记向量做一次扫描,共需$(M-1)\times k^2$大小的存储空间来保存OVERLAP矩阵.整个匹配过程是快速和高效的.

在个体学习器的结合阶段,本文采用的是一种基于权值的选择性投票策略.在用于投票的个体标记向量的权值计算上,受Strehl和Ghosh工作[5]的启发,本文认为聚类标记向量间的互信息在某种程度上能够刻画聚类个体间的紧密程度,因此,利用互信息来表示个体标记向量的权值将有助于得到更好的集成结论.假设通过k均值聚类器对大小为n的向量集进行聚类,分别得到具有k个聚类的标记向量$\lambda^{(a)}$和$\lambda^{(b)}$,用聚类标记$\{C_1^{(a)},C_2^{(a)},\ldots,C_k^{(a)}\}$和$\{C_1^{(b)},C_2^{(b)},\ldots,C_k^{(b)}\}$来表示.假设聚类标记$C_i^{(a)}$中含有$n_i$个元素,聚类标记$C_j^{(b)}$中有$n_j$个元素,其中$C_i^{(a)}$和$C_j^{(b)}$中的相同元素有$n^{ij}$个.那么,互信息可定义如下:

$$\Phi^{NMI}\left(\lambda^{(a)},\lambda^{(b)}\right)=\frac{2}{n}\sum_{i=1}^{k}\sum_{j=1}^{k}n^{ij}\log_{k^2}\left(\frac{n^{ij}n}{n_in_j}\right) \tag{1}$$

对于个体标记向量,其平均互信息可用下式表示:

$$\beta_m=\frac{1}{t-1}\sum_{l=1,l\neq m}^{t}\Phi^{NMI}(\lambda^{(m)},\lambda^{(l)})\ (m=1,2,...,t) \tag{2}$$

当β_m越大时,标记向量$\lambda^{(m)}$所包含的其他标记向量所不具备的统计信息也就越少.所以,个体标记向量的权值可定义如下:

$$w_m=\frac{1}{\beta_mZ}\ (m=1,2,...,t) \tag{3}$$

其中,Z用于将权值规范化,使聚类标记的权值满足:

$$w_m>0(m=1,2,...,t)\text{ and }\sum_{m=1}^{t}w_m=1 \tag{4}$$

本文将前面所定义的个体标记向量的权值作为选择聚类集成个体标记向量的依据.当标记向量的权值低于某个预设阈值如$1/T$(T为集成中个体标记向量的个数)时,该个体标记向量将不参加最后的结论合成.最后,将挑选出的个体标记向量再基于权值进行投票.

基于Bagging的选择性聚类集成算法的伪码描述如下.

Input: number of cluster centers k, data set S, number of bootstap sampling T

Output: cluster label $C^*(x)$ for data object x

For t =1 to T Do

S_t = bootstrap sample from S

$\lambda^{(t)}$ = kmeans(k,S_t)

/* $\lambda^{(t)}$ is represented as $\{C_1^{(t)},C_2^{(t)},\ldots,C_k^{(t)}\}$ */

/* for data object x, its cluster label determined by $\lambda^{(t)}$ is $\lambda^{(t)}(x)\in\{C_1^{(t)},C_2^{(t)},\ldots,C_k^{(t)}\}$ */

End of For

$\lambda^{(baseline)}$ = randomly selected from $\{\lambda^{(t)}\}$,t=1,…,T

Delete $\lambda^{(baseline)}$ from $\{\lambda^{(t)}\}$, t=1,…,T

For each $\lambda^{(t)}$ in $\{\lambda^{(t)}\}$Do

For i=1 to k, j=1 to k Do

OVERLAP(i,j)=Count($C_i^{(t)}$,$C_j^{(baseline)}$)

End of For

/* OVERLAP is a $k\times k$ matrix; Count(A, B) counts the number of same */

/* elements labeled by A and B*/

$\Gamma=\varnothing$

While $\Gamma\neq\{C_1^{(baseline)}, C_2^{(baseline)},\ldots,C_k^{(baseline)}\}$ Do

(u,v)=argmax(OVERLAP(i,j)) /* OVERLAP(u, v) is the biggest element */

$C_v^{(t)}=C_u^{(baseline)}$ /* align $C_v^{(t)}$ to $C_u^{(baseline)}$ */

Delete OVERLAP(u,*)

Delete OVERLAP(*,v)

$\Gamma=\Gamma\cup\{C_v^{(t)}\}$

End of While

End of For

For t=1 to T Do

Calculate w_t for each $\lambda^{(t)}$

End of For

$$C^*(x)=\underset{(1\le t\le T)\wedge(w_t>1/T)}{\arg\max}\ w_t\lambda^{(t)}(x)$$

3 实验测试

本文除对基于 Bagging 的选择性聚类集成(sel-b-voting)算法进行了实验测试,还对基于简单投票的聚类集成(即随机初始化聚类中心产生个体标记向量,再进行简单投票)和基于加权投票的聚类集成(即随机初始化聚类中心产生个体标记向量,并根据相互间的互信息计算权值,然后进行基于权值投票)做了实验测试.

我们采用 UCI 机器学习数据库[7]中的 10 个数据集对上述聚类集成方法进行了实验.在这 10 个数据集中,除了类别属性以外,其余属性均为数值属性.其中,Image Segmentation 数据集中由于存在一列常数属性,对以后的聚类过程没有帮助而被删除.这 10 个数据集的具体信息见表 1.

当数据有分类信息时,可认为该分类信息在一定程度上表达了数据的一些内部分布特性,如果该分类信息没有被聚类过程所利用,则可以用它来评价聚类效果.在 Modha 和 Spangler[8]的工作中就是利用数据的分类信息来评价聚类结果的好坏.如果标记向量中某聚类标记和类别属性中某已知类别所覆盖的相同对象个数最多,则将该聚类标记对应为这个已知类别.这样,多个聚类标记可以对应同一个类别,而一个聚类标记则不能对应多个类别.通过这样的匹配,聚类结果就可以利用分类信息来进行评价.

Table 1 Data sets used in the experiments

表 1 实验所用数据集

Data set	No.attributes	No.classes	No.instances
Image segmentation	18	7	2 310
Ionosphere	34	2	351
Iris	4	3	150
Liver disorder	6	2	345
Page bocks	10	5	5 473
Vehicle	18	4	846
Waveform-21	21	3	5 000
Waveform-40	40	3	5 000
Wine	13	3	178
Wpbc	33	2	198

假设大小为 n 的向量集中存在的已知类别标记为 C,可用 $\{C_1,C_2,\ldots,C_h\}$ 来表示.通过 k 均值聚类器对该向量集进行聚类所得到的结果为一个具有 k 个聚类标记的标记向量 λ,可用 $\{\lambda_1,\lambda_2,\ldots,\lambda_k\}$ 来表示.通过上述的匹配过程,聚类标记向量中的每一个 λ_i 将对应为类别标记中的某一个 C_j.假设 α_i 为 λ_i 中被正确分类为对应类别 C_i 的示例个数,那么该聚类标记向量的结果可用 Micro-precision 进行衡量,具体的定义公式如下:

$$\text{micro-p}=\frac{1}{n}\sum_{i=1}^{k}a_i \tag{5}$$

Micro-precision 的值越大,则表示聚类的标记向量越好.需要注意的是,这种评价策略只能用于比较能产生固定聚类个数的聚类器性能,而不能用于评价聚类个数产生具有不确定性的聚类器.因此,在本节的实验中,k 均值聚类器的聚类个数固定设置为其已知的类别数.

实验中,k 均值聚类器的最大迭代步数设为 100,误差停止阈值为 1e-5.对于每个数据集,用上述所提到的 5 种聚类集成算法构造大小为 5,8,13,20 和 30 规模的集成.其中对于每一种测试都重复 10 次实验,然后记录下平均的 Micro-precision 和标准偏差.同时,单个 k 均值聚类器的结果也记录了下来,以便于和集成结果进行比较.

表 2 给出了在显著程度 0.05 时的双边 t 检验结果,其中 voting 表示基于简单投票的聚类集成算法,w-voting 表示基于加权投票的聚类集成,sel-b-voting 表示基于 Bagging 的选择性聚类集成.“win”表示聚类集成结果显著地比单个 k 均值聚类器的结果“好”,“loss”表示聚类集成结果显著地比单个 k 均值聚类器的结果“差”,而“tie”则表示聚类集成结果和单个 k 均值聚类器的结果没有显著差异.

Table 2 Results of pairwise two-tailed t tests with significance level 0.05

表 2 显著程度 0.05 时双边 t 检验的结果

Ensemble size	Voting win/tie/loss	W-Voting win/tie/loss	Sel-B-Voting win/tie/loss
5	1/7/2	2/6/2	3/5/2
8	0/8/2	3/6/1	4/5/1
13	2/6/2	2/6/2	6/3/1
20	1/6/3	3/4/3	5/4/1
30	1/5/4	3/4/3	5/4/1

由表 2 可以看出,基于简单投票的聚类集成在性能上比单个 k 均值聚类器要差;基于加权投票的聚类集成和单个 k 均值聚类器所得到的结果相差不大;而基于 Bagging 的选择性聚类集成则比单个 k 均值聚类器要好.特别是当集成的大小为 13 的时候,所有 10 个数据集中,选择性聚类集成所得到的结果仅在 1 个数据集上比单个 k 均值聚类器所得到的结果要差,而在 6 个数据集上,其结果比单个 k 均值聚类器所得到的结果要好.这说明选择性集成方法在非监督学习领域特别是对于聚类问题同样有效.

表 3 给出了不同集成规模下,选择性聚类集成实际使用的个体聚类器数目占所有个体聚类器数目的百分比.由表 3 可以看出,基于 Bagging 的选择性聚类集成仅使用了 28%~40%的个体聚类器参与构造最后的集成.

从以上的实验数据分析可以看出,选择性聚类集成仅利用了少数个体聚类器就达到了比利用所有个体聚类器更好的效果.

同时,实验中对于每一个数据集都构造了规模不等的聚类集成.为了解聚类集成的结果是否会受到集成规模变化的影响,本文也记录了相关的实验结果.图 1 给出了聚类集成的性能随着集成规模的大小而变化的关系图.其中,voting 表示基于简单投票的聚类集成算法,w-voting 表示基于加权投票的聚类集成,sel-b-voting 表示基

于 Bagging 的选择性聚类集成.

Table 3 The average percentage of component clusterers selected by the proposed algorithm under different ensemble sizes

表 3 在不同集成规模下,本文方法所选用的个体聚类器的平均百分比

Ensemble size	Percentage of component clusterers selected (%)
5	39.2 (1.96/5)
8	38.1 (3.05/8)
13	33.7 (4.38/13)
20	31.8 (6.36/20)
30	28.0 (8.41/30)

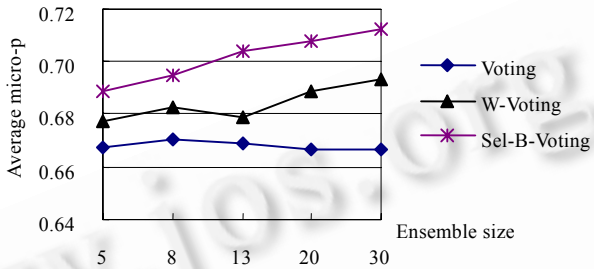

Fig.1 The influence of ensemble size on the performance of clusterer ensemble

图 1 集成规模变化对聚类集成的性能的影响

从图 1 可以看出,基于 Bagging 的选择性聚类集成随着集成规模的增大,其聚类性能也有所提高.而对基于简单投票的聚类集成,当集成的规模增大时,其性能反而有下降的趋势.产生这种现象的原因可能是由于随着集成规模的增大,将产生更多的个体标记向量覆盖整个向量空间,而并不是所有的标记向量都有助于提高集成的性能,其中某些个体可能存在着严重的误导性.图 1 也验证了 Zhou 等人[9]的结论,即当拥有一组个体学习器时,进行选择性集成比用所有个体集成更好.需要注意的是,图中横轴显示出的是个体聚类器的总数,由于基于 Bagging 的选择性聚类集成算法仅对挑选出的个体聚类器进行集成,其实际规模要远小于图 1 中显示的值.

4 结束语

本文提出了基于 Bagging 的选择性聚类集成算法.实验结果表明,基于 Bagging 的选择性聚类集成算法能有效地利用集成学习技术来提高聚类性能.由于选择性聚类集成只利用到了部分而不是所有聚类个体,这就降低了计算开销和存储开销.另外,由于该算法利用互信息计算权值的过程只对 k 均值聚类器的结果标记向量进行分析,这与选择什么样的聚类分析算法无关,所以本文提出的基于 Bagging 的选择性聚类集成算法同样也能适用于 k 均值外的其他聚类器.本文使用的个体聚类器均采用欧氏距离,由于不同的相似度距离对聚类结果有不同程度的影响,因此也可以尝试使用多种距离度量的个体聚类器进行集成.

值得注意的是,在本文工作进行时,将集成学习技术用于聚类的工作还非常少[5,9],但到目前已经出现了很多此类工作[10−12],这说明利用集成学习技术来改善聚类性能是一个新兴的研究热点.事实上,集成学习技术除了被用于监督学习和非监督学习,还被用于多示例学习[13,14].这一方面说明集成学习的效用已经受到了广泛的认可,其适用范围正逐渐扩大;另一方面也说明集成学习的研究空间还很大,尤其是将集成学习技术应用到监督学习之外的场合,还有很多重要的问题需要研究.

References:

[1] Estivill-Castro V. Why so many clustering algorithms—A position paper. SIGKDD Explorations, 2002,4(1):65−75.

[2] Dietterich TG. Machine learning research: Four current directions. AI Magazine, 1997,18(4):97−136.

[3] Breiman L. Bagging predicators. Machine Learning, 1996,24(2):123−140.

[4] Zhou ZH, Wu J, Tang W. Ensembling neural networks: Many could be better than all. Artificial Intelligence, 2002,137(1−2): 239−263.

[5] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse framework for combining partitionings. In: Dechter R, Kearns M, Sutton R, eds. Proc. of the 18th National Conf. on Artificial Intelligence. Menlo Park: AAAI Press, 2002. 93−98.

[6] MacQueen JB. Some methods for classification and analysis of multivariate observations. In: LeCam LM, Neyman J, eds. Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967,1:281−297.

[7] Blake C, Keogh E, Merz CJ. UCI Repository of machine learning databases. Irvine: Department of Information and Computer Science, University of California, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html

[8] Modha DS, Spangler WS. Feature weighting in *k*-means clustering. Machine Learning, 2003,52(3):217−237.

[9] Zhou ZH, Tang W. Clusterer ensemble. Technical Report, Nanjing: AI Lab., Department of Computer Science & Technology, Nanjing University, 2002.

[10] Fern XZ, Brodley CE. Random projection for high dimensional data clustering: A cluster ensemble approach. In: Fawcett T, Mishra N, eds. Proc. of the 20th Int'l Conf. on Machine Learning. Menlo Park: AAAI Press, 2003. 186−193.

[11] Fern XZ, Brodley CE. Solving cluster ensemble problems by bipartite graph partitioning. In: Greiner R, Schuurmans D, eds. Proc. of the 21st Int'l Conf. on Machine Learning. 2004. http://www.aicml.cs.ualberta.ca/ banff04/icml/pages/proceedings.htm

[12] Topchy A, Jain AK, Punch W. A mixture model for clustering ensembles. In: Berry MW, Dayal U, Kamath C, Skillicorn DB, eds. Proc. of the 4th SIAM Int'l Conf. on Data Mining. Philadelphia: SIAM, 2004. http://www.siam.org/meetings/sdm04/proceedings/index.htm

[13] Zhou ZH, Zhang ML. Ensembles of multi-instance learners. In: Lavrač N, Gamberger D, Blockeel H, Todorovski L, eds. Lecture Notes in Artificial Intelligence 2837, Berlin: Springer-Verlag, 2003. 492−502.

[14] Xu X, Frank E. Logistic regression and boosting for labeled bags of instances. In: Dai H, Srikant R, Zhang C, eds. Lecture Notes in Artificial Intelligence 3056, Berlin: Springer-Verlag, 2004. 272−281.