

引言

- 实际应用中，标记会因为人为或设备的原因缺失或无法获取；
- 不完整的标记空间会导致特征和标记集相关性度量不准确，并且在特征选择过程中会丢失一些有价值特征；
- 基于AP聚类的填补方法聚类结果比较稳定，且互信息可以有效地进行特征选择。

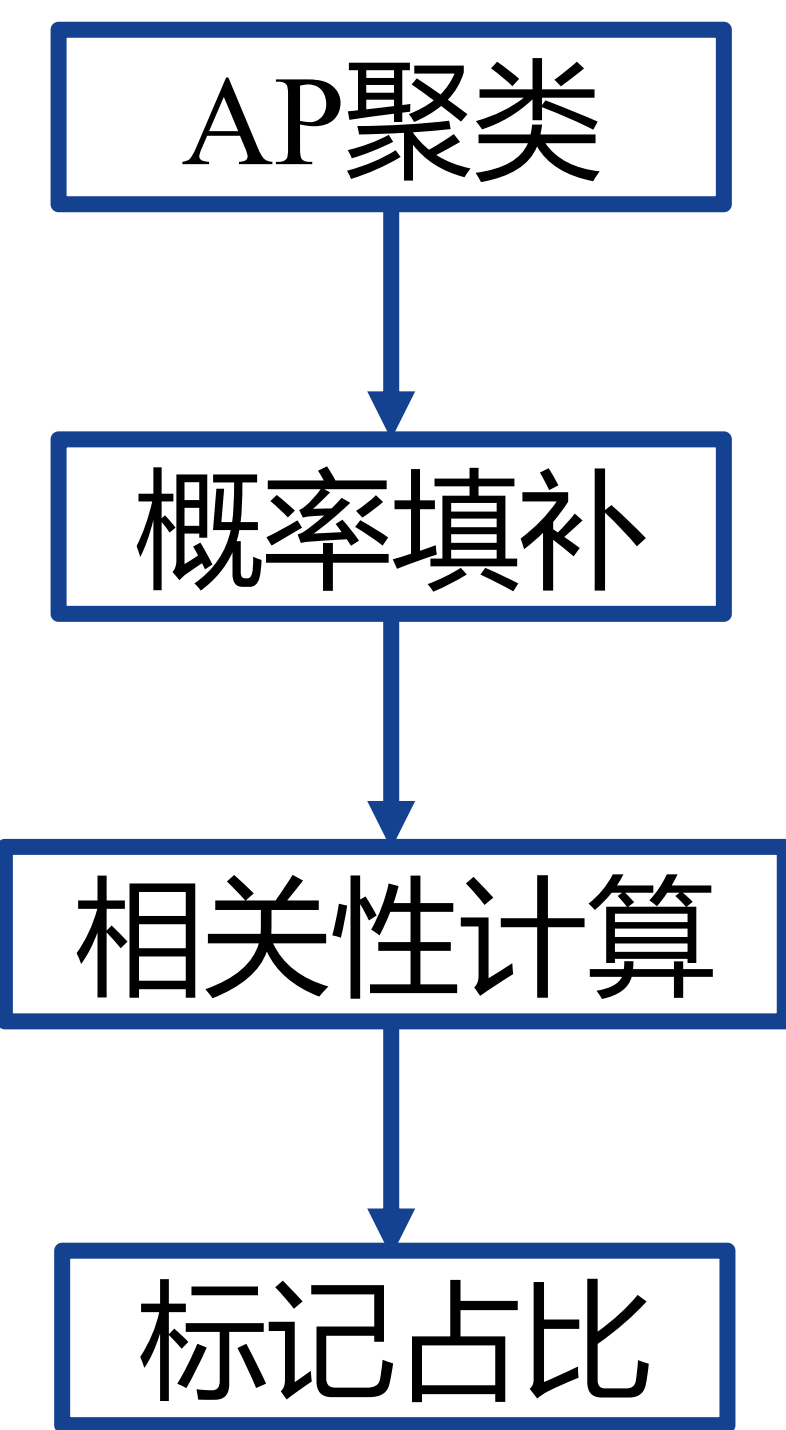
方法

- 本文借鉴AP聚类算法的优势，结合原有完整标记信息和样本相似性，构建概率填补公式，补全缺失标记；

$$s(x_i, x_k) = -\left\|x_i - x_k\right\|^2$$
$$pre_label = \sum \frac{s(x_i, x_m)}{s} \cdot x_m(l_j)$$

- 引入标记占比改进互信息公式，度量特征和标记集相关性，选择最优特征子集。

$$R_{pi, L} = \sum_{j=1}^L I(p_i; l_j) \cdot W(l_j)$$

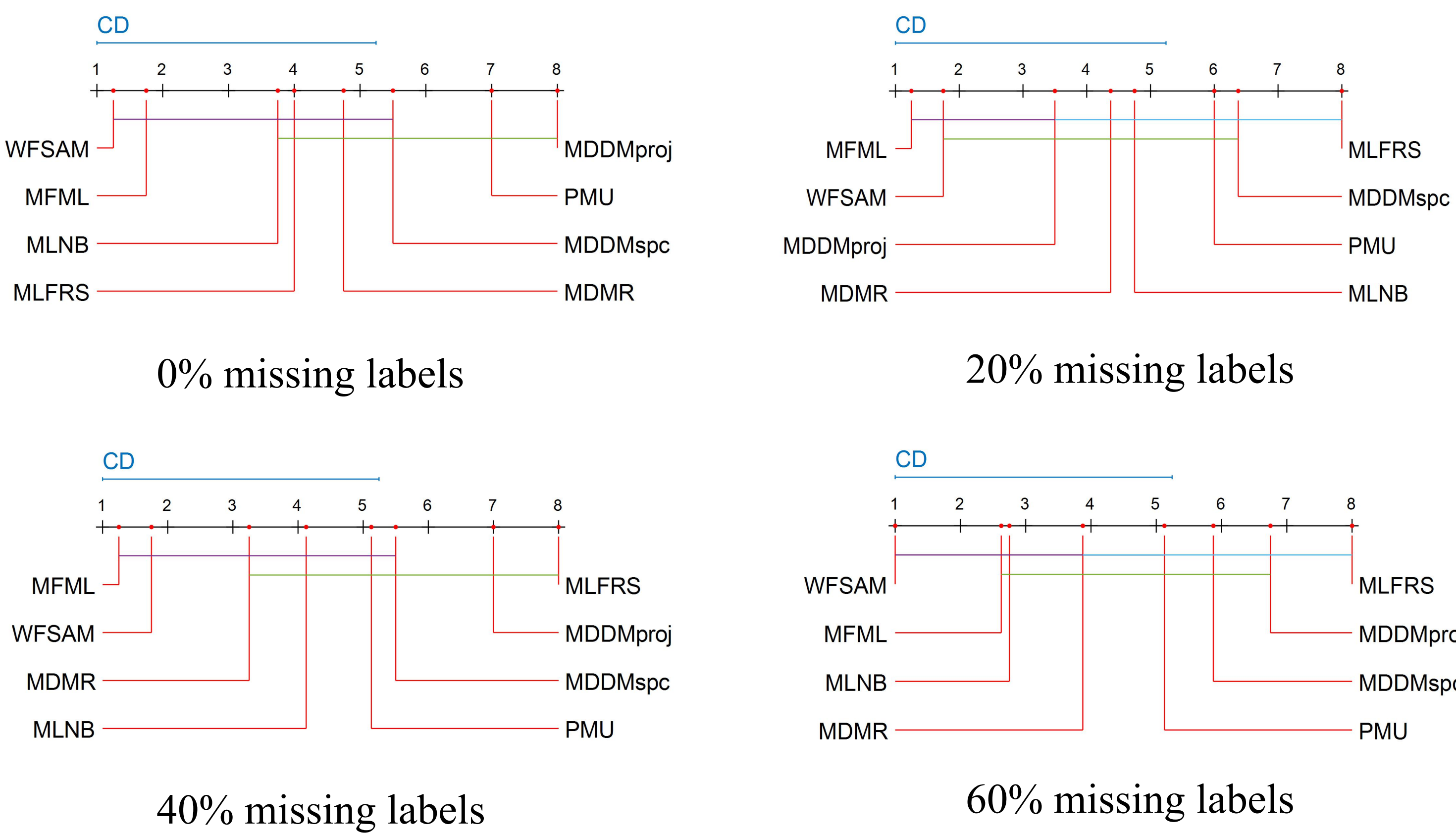
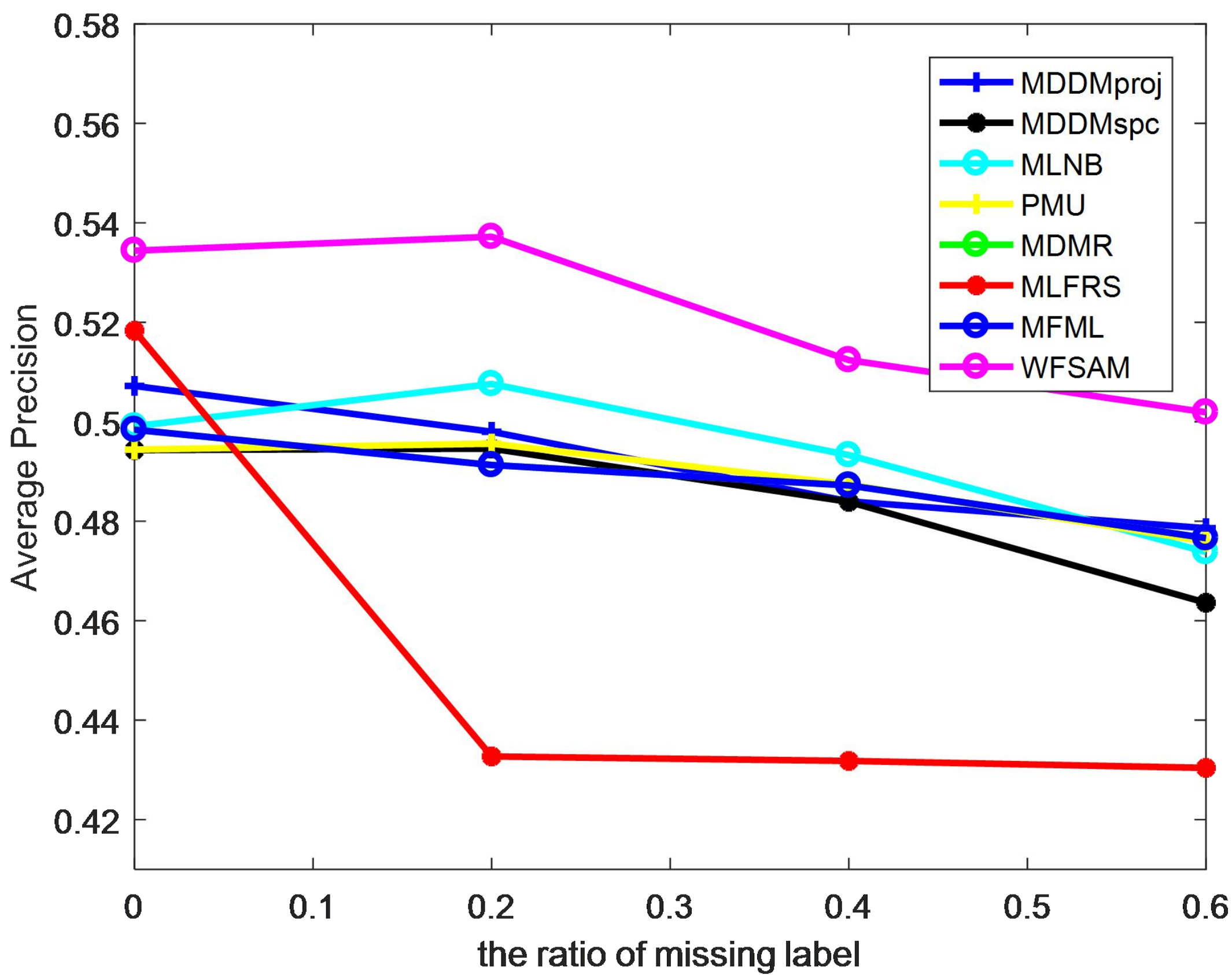


实验

- 分类器：ML-KNN (s=1 , n=10) ；
- 评价指标：AP , RL , CV , OE ；
- 特征子集：特征排序的前30% ；
- 标记缺失比率：0% , 20% , 40% , 60% ；
- 比较算法：MFML、PMU、MDMR、MLNB、MDDMspc、MDDMproj和MLFRS

数据集	实例	特征	标记
Arts	5000	462	26
Computers	5000	681	33
Entertainment	5000	640	21
Enron	1702	1001	53
Recreation	5000	606	22
Science	5000	743	40

结果



结论

- 当标记缺失比率为0%，即具有完整标记空间时，本文所提WFSAM算法分类性能最优；
- 当标记缺失比率为20%、40%和60%时，虽然在Computers和Enron数据集上表现不佳，但在其他数据集上分类性能，以及所有数据集的平均性能而言，WFSAM算法性能最佳；
- 对所有比较算法进行统计分析，在不同标记缺失比率下，WFSAM与MFML算法分类效果大致相同，但优于其余6种比较算法。