符号属性数据的半监督聚类与属性选择

王文涛

2014 年 6 月 10 日

聚类

聚类问题就是在没有任何数据的先验信息下对数据进行聚类分析，它是一种有效的分析数据结构的手段。

- 基于划分，基于层次，基于密度，基于网格

聚类

聚类问题就是在没有任何数据的先验信息下对数据进行聚类分析，它是一种有效的分析数据结构的手段。

- 基于划分，基于层次，基于密度，基于网格
- 符号属性数据聚类
  - 基于类型转换
  - 基于概率统计
  - 基于相异测度

属性选择

属性选择是指在初始属性集中选择出一个属性子集，可以像属性全集一样用来正确区分数据集中的每个数据对象。

- 过滤模型、封装模型和混合模型

属性选择

属性选择是指在初始属性集中选择出一个属性子集，可以像属性全集一样用来正确区分数据集中的每个数据对象。

- 过滤模型、封装模型和混合模型
- 属性评价方法：距离度量、信息度量、依赖性度量

属性选择

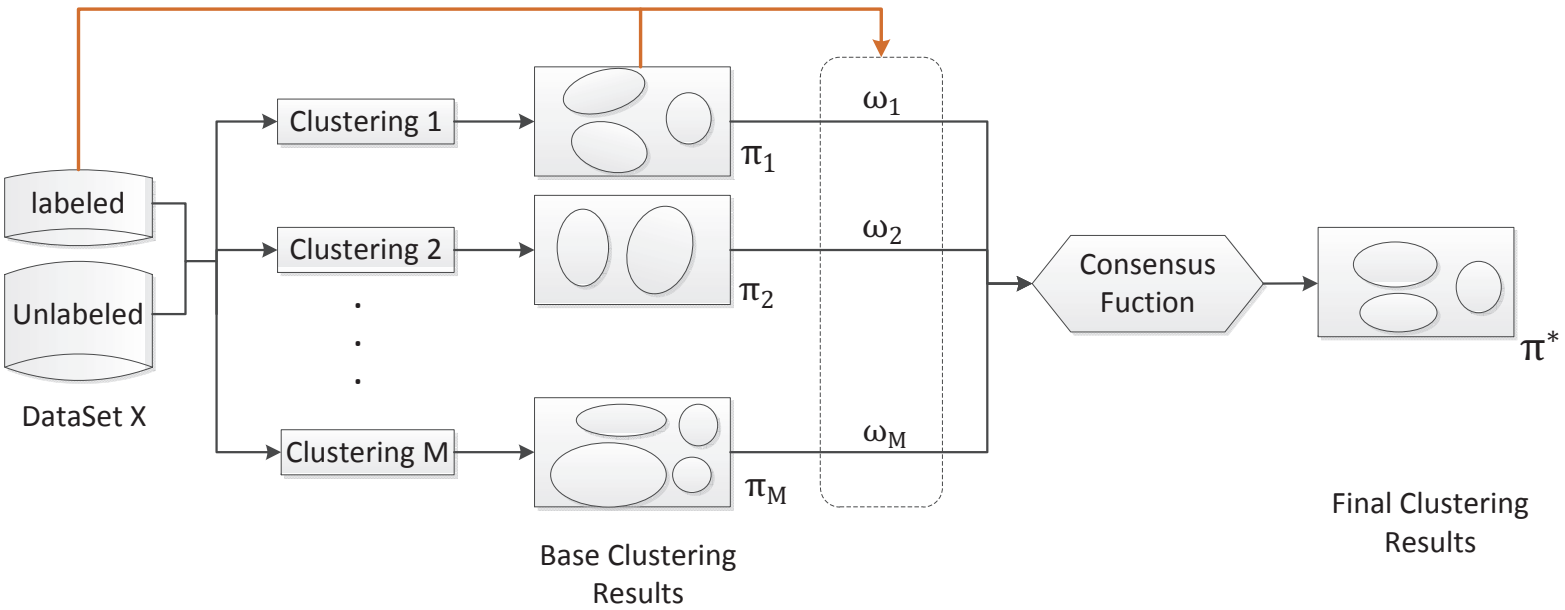
属性选择是指在初始属性集中选择出一个属性子集，可以像属性全集一样用来正确区分数据集中的每个数据对象。

- 过滤模型、封装模型和混合模型
- 属性评价方法：距离度量、信息度量、依赖性度量
- 搜索策略：启发式、穷尽式、随机式

符号属性半监督学习

针对无监督聚类和属性选择已经成为机器学习领域中的重要的研究方向，但是，半监督聚类和属性选择研究得相对较少，特别是符号属性数据，因此符号属性数据的半监督聚类和属性选择具有研究价值。

权值投票的半监督聚类集成



权值投票的半监督聚类集成算法

- 生成不同的聚类结果：k-Modes 算法初始点不同

权值投票的半监督聚类集成算法

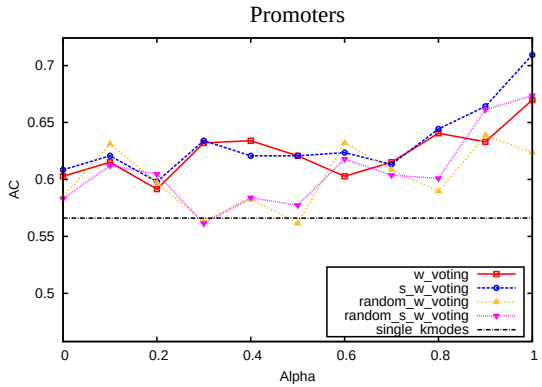
- 生成不同的聚类结果：k-Modes 算法初始点不同
- 权值产生：有监督部分和无监督部分 (NMI)

权值投票的半监督聚类集成算法

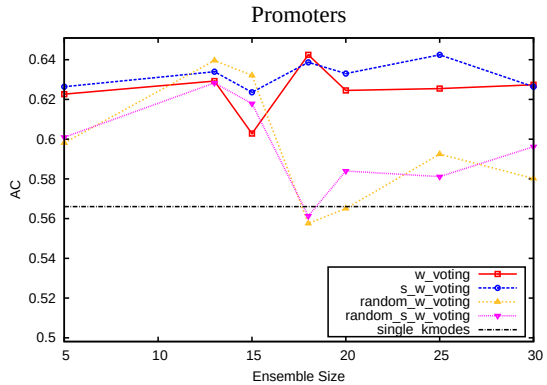
- 生成不同的聚类结果：k-Modes 算法初始点不同
- 权值产生：有监督部分和无监督部分 (NMI)

权值投票的半监督聚类集成算法

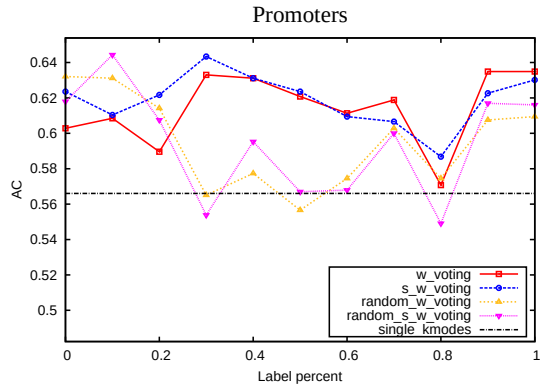
- 生成不同的聚类结果：k-Modes 算法初始点不同
- 权值产生：有监督部分和无监督部分 (NMI)
- 一致性函数：标签对齐，权值投票
  - W_Voting
  - S_W_Voting
  - Random_W_Voting
  - Random_S_W_Voting



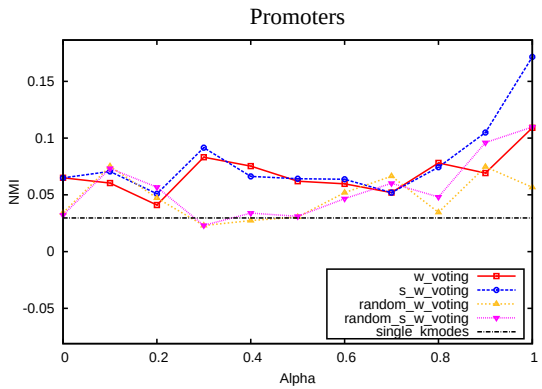
(a) ACC-Alpha



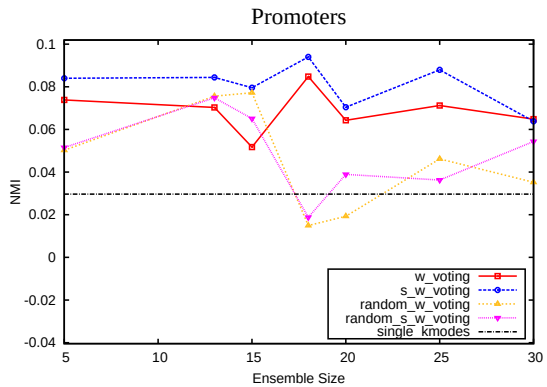
(b) ACC-Ensize



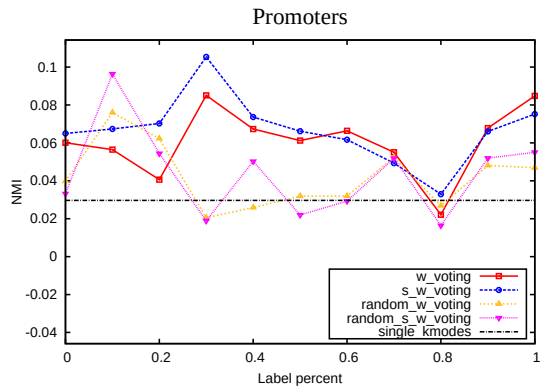
(c) ACC-Lpercent



(d) NMI-Alpha

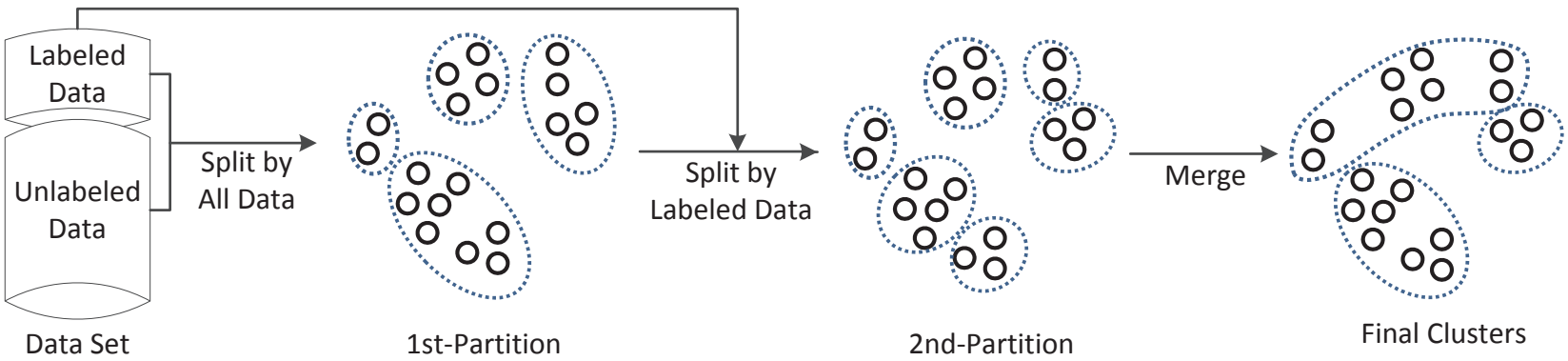


(e) NMI-Ensize



(f) NMI-Lpercent

基于分裂重组合的半监督聚类

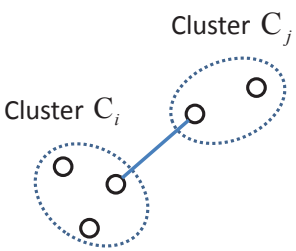


分裂组合策略

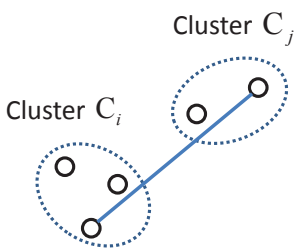
- 分裂策略：属性等价关系与类标等价关系形成划分

分裂组合策略

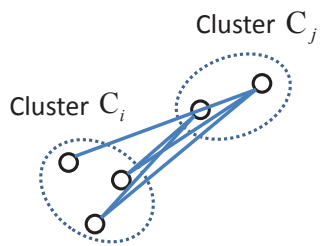
- 分裂策略：属性等价关系与类标等价关系形成划分
- 组合策略：



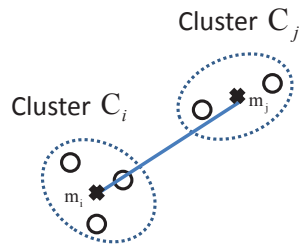
(k) Single



(l) Complete



(m) Average

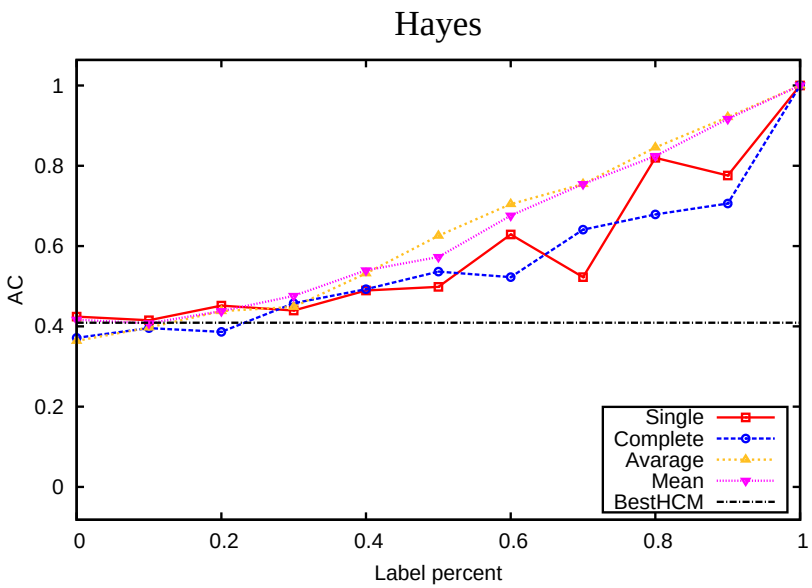


(n) Mean

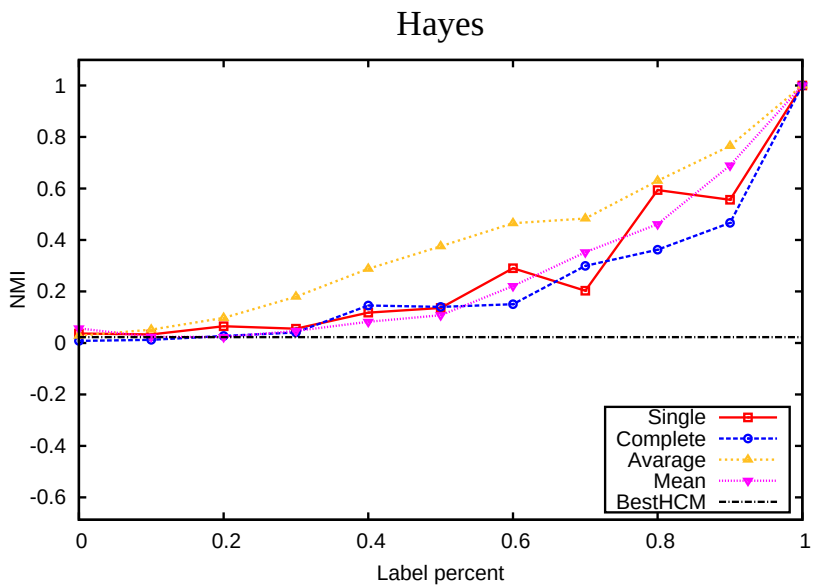
对象间距离函数

假设数据集 $X = X^L \bigcup X^U \in R^m$，其中 X^L 为有类标数据，X^U 为无类标数据，m 为符号属性的个数，则 $x_i, x_j \in X$ 的半监督差异测度定义为

$$d(x_i, x_j) \begin{cases} -m & x_i \in X^L \wedge x_j \in X^L \wedge d_i = d_j \\ m & x_i \in X^L \wedge x_j \in X^L \wedge d_i \neq d_j \\ \sum_{l=1}^{m} \delta(x_{il}, x_{jl}) & otherwise \end{cases} \quad (1)$$



(o) ACC-Lpercent



(p) NMI-Lpercent

最小冗余最大相关半监督属性选择

- 属性评价：

  假设数据集 $X = X^L \bigcup X^U$，$A = \{a_1, a_2, \ldots, a_m\}$ 为 m 个属性的集合，对于有类标数据 X^L，d 为决策属性，属性 a_i 的相关性与冗余性定义为

$$D(a_i) = MI^L(a_i, d)$$

$$R(a_i, S_m) = MI^U(a_i, S_m) + \sum_{a_j \in S_m} \frac{MI(a_i, a_j)}{|S_m|} \tag{2}$$

最小冗余最大相关半监督属性选择

- 属性评价：

  假设数据集 $X = X^L \bigcup X^U$，$A = \{a_1, a_2, \ldots, a_m\}$ 为 m 个属性的集合，对于有类标数据 X^L，d 为决策属性，属性 a_i 的相关性与冗余性定义为

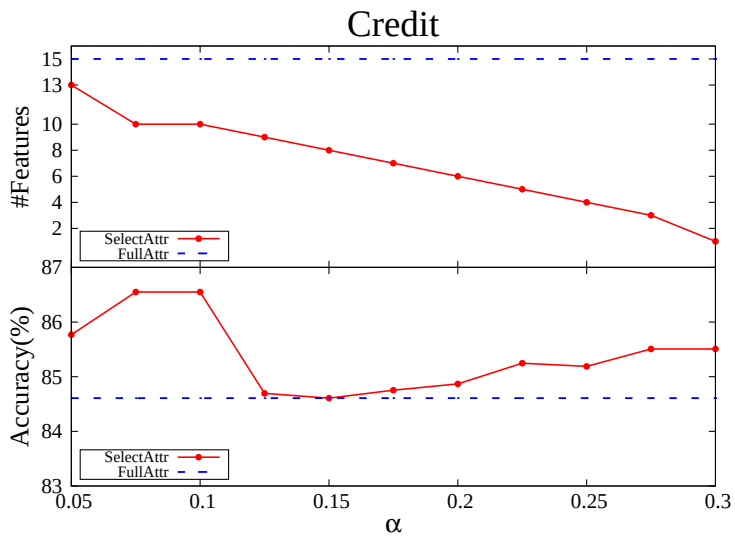
$$D(a_i) = MI^L(a_i, d)$$

$$R(a_i, S_m) = MI^U(a_i, S_m) + \sum_{a_j \in S_m} \frac{MI(a_i, a_j)}{|S_m|} \tag{2}$$

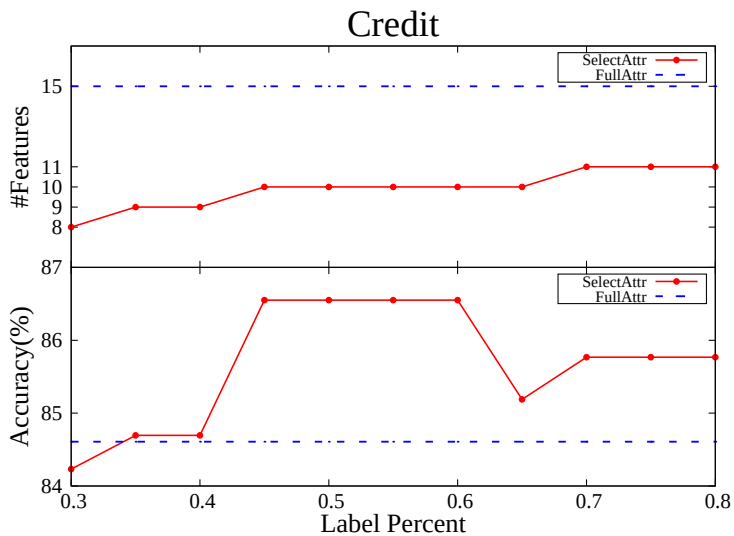
- 搜索策略：贪心式算法优化的约束条件为：

$$\max_{a_j \in \{A - S_{m-1}\}} \left[D(a_j) - R(a_j, S_{m-1}) \right] \tag{3}$$

- 半监督属性选择分类效果



(q) CART-Alpha



(r) CART-Lpercent

正域及依赖度

假设给定一个知识表达系统 $S =< U, A, V, f >$，$\forall X \subseteq U$ 和 $B \subseteq A$，子集 X 关于 B 的下近似和上近似分别是

$$\begin{aligned} \underline{B}(X) &= \{x | \forall x \in U, [x]_B \subseteq X\} \\ \overline{B}(X) &= \{x | \forall x \in U, [x]_B \cap X \neq \varnothing\} \end{aligned} \tag{4}$$

$POS_B(X) = \underline{B}(X)$ 称为 X 关于 B 的正域。

$$\gamma_B(D) = \frac{POS_B(D)}{|U|} = \frac{\left| \bigcup_{X \in U/D} \underline{B}(X) \right|}{|U|} \tag{5}$$

为 D 依赖于 B 的程度，记作 $B \Rightarrow_k D$

耦合依赖度的半监督属性选择算法

- 属性评价：

  假设给定一个知识表达系统 $S =< U, A, V, f >$， $a_i, a_j \subseteq A$，耦合依赖度，相关性及冗余性定义为：

$$daulDep(a_i, a_j) = \frac{POS_{a_i}(a_j) + POS_{a_j}(a_i)}{2|U|} \tag{6}$$

$$Rel(a_i, d) = daulDep^L(a_i, d) \tag{7}$$

$$Red(a_i, a_j) = daulDep(a_i, a_j) \tag{8}$$

耦合依赖度的半监督属性选择算法

- 属性评价：

  假设给定一个知识表达系统 $S =< U, A, V, f >$， $a_i, a_j \subseteq A$，耦合依赖度，相关性及冗余性定义为：

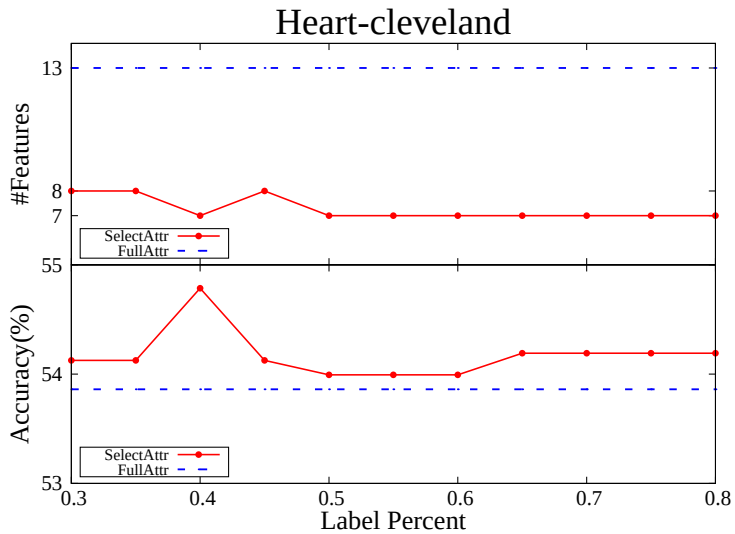
$$daulDep(a_i, a_j) = \frac{POS_{a_i}(a_j) + POS_{a_j}(a_i)}{2|U|} \tag{6}$$

$$Rel(a_i, d) = daulDep^L(a_i, d) \tag{7}$$

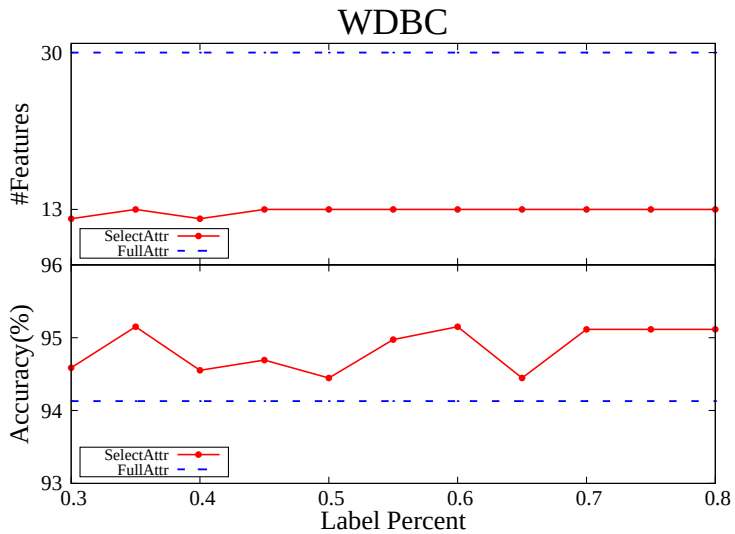
$$Red(a_i, a_j) = daulDep(a_i, a_j) \tag{8}$$

- 搜索策略：如果属性的相关性大于其对于已选属性中每个属性的冗余性，我们就将它加入的属性选择子集中，反之，就将它丢弃。

- 半监督属性选择分类效果

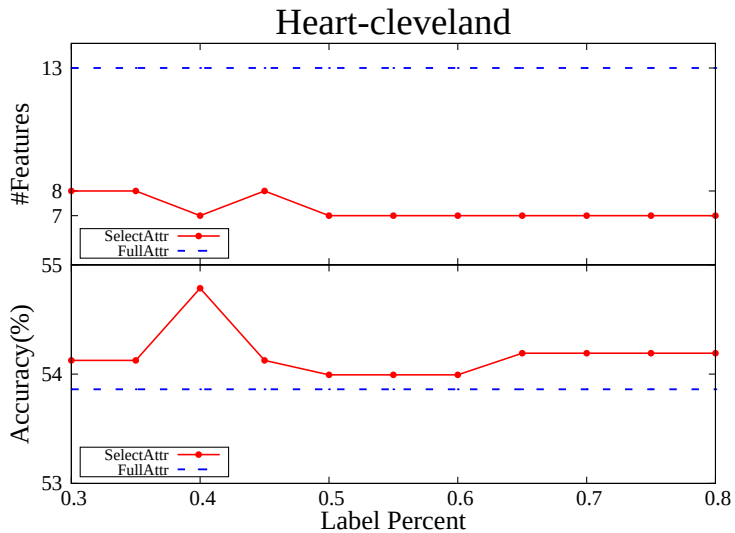


(s) CART-Lpercent

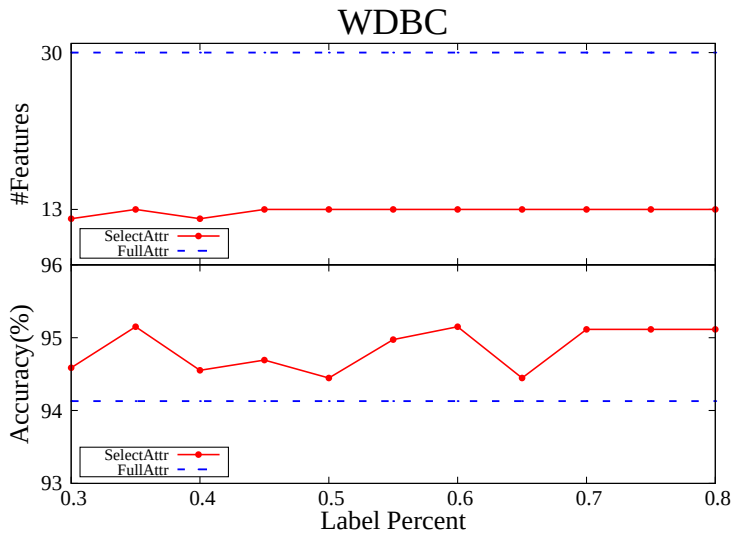


(t) CART-Lpercent

- 半监督属性选择分类效果



(u) CART-Lpercent



(v) CART-Lpercent

- 半监督聚类与属性选择结合验证：

  实验结果表明经过 SemiMRMR 算法和 DualPOS 算法属性选择后的数据大都能保持属性全集上的半监督聚类效果。

Thank you!

浙江大学计算机学院

E-mail:wwtzju@qq.com