A NOVEL MULTI-VIEW ENSEMBLE CLUSTERING FRAMEWORK FOR CANCER SUBTYPE DISCOVERY

Michael G. Schimek[1], Bastian Pfeifer[2] and Marcus D. Bloice[3]

[1] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria, (e-mail: michael.schimek@medunigraz.at)

[2] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria, (e-mail: bastian.pfeifer@medunigraz.at)

[3] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria, (e-mail: marcus.bloice@medunigraz.at)

ABSTRACT: Multi-view clustering methods are essential for the stratification of patients into sub-groups of similar molecular characteristics. Recently, a wide range of methods has been developed for this purpose. However, due to the high diversity of cancer-related data, a single method may not perform sufficiently well in all instances. We present a multi-view hierarchical ensemble clustering framework of methods. We apply and validate it on real-world multi-view cancer patient data. Our approach outperforms the current state-of-the-art in all but one case. It is integrated into our Python package *Pyrea* [https://github.com/mdbloice/Pyrea].

KEYWORDS: multi-view clustering, ensemble clustering, hierarchical clustering, multi-omics, disease subtyping

1 Introduction

Multi-view data contain information relevant for the identification of patterns or clusters that allow us to specify groups of subjects or objects. This presentation is based on (Pfeifer *et al.*, 2023) with a focus on patients for which we have bio-medical and/or clinical observations describing their characteristics obtained from various diagnostic procedures or different molecular technologies. The different types of subject characteristics constitute views related to the patients of interest. Integrative clustering of these views facilitates the detection of patient groups, with the advantage of improved clinical diagnostic and treatment schemes.

Simple integration of single view clustering results is not appropriate for the diversity and complexity of available medical information. Even state-of-the-art multi-view approaches have their limitations, although ensemble clustering has the potential to overcome some of them (Alqurashi & Wang, 2019).

Data views can stem from highly heterogeneous input sources. Therefore, each view needs to be clustered with the most adequate strategy. Multi-view clustering methods are widely applied within the bio-medical domain, where often molecular data are retrieved from different biological layers for the same set of patients. Those clusters inferred from these multi-omics observations facilitate the stratification of cancer patients into sub-groups, providing a useful tool towards precision medicine.

There are two basic types of a multi-view clustering integration, one horizontal and the other vertical (Richardson *et al.*, 2016). Horizontal integration is the aggregation of homogeneous data views, while vertical integration entails the joint analysis of heterogeneous data views from the same group of patients. When data are highly diverse with respect to their probability distributions, problems can arise in vertical integration. Simple data concatenation and the application of single-view methods are most likely to produce biased results.

Clustering ensembles and multi-view clustering methods should provide more robust and accurate clustering results compared to an individual clustering algorithm. A wide range of multi-view clustering methods has been developed, for instance (Xue *et al.*, 2019), (Liu *et al.*, 2021), and (Yang *et al.*, 2022). Other recent approaches, e. g. (Rappoport & Shamir, 2019), (John *et al.*, 2020), and (Pfeifer & Schimek, 2021), have specialised in biomedical applications such as disease subtype detection. However, only a few contributions have investigated the possibility of combining the strengths of both ensemble clustering and multi-view clustering to further improve consistency and accuracy. Here, in contrast to the above-mentioned as well as many other methods, we aim at a generic theoretical and practical framework to enhance flexible ensemble-based multi-view clustering. Our framework is flexible with regard to those clustering techniques that are most suitable for the considered data. Furthermore, the framework allows to construct arbitrarily complex ensemble architectures.

2 The ensemble architecture and proposed methodology

Each view $V \in \mathbb{R}^{n \times p}$ is associated with a specific clustering method c, where n is the number of samples and p is the number of predictors. In total let us have N data views. An ensemble, called $\mathcal{E}$, can be modelled using a set of views $\mathcal{V}$ and an associated fusion algorithm f. Let us have $\mathcal{V} \leftarrow\!\shortmid \{(V \in \mathbb{R}^{n \times p}, c)\}$, $\mathcal{E}(\mathcal{V}, f) \mapsto \widetilde{V} \in \mathbb{R}^{n \times n}$, and $\mathcal{V} \leftarrow\!\shortmid \{(\widetilde{V} \in \mathbb{R}^{n \times n}, c)\}$. From these equations we can see that a specified ensemble $\mathcal{E}$ creates a view $\widetilde{V} \in \mathbb{R}^{n \times n}$ which again can

be used to specify $\mathcal{V}$, including an associated clustering algorithm c. With this concept it is possible to stack layer-wise views and ensembles into arbitrarily complex ensemble architectures. It should be noted, however, that the resulting view of a specified ensemble $\mathcal{E}$ forms an affinity matrix of dimension $n \times n$, and thus only those clustering methods which are compatible with an affinity or distance matrix as input are applicable. The data views are clustered with up to N different hierarchical clustering methods $hc_1, hc_2, \ldots, hc_N$, where N is the number of views. The best combination of clustering methods is inferred by a genetic algorithm, where the silhouette coefficient is adopted as a fitness function. For technical details see (Pfeifer & Schimek, 2021). The Parea$_{hc}$ ensemble approach comprises two different strategies: Parea$^1_{hc}$ is limited to the application of two selected hierarchical clustering methods, while Parea$^2_{hc}$ allows for a variation of hierarchical clustering methods in the data fusion process. Based on machine learning benchmark data sets, a comparison with state-of-the-art methods, such as multi-view spectral clustering and multi-view k-means clustering, was carried out in support of the described approach.

3 Multi-omics clustering for disease subtype discovery

We applied our methodology to a set of real patient data, often used as benchmark data (Rappoport & Shamir, 2018), of seven different cancer types, namely glioblastoma multiforme (GBM), kidney renal clear cell carcinoma (KIRC), liver hepatocellular carcinoma (LIHC), skin cutaneous melanoma (SKCM), ovarian serous cystadenocarcinoma (OV), sarcoma (SARC), and acute myeloid leukemia (AML), aiming at the externally known survival outcome. The Parea$_{hc}$ ensemble approach was studied on multi-omics data, including gene expression (mRNA), DNA methylation, and micro-RNA. Parea$_{hc}$ was compared with SNF (Wang *et al.*, 2014), NEMO (Rappoport & Shamir, 2019), HCfused (Pfeifer & Schimek, 2021), and PINSplus (Nguyen *et al.*, 2019). It is important to mention that the cancer patients were exclusively clustered based on their genomic footprints.

The survival data of all patients were used for the validation of the obtained patient clusters. For the quantification of differences between the studied methods, the Cox log-rank test was applied. The obtained p-values are displayed in Table 1. Our Parea$_{hc}$ ensembles outperform the alternative approaches in almost all cases. SKCM is the only cancer type for which HCfused achieved a superior result. Notably, the spectral-based clustering methods NEMO and SNF performed poorly.

Table 1. *Survival analysis of TCGA cancer group clusters*

Cancer type	**Sample size**	SNF	PINSplus	NEMO	HCfused	Parea^{1}_{hc}	Parea^{2}_{hc}
GBM	538	0.1304	0.2223	**0.0347**	0.0997	**0.0447**	**0.0347**
KIRC	606	0.3962	0.4005	0.3464	0.0561	**0.0137**	**0.0400**
LIHC	423	0.5357	0.6731	0.4354	0.2062	**0.0334**	**0.0436**
SKCM	473	0.5153	0.3956	0.4565	0.0699	0.1677	0.1629
OV	307	0.4042	0.5300	0.3593	0.2594	0.1685	0.2870
SARC	265	0.1622	0.2024	0.0979	**0.0408**	**0.0076**	**0.0109**
AML	173	0.0604	0.1973	**0.0440**	0.1148	**0.0167**	0.0502

Results based on (Pfeifer *et al.*, 2023): Median p-values of the Cox log-rank test. Significant results ($\alpha = 0.05$) for the separation of patient cluster survival curves in **bold**.

References

ALQURASHI, TAHANI, & WANG, WENJIA. 2019. Clustering ensemble method. *International Journal of Machine Learning and Cybernetics*, **10**(6), 1227–1246.

JOHN, CHRISTOPHER R, WATSON, DAVID, BARNES, MICHAEL R, PITZALIS, COSTANTINO, & LEWIS, MYLES J. 2020. Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, **36**(4), 1159–1166.

LIU, JIANLUN, TENG, SHAOHUA, FEI, LUNKE, ZHANG, WEI, FANG, XIAOZHAO, ZHANG, ZHUXIU, & WU, NAIQI. 2021. A novel consensus learning approach to incomplete multi-view clustering. *Pattern Recognition*, **115**, 107890.

NGUYEN, HUNG, SHRESTHA, SANGAM, DRAGHICI, SORIN, & NGUYEN, TIN. 2019. PINSplus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, **35**(16), 2843–2846.

PFEIFER, BASTIAN, & SCHIMEK, MICHAEL G. 2021. A hierarchical clustering and data fusion approach for disease subtype discovery. *Journal of Biomedical Informatics*, **113**, 103636.

PFEIFER, BASTIAN, BLOICE, MARCUS, D., & SCHIMEK, MICHAEL G. 2023. Parea: Multi-view ensemble clustering for cancer subtype discovery. *Journal of Biomedical Informatics*, **143**, 104406.

RAPPOPORT, NIMROD, & SHAMIR, RON. 2018. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, **46**(20), 10546–10562.

RAPPOPORT, NIMROD, & SHAMIR, RON. 2019. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, **35**(18), 3348–3356.

RICHARDSON, SYLVIA, TSENG, GEORGE C, & SUN, WEI. 2016. Statistical methods in integrative genomics. *Annual Review of Statistics and its Application*, **3**, 181–209.

WANG, BO, MEZLINI, AZIZ M, DEMIR, FEYYAZ, FIUME, MARC, TU, ZHUOWEN, BRUDNO, MICHAEL, HAIBE-KAINS, BENJAMIN, & GOLDENBERG, ANNA. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, **11**(3), 333–337.

XUE, ZHE, DU, JUNPING, DU, DAWEI, & LYU, SIWEI. 2019. Deep low-rank subspace ensemble for multi-view clustering. *Information Sciences*, **482**, 210–227.

YANG, MOUXING, LI, YUNFAN, HU, PENG, BAI, JINFENG, LV, JIANCHENG, & PENG, XI. 2022. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**(1), 1055–1069.