

可解释的序列分类与聚类算法

开题报告

董俊杰

22217058

软件学院, 大连理工大学

指导教师: 何增有

November 9, 2023

① 研究介绍

研究背景
研究动机

② 相关工作

序列的分类与聚类

③ 研究路线

序列分类
序列聚类

④ 时间规划

⑤ 参考文献

参考文献

大纲

① 研究介绍

研究背景

研究动机

② 相关工作

序列的分类与聚类

③ 研究路线

序列分类

序列聚类

④ 时间规划

⑤ 参考文献

参考文献

离散序列

Definition 1.1: 离散序列

离散序列是指由离散符号组成的列表，如 DNA 序列、蛋白质序列、文本序列、行为模式序列等。

Example 1.1: 人类行为数据库

A = 饮酒, B = 服用, C = 睡觉, D = 锻炼

- ACACACACAAC → 生存
- BCDCAC → 生存
- ABCDAC → 死亡
- DABCD → 死亡
- BCDCA → 生存

重要模式: $\langle AB \rangle$

Example 1.2: 基因数据库

A = 腺嘌呤, C = 胞嘧啶, G = 鸟嘌呤, T = 胸腺嘧啶

- ACGTACGTACGT → 基因
- ACGTACG → 致病基因
- TACGTAC → 致病基因
- GCTAC → 基因

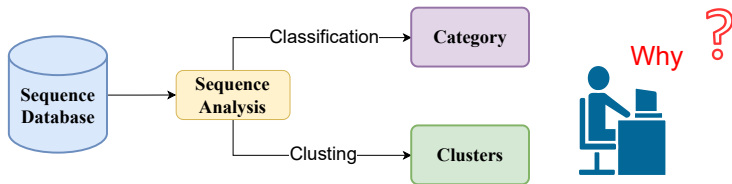
重要模式: $\langle AC \rangle$

分析序列可以帮助我们理解序列的功能，从而做出科学决策或者发现核心规律。

研究动机

两种主要分析策略:

- 序列分类
- 序列聚类



E.g., Why is this gene belong to this famliy?

Figure 1: 可解释的算法旨在提供尽可能透明的决策流程，这对特定科学研究中**至关重要**

然而，现有框架**存在以下问题**:

分类:

- CNN, LSTM, Transformer 等算法**不具备可解释性**
- 从模式出发的算法 **性能不足**

聚类:

- 各特征之间的复杂耦合关系**难以解释**
- **没有**可解释的序列聚类算法

大纲

① 研究介绍

研究背景

研究动机

② 相关工作

序列的分类与聚类

③ 研究路线

序列分类

序列聚类

④ 时间规划

⑤ 参考文献

参考文献

序列的分类与聚类

现有的序列分类与聚类算法主要分支:

- 基于模式的序列分类:

- ① 从连续子串出发

- 直接使用
- 筛选^[1]

- ② 将子序列作为模式^[2]

- 提升模式质量: 减少冗余、提高精度^[3-9]
- 保证判别力: 与分类器耦合挖掘^[10-13]

- ③ 字符串核方法^[14-15]

- ④ 序列映射^[16-17]

缺点:

- 难以挖掘组合型判别模式
- 字符串核与序列映射方法可解释性差

- 序列聚类:

- ① 模式转特征向量

- 转换后可通用的使用数据聚类方法^[16,18-21]

- ② 基于模型的方法

- MM^[22-25]、HMM^[26-27]

- ③ 基于层次的方法

- 编辑距离^[28]
- Jaccard 相似度等^[29-30]

- ④ 直接划分

- 在原始空间直接划分^[31-32]

- 算法无法直接提供可解释的决策流程
- 簇划分结果可信度差

大纲

① 研究介绍

研究背景

研究动机

② 相关工作

序列的分类与聚类

③ 研究路线

序列分类

序列聚类

④ 时间规划

⑤ 参考文献

参考文献

序列分类：现有方法精度不高

在序列分类中，从模式出发是最具有可解释性的策略，

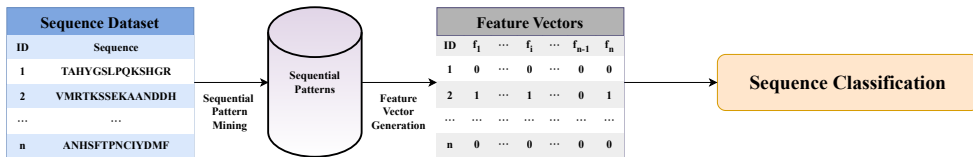


Figure 2: 经典策略：从序列模式出发的序列分类方法

- 1 挖掘序列模式过程中很难找到具有高判别力的模式组合
- 2 即使和分类器耦合挖掘，也只能启发式的挖掘部分组合型模式
- 3 利用存在性函数度量转换特征，对模式的利用率低

需要挖掘高判别力的序列模式集！

序列分类：借助自动编码器

- 神经网络可以自动学习一组用于分类的模式组合，但是这些模式是难以解释的
 - CNN 的权重对人来说只是一组数字，很难解释其含义
 - 即使经过量化，依然是一组 1 和-1 或者 0 和 1 的组合，难以解释
- 然而可以理论上可以通过特殊的量化函数将其转换为可解释的模式集合
 - 量化策略可以将 CNN 的权重转化为序列模式
 - 通过优化 CNN 权重寻找具有判别力的序列模式组合

核心路线：

寻找可解释的 CNN 量化策略，通过优化 CNN 权重寻找具有判别力的序列模式组合。

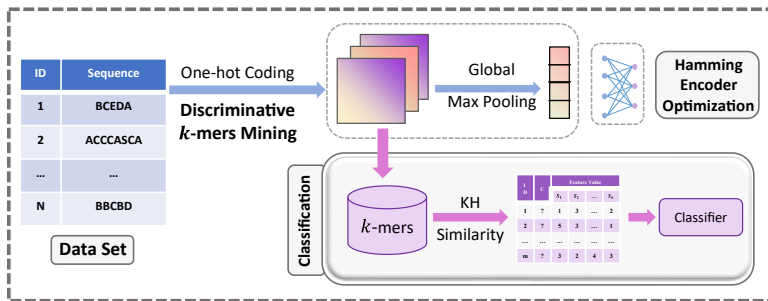


Figure 3: 从可解释的 CNN 权重出发寻找判别模式的算法框架

序列分类：带量化的 1DCNN 编码器

一个训练完的 $sign$ 量化 (+1, -1) 的 1DCNN 预测时前向计算如下:

$$sign(x) = \begin{cases} +1, & \text{if } x \geq 0 \\ -1, & \text{otherwise.} \end{cases} \quad (1)$$

① 将序列编码为 one-hot 向量 $\mathbf{a}$:

$$\mathbf{a}_{i,j} = \begin{cases} 1, & \text{if } s_j \text{ is the } i\text{-th item of } I, \\ 0, & \text{Otherwise.} \end{cases} \quad (2)$$

② 权重为 $\mathbf{w}$ 的 1D 卷积层:

$$f_{\mathbf{w}_b}(\mathbf{a}) = sign(\mathbf{w}) \otimes \mathbf{a}. \quad (3)$$

③ 通过最大池化层得到特征向量 $\mathbf{z}$:

$$\mathbf{z}_i = \max_j (f_{\mathbf{w}_b}(\mathbf{a})_{i,j}). \quad (4)$$

④ 通过全连接层得到最终的分类结果:

$$\mathbf{y} = \mathbf{z}\mathbf{w}_{fnn}^T. \quad (5)$$

序列分类：可解释量化的 Hamming 编码器

判别模式集合的挖掘过程伪代码如下：

Algorithm 1 Hamming Encoder Mining

Input: A sequence database D , initial weight $\mathbf{w}^1$, number of epochs e_{max} , batchsize N , itemset I .

Output: Discriminative k -mer set P .

1: $t \leftarrow 1$ ▷ Step

2: $\mathbf{a} \leftarrow \text{one-hot}(I, D)$

3: for $e = 1 \dots e_{max}$ do ▷ Epochs

4: for $i = 1 \dots n$ do ▷ For each batch

5: //Forward pass

6: $\mathbf{w}_{bcnn}^t \leftarrow \mathbf{Q}(\mathbf{w}_{cnn}^t)$; ▷ Binarized weights

7: $\mathbf{z} \leftarrow \mathbf{w}_{bcnn}^t \otimes \mathbf{a}_i$; ▷ Convolution

8: $\mathbf{k} \leftarrow \text{GlobalMaxPooling}(\mathbf{z})$

9: $\mathbf{y} \leftarrow \mathbf{k} \mathbf{w}_{fnn}^T$ ▷ FNN layer

10: $\mathcal{L} \leftarrow \text{loss}(D[i]; \mathbf{y})$; ▷ Compute loss

11: //Backward pass

12: $g_{\mathbf{w}_{cnn}} \leftarrow \text{backward}(\mathcal{L})$

13: $\mathbf{w}_{cnn}^{t+1} \leftarrow \text{update}(\mathbf{w}_{cnn}^t, g_{\mathbf{w}_{cnn}})$; ▷ Optimize

14: end for

15: end for

16: $P \leftarrow \text{inverse one-hot}(I, \mathbf{w}_{bcnn})$ ▷ Extract k -mers

17: Output: P

核心二值化策略：

$$\mathbf{Q}(\mathbf{w})_{i,j} = \begin{cases} 1, & \text{if } i = \arg \max_k \mathbf{w}_{k,j} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

序列分类: Hamming 编码器

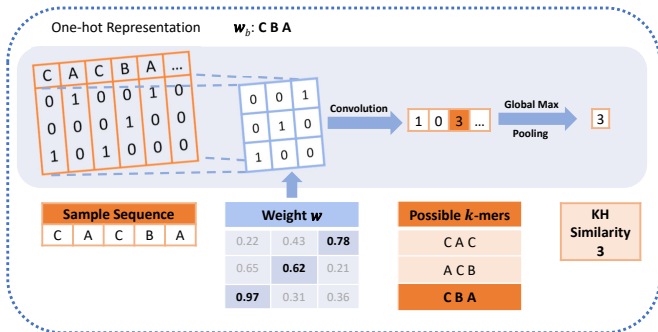


Figure 4: Hamming 编码器的计算细节

序列相似度与编码器输出的等价性关系:

$$\text{Sim}(p, s) = k - \min_{t \in S_k(s)} \{H(t, p)\}, H(t, p) = \sum_{i=1}^k (t[i] \oplus p[i]), \quad (7)$$

$$\text{Sim}(p, s) = \max_j (f_{w_b}(\mathbf{a})_{i,j}), f_{w_b}(\mathbf{a}) = \mathbf{w}_b \otimes \mathbf{a}. \quad (8)$$

序列分类: Hamming 编码器的端到端性能比较

Table 1: Hamming 量化与基准 CNN、Heaviside、Sign 量化的分类准确率比较

Datasets	Hamming	Baseline CNN	Heaviside	Sign
aslbu	0.633	0.446	0.524	0.552
auslan2	0.277	0.233	0.277	0.275
context	0.933	0.943	0.941	0.943
epitope	0.896	0.945	0.935	0.945
gene	1.000	1.000	1.000	1.000
pioneer	0.985	0.891	0.785	0.830
question	0.954	0.975	0.950	0.948
robot	0.904	0.974	0.967	0.969
skating	0.351	0.351	0.345	0.336
reuters	0.979	0.976	0.970	0.977
unix	0.922	0.932	0.932	0.932
Average Acc	0.803	0.788	0.784	0.790

序列分类：Hamming 编码器挖掘的模式质量评估

Table 2: Hamming 编码器与其他特征提取方法的平均准确率比较

Method	Hamming Encoder	MiSeRe	Sqn2VecSEP	Sqn2VecSIM	iBCM	Classifier	SCIP
SVM	0.810	0.782	0.745	0.757	0.763	SCII_HAR	0.681
NB	0.738	0.721	0.701	0.713	0.721	SCII_MA	0.678
KNN	0.792	0.777	0.777	0.776	0.707	SCIS_HAR	0.719
DT	0.765	0.764	0.659	0.658	0.738	SCIS_MA	0.720

序列聚类：构建一个小而精的模式决策树

在数值聚类领域，已经有多种方法被提出来构建可解释的聚类模型^[33]：

① 模糊规则方法^[34-36]

Rule R_i : IF x_1 is ϕ_{11} and ... and x_n is ϕ_{jj} THEN cluster c_i .

- 自适应网格方法^[37]
- 固定网格方法^[38-39]

② 多面体方法

类似 SVM，通过构建一个多面体来划分簇

- 矩形方法^[37-38]
- 多面体方法^[39]

③ 阈值树方法^[40-46]

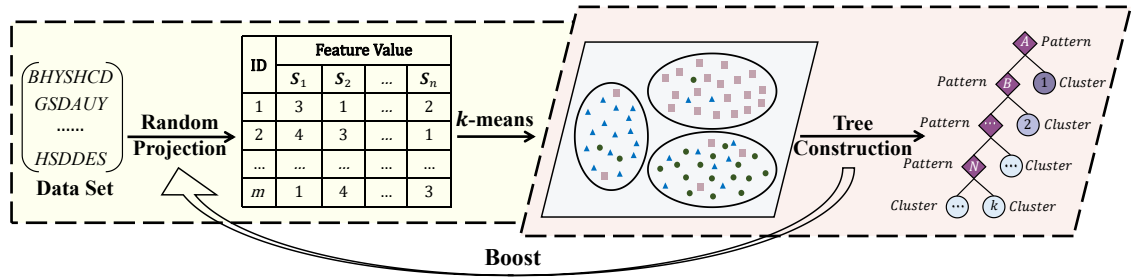
有研究表明，可以通过 $k-1$ 个阈值来划分 k 个簇，这样可以获得一个精简而直观的聚类决策树

- 直接构建^[40-42]
- 基于伪标签^[43-46]

核心路线：

挖掘 $k-1$ 个 **Top-1** 簇上的判别模式，构建小而精的可解释序列聚类树。

序列聚类：可解释的序列聚类树的构建策略



构建流程:

- 1 快速随机投影
- 2 伪标签获得
- 3 判别模式发现
- 4 裂分模式树

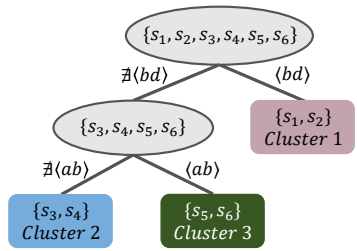


Figure 5: 可解释模式树的构建策略以及示例

序列聚类：对比性能比较

Table 3: ISCT 与其他聚类算法的性能比较

Datasets	Evaluation	ISCT	HC	FB-LL	MCSC
Average (Rank)	Purity	0.584 (2.5)	0.421 (6.43)	0.499 (4.43)	0.474 (4.64)
	NMI	0.33 (2.14)	0.077 (6.07)	0.2 (3.43)	0.081 (5.5)
	F1-score	0.525 (2.5)	0.478 (2.93)	0.476 (3.21)	0.367 (5.79)
Datasets	Evaluation	k-Median	MinDL	kkmeans	SGT
Average (Rank)	Purity	0.52 (3.71)	0.567 (3.21)	0.577 (2.71)	/
	NMI	0.185 (4)	0.151 (3.79)	0.274 (2.93)	/
	F1-score	0.455 (4.14)	0.303 (5.93)	0.496 (3.5)	/

大纲

① 研究介绍

研究背景
研究动机

② 相关工作

序列的分类与聚类

③ 研究路线

序列分类
序列聚类

④ 时间规划

⑤ 参考文献

参考文献

- ① 第 1 个月—第 3 个月：阅读相关文献，学习基于特征的序列分类方法以及现有的可解释性分类方法，编写代码并验证。学习现有的应用与数据挖掘的二值化策略，探究如何约束并限制 CNN 权重的学习序列模式，从而保证该权重可以被“翻译”为序列模式。
- ② 第 3 个月—第 6 个月：建立序列模式和在 1DCNN 中前向计算的输出之间的关系，从而可以保证得到的序列模式在转化为特征向量后依然具有高判别能力并且提高 1DCNN 的可解释性。完成论文的初稿。
- ③ 第 6 个月—第 9 个月：对其他的序列聚类算法进行调研。研究树结构的序列聚类算法，探究如何将序列模式转换为树结构，并给出一个人类可以理解的具有高度可解释性的聚类结果。
- ④ 第 9 个月—第 12 个月：编写代码实现并实验验证对比，评估各个算法的性能表现，并分析其优缺点。并且对算法的可解释性进行分析，探究其背后的原理和特点。最后，撰写论文，详细总结本研究的贡献和创新点，并提出未来的研究方向和挑战。

大纲

① 研究介绍

研究背景

研究动机

② 相关工作

序列的分类与聚类

③ 研究路线

序列分类

序列聚类

④ 时间规划

⑤ 参考文献

参考文献

参考文献 I

- [1] CHUZHANOVA N A, JONES A J, MARGETTS S. Feature selection for genetic sequence classification.[J]. *Bioinformatics*, 1998, 14(2).
- [2] LESH N, ZAKI M J, OGIHARA M. Mining features for sequence classification[C]. in: *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1999: 342-346.
- [3] JI X, BAILEY J, DONG G. Mining Minimal Distinguishing Subsequence Patterns with Gap Constraints[J]. *Knowledge and Information Systems*, 2007, 11(3): 259-286.
- [4] PEI J, HAN J, WANG W. Constraint-based sequential pattern mining: the pattern-growth methods[J]. *Journal of Intelligent Information Systems*, 2007, 28(2): 133-160.
- [5] SRIKANT R, AGRAWAL R. Mining sequential patterns: Generalizations and performance improvements[C]. in: *Proceedings of the Advances in Database Technology—EDBT'96: 5th International Conference on Extending Database Technology*. 1996: 1-17.
- [6] ZHOU C, CULE B, GOETHALS B. Pattern Based Sequence Classification[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(5): 1285-1298.
- [7] EXARCHOS T P, TSIPOURAS M G, PAPALOUKAS C, et al. A two-stage methodology for sequence classification based on sequential pattern mining and optimization[J]. *Data & Knowledge Engineering*, 2008, 66(3): 467-487.
- [8] HE Z, ZHANG S, WU J. Significance-Based Discriminative Sequential Pattern Mining[J]. *Expert Systems with Applications*, 2019, 122: 54-64.

参考文献 II

- [9] LO D, CHENG H, HAN J, et al. Classification of software behaviors for failure detection: a discriminative pattern mining approach[C]. in: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2009: 557-566.
- [10] HE Z, WU Z, XU G, et al. Decision Tree for Sequences[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(1): 251-263.
- [11] IFRIM G, WIUF C. Bounded coordinate-descent for biological sequence classification in high dimensional predictor space[C]. in: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2011: 708-716.
- [12] IFRIM G, BAKIR G, WEIKUM G. Fast logistic regression for text categorization with variable-length n-grams[C]. in: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2008: 354-362.
- [13] OKANOHARA D, TSUJII J. Text categorization with all substring features[C]. in: *Proceedings of the 2009 SIAM International Conference on Data Mining*. 2009: 838-846.
- [14] LESLIE C, ESKIN E, WESTON J, et al. Mismatch string kernels for SVM protein classification[C]. in: *NIPS*: vol. 15. 2002: 1441-1448.
- [15] LODHI H, SAUNDERS C, SHAW-ETAYLOR J, et al. Text classification using string kernels[J]. *Journal of Machine Learning Research*, 2002, 2(Feb): 419-444.

参考文献 III

- [16] RANJAN C, EBRAHIMI S, PAYNABAR K. Sequence graph transform (SGT): a feature embedding function for sequence data mining[J]. Data Mining and Knowledge Discovery, 2022, 36(2): 668-708.
- [17] NGUYEN D, LUO W, NGUYEN T D, et al. Sqn2vec: Learning sequence representation via sequential patterns with a gap constraint[C]. in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2018: 569-584.
- [18] YUAN L, WANG W, CHEN L. Two-stage pruning method for gram-based categorical sequence clustering[J]. International Journal of Machine Learning and Cybernetics, 2019, 10: 631-640.
- [19] GURALNIK V, KARYPIS G. A scalable algorithm for clustering sequential data[C]. in: Proceedings of the 2001 IEEE International Conference on Data Mining. 2001: 179-186.
- [20] LI C, DONG X, LIU W, et al. Ssrdivis: Interactive visualization for event sequences summarization and rare detection[J]. Journal of Visualization, 2020, 23: 171-184.
- [21] LI C, YANG Q, WANG J, et al. Efficient mining of gap-constrained subsequences and its various applications [J]. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1): 1-39.
- [22] SOCIETY T X, WANG S, JIANG Q, et al. A novel variable-order Markov model for clustering categorical sequences[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(10): 2339-2353.
- [23] XIONG T, WANG S, JIANG Q, et al. A new markov model for clustering categorical sequences[C]. in: Proceedings of the 2011 IEEE 11th International Conference on Data Mining. 2011: 854-863.

参考文献 IV

- [24] YANG J, WANG W. CLUSEQ: Efficient and effective sequence clustering[C]. in: Proceedings of the 19th International Conference on Data Engineering. 2003: 101-112.
- [25] CADEZ I, HECKERMAN D, MEEK C, et al. Model-based clustering and visualization of navigation patterns on a web site[J]. Data Mining and Knowledge Discovery, 2003, 7: 399-424.
- [26] SMYTH P. Clustering sequences with hidden Markov models[J]. Advances in Neural Information Processing Systems, 1996, 9.
- [27] BICEGO M, MURINO V, FIGUEIREDO M A. Similarity-based clustering of sequences using hidden Markov models[C]. in: Proceedings of the Third International Conference on Machine Learning and Data Mining in Pattern Recognition. 2003: 86-95.
- [28] BOSE R J C, VAN DER AALST W M. Context aware trace clustering: Towards improving process mining results[C]. in: Proceedings of the 2009 SIAM International Conference on Data Mining. 2009: 401-412.
- [29] RAMONI M, SEBASTIANI P, COHEN P. Bayesian clustering by dynamics[J]. Machine Learning, 2002, 47: 91-121.
- [30] OH S J, KIM J Y. A hierarchical clustering algorithm for categorical sequence data[J]. Information Processing Letters, 2004, 91(3): 135-140.
- [31] CHEN Y, XU P, REN L. Sequence synopsis: Optimize visual summary of temporal event data[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 24(1): 45-55.

参考文献 V

- [32] DINU L P, IONESCU R T. Clustering based on median and closest string via rank distance with applications on DNA[J]. *Neural Computing and Applications*, 2014, 24: 77-84.
- [33] YANG H, JIAO L, PAN Q. A survey on interpretable clustering[C]. in: *Proceedings of the 40th Chinese Control Conference (CCC)*. 2021: 7384-7388.
- [34] WANG X, LIU X, ZHANG L. A rapid fuzzy rule clustering method based on granular computing[J]. *Applied Soft Computing*, 2014, 24: 534-542.
- [35] MANSOORI E G. FRBC: A fuzzy rule-based clustering algorithm[J]. *IEEE Transactions on Fuzzy Systems*, 2011, 19(5): 960-971.
- [36] HSIEH C H, YAN C H, MAO C H, et al. GMiner: Rule-based fuzzy clustering for Google Drive behavioral type mining[C]. in: *Proceedings of the 2016 International Computer Symposium (ICS)*. 2016: 98-103.
- [37] PELLEG D, MOORE A W. Mixtures of Rectangles: Interpretable Soft Clustering[C]. in: *Proceedings of the International Conference on Machine Learning*. 2001: 401-408.
- [38] CHEN J, CHANG Y, HOBBS B, et al. Interpretable clustering via discriminative rectangle mixture model[C]. in: *Proceedings of the 2016 IEEE 16th International Conference on Data Mining*. 2016: 823-828.
- [39] LAWLESS C, KALAGNANAM J, NGUYEN L M, et al. Interpretable clustering via multi-polytope machines[C]. in: *Proceedings of the AAAI Conference on Artificial Intelligence*: vol. 36: 7. 2022: 7309-7316.

参考文献 VI

- [40] BASAK J, KRISHNAPURAM R. Interpretable hierarchical clustering by constructing an unsupervised decision tree[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(1): 121-132.
- [41] FRAIMAN R, GHATTAS B, SVARC M. Interpretable clustering using unsupervised binary trees[J]. *Advances in Data Analysis and Classification*, 2013, 7: 125-145.
- [42] GHATTAS B, MICHEL P, BOYER L. Clustering nominal data using unsupervised binary decision trees: Comparisons with the state of the art methods[J]. *Pattern Recognition*, 2017, 67: 177-185.
- [43] MOSHKOVITZ M, DASGUPTA S, RASHTCHIAN C, et al. Explainable k-means and k-medians clustering [C]. in: *Proceedings of International Conference on Machine Learning*. 2020: 7055-7065.
- [44] GAMLATH B, JIA X, POLAK A, et al. Nearly-tight and oblivious algorithms for explainable clustering[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 28929-28939.
- [45] BANDYAPADHYAY S, FOMIN F, GOLOVACH P A, et al. How to Find a Good Explanation for Clustering? [C]. in: *Proceedings of the AAAI Conference on Artificial Intelligence*: vol. 36: 4. 2022: 3904-3912.
- [46] MAKARYCHEV K, SHAN L. Explainable k-means: don' t be greedy, plant bigger trees![C]. in: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. 2022: 1629-1642.

Thanks!