N° d'Ordre : 2636

EDSPIC :

Université Sorbonne Paris Nord

ÉCOLE DOCTORALE GALILÉE

Université Sidi Mohamed Ben Abdellah

CENTRE D'ÉTUDES DOCTORALES SCIENCES ET TECHNOLOGIES

Joint Ph.D. Thesis

by

Fatima Ezzahraa Ben Bouazza

for the dgree of

Doctor of Computer Science

Multi-Models clustering through Optimal Transport theory

defended on 26th november 2020 in front the following jury :

Thesis supervisors :

Younès Bennani	Professor, Université Sorbonne Paris Nord
Abdelfattah Touzani	Professor, Université Sidi Mohamed Ben Abdellah

Reporters :

Rosanna Verde	Professor, Seconda Universita di Napoli
Ahmed Youssfi	Professor, ENSA-Fès
Noura Yousfi	Professor, Université Hassan II

Examiners :

Guénaël Cabanes	Associate Professor, Université Sorbonne Paris Nord
Stefano Guerrini	Professor, Université Sorbonne Paris Nord
Ievgen Redko	Associate Professor, Université de Saint-Etienne
Souad Wardi	Professor, Université Sidi Mohamed Ben Abdellah

“*Intelligence is what makes us human, and AI is an extension of that quality.*”

Yann LeCun

“*AI is more artificial than intelligent, like an illusion game. Wouldn’t it be more reasonable to associate AI with the term Advanced Informatics rather than Artificial Intelligence?*”

Younès Bennani

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Younès Bennani for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study. I would also like to my deepest recognition to my second supervisor Prof. Abdelfettah Touzani for his priceless advises and guidance,

Besides my supervisors, I would like to thank the rest of my thesis committee: Prof. Stefno Guerrini to have accepted to preside over my thesis defence, Prof. Rosana Verdé, Prof. Noura Yousfi, and Prof. Ahmed Youssfi for their insightful reports, and finally, Prof. Guenal Cabanes, Prof. Ievgen Redko and Prod Souad Wardi for their gracious comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thanks also go research team: ADA-team where i had the chance to work with different members and make a new friend who gave a pleasant taste to this adventure. Without their precious support it would not be possible to conduct this research.

A very special thanks to my friends in France, Sarah Zouinina, Kouatar Benlamine, for their supports and help during my thesis. Also, I would like to thank Mouard El Hamri, Nistor Grozavu, Bsarab Matei, Issam falih, Guaneal Cabanes, Parisa Rastin who were always helpful in various ways.

I would like to thank all my family members especially my father, my step mother, my brother and sister for supporting me spiritually throughout writing this thesis and my life in general.

Last but not the least, I would like to express my heartfelt gratitude to my beloved husband Mohamed Abdelali Megane for sharing with me this adventure since the beginning, for his sacrifices, help, encouragements, patience, and support especially during the most difficult time in my thesis.

Résumé

Le travail de recherche exposé dans cette thèse concerne le développement d'approches de clustering multi-modèles à partir de données distribuées. Nous travaillons sur deux volets principaux en apprentissage automatique multi-modèles. Le premier concerne le clustering multi-vues où nous visons à apprendre un modèle optimal global à partir des modèles locaux des différentes vues. Le second volet porte sur le clustering collaboratif où l'idée principale est de trouver un modèle d'échange de connaissances entre les différents collaborateurs afin d'améliorer leurs propres performances.

Pour le clustering multi-vues, nous avons proposé deux approches basées sur la théorie du transport optimal. La première approche (PCA) consiste à trouver un modèle de consensus à partir des modèles locaux, en projetant les distributions des différentes vues sur un espace global. La seconde approche (CNR) vise à apprendre une nouvelle représentation consensuelle à partir de la représentation locale des distributions.

Dans le cas du clustering collaboratif, nous introduisons une nouvelle approche basée sur la théorie du transport optimal (Co-OT) qui vise à améliorer le mécanisme de la collaboration et la manière de transporter les informations entre les collaborateurs avec un coût minimum. Pour ce faire, nous proposons une fonction objective pour la collaboration basée sur la distance de Wasserstein. Nous proposons une solution pour choisir les meilleurs collaborateurs par comparaison de la distribution locale des prototypes et l'analyse de la diversité entre les collaborateurs. Pour une autre approche dans ce cadre, nous présentons un nouveau modèle de collaboration guidé par la sélection des caractéristiques, où l'idée principale est de choisir les caractéristiques qui donnent la meilleure représentation pour chaque collaborateur et garantissant la meilleure communication entre eux, tout en préservant la confidentialité des données de chaque collaborateur. Cette dernière approche collaborative est aussi développée dans le cadre de la théorie du transport optimal. Enfin, plusieurs expériences approfondies sur de multiples ensembles de données réelles sont proposées pour évaluer les approches développées et démontrer leur utilité et efficacité dans le cas des données distribuées.

Abstract

The research work presented in this Ph.D. thesis concerns the development of multi-models clustering approaches based on distributed data. We are working on two main aspects of multi-models machine learning. The first concerns multi-view clustering where we aim to learn an optimal global model from the local models of the different views. The second is collaborative clustering, where the main idea is to find a model for knowledge exchange so that each collaborators can improve their own model.

For multi-view clustering, we proposed two approaches based on the optimal transport theory. The first approach (PCA) consists in finding a consensus model from local models, by projecting the distributions of the different views on the global space. The second approach (CNR) aims at learning a new consensus representation from the local representation of the distributions.

In the case of collaborative clustering, we introduce a new approach based on optimal transport theory (Co-OT) which aims to improve the mechanism of collaboration and the way of transporting information between collaborators with minimum cost. To do so, we propose an objective function for collaboration based on the Wasserstein distance. We offer a solution for selecting the best collaborators by comparing the local distribution of prototypes and analysing the diversity between collaborators. For another approach within this framework, we present a new model of feature-driven collaboration, where the main idea is to choose the features that give the

best representation for each collaborator and guarantee the best communication between them, while preserving the confidentiality of each collaborator’s data. This last collaborative approach is also developed within the framework of the theory of optimal transport. Finally, several in-depth experiments on multiple real data sets are proposed to evaluate the approaches developed and demonstrate their usefulness and effectiveness in the case of distributed data.

Contents

Acknowledgements iii

Contents ix

List of Figures xiii

List of Tables xv

Avant Propos 1

Introduction 7

1 Optimal transport Theory 13

1.1 Optimal transport 13

1.2 Monge Problem 15

1.3 Kantorovich Problem 17

1.4 Wasserstein distances. 19

1.4.1 Wasserstein Barycenters. 20

1.5 Sinkhorn's Algorithm 21

1.6 Summary 25

2 Multi-Models clustering 27

2.1 Introduction 27

2.2 Distributed Data Clustering 28

2.3 Multi-view Clustering 29

2.3.1 Fundamental concept of Multi-view clustering 29
2.3.2 Multi-view clustering methods 30
2.4 Collaborative clustering 36
2.4.1 Fundamental concept of collaborative clustering 37
2.4.2 Collaborative clustering methods 39
2.5 Summary 40

3 Multi-view Clustering through Optimal transport 43

3.1 Introduction 43
3.2 Consensus clustering 44
3.3 Multi-view Consensus Clustering through Optimal Transport 47
3.3.1 Motivations 47
3.3.2 Proposed approaches 49
Consensus Projection approach 52
Consensus with new representation 53
3.4 Experimental validation 53
3.4.1 Experimental protocol 54
3.4.2 Experiment results 55
3.5 Summary and discussion 59

4 Collaborative clustering through Optimal Transport 61

4.1 Introduction 61
4.2 Prototype based Collaboration 63
4.3 Proposed approach 66
4.3.1 Motivation 66
4.3.2 Collaborative Learning algorithms 67
Local step 68
Global step 70
4.4 Experiments 73

4.4.1 Setting . . . 73

Data-sets . . . 73

Data-set splitting . . . 74

Quality measures . . . 76

Computation tools . . . 77

4.4.2 Results and discussion . . . 77

Vertical Collaboration case . . . 78

Horizontal collaboration case . . . 82

Comparison with other collaborative approaches . . . 87

4.5 Conclusion . . . 96

5 Subspace guided collaborative clustering based on optimal transport 99

5.1 Introduction . . . 99

5.2 Feature selection . . . 101

5.3 Proposed approach . . . 103

5.3.1 Motivation . . . 103

5.3.2 Proposed algorithm . . . 104

5.4 Experiments Results . . . 107

5.4.1 Protocol experiments . . . 107

5.4.2 Experiments Results . . . 108

5.5 Summary and discussion . . . 115

Conclusion and perspectives 117

5.6 Conclusion . . . 117

5.7 Future perspectives . . . 119

A UIC Data sets 121

B Main publications 125

List of Figures

2.1 General procedure of Co-training clustering 31
2.2 General process of Multi-Kernel Clustering 32
2.3 General process of graph-based clustering. 34
2.4 General mechanism of Multi-subspace clustering 35
2.5 General mechanism of Multi-view Multi task clustering 36
2.6 General mechanism of collaborative clustering framework 38
2.7 General mechanism of Horizontal collaboration clustering framework 40
2.8 General mechanism of Vertical collaboration clustering framework 41
3.1 Friedman test for comparing multiple approaches over multiple data sets 56
4.1 The mechanism of the proposed approach where the local step consist to find the centroids for each collaborator, the collaborative step consist to find the best collaborator by computing a transport plan between the local distribution and the final step consist to influence the local distribution by the distant distribution prototype. 66
4.2 Splitting of the Vertical collaboration 75
4.3 Splitting of the Horizontal collaboration 75
4.4 Sensitivity Box-Whiskers plots for the vertical collaboration case . . 81
4.5 Sensitivity Box-Whiskers plots for the horizontal collaboration case 86

4.6 Friedman and Nemenyi test for comparing multiple approaches over multiple data sets: Approaches are ordered from left (the best) to right (the worst) 94

5.1 Feature selection approaches for clustering 103

5.2 Histogram that compares *Silhouette* index values for 10 collaborators for each data sets. 109

5.3 Histogram that compares Adjusted Rand Index (ARI) values for 10 collaborators for each data sets. 112

5.4 Sensitivity Box plost CO-OT and CO-FS-OT comparaison 114

List of Tables

3.1 Notations 49

3.2 Some Characteristics of used Real-World Datasets 55

3.3 Clustering performance on seven real-world data sets by $DavisBouldin$ Index DB 56

3.4 Clustering performance on seven real-world data sets by Normalized Rand Index R_n 57

4.1 Some characteristics of the experimental real-world data-sets 74

4.2 Values of the different quality indexes before and after a vertical collaboration for each collaborator built from the Spambase data set. 78

4.3 Average values ($\pm CI_{95\%}$) of the different quality indexes before and after the vertical collaboration for each data set over 20 executions. . 79

4.4 Values of the different quality indexes before and after the horizontal collaboration for each collaborator built from the Spambase data set. 84

4.5 Average values ($\pm CI_{95\%}$) of the different quality indexes before and after the horizontal collaboration for each data set over 20 executions. 85

4.6 Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the $Silhouette$ index for the Glass data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions. 88

4.7 Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the $Silhouette$ index for the Spambase data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions. 89

4.8 Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the silhouette index for the Waveform-noise data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions. . . . 90

4.9 Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the silhouette index for the WDBC data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions. 92

4.10 Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the silhouette index for the Wine data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions. 93

4.11 Comparison of DB index Between SOM, GTM and OT based approaches on different data-sets for two collaborators 94

5.1 Average values ($\pm CI_{95\%}$) of Silhouette index for each data set over 20 executions. 110

5.2 Average values ($\pm CI_{95\%}$) of ARI index for each data set over 20 executions. 111

Avant propos

Le transport optimal devient progressivement un outil mathématique puissant et indispensable pour comparer des mesures de probabilité qui, en apprentissage automatique, prennent la forme de nuages de points, d'histogrammes, de caractéristiques ou plus généralement de données à comparer avec des densités de probabilité et des modèles génératifs. Pour cette théorie du transport optimal on peut se référer à un travail développé par Monge, puis par Kantorovich et Danzig à la naissance de la programmation linéaire. La théorie mathématique de l'OT a produit plusieurs développements importants depuis les années 90, couronnés par la médaille Fields de Cédric Villani en 2010. Les applications de cette théorie s'étendent maintenant à d'autres domaines, y compris les applications récentes à l'apprentissage automatique, car il peut y traiter des problèmes de prédiction structurelle qui impliquent des histogrammes, et l'estimation de modèles génératifs de grande taille.

Un nombre important de nouveaux algorithmes de clustering ont été mis au point ces dernières années, et les méthodes existantes ont également été modifiées et améliorées. Cette abondance de méthodes peut s'expliquer par la difficulté de proposer des méthodes génériques qui s'adaptent à tous les types de données disponibles. En effet, chaque méthode a un biais induit par l'objectif choisi pour créer les clusters. Ainsi, deux méthodes différentes peuvent proposer des résultats de clustering très différents à partir des mêmes données. De plus, le même algorithme peut fournir des résultats différents selon son initialisation ou ses paramètres.

Pour apporter des solutions à ce problème, certaines méthodes proposent d'utiliser plusieurs résultats de clustering de modèles différents pour mieux refléter la diversité potentielle des résultats. Ces approches tirent parti des informations fournies par les différents modèles d'une manière sensiblement différente.

L'analyse de ces différentes sources distribuées nécessite des techniques de clustering distribué pour trouver des modèles globaux représentant l'ensemble des informations. La transmission de l'ensemble des données locales est souvent difficile pour des contraintes de bande passante, de respect de la confidentialité et de sécurité. Les algorithmes de clustering traditionnels, qui exigent l'accès à des données complètes, ne sont pas appropriés pour les applications distribuées. Il est donc nécessaire de disposer d'algorithmes de clustering distribué afin d'analyser ces informations localement et de les échanger de manière optimale.

L'échange des informations entre les différents modèles conduit à deux types de résultats : un seul partitionnement de données, ou un ensemble de résultats de clustering. Le premier cas, est l'approche apprentissage d'ensemble. C'est le plus étudié à l'heure actuelle, et nécessite la mise en œuvre de techniques de fusion ou de combinaison des résultats des différents clustering. Le second cas représente l'approche apprentissage collaboratif. Ce sont des méthodes dites de clustering multi-objectifs qui consistent à optimiser simultanément plusieurs critères plutôt que de donner un consensus de résultats.

Dans cette thèse, nous discutons de la manière de transformer l'apprentissage collaboratif et l'apprentissage d'ensemble en un problème de transport optimal. Plus spécifiquement, en apprentissage d'ensembles, nous visons à transporter toutes les informations que nous avons obtenu dans chaque site pour former un modèle consensuel qui présente une synthèse des informations des différents sites.

Nous discuterons aussi de la manière d'utiliser la théorie du transport optimal en apprentissage collaboratif, non seulement en optimisant le transfert des informations entre les différents collaborateurs, mais nous montrons aussi comment le transport optimal peut aider chaque collaborateur à choisir la bonne collaboration, en conséquence on pourra établir un ordre de collaboration optimal entre les différents collaborateurs.

Organisation de la thèse

Ce manuscrit est organisé en cinq chapitres principaux encadrés par une introduction et une conclusion.

Chapitre 1 : Théorie du transport optimal

Dans ce chapitre, nous présentons le formalisme de la théorie du transport optimal. Nous introduisons l'idée de base du transport optimal en partant du problème de Monge et la définition de sa relaxation, ainsi que ce qui a été fait par Kantorovish et a conduit à définir une distance de transport optimal appelée la *distance de Wasserstein* . Cette distance permet de comparer les distributions et crée un plan de transport optimal qui garantit le transport des distributions avec le coût le plus bas possible tout en conservant la masse de l'ensemble de la distribution. Toujours dans ce contexte, nous avons introduit la distance de Wasserstein régularisée en ajoutant un terme de pénalité entropique, ce qui permet de travailler sur des ensembles de données de haute dimension, de garantir une solution unique du problème de transport, et également utiliser l'algorithme de Sinkorn-Knopp qui a prouvé de meilleurs résultats par rapport à l'algorithme classique.

Chapitre 2 : Le clustering Multi-modèles

Dans ce chapitre, nous introduisons la notion de données distribuées et la méthode d'ensembles de clustering de ce type de données. Dans ce contexte, nous avons introduit un des cadres les plus importants de l'apprentissage multi-modèles, qui est le clustering multi-vues avec son concept fondamental. Nous avons ensuite fait le point sur les méthodes existantes du clustering multi-vues et résumé leurs points forts et leurs points faibles. Toujours dans le cadre de l'apprentissage à partir de données distribuées, nous introduisons la notion d'apprentissage collaboratif et ses applications. Nous détaillons le concept fondamental de ce paradigme d'apprentissage, puis nous discutons des différentes méthodes d'apprentissage collaboratif. Le chapitre se termine par une discussion sur le clustering multi-modèles et sur la façon dont les méthodes de clustering classiques peuvent être étendues pour ce type de données tout en assurant une meilleure qualité de clustering.

Chapitre 3 : Le clustering Multi-vues basé sur la théorie du Transport Optimal

Dans ce chapitre, nous présentons un nouveau cadre de clustering multi-vues formalisé dans la théorie du transport optimal, où l'idée principale est d'apprendre des modèles locaux à partir de chaque vue en se basant sur le clustering à base de Sinkhorn, et de chercher un clustering d'ensemble de toutes les vues pour obtenir un modèle de consensus. Pour ce faire, nous proposons deux approches : une approche de projection consensuelle (CPA : Consensus Projection Approach) qui consiste à apprendre un consensus sur l'espace original, et un consensus avec une nouvelle représentation

(CNR : Consensus with New Representation) où l'idée principale est de s'appuyer sur une nouvelle distribution consensuelle apprise à partir de la distribution des différentes vues. Les deux approches sont basées sur la distance de Wsserstein régularisée et les barycentres Wasserstein. Les deux approches sont comparées à l'approche des vues uniques et à l'approche classique du clustering multi-vues basé sur les k-means, à travers un ensemble d'expériences qui mettent en évidence l'efficacité de nos méthodes. Le chapitre se termine par une discussion et une interprétation des résultats afin d'aborder de nouvelles perspectives dans ce domaine de recherche.

Chapitre 4 : Le clustering collaboratif dans le cadre du transport optimal

Dans ce chapitre, nous avons présenté un nouveau cadre de clustering collaboratif basé sur la théorie du transport optimal, où l'idée principale est d'améliorer le mécanisme de collaboration en utilisant le principe de transport des connaissances entre les collaborateurs avec le coût le plus bas possible. Dans ce chapitre, nous avons proposé une nouvelle fonction objective de clustering collaboratif basée sur la distance de Wassesretin. Nous avons également étudié comment choisir les meilleurs collaborateurs en fonction de leur diversité. L'algorithme proposé (Co-Sin-OT : Collaborative Clustering based on Sinkhorn and Optimal Transport) s'est avéré de meilleure qualité comparé au clustering collaboratif classique. De plus, l'approche a prouvé son adaptabilité à différents types de modèles locaux comme (Co-SOM-OT). Ce chapitre présente une série d'expériences dans lesquelles nous comparons l'approche proposée aux méthodes classiques basées sur des prototypes, et nous prouvons l'adaptabilité de notre algorithme à différents types de modèles de base.

Chapitre 5 : Le clustering collaboratif basé sur le transport optimal et guidé par la sélection de caractéristiques

Dans ce chapitre, nous présentons un nouveau modèle de collaboration guidé par la sélection des caractéristiques (Co-FS-OT : Collaboration guided by Feature Selection and Optimal Transport), où l'idée principale est de choisir les caractéristiques qui donnent la meilleure représentation pour chaque collaborateur et garantissent des échanges pertinents entre eux, tout en préservant la confidentialité des données de chacun. Le clustering collaboratif sera développé dans le cadre de la théorie du transport optimal. En effet, cette théorie offre un formalisme très adapté à la collaboration entre les membres d'un ensemble de collaborateurs. Le chapitre comprend des expériences approfondies sur de multiples ensembles de données afin d'évaluer l'approche proposée et de démontrer son utilité. Les propositions de ce chapitre peuvent être présentées comme des extensions des approches multi-vues développées dans le chapitre 4.

Conlusion et perspectives

Dans ce chapitre, nous résumons les principaux résultats présentés dans cette thèse. En outre, une discussion sur les orientations futures de chacune des contributions proposées est présentée, ainsi que les questions de recherche qui restent ouvertes.

Introduction

Optimal transport is gradually becoming a powerful and indispensable mathematical tool for comparing probability measurements, which, in automatic learning, take the form of point clouds, histograms, characteristics or more generally data to be compared with probability densities and generative models. For this theory of optimal transport, we can refer to a work developed by Monge, then by Kantorovich and Danzig at the birth of linear programming. The mathematical theory of EO has produced several important developments since the 1990s, crowned by the Fields Medal of Cédric Villani in 2010. Applications of this theory are now extending to other fields, including recent applications to machine learning, where it can deal with structural prediction problems involving histograms, and the estimation of large generative models.

A significant number of new clustering algorithms have been developed in recent years, and existing methods have been modified and improved. This abundance of methods can be explained by the difficulty of proposing generic methods that adapt to all types of available data. Indeed, each method has a bias induced by the objective chosen to create the clusters. Thus, two different methods can propose very different clustering results based on the same data. Moreover, the same algorithm can provide different results depending on its initialization or parameters.

To provide solutions to this problem, some methods propose to use several clustering results from different models to reflect better the potential diversity of results. These approaches take advantage of the information provided by the different models in significantly different ways.

The analysis of these different distributed sources requires distributed clustering techniques to find global models that represent all the information. The transmission of all local data is often difficult due to bandwidth, confidentiality and security constraints. Traditional clustering algorithms, that require access to complete data, are not appropriate for distributed applications. Therefore, distributed clustering algorithms are needed to analyze this information locally and exchange it in an optimal way.

The exchange of information between the different models leads to two types of results: a single data partitioning, or a set of clustering results. The first case is the ensemble learning approach. This is the most studied at present, and requires the implementation of techniques for merging or combining the results of the different clustering. The second case represents the collaborative learning approach. These are so-called multi-objective clustering methods that consist in simultaneously optimizing several criteria rather than giving a consensus of results.

In this thesis, we discuss how to turn collaborative and whole-group learning into an optimal transport problem. More specifically, in ensemble learning, we aim to transport all the information we have obtained to each site to form a consensual model that presents a synthesis of information from the different sites.

We will also discuss how to use optimal transport theory in collaborative learning, not only by optimizing the transfer of information between different collaborators, but we will also show how optimal transport can help each collaborator to choose the right collaboration, because of which an optimal order of collaboration between the different collaborators can be established.

Overview of the thesis

Chapter 1: Optimal transport theory

In this chapter we present the formalism of the optimal transport theory. We introduce the basic idea of optimal transport starting from Monge's problem and the definition of its relaxation, as well as what was done by Kantorovish and led to the definition of an optimal transport distance called the *Wasserstein distance*. This distance allows the distributions to be compared and creates an optimal transport plan that guarantees the transport of the distributions with the lowest possible cost while maintaining the mass of the entire distribution. Also in this context, we have introduced the regularised Wasserstein distance with the addition of an entropy penalty term, which makes it possible to work with high-dimensional datasets, to guarantee a unique solution to the transport problem, and also to use the Sinkorn-Knopp algorithm, which has proven better results compared to the classical algorithm.

Chapter 2: Multi-Models clustering

In this chapter, we introduce the notion of distributed data and the method of clustering sets for this type of data. In this context, we have introduced one of the most important frameworks of multi-model learning, which is multi-view clustering with its fundamental concept. We then reviewed the existing methods of multi-view clustering and summarized their strengths and weaknesses. Also in the context of learning from distributed data, we introduce the notion of collaborative learning and its applications. We detail the basic concept of this learning paradigm and then discuss the different methods of collaborative learning. The chapter ends with a discussion on multi-models clustering and how classical clustering methods can be extended for this type of data while ensuring better quality clustering.

Chapter 3: Multi views clustering through optimal transport

In this chapter, we present a new framework for multi-view clustering formalized in optimal transport theory, where the main idea is to learn local models from each view based on Sinkhorn-based clustering, and to look for an overall clustering of all views to obtain a consensus model. To do this, we propose two approaches: a Consensus Projection Approach (CPA) where the main idea is to learn a consensus on the original space and a Consensus with New Representation (CNR) where the main idea is to rely on a new consensus distribution learned from the distribution of the different views. Both approaches are based on the regularized Wasserstein distance and the Wasserstein barycenter. The two approaches are compared to the single-view approach and to the classical multi-view clustering approach based on k-means, through a set of experiments that demonstrate the efficiency of our methods. The chapter ends with a discussion and interpretation of the results in order to address new perspectives in this field of research.

Chapter 4: Collaborative clustering through Optimal Transport

In this chapter, we have presented a new collaborative clustering framework based on optimal transport theory, where the main idea is to improve the collaboration mechanism by using the principle of transporting knowledge between collaborators at the lowest possible cost. In this chapter, we proposed a new objective collaborative clustering function based on the Wasserstein distance. We also looked at how to select the best collaborator based on their diversity. The proposed algorithm (Co-Sin-OT: Collaborative Clustering based on Sinkhorn and Optimal Transport) proved higher quality compared to traditional collaborative clustering. Moreover, the approach proved its adaptability to different types of local models such as (Co-SOM-OT). This chapter presents a series of experiments in which we compare the proposed approach with classical prototype-based methods, and prove the adaptability of our algorithm to different types of basic models.

Chapter 5: Subspace guided collaborative clustering based on optimal transport

In this chapter we present a new model of collaboration guided by Feature Selection (Co-FS-OT: Collaboration guided by Feature Selection and Optimal Transport), where the main idea is to choose the features that give the best representation for each collaborator and guarantee relevant exchanges between them, while preserving the confidentiality of each one's data. Collaborative clustering will be developed within the framework of the theory of optimal transport. Indeed, this theory offers a formalism that is highly adapted to collaboration between members of a group of collaborators. The chapter includes in-depth experiments on multiple data sets in order to evaluate the proposed approach and demonstrate its usefulness. The proposals in this chapter can be presented as extensions of the multi-view approaches developed in the previous chapter 4.

Conclusion and perspectives

In this chapter, we summarize the main results presented in this thesis. In addition, a discussion on the future directions of each of the proposed contributions is presented, as well as the research questions that remain open.

1 Optimal transport Theory

1.1 Optimal transport

Optimal transportation (OT) theory was initiated in 1781 by G. Monge in [40] to study some problem of resource allocation, more than a century after the Monge's work, [34] proposed to relax the Monge's setting and show, under suitable hypothesis, that the relaxed problem admits an optimal solution which can be obtained by linear programming. In the last years Cuturi [16] who introduced an algorithm based on entropy regularization to solve and found the optimum.

Our work is basically discrete and we present briefly below the Monge and Kantorovich approaches in the discrete setting. We also recall the regularization proposed by Cuturi. The familiar reader with these subjects could skip this Section and start with our proposed approach for collaborative learning. Also the on the general continuous case we refer the interested reader to the Villani's monograph [55].

ParagraphBackground and notations A probability vector is any element $a \in \Sigma_n$ that belongs to the probability simplex defined as follows:

$$\Sigma_n \overset{\text{def.}}{=} \left\{ a \in \mathbb{R}^n_+; \quad \sum_{i=1}^{n} a_i = 1 \right\}. \tag{1.1}$$

Let $\mathcal{X}$ be a measurable space, a discrete probability measure with weights $a \in \Sigma_n$ and locations $x_1, \dots, x_n \in \mathcal{X}$ is defined as:

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i} \tag{1.2}$$

where $\delta(x)$ is the Dirac distribution at position x, intuitively a unit of mass which is infinitely concentrated at location x. We denote by $\mathcal{M}(\mathcal{X})$ the set of probability measures on the space $\mathcal{X}$. Moreover, if each of the "weights" described in vector a is positive itself, the corresponding probability measure is called positive. We denote $\mathcal{M}_+(\mathcal{X})$ the set of all positive measures on $\mathcal{X}$.

Let $\mathcal{X}$ be a measurable space and $\mathcal{Y}$ another measurable space. For some continuous map $T : \mathcal{X} \to \mathcal{Y}$, we define the pushforward operator $\mathrm{T}_\sharp : \mathcal{M}(X) \to \mathcal{M}(Y)$ as follows:

$$\mathrm{T}_\sharp \alpha \overset{\text{def.}}{=} \sum_i a_i \delta_{\mathrm{T}(x_i)}. \tag{1.3}$$

For discrete measures (1.2), the pushforward operation consists simply in moving the positions of all the points in the support of the measure. Intuitively, a measurable map $T : \mathcal{X} \to \mathcal{Y}$, can be interpreted as a function "moving" a single point from a measurable space to another. The more general extension $T_\sharp$ can now "move" an entire probability measure on $\mathcal{X}$ towards a new probability measure on $\mathcal{Y}$. The operator $T_\sharp$ "pushes forward" each elementary mass of a measure α on $\mathcal{X}$ by applying the map T to obtain then an elementary mass in $\mathcal{Y}$, to build on aggregate a new measure on $\mathcal{Y}$ written $T_\sharp \alpha$. Note that such a push-forward $\mathrm{T}_\sharp : \mathcal{M}_+^1(\mathcal{X}) \to \mathcal{M}_+^1(\mathcal{Y})$ is a linear operator between measures.

Definition 1.1.1 *Push-forward We denote $\mathcal{M}_+(\mathcal{X})$ the set of all positive measures on $\mathcal{X}$. The set of probability measures is denoted $\mathcal{M}_+^1(\mathcal{X})$, which means that any $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ is positive, and that $\alpha(\mathcal{X}) = \int_{\mathcal{X}} x d\alpha = 1$. For $T : \mathcal{X} \to \mathcal{Y}$, the push forward measure $\beta = T_\sharp \alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$ says that for any* *measurable set $B \subset Y$, one has*

$$\beta(B) = \alpha(\{x \in \mathcal{X}; \quad T(x) \in B\}). \tag{1.4}$$

Note that $T_\sharp$ preserves the positivity and also the total mass, so that if $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ then $T_\sharp \alpha \in \mathcal{M}_+^1(\mathcal{Y})$.

1.2 Monge Problem

Mines produce ressources across a country and factories consume ressources across a country. We assume that there exists a local cost for distributing one ressource from a mine to a factory. The Monge's problem consists to find the least costly transportation plan from mines into factories. Let $\mathcal{X}$ $\{x_1, \dots, x_n\}$ be the source space (the mines space) and $\mathcal{Y} \overset{\text{def.}}{=} \{y_1, \dots, y_m\}$ be the target space (the space of factories). For two discrete probability measures :

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_{j=1}^{m} b_j \delta_{y_j} \tag{1.5}$$

the Monge problem [40] seeks for a map T that associates to each point x_i a single point y_j, and which must push the mass of α toward the mass of β, which is to say that such a map $T : \{x_1, \dots, x_n\} \to \{y_1, \dots, y_m\}$ must verify that

$$\forall j \in 1, \dots, m \quad b_j = \sum_{i:\mathrm{T}(x_i) = y_j} a_i \tag{1.6}$$

which we write in compact form as $\mathrm{T}_\sharp \alpha = \beta$. Here $\mathrm{T}_\sharp$ is the pushforward operator. For discrete measures (1.2), $\mathrm{T}_\sharp$ consists simply in moving the positions of all the points in the support of the measure

$$\mathrm{T}_\sharp\alpha \overset{\text{def.}}{=} \sum_i a_i \delta_{\mathrm{T}(x_i)}. \tag{1.7}$$

The map T should minimize some transportation cost, which is parameterized by a function $c(x,y)$ defined for points $(x,y) \in \mathcal{X} \times \mathcal{Y}$

$$\min_T \left\{ \sum_i c(x_i, \mathrm{T}(x_i)); \quad \mathrm{T}_\sharp\alpha = \beta \right\}. \tag{1.8}$$

Such a map between discrete points can be of course encoded, assuming all x's and y's are distinct, using indices $\sigma : \{1, \dots, n\} \to \{1, \dots, m\}$ so that $j = \sigma(i)$, and the mass conservation is written as

$$\sum_{i \in \sigma^{-1}(j)} a_i = b_j. \tag{1.9}$$

Let us assume $n = m$. The Monge's problem could be recasted as follows: given a cost matrix $(\mathbf{C}_{i,j})_{i,j \in \{1,\dots,n\}}$, the optimal assignment problem seeks for a bijection σ in the set $Perm(n)$ of permutations of n elements solving the problem:

$$\min_{\sigma \in Perm(n)} \frac{1}{n} \sum_{i=1}^{n} \mathbf{C}_{i,\sigma(i)}. \tag{1.10}$$

One could naively evaluate the cost function above using all permutations in the set $Perm(n)$. However, that set has size $n!$, which is gigantic even for small n. We also note that the optimal assignment problem may have several optimal solutions. If all weights are uniform, that is $a_i = b_j = 1/n$, then the mass conservation constraint implies that T is a bijection, such that $T(x_i) = y_{\sigma(i)}$, and the Monge problem is equivalent to the optimal matching problem (1.10) where the cost matrix is

$$\mathbf{C}_{i,j} \overset{\text{def.}}{=} c(x_i, y_j). \tag{1.11}$$

When $n \neq m$, note that Monge maps may not even exist between an empirical measure to another. This happens when their weight vectors are not compatible, which is always the case when the target measure has more points than the source measure.

Monge problem (1.8) is extended to the setting of two arbitrary probability measures (α, β) on two spaces $(\mathcal{X}, \mathcal{Y})$ as finding a map $T : \mathcal{X} \to \mathcal{Y}$ that minimizes

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x)) \delta\alpha(x); \quad \mathrm{T}_\sharp \alpha = \beta \right\} \tag{1.12}$$

The constraint $T_\sharp \alpha = \beta$ means that T pushes forward the mass of α to β, and makes use of the push-forward operator (1.4).

1.3 Kantorovich Problem

In practical settings, the assignment problem (1.10) has several limitations since it is formulated as a permutation problem, it can only be used to compare two probability vectors of the *same* size. Additionally, the assignment Problem (1.8) is combinatorial and non-convex. Both are therefore difficult to solve in their original formulation.

The approach of Kantorovich [34] is to relax the nature of transportation, In Monge's problem a source point x_i can only be transported to one and one location $T(x_i)$ only. Kantorovich proposes a"mass splitting" strategy, i.e. the mass at any source point x_i be potentially dispatched across several target locations y_j . In other words Kantorovich moves away from the idea that mass transportation should be "deterministic" to consider instead a "probabilistic" (or "fuzzy") transportation. This flexibility is encoded using, in place of a permutation σ or a map T, a coupling matrix $P \in \mathbb{R}_+^{n \times m}$, where $P_{i,j}$ describes the amount of mass flowing from point x_i towards points y_j.

Admissible couplings admit a far simpler characterization than Monge maps:

$$\mathcal{U}(a,b) \overset{\text{def.}}{=} \{P \in \mathbb{R}_+^{n\times m}; \quad P\mathbf{1}_m = a \quad \text{and} \quad P^T\mathbf{1}_n = b\} \tag{1.13}$$

where we used the following matrix-vector notation

$$P\mathbf{1}_m = \left(\sum_j P_{i,j}\right)_i \in \mathbb{R}^n \quad \text{and} \quad P^T\mathbf{1}_n = \left(\sum_i P_{i,j}\right)_j \in \mathbb{R}^m. \tag{1.14}$$

The set of matrices $\mathbf{U}(a,b)$ is bounded, defined by $n+m$ equality constraints, and therefore a convex polytope (the convex hull of a finite set of matrices).

Additionally, whereas the Monge formulation was intrisically asymmetric, Kantorovich's relaxed formulation is always symmetric, in the sense that a coupling P is in $\mathcal{U}(a,b)$ if and only if P^T is in $\mathcal{U}(b,a)$. Kantorovich's optimal transport problem writes:

$$\mathrm{L}_C(a,b) \overset{\text{def.}}{=} \min_{P\in\mathcal{U}(a,b)} \langle C,\, P\rangle \overset{\text{def.}}{=} \sum_{i,j} C_{i,j}P_{i,j}. \tag{1.15}$$

This is a linear program and as is usually the case with such programs, its solutions are not necessarily unique.

Remark 1 (Kantorovich formulation for arbitrary measures.) *The definition of* L_C *in Equation* (1.15) *can be extended to arbitrary measures by considering couplings* $\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})$ *which are joint distributions over the product space. The discrete case is a special situation where one imposes this product measure to be of the form* $\pi = \sum_{i,j} P_{i,j}\delta_{(x_i,y_j)}$. *In the general case, the mass conservation constraint* (1.13) *should be rewritten as a marginal constraint on joint probability distributions*

$$\mathcal{U}(\alpha,\beta) \overset{\text{def.}}{=} \left\{\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}); \quad P_{\mathcal{X}\sharp}\pi = \alpha \quad \text{and} P_{\mathcal{Y}\sharp}\pi = \beta\right\} \tag{1.16}$$

Here $P_{\mathcal{X}\sharp}$ and $P_{\mathcal{Y}\sharp}$ are the push-forward (see Definition 1.1.1) by the projections $P_{\mathcal{X}}(x,y) = x$ and $P_{\mathcal{Y}}(x,y) = y$.

Using (1.4), these marginal constraints are equivalent to imposing that $\pi(A \times \mathcal{Y}) = \alpha(A)$ and $\pi(\mathcal{X} \times B) = \beta(B)$ for sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$. The Kantorovich problem (1.15) is then generalized as

$$\mathcal{L}_c(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\delta\pi(x,y). \tag{1.17}$$

This is an infinite-dimensional linear program over a space of measures.

1.4 Wasserstein distances.

An important feature of OT is that it defines a distance between probability measures as soon as the cost matrix satisfies certain suitable properties. Indeed, OT can be understood as a canonical way to lift a ground distance between points to a distance between probability measures.

We first consider the case where the “ground metric” matrix C is fixed, representing substitution costs between points, and shared across several measures we would like to compare. Note that OT provides a meaningful distance between probability measures supported on these points as in the following definition.

Definition 1.4.1 *We suppose $n = m$, and that for some $p \geq 1$, $C = D^p = (D_{i,j}^p)_{i,j} \in \mathbb{R}^{n\times n}$ where $D \in \mathbb{R}_+^{n\times n}$ is a distance on $\{1,\ldots,n\}$,* i.e. *: $D \in \mathbb{R}_+^{n\times n}$ is symmetric, has null diagonal and satisfies the triangle inequality. Then*

$$W_p(a,b) \overset{\text{def.}}{=} \mathrm{L}_{D^p}(a,b)^{1/p} \tag{1.18}$$

(note that W_p depends on D) defines the p-Wasserstein distance on σ_n, i.e. *W_p is symmetric, positive, $W_p(a,b) = 0$ if and only if $a = b$, and it satisfies the triangle inequality .*

The Wasserstein distance W_p has many important properties, the most important one being that it is a weak distance, *i.e.* it allows to compare singular distributions (for instance discrete ones) and to quantify spatial shift between the supports of the distributions. In particular, "classical" distances, e.g. L^2 or l^2, are not even defined between discrete distributions.

1.4.1 Wasserstein Barycenters.

Given input probability measures $\{b_s\}_{s=1}^S$, where $b_s \in \Sigma_{n_s}$, and weights $\lambda \in \Sigma_S$, a Wasserstein barycenter is computed by minimizing

$$\min_{a\in\Sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}_{C_s}(a, b_s) \tag{1.19}$$

where the cost matrices $C_s \in \mathbb{R}^{n\times n_s}$ need to be specified. A typical setup is "Eulerian", so that all the barycenters are defined on the same grid, $n_s = n$, $C_s = C = D^p$ is set to be a distance matrix, so that one solves

$$\min_{a\in\Sigma_n} \sum_{s=1}^{S} \lambda_s W_p^p(a, b_s). \tag{1.20}$$

This barycenter problem (1.19) was introduced by [2] following earlier ideas of [1]. They proved uniqueness of the barycenter for $c(x,y) = \|x-y\|^2$ over $\mathcal{X} = \mathbb{R}^d$, if one of the input measure has a density with respect to the Lebesgue measure. The barycenter problem for probability measures (1.19) is in fact a linear program, since one can look for the S couplings $(P_s)_s$ between each input and the barycenter itself

$$\min_{a\in\Sigma_n,(P_s\in\mathbb{R}^{n\times n_s})_s} \left\{ \sum_{s=1}^{S} \lambda_s \langle P_s,\, C_s\rangle ; \quad \forall s, P_s^\top \mathbf{1}_{n_s} = a, P_s^\top \mathbf{1}_n = b_s \right\}. \tag{1.21}$$

Although this problem is an LP, its scale forbids the use generic solvers for medium scale problems. In order to fix this problem one can use entropic smoothing and approximate the solution of (1.19). We describe this regularization in the next paragraph.

1.5 Sinkhorn's Algorithm

In this section we recall a family of numerical scheme to approximate solutions to Kantorovich formulation of optimal transport. It operates by adding an entropic regularization penalty to the original problem. The main advantages is that the minimization of the regularized problen can be solved using a simple alternate minimization scheme; that scheme translates into iterations that are simple matrix products the resulting approximate distance is smooth with respect to input probability measure weights and positions of the Diracs.

Entropic regularization The discrete entropy of a coupling matrix is defined as

$$\mathrm{H}(P) \overset{\text{def.}}{=} -\sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1), \tag{1.22}$$

with an analogous definition for vectors, with the convention that $\mathrm{H}(a) = -\infty$ if one of the entries a_j is 0 or negative. Note that H is 1-strongly concave function.

Let us define:

$$\mathrm{L}_{\mathrm{C}}^{\epsilon}(a,b) \overset{\text{def.}}{=} \min_{P\in\mathcal{U}(a,b)} \langle P,\, C\rangle - \epsilon \mathrm{H}(P). \tag{1.23}$$

Since the objective is a ϵ-strongly convex function, problem (1.23) has a unique optimal solution [12]. Defining the Kullback-Leibler divergence between couplings as

$$\mathrm{KL}(P|K) \overset{\text{def.}}{=} \sum_{i,j} P_{i,j} \log\left(\frac{P_{i,j}}{K_{i,j}}\right) - P_{i,j} + K_{i,j}, \tag{1.24}$$

the unique solution P_ϵ of (1.23) is a projection into $\mathcal{U}(a,b)$ of the Gibbs kernel associated to the cost matrix C as

$$K_{i,j} \overset{\text{def.}}{=} e^{-\frac{C_{i,j}}{\epsilon}} \tag{1.25}$$

Indeed one has that using the definition above

$$P_\epsilon = \mathbf{Proj}^{\mathrm{KL}}_{\mathcal{U}(a,b)}(K) \overset{\text{def.}}{=} \underset{P \in \mathcal{U}(a,b)}{\operatorname{argmin}} \mathrm{KL}(P|K). \tag{1.26}$$

Remark 2 (Convergence with ϵ) *The unique solution P_ϵ of (1.23) converges to the optimal solution with maximal entropy within the set of all optimal solutions of the Kantorovich problem, namely*

$$P_\epsilon \overset{\epsilon \to 0}{\longrightarrow} \underset{P}{\operatorname{argmin}} \left\{ -\mathrm{H}(P), \quad P \in \mathcal{U}(a,b), \quad \langle P, \mathbf{C} \rangle = \mathrm{L}_{\mathrm{C}}(a,b) \right\}. \tag{1.27}$$

Formula (1.27) states that for low regularization, the solution converges to the maximum entropy optimal transport coupling. In sharp contrast, (1.26) shows that for large regularization, the solution converges to the coupling with maximal entropy between two prescribed marginals a, b, namely the joint probability between two independent random variables with prescribed distributions. A refined analysis of this convergence is performed in [12], including a first order expansion in ϵ (resp. $1/\epsilon$) near $\epsilon = 0$ (resp $\epsilon = +\infty$).

The solution of (1.19) to define Wasserstein barycenter of S probability measures is to approximate (1.19) by entropic regularization. Let us define:

$$\min_{a \in \sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}^{\epsilon}_{C_s}(a, b_s) \tag{1.28}$$

for some $\epsilon > 0$. This is a smooth convex minimization problem, which can be tackled using gradient descent [16]. A simple but effective approach, as remarked in [4] is to rewrite (1.28) as a (weighted) KL projection problem

$$\min_{(P_s)_s} \left\{ \sum_s \lambda_s \mathrm{KL}(P_s | K_s), \quad \forall s, \quad {P_s}^T \mathbf{1}_m = b_s, P_1 \mathbf{1}_1 = \cdots = P_S \mathbf{1}_S \right\}. \tag{1.29}$$

where we denoted $K_s \overset{\text{def.}}{=} e^{-C_s/\epsilon}$. Here, the barycenter a is implicitly encoded in the row marginals of all the couplings $P_s \in \mathbb{R}^{n \times n_s}$ as $a = P_1 \mathbf{1}_1 = \ldots = P_S \mathbf{1}_S$. The optimal couplings $(P_s)_s$ solving (1.29) are computed in scaling form as

$$P_s = \mathrm{diag}(u_s) K \,\mathrm{diag}(v_s). \tag{1.30}$$

The scalings factors v_s, u_s and a are sequentially computed. A way to derive these iterations is to perform alternate minimization on the variables of a dual problem. As detailed in [4], one can also obtain these iterations as a special case of the generalized Sinkhorn detailed below.

The solution of (1.28) has a specific form, which can be parameterized using $n + m$ variables. That parameterization is therefore essentially dual, in the sense that a coupling P in $\mathcal{U}(a, b)$ has nm variables but $n + m$ constraints. More precisely, the solution to (1.28) is unique and has the matrix form

$$P = \mathrm{diag}(u) K \,\mathrm{diag}(v) \tag{1.31}$$

for two (unknown) scaling variable $(u, v) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$. Therefore u, v must satisfy the following non-linear equations which correspond to the mass conservation constraints inherent to $\mathcal{U}(a, b)$,

$$\operatorname{diag}(u) K \operatorname{diag}(v) \mathbf{1}_m = a \quad \text{and} \operatorname{diag}(v) K^{\mathrm{T}} op \operatorname{diag}(u) \mathbf{1}_n = b, \tag{1.32}$$

This problem is known in the numerical analysis community as the matrix scaling problem. An intuitive way to try to solve these equations is to solve them iteratively, by modifying first u so that it satisfies the left-hand side of Equation (1.32) and then v to satisfy its right-hand side. These two updates define Sinkhorn's algorithm:

$$u^{(l+1)} \overset{\text{def.}}{=} \frac{a}{Kv^{(l)}} \quad \text{and} v^{(l+1)} \overset{\text{def.}}{=} \frac{b}{K^T u^{(l+1)}}, \tag{1.33}$$

initialized with an arbitrary positive vector $v^{l+1} = \mathbf{1}_m$. The division operator used above between two vectors is to be understood entry-wise. It turns out however that these iterations converge and all result in the same optimal coupling $\operatorname{diag}(u) K \operatorname{diag}(v)$.

Remark 3 (General formulation) *One can consider arbitrary measures by replacing the discrete entropy by the relative entropy with respect to the product measure* $\delta\alpha \otimes \delta\beta(x, y) \overset{\text{def.}}{=} \delta\alpha(x)\delta\beta(y)$, *and propose a regularized counterpart to* (1.17) *using*

$$\mathcal{L}_c^\epsilon(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha, \beta)} \int_{X \times Y} c(x, y) \delta\pi(x, y) + \epsilon \operatorname{KL}(\pi | \alpha \otimes \beta) \tag{1.34}$$

where the relative entropy is a generalization of the discrete Kullback-Leibler divergence (1.24)

$$\begin{aligned} \operatorname{KL}(\pi | \xi) \overset{\text{def.}}{=} \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{\delta\pi}{\delta\xi}(x, y) \right) \delta\pi(x, y) + \\ \int_{\mathcal{X} \times \mathcal{Y}} (\delta\xi(x, y) - \delta\pi(x, y)), \end{aligned} \tag{1.35}$$

and by convention $\mathrm{KL}(\pi|\xi) = +\infty$ *if* π *does not have a density* $\frac{\delta\pi}{\delta\xi}$ *with respect to* ξ*. It is important to realize that the reference measure* $\alpha \otimes \beta$ *chosen in* (1.34) *to define the entropic regularizing term* $\mathrm{KL}(\cdot \mid \alpha \otimes \beta)$ *plays no specific role, only its support matters.*

Formula (1.34) *can be re-factored as a projection problem*

$$\min_{\pi \in \mathcal{U}(\alpha,\beta)} \mathrm{KL}(\pi|\mathcal{K}) \tag{1.36}$$

where $\mathcal{K}$ *is the Gibbs distributions* $\delta\mathcal{K}(x,y) \overset{\text{def.}}{=} e^{-\frac{c(x,y)}{\epsilon}}\delta\mu(x)\delta\nu(y)$*. This problem is often referred to as the "static Schrödinger problem" [44], since it was initially considered by Schrödinger in statistical physics [44]. As* $\epsilon \to 0$*, the unique solution to* (1.36) *converges to the maximum entropy solution to* (1.17) *[2].*

1.6 Summary

Genarlly speaking, Optimal Transport has been widely successful throughout the years, with Nobel Loreate in economics for Leonid Kantorovich in 1975 and the Fields Medalists for Cedric Villani in 2010, and Alessio Figalli in 2018.

Given the regularized version of the Wasserstein distance, the Optimal Transport theory has been very useful recently especially in machine learning such as domain adaptation [13] metric learning [15], clustering [16] and multi-level clustering [30]. The particularity about this distance is that it takes into account the geometry of the data using the distance between the samples, which explains its efficiency. On the other hand, in term of computation, the success of this distance also comes from the work of Cuturi [14].

I started investigating this theory and its numerical application, and have been witness of an explosion of interest towards this topic, especially in computer science filed. Although the explosion of the application based on optimal transport is not nearly dramatic as the explosion towards deep learning, but still gives birth to a very pertinent application and result.

2 Multi-Models clustering

2.1 Introduction

Clustering is one of the main exploratory task in Machine Learning that interests many researchers. There are a huge number of existing clustering algorithms that can give very different results with the same data, and choosing between several clustering results is often problematic. This problem can only be solved by asking an expert to choose the most adapted method and the parameters that will work best for a specific data set. This very difficult task can have a heavy influence on the results. Making this kind of decision requires a deep knowledge of both the data to be analyzed, and a large amount of algorithms that are available. Furthermore, even with a good expert having a decent knowledge of both the data and the algorithms, it is still difficult to make the right choices when it comes to clustering.

On the other hand, with the development of hardware technology, a huge amount of data represented in different views and different structures has been generated in real word applications. This kind of data considered as a new challenge to develop the existing clustering algorithms designed for a single view data to be more adaptable to multi-view data or distributed data.

2.2 Distributed Data Clustering

With the huge development of technology, computing environments have been evolving towards dynamic, interacted, and distributed data on multi sources that contain massive amounts of different type of data either spatially or temporally. As an example of this computing environments is the grid-based or cloud-based where the data is distributed across multi sources. Thus, Communication between these sources and the computing locations is necessary. The first intuition is to centralize the all the data in a single view and apply the classical algorithms. However, this may be possible for two principal reasons:

1. The volume of the data. With the hardware technology, most of the data sets are very massive that can be stored in a single computer and may need a splitting to be processed with the classical algorithms.

2. Privacy and security issues. In most of the cases, it may not be possible to share or combine the data distributed across different sources simply because some information in some sources are not meant to be read or shared outside the storage site.

To illustrate the case of distributed data, we present the following situation, where we have several organizations or companies who have a collection of data sets that could concern the same or different customers. This could be data describing customers of banking institutions, state organizations and hospitals with medical information records, etc. Imagine that all these organizations are dealing with the same individuals but every organization may have different characteristics and descriptors for these individuals linked to their activities. All these organizations may want to explore data mining algorithms on there one data set. On the other hand, they also recognize that, as they are other data sets containing information about the same individuals, it would be advantageous to learn about the dependencies that they have so that they could reveal a macro-picture. However, due to ethical consideration and privacy issues, these organizations are forbidden to share their data sets. Which prevents the experts to combine all these data sets into a single view and carrying out different algorithms of classical clustering. For example, the confidentiality requirements in medical records of patients could deny the access into their personal information, and security issues in banking organization forbid to share customer's information. Besides, experts may hesitate about losing the real structure of the data by adding more information and characteristics.

To point out the mismatch, a large number of algorithms have been proposed in this research where privacy, security, and confidentiality are preserved and maintained during all the process while sharing information about some segment or prototype are involved. These studies main to cluster each data set separately while being comfortable to share the partial results of the local clustering algorithms. Thus, each of the companies could then use the results of the clustering learned from other one to complete their own finding.

In the next section we will present one of the popular solutions developed to seek the clustering of distributed data.

2.3 Multi-view Clustering

In this section, we will introduce the fundamental concept of Multi-view clustering and present the different algorithms represented in this research area.

2.3.1 Fundamental concept of Multi-view clustering

One of the common issues on the distributed data is the huge amount of the redundant characteristics available to describe the same objects especially when these objects are represented in different sites. The problem could be more complicated when the objects are described by different categories of features that could be discrete, continuous, texts or intervals. Nevertheless, when a single view is containing different categories of attributes, it is more complicated to define a similarity function and require a deep knowledge of the field or we may have biases models.

We distinct 2 fundamental concepts of Multi-view clustering, complementary and consensus principles.

Complementary principle The complementary principal in multi-view clustering aims to describe the same data in different views that complete each other, in other word each view my contains knowledge that other views doesn't have. More specifically, we don't have redundant views. In this framework we aim to build a custom distance that work will combine all the similarities in each view to achieve an accurate evaluation of data.

Consensus principle The main idea of this framework of Multi-view clustering is to maximize the agreement across multiple distinct views [48]. We may run different clustering algorithms adapted to each kind of features and then aggregate the results to obtain an ensemble model which contains all the information resulting from the views.

2.3.2 Multi-view clustering methods

Co-training clustering In [7] they introduced a method of co-training based on two hypotheses trained on distinct views, the training algorithm aims to increase the global training of two classifiers with the highest confidence examples from unlabeled data, this method requires the independence of the views.

Co-training algorithms attend to find a global agreement across two views in order to find a global consensus of all the views. These algorithms are mainly based on finding a consistency between views based on a prior information on their local representation. Thus, the consistency could be built through exchanging information. Having the best results using the Co-training algorithms depends on three conditions:

1. Sufficiency: Where each view may have the ability to learn tasks.
2. Compatibility: The explored information in each view should be compatible with the objective functions, and export the same perfections for co-occurring feature.
3. Conditional independence: The considered views must provide learning labels that verify the conditional independence. However, In practice it is usually too hard to verify this condition in real data sets.

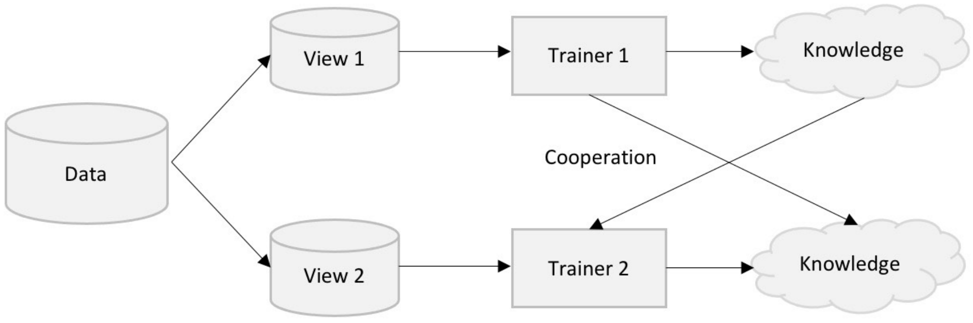


FIGURE (2.1) General procedure of Co-training clustering

In unsupervised learning, Bickel and Scheffer [5] had studied the problem where the set of attributes can be split randomly into tow subsets, based on the idea of Co-training, two proposed methods seek to optimize the agreement between the views. These methods were designed for text data. The first method is based on EM algorithm that alternates between views in order to find an agreement as showed in figure 2.1. The second method used agglomeration algorithm. Unlike the agglomeration algorithm, which gives poorer results, the EM algorithm proved positive results in Multi-view clustering. Although the Co-training style had given many extended methods in machine learning. But still, very limited to the kind of the data presented nowadays.

Multi-Kernel clustering was developed to resolve the problem of massive data [52]. The general procedure of Multi-kernel learning aims to combine pre-learned kernels that represent the views in order to get a unified optimal kernel. The figure 2.2 shows general mechanism of how we can unify the kernel of global views, by combining the kernels learned in each view.

One of the most important challenges is to choose the right Kernel function e.g., Linear kernel, Polynomial, Or Gaussian, and how to estimate the wide combination of the kernels. Another estimation Multi-view Kernel clustering was investigated, based on weighted combination of all the kernels, these weights are proportional to the quality of views [53].

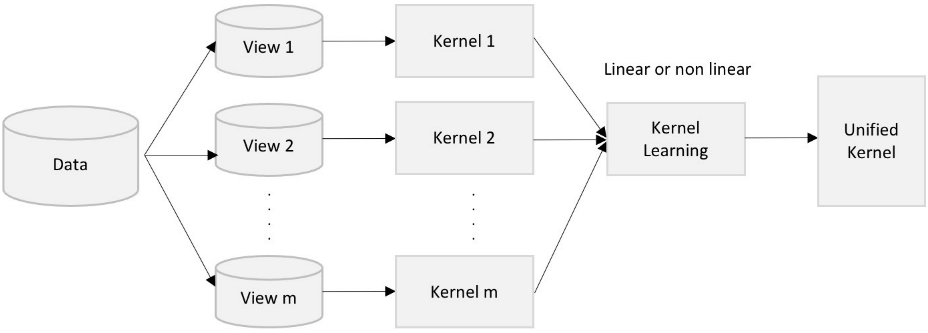


FIGURE (2.2) General process of Multi-Kernel Clustering

The Multi-Kernel clustering. revealed significant results especially when some or all the views produced incomplete data [37], where the main idea is to complete the

Kernels of incomplete views by optimizing the alignment of shared instances of those views.

Multi-view graph clustering is a wide domain where we introduce the graphs (or networks) to represent the relationships between objectives. The main idea behind is to analyze the data objects using many graphs, where each graph capture a distinct view of data objects. Thus these graphs will reinforce each other to perform a unified graph which represents a global view [68].

Many studies had been done to improve the Mutli-view graph based methods; nonetheless, these methods still suffer from one of its biggest limitations, which is a hypothesis that must be considered to perform the graph-based algorithms. The hypothesis is that the same data is available in different views in order to guarantee the one to one relationship between data objects. Hypothesis that is difficult to ensure in real data sets.

To resolve this paradigm, a proposed extension of Multi-view graph based algorithm, where they use an improved networks based on regularized graph clustering, and aims to create multiple relationships for each data objects. This regularization can ensure an associated weight to each graph [10]. In this framework we refer also to the Multi-view Spectral clustering, which is based on a classical clustering methods, where the basic idea is to learn a consistent similarity matrix, that will be normalized and finally get its eigenvectors (characteristics of the vectors) [32]. This method is used to minimize the disagreement between the graphs of each view based on a disagreement criterion.

The figure 2.3 exposes the process using graph based clustering in Multi-view clustering to get a global graph which represents all the views and build from the fusion of the graphs that represent the views.

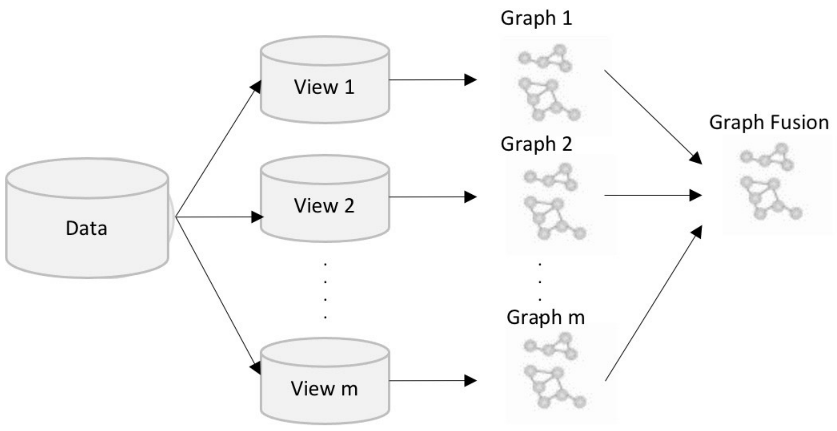


FIGURE (2.3) General process of graph-based clustering.

Multi-subspace clustering aims learn a common representation of all the views of the data, from the sub-spaces created from the data. The main idea of this method is to find a unified representation despite the high dimension [9]. We distinct inf this task two principal procedure:

1. This procedure consists to learn sub-spaces representation of each view in the first step, while the second step aims to learn a unified representation from the sub-spaces, and then applied some clustering algorithm.
2. This procedure aims to introduce a latent space in which compressed the data by representing the closer points on real space similar point, After getting the sub spaces so we can avoid the curse of dimension and finally fit clustering model to produce optimal global clusters.

The figure 2.4 represents the different type unified representation where the first one is built only from the sub-spaces learned from the views, and the second one introduce a latent space before getting a unified representation.

Multi-task Multi-view clustering is a field based on enhancing the quality of clustering by exploiting the consistent and the complementary between the views [70].

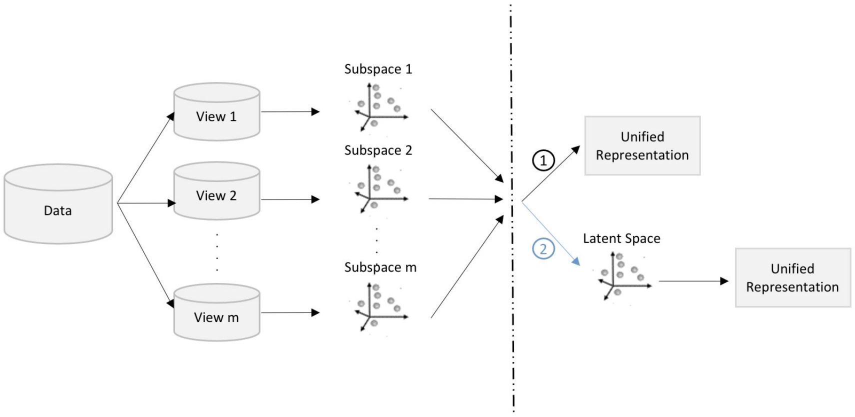


FIGURE (2.4) General mechanism of Multi-subspace clustering

The main idea behind is to improve local task by transferring knowledge across related task. However the main challenge in multi task clustering is to define the relationship between the multiple views, and also how to guarantee an optimal transfer of the knowledge between tasks. We distinct in this frame work two principal procedure:

1. The first category of assumes that the tasks are completely related, and the consensus can be learned involving sub-space shared by multiple connected tasks, or by learning a kernel space in which the distributions of the related tasks are close to each other and the geometric structures of the original data are preserved. Although many works had been done to improve this kind of methods, but they still suffer from the condition that the data may be verified to use these methods which is the label spaces among the tasks must be the same which is in general not the case in real data sets [28].

2. The second category is designed to partially related data, They are based on an improved discriminative methods adaptable to Multi-task clustering. For example, and improved Bregman iterations for Multi-task clustering update the clusters, and learn the relationships between the clusters of different tasks while every iteration boosts the other one. Many studies had been proposed to improve these methods. However, up to know most of the methods requires that all the tasks must have the same labels, and the same number of clusters,

which very difficult to ensure in Multi-view clustering due to the diversity of the views [69].

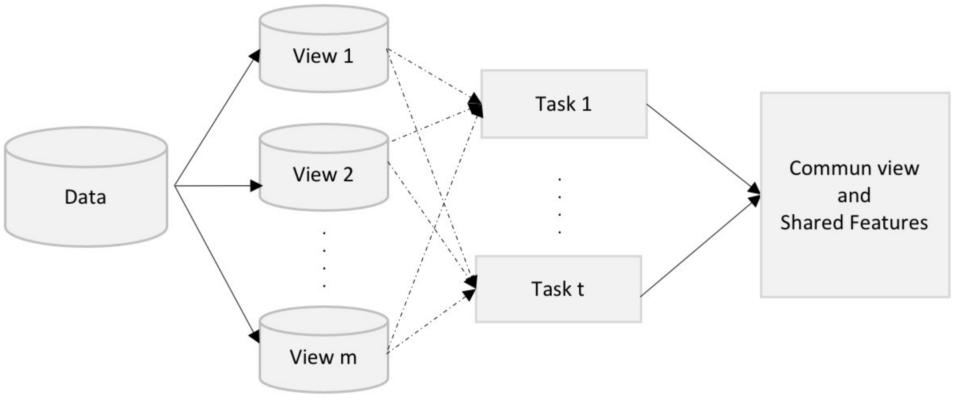


FIGURE (2.5) General mechanism of Multi-view Multi task clustering

The figure 2.5 shows the general mechanism of how we can use the Multi-task Clustering to improve the Multi-view clustering, and get a unified consensus build from the relationships learned between the tasks.

2.4 Collaborative clustering

In the previous section we presented several algorithms that aims basically to learn from learners, where the basic idea is aggregates the local models of each view to build an optimal model consensus model that captures all the views, While in this section, we will present a recent framework of learning from other learners, where the main idea is to find local minimum for each views or sub-sets, by learning from other views (or sub-sets).

2.4.1 Fundamental concept of collaborative clustering

Collaborative clustering was originally thought to be applied to the specific case of distributed data clustering where the main idea make different models work together in order to increase the clustering quality [41]. Later many studies have been proposed to extend the collaborative clustering into large data coming from different learners and could either char the same characteristics or captured from different views.

The fundamental concept of collaborative clustering can be defined as a collective scheme to that seeks to creates links between different sites so they can communicate at some level in order to help each other to increase there models.

Generally speaking, collaborative clustering method follow two principal steps 2.6:

1. **Local step**: this step takes place at the local level, where each sites process a clustering algorithm to produce a local model the new partition of the data of this set independently of the other sites. It must be mentioned that although at this stage we can use any clustering algorithm but the number of clusters must be the same for all the data sets.

2. **Global step**: This step consists to question the partition of the data that we already obtained in the local step in each site. This can be done by questioning the distant collaborators about there results, and ask them to exchange the knowledge in order to improve the local quality and guarantee better clustering of the data.

In this research area, many challenges had been exposed during the development of this framework, The first challenge is how to exchange this knowledge between collaborators, and what knowledge can be beneficial for the distant collaborator while preserving the privacy. In addition choosing the right collaborator still up to now a very important field to investigate, this is due to the complexity of this chose which

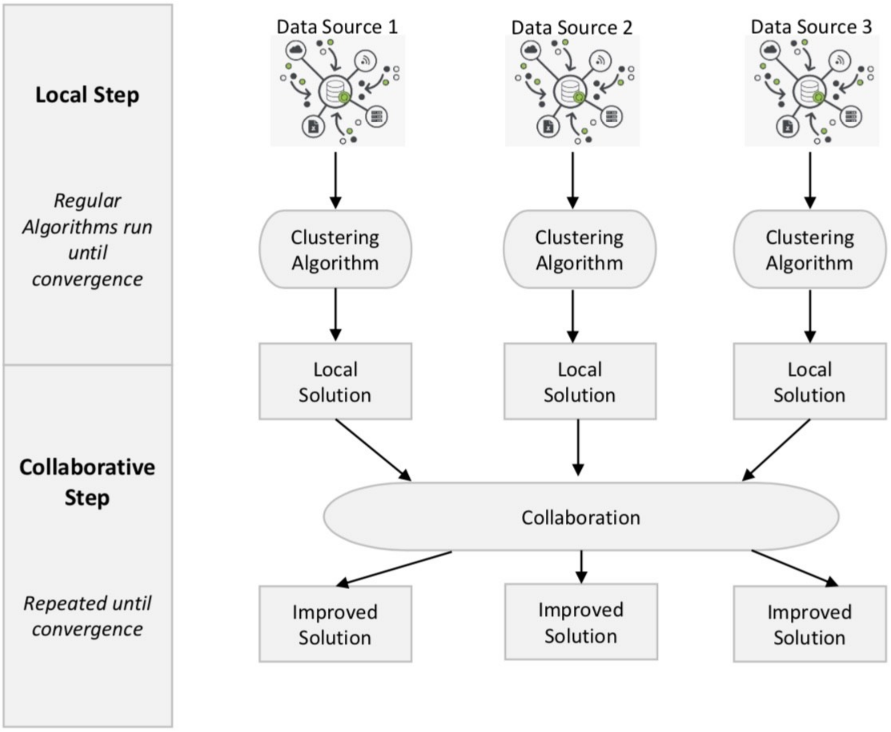


FIGURE (2.6) General mechanism of collaborative clustering framework

depends on diversity between the representation and quality of each collaborator. More precisely, Although a distant collaborators can dispose of a better partition of data a better clustering but still, the exchange with him can brought negative results, and we can even completely lost the origin representation of the origin collaborator, if the diversity between these collaborators is very high. This balance between the quality and diversity can affects not only the quality of clustering, but also the order of the collaborations. Another challenge in collaborative clustering is how to stop the collaboration and which stopping criteria must be used.

2.4.2 Collaborative clustering methods

In this section we present different manipulation of collaborative clustering where the collaboration could be done between samples that share the same patterns, or between different views that capture different characteristics. Or an hybrid model that combine between different kind of sub-sets.

Horizontal collaboration In this situation we deal with sets which contain the same instances but represented with different features. The collaboration is said horizontal because the splitting of the data is done horizontally, where we applied clustering algorithm on each view.Thus the exchange of the knowledge is done horizontally as represented in 2.7. This kind of collaboration is pretty interesting for Multi-view, Multi scale learning. Moreover, it can help the manipulation of a high dimension data.

Vertical collaboration refers the situation where the data sets is represented by the same patterns but different instances. In this case the collaborators contain different objects, but the same features. The collaboration is said " vertical collaboration" because the data set is split is alongside the instances see figure 2.8. The kind of collaboration is closer to transfer leaning idea, where the exchange of the knowledge is done between different sample represented by the same characteristics.

Hybrid Collaboration in this third situation we deal with a collaboration, where both the vertical and the horizontal families are combined, and used at the same time. Given the example of instances that are captured from different views that represent different patterns, thus using some feature selection algorithm we can rise a common subset of the date which is positioned in the same feature space. Hence, the name of Hybrid collaboration. However, the results in in Hybrid collaboration still very

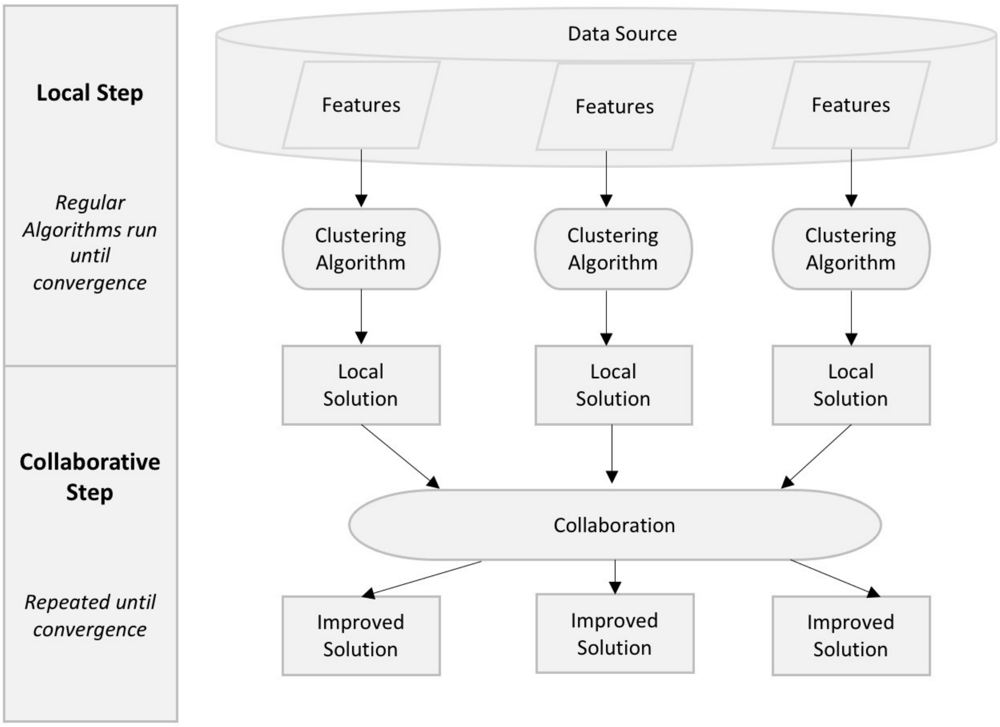


FIGURE (2.7) General mechanism of Horizontal collaboration clustering framework

limited, due to the lack of a common ground between collaborators in order to build the link to exchange the information between collaborators.

2.5 Summary

In this chapter, we have discussed the paradigm of how to make several classical clustering algorithms work together in order to address to difficult data sets, which is either distributed on multiple sets, or have a high dimension and need to be split.

To resolve this problematic, there is two principal framework, The first one consists to find a global maximum that combines all the the algorithms in effort to find a global

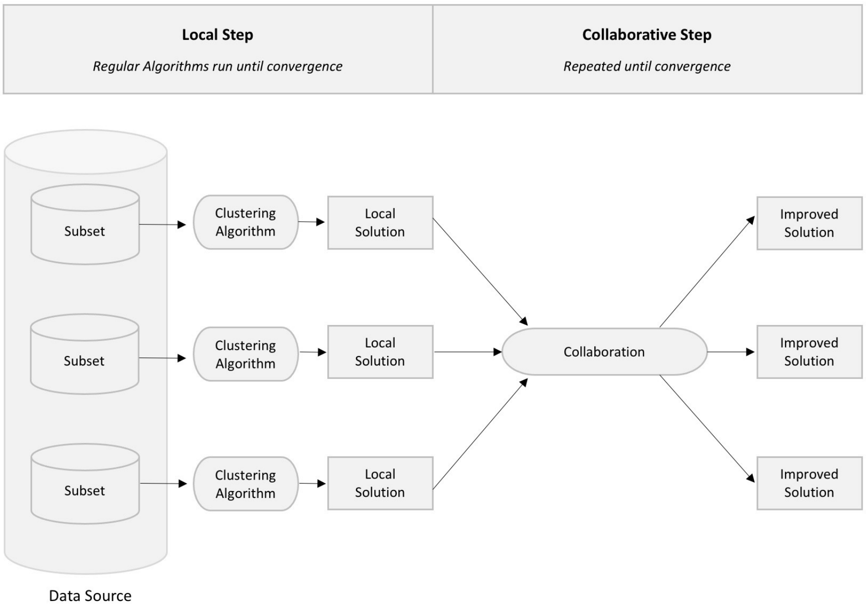


FIGURE (2.8) General mechanism of Vertical collaboration clustering framework

optimum of all the models that work together. The second framework, intends to find local optimum for each sets.

We have discussed different method of multi-views clustering that belong to the first framework, where the main goal is to create the transition from multi-view clustering to a consensus clustering, where the views can be build from a splitting of the data, alongside the features, and the consensus is unified view that combines all the knowledge coming from different views for the purpose to find a global optimal model that guarantee to a better clustering of the data sets.

In the second section, we put the lights on a very important research area, that belong to the second framework, which is the collaborative clustering. The main idea behind this approach is to create interaction between the different sources of information that can share either the same instances captured from different views, or different samples represented in the same feature space. the main goad of this algorithm is to exchange the knowledge between different sets, so that the local quality of clustering of each sets increased regarding the privacy of local information of each sets.

In this thesis, we will tackle this frameworks by introducing the optimal transport theory, in the purpose to propose new approaches of Multi-view clustering and collaborative clustering higher quality, better convergence and well defended formalism. In chapter 3, we introduce the optimal transport theory in the multi-view clustering, where we proposed two main approaches that aim to transport the local distribution of the views intending to learn a consensus distribution using with higher quality of clustering and better representation of the data sets. In chapter 4, we introduce a new formalism of collaborative clustering based on optimal transport theory where we aim to develop choice criteria based on the comparison of the data distribution in each set without sharing the information of the data. In addition we develop a stopping mechanism that intending to avoid negative collaboration. In chapter 5, we improved the collaborative clustering based on optimal transport theory by introducing a feature selection algorithm that aim to improve the interaction between the collaborators that can be seen as multi-level collaboration, in order to avoid the negative collaboration and ensure better quality of clustering.

3 Multi-view Clustering through Optimal transport

3.1 Introduction

Unsupervised clustering methods have become recently more and more popular because of their ability to cluster unlabeled data which was very difficult to realize for human being. A significant number of new clustering algorithms have been developed in recent years, and existing methods have also been modified and improved. This abundance of methods can be explained by the difficulty of proposing generic methods that adapt to all types of data available. Indeed, each method has a bias induced by the objective chosen to create the clusters. Therefore, two different methods can offer very different clustering results from the same data. In addition, the same algorithm can provide different results depending on its initialization or its parameters.

To solve this problem, some methods propose to use several different clustering results to better reflect the potential diversity of the results. These approaches take advantage of the information provided by the different results in a significantly different way. multi-view-clustering is one of this strong methods,the aim of this method is to form a consistent clusters of similar subjects by combining the multi-view feature information instead of the classical clustering method that use only a single set of features or one information of the subjects.

The important of this method is in its diversity of the features, this diversity guarantee not only a consistent regrouping but also more precision in therm of the interpretation of clusters, thanks to the cluster representative witch includes all information and the precision of all the views.

We propose a new method of multi-view clustering based on theory. More specifically, this new method aims to learn a new data structure from distribution's data on each view in order to increase the quality and richness of the cluster. thanks to optimal transport theory which not only allows us to compare the distribution but also to transport information to form a better consensus.

The rest of the chapter will be organized as follows: we will present the most important related work to multi-view clustering, then we will briefly introduce some theoretical background of OT in section 3 and proceed to our proposed approach in section 3 and finally evaluate our approach on synthetic data and real data.last section concludes the chapter and gives a couple of hints for possible research.

3.2 Consensus clustering

The idea of combining multiple models came to resolve the paradigm of distributed data, where the main idea is to get a unified views of the global data sets, while ensuring the higher quality and better visualization of the data. This fusion of information must done respecting many properties in order to enhance the quality, to avoid losing the knowledge learned from the models. To do so, we remind the most important conditions to checked:

- Robustness: the combination must performs better models than single clustering algorithms.

- Consistency: the learned model in ensemble clustering must avoid losing the real structure of the data.
- Novelty: the ensemble models must guarantee finding better results which is not attainable with single clustering algorithm.
- Stability : we must obtain lower sensitivity to noise and outliers.

The challenge in ensemble clustering is how to construct an appropriate consensus function. Mainly the consensus function must contains two principle part : the first one must be capable to detect the best clustering model for each sets,and the second part intends to improve these model by finding the right fusion, in order to learn a better model than a single view. In the stat of the art, there is two main consensus function approaches: the objective co-occurrence and the median partition.

The idea behind the first approach consists the find the best cluster label in the consensus partition that must associate to each object. In this approach the consensus is obtained through a voting process among each objects, where we analyze how many time the object belongs to the same cluster, so it can be associated to in the consensus partition [3].

The second approach, the consensus model is obtained based on an optimization problem regrading the local models. The main idea is to maximize the similarity between the local partition where the main challenge is how to chose an appropriate measure of similarity, where either it can a counting pairs measure that measure the agreement between a two pairs objects [47]. Or set matching measures, which are based on set cardinally comparison [39]

On the other hand, the transition from Multi-view learning to a consensus clustering, was introduced in [67] by Yarowsky, were it is applied for word sense disambiguation. Mainly, this approach is based on a voting consensus which is represented by two different classifiers: the local context of a word as first view and the senses of the other occurrences of that word as a second view. In [7], the authors introduced an approach of co-training based on two hypotheses trained on distinct views. The algorithm aims to improve the global learning quality of two classifiers with the highest confidence instances from unlabeled data. This approach requires that the views are independent.

In the same context, in [8], a co-EM algorithm was presented as a multi-view consensus version of the Expectation-Maximization algorithm for semi-supervised learning. In 2004, Bickel and Scheffer [5] studied the problem where the set of attributes can be split randomly into two subsets. These approaches seek to optimize the agreement between the views. The authors described two different algorithms, an EM-based algorithm which gives very significant results and an agglomerative multi-view algorithm which seems to be less efficient than single view approaches.

To summary, the between multi-view clustering is considered as basic task for several subsequent analyses in machine learning, in particular for ensemble clustering [54], also named aggregation clustering. The aim of ensemble clustering is to bring all the clusters information from different sources of the same data-set, or from different runs of the same clustering algorithm, so as to form a consensus clustering that includes all the information. This approach becomes a framework from multi-view clustering when it is applied to a clustering with a multi-view description of the data [51,63].

In what next, we propose a Multi-view consensus clustering based on optimal transport theory (see chapter 1) where the main idea to use this proposed distance that balance between a good transportation of the knowledge, and optimum transportation cost.

3.3 Multi-view Consensus Clustering through Optimal Transport

In this section we show how the multi-view-clustering can be solved using optimal transport theory, and how to ensemble all the information from all the views to form a consensus in optimal way.

3.3.1 Motivations

In order to justify our approach from theoretical point of view, we will explain the fundamentals of multi-view-clustering and how it can be transformed into an optimal transport problem. Multi-view clustering can be divided into two steps, the local step which consist basically to get a better cluster in each view, and global step which consist to aggregate this information- *centroids of clusters*-to from a consensus representing all the view at the same time.
We consider $X = \{x^1, x^2, .., x^r\}$ with r multiple views. where $x^v = \{x_1^v, x_2^v, .., x_n^v\}$ with n points in Ω in the vth view, in general way the methods existing to from a unified view or a consensus is maximized some objective function that combines the basic partitioning $H = \{h_1, h_2, .., h_r\}$ given by some algorithm to find a consensus partitioning h, where the choice of the utility function U is very important.

$$\Gamma(h, H) = \sum_{i=1}^{r} w_i U(h, h_i) \tag{3.1}$$

where $\Gamma : \mathbb{Z}_+^n \times \mathbb{Z}_+^{nr} \mapsto \mathbb{R}$ is the consensus function, and $U : \mathbb{Z}_+^n \times \mathbb{Z}_+^r \mapsto \mathbb{R}$ the utility function with $\sum_i^r w_i = 1$. This problem can be transformed to a minimization problem without changing its nature by using different distances like Mirkin distance-[39]. Moreover, [60] they proved that the consensus problem is equivalent to K-means problem under some assumption defined in the following definition.

Definition 3.3.1 *[60] A utility function U is a K-means consensus clustering utility function, if $\forall \quad H = \{h_1, \ldots, h_r\}$ and $K \geq 2$, there exists a distance f such that*

$$\max_{h \in H} \sum_{i=1}^{r} w_i U(h, h_i) \Leftrightarrow \min_{h \in H} \sum_{k=1}^{K} \sum_{x_l \in C_k} f(x_l^{(b)}, c_k) \tag{3.2}$$

holds for any feasible region H.
Where $X^b = \{x_l^b \mid 1 \leq l \leq n\}$ be a binary data set derived from the set of r basic partitioning H as follow:
$x_l^{(b)} =< x_{l,1}^{(b)}, \ldots, x_{l,i}^{(b)}, \ldots, x_{l,r}^{(b)} >$, *with* $x_{l,i}^{(b)} =< x_{l,i1}^{(b)}, \ldots, x_{l,ij}^{(b)}, \ldots, x_{l,iK}^{(b)} >$
and $x_{l,ij} = \begin{cases} 1 & L_{h_i}(x_l) = j \\ 0 & otherwise \end{cases}$ *and* c_k *is the centroids of the cluster*

Based on the idea that consensus clustering can be seen as a K-means problem, we can transform it to an optimal transport problem based on the *Wasserstein* distance. more specifically, we can see each view as a distribution set that we can assemble to form an optimum consensus represented in the global step.

We will detail the proposed approach and how we will improve the consensus clustering using the optimal transport theory. we will also propose different types of consensus that will be validated by tests on several data sets to evaluate each method

Notations	
X	the global data such that $x_i \in \mathbb{R}^d$
μ	the distribution $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$
X_v	the view v such that $x_i \in \mathbb{R}^{d_v}$ with $d_v < d$
μ^v	the distribution of the view v $\mu^v = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i^v}$
d_v	the dimension of the view v
c_j^v	the centroid j in the view v between the data and the centroids c_j^v
ν^v	the distributions of the centroids $\frac{1}{k_v}\sum_{j=1}^{k_v}\delta_{c_j^v}$
$L^v = \left\{l_{ij}^v\right\}$	the optimal transport matrix of the view v
c_k	the centroids of the the consensus cluster
$L = \{l_{ik}\}$	the optimal transport matrix between the centroids of each view and the consensus centroids
ν	the distribution of the consensus centroids $\frac{1}{K}\sum_{k=1}^{K}\delta_{c_k}$
Π	the optimal transport matrix between x_i and c_k

TABLE (3.1) Notations

3.3.2 Proposed approaches

We remind the Wasserstein distance defined in chapter 1 (see. definition 1.19) That allows to define distance between distribution. The Wasserstein distance has been very useful recently especially in machine learning such as domain adaptation [13] metric learning [15], clustering [16] and multi-level clustering [30]. The particularity about this distance is that it takes into account the geometry of the data using the distance between the samples, which explains its efficiency. On the other hand, in term of computation, the success of this distance also comes from the work of Cuturi [14], who introduced an algorithm based on entropy regularization, as presented in the next section.

Even though the Wasserstein distance has known very significant successes, in term of computation the objective function has always suffered from a very slow convergence, especially in high dimension, which lead to the idea of proposing a smoothed objective function by adding a term of entropic regularization, introduced in [45] and applied to the optimal transport problem in [14] in order to speed up the convergence and improve the stability.

Let $X = \{x_1, x_2, \dots, x_N\}, x_n \in \Omega \subset \mathbb{R}^{d\times 1}, 1 \le n \le N$ be our data set made of d numerical attributes. Let $X^v = \left\{x_1^v, x_2^v, \dots, x_n^v\right\}, x_n^v \in \mathbb{R}^{d_v \times 1}, d_v < d$, the subset of attributes processed by the view v.

```
Algorithm 1: Local view algorithm
Input  : for the data of the $vth$ view $v$, $X^v = \left\{x_i^v\right\}_{i=1}^n \in \mathbb{R}^{d_v}$ such that $d_v < d$ and
         $k_v$ the number of clusters
         the entropic constant $\lambda$
Output : The OT matrix $L^v = \left\{l_{ij}^v\right\}$ and the centroids $c_j^v$
Initialize $k_v$, random centroids $c^v(0)$ with the distribution $\nu^v = \frac{1}{k_v}\sum_{j=1}^{k_v}\delta_{c_j}$ ;
t=0;
while not converge do
    Compute the OT matrix $L^v = \left\{l_{ij}^v\right\} 1 \le i \le n, 1 \le j \le k_v;$
        $$L^v = \min W_{2,\lambda}^2(\mu^v, \nu^v);$$
    Update the distribution centroids $c_j^v$:
        $$c_j^v = \sum_i l_{ij} x_i^v \quad 1 \le j \le k_v;$$
return $\{L^v\}$ and $\left\{c_j^v\right\}_{j=1}^{k_v}$
```

Local step

We consider the empirical measure of the data: $\mu^v = \frac{1}{n}\sum_{i=1}^n \delta_{x_n^v}$, which represent the data of each view v $x_n^v, 1 \le i \le n$ in the view v is uniformly distributed over the view v.

We seek to find a discrete probability measure $\nu^v \in \Sigma_{k_v}$ which is an approximation of μ^v defined by k_v centroids $C^v = \left\{c_1^v, c_2^v, \ldots, c_{k_v}^v\right\}$. To this end we compute the optimal transport of μ^v from Σ_n to Σ_{k_v}, we therefore define ν^v as the solution of the following problem:

$$\underset{L^v \in \Pi(\mu^v, \nu^v), C^v}{\operatorname{argmin}} \quad W_{2,\lambda}^2(\mu^v, \nu^v) \tag{3.3}$$

We should notice that in [16], when $d = 1$ and $p = 2$ and without constraints on the weight over Σ_{k_v}, this problem is equivalent to Lloyd's algorithm. In what follows we consider $p = 2$. In order to solve the optimization problem (3.3) we proceed similarly to k-means clustering. The local step for the view v clustering itteratively alternates between the assignment of each data to the nearest centroid and the optimization of the centroids $C^v = \left\{c_1^v, c_2^v, \ldots, c_{k_v}^v\right\}$.

The algorithm 1 describe how we cluster the data locally, it is an alternation between the computation sinkhorn optimal transport algorithm to assign each data point to its nearest centroid and the update of the centroids distribution to be the average weighted to the data point assigned to it. It should be noted that algorithm 1 is equivalent to K-means,but it allows a soft assignment instead of the hard one which means that $l_{ij}^v \in [0, \frac{1}{n}]$ and also the term of regularization $-\frac{1}{\lambda}H(\gamma)$ will guarantee a solution with higher entropy, which means that the point will be more uniformly assigned to the clusters.

Global step

We aim on the global step to ensemble all the information that we already get in each view to form a consensus cluster for all data. We distinguish two approaches based on optimal transport theory: projection approach, consensus with new representation.

Consensus Projection approach

Consensus Projection approach (CPA) consist to project structure of the cluster of each view on the global space, the idea behind this projection is to visualize the structure of each view in the global space to enrich the information on the data so we can increase the quality of the consensus clustering. More precisely, it is a kind of super clustering to obtain more precise prototypes that contain more information about the data. As a result,the matrices of the partition resulting from the transportation of the data will be more complete and more precise and will guarantee a higher quality.

```
Algorithm 2: Consensus with projection (CPA)
Input  : Data $X = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$ represented by $\nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$and $L_v$ and
         $\left\{c_j^v\right\}_{j=1}^{k_v}$ represented by the distribution $\nu^v$, $1 \leq v \leq r$
         The number of the cluster $K$
         The entropic constant $\lambda$
Output : the consensus centroids $c_k$ and $\Pi = \{\pi_{ik}\}$ the OT matrix for all the data
Initialize $K$ centroids $c_k(0)$ with the distribution $\nu = \frac{1}{K}\sum_{k=1}^{K}\delta_{c_k}$;
t=0;
compute the matrix centroids $C = \{c_k\}$ in the global space:
    $$C = L^v X \quad for \quad 1 \leq v \leq r$$
while not converge do
    Compute the OT matrix between the views and the consensus
     $L = \left\{l_{jk}\right\} 1 \leq j \leq k_v, 1 \leq k \leq K, 1 \leq v \leq r;$
        $$L = \text{argmin}\, W_{2,\lambda}^2(\nu^v, \nu);$$
     Update the consensus centroids $c_k$;
        $$c_k = \sum_{i=1}^{n} l_{ik}.x_i \quad 1 \leq k \leq K;$$
compute the OT between the data and the consensus centroids $c_k$;
    $$\pi_{ik} = \underset{k}{\text{argmin}}\, W_{2,\lambda}^2(\mu, \nu)$$
return $c_k$ and $\Pi = \{\pi_{ik}\}$
```

In algorithm 2 we explain the mechanism of this method, the first step of the algorithm consist to project the centroids of each view in the global space of the data and then re-cluster the projection of the centroids in the global space using Sinkhon -means algorithm based on Wasserstein distance to obtain a new prototypes c_k, the last step of the algorithm aim to transport the instances the new prototypes.

Consensus with new representation

Consensus with new representation (CNR) aims to ensemble the structure obtained in each view in order to rebuild a new representation of the data based on partition matrix.This representation gives the posterior probability of belonging of each point to each cluster of each view, in other words it is a kind of superposition of all the views which allows to form a consensus clustering of all information that we already got from each view.

Algorithm 3 explains the process this method, the first step is to concatenate all the matrix partition and re-cluster it using $Sinkhorn$ algorithm to obtain a better partition of the data and centroids that contains information emerge from each view.

3.4 Experimental validation

In this section we will evaluate the approaches and test them on several data sets,we will also compare it to classical consensus K-means clustering algorithm and the clustering of single view.

Algorithm 3: Consensus with new representation (CNR)

```
Input  : $L_v$ and $\left\{c_j^v\right\}_{j=1}^{k_v}$ represented by the distribution $\nu^v$, $1 \leq v \leq r$
         The number of the cluster $K$
         The entropic constant $\lambda$
Output : The consensus centroids $c_k$ and $\Pi = \{\pi_{ik}\}$ the OTmatrix for all the data
Initialize $K$ centroids $c_k(0)$ with the distribution $\nu = \frac{1}{K}\sum_{k=1}^{K}\delta_{c_k}$;
t=0;
Compute the new representation matix of the data;
        $X = concat(L_v) \quad for \quad 1 \leq v \leq r$
while not converge do
    Compute the OT matrix $\Pi = \{\pi_{ik}\}$ ;
        $\Pi = \operatorname{argmin} W_{2,\lambda}^2(\mu,\nu);$
    Update the consensus centroids $c_k$;
        $c_k = \sum_{i=1}^{n} \pi_{ik}.x_i \quad 1 \leq k \leq K;$
return $c_k$ and $\Pi = \{\pi_{ik}\}$
```

3.4.1 Experimental protocol

In order to test experimentally the proposed algorithm, we first proceeded with a data pre-processing in order to create the local views. Mainly, we split the data into 5 views where each view represents the data in different feature space. Then we cluster the data locally using algorithm 1. After that, we applied the consensus proposed algorithms on Multiple views for each data set, we evaluate the approaches comparing to a single views clustering using an unsupervised clustering [17], and we compare it to a classical method of consensus clustering based on $k-$means.

We mentioned that the used data sets are available ON UCI Machine Learning Repository [21].

Datasets	#instances	#Attributes	#Classes
Dermatology	358	33	6
Ecoli	332	7	6
Iris	150	4	3
PenDigits	10992	16	10
Satimage	4435	36	6
WDBC	699	9	2
Wine	178	13	3

TABLE (3.2) Some Characteristics of used Real-World Datasets

3.4.2 Experiment results

In this section we will evaluate the approaches and test them on several data sets represented in the previous section see table 3.2, we will also compare it to classical consensus K-means clustering algorithm and the clustering of single view. Table 2 and Table 3 summarize the clustering results of the proposed approaches (**CNR** and **CPA**) and other methods in terms of DB and R_n respectively. As can be seen, our approaches generally performs best on all the data sets by both metrics.

In table 3.3 we evaluate the two approaches (CNR and CPA), compared to a Single View approach (SVA) using $Davies - Bouldin$ index [17]

$$DB = \frac{1}{K}\sum_{k=1}^{K}\max_{k\neq k'}\frac{\Delta_n(c_k)+\Delta_n(c_{k'})}{\Delta(c_k,c_{k'})} \tag{3.4}$$

where K is the number of clusters, Δ_n is the average distance of all elements from the cluster C_k to their cluster centre c_k, $\Delta(c_k, c_{k'})$ is the distance between clusters centres c_k and $c_{k'}$. This index well evaluates the quality of unsupervised clustering because it's based on the ratio of the sum of within-clusters scatter to between-clusters separation. More the value of DB is lower, means that we have a better cluster.

Datasets	CNR	CPA	SVA
Dermatology	1.926	**1.194**	1.310
Ecoli	**1.145**	1.405	1.236
Iris	**0.893**	0.908	0.915
PenDigits	**1.136**	1.334	1.257
Satimage	**1.011**	1.221	1.274
WDBC	1.735	**1.727**	1.742
Wine	1.308	**0.556**	**0.556**

TABLE (3.3) Clustering performance on seven real-world data sets by $DavisBouldin$ Index DB

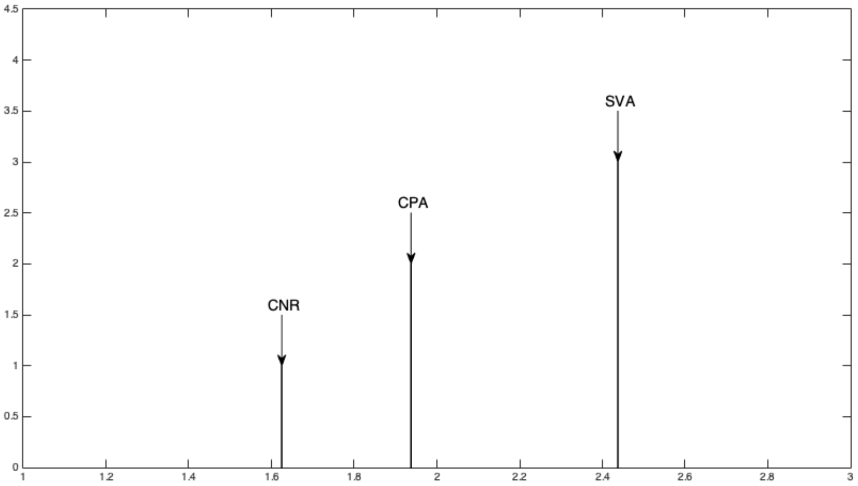


FIGURE (3.1) Friedman test for comparing multiple approaches over multiple data sets

As we can see in table 3.3 the **CNR** obtained better values for several data sets, this is explained by the fact that this method makes the clustering on the structures of each view which guarantee the improvement of the quality of the global cluster.

while the method of **CPA** do a kind of forcing assignment of the instances to the centoirds obtained from the consensus clustering.To better analyze the DB values,we propose in figure 3.1, the critical diagram represents a projection of the average ranks methods on enumerated axis. The methods are ordered from left (the best) to right (the worst), in our case, the method **CNR** is the best and the worst is **SVA**. Approach **CNR** outperforms the other proposed techniques since it gives a new representation for data including all the structures of all the views.

Data sets	CNR	CPA	SVA	KKC
Dermatology	**0.4608**	0.1202	0.1472	0.0352
Ecoli	0.2822	0.3447	0.3389	**0.5065**
Iris	0.4423	0.4605	0.4491	**0.7352**
PenDigits	0.5064	**0.6039**	0.5356	0.5347
Satimage	0.3576	**0.4702**	0.4679	0.4501
WDBC	0.3315	**0.7374**	0.6891	0.0556
Wine	**0.2264**	0.2149	0.2149	0.1448
Score	5.2074	**5.5170**	5.3651	4.6340

TABLE (3.4) Clustering performance on seven real-world data sets by Normalized Rand Index R_n

In table 3.4 we validate our approaches comparing the classical consensus based on k-means by the normalized Rand index R_n [61] which measures the agreement between two partitions: one given by the clustering process and the other defined by external criteria . The values of R_n is included between $[0, 1]$, when the value is close to 1, the quality of the cluster is much better.As we see, the highest score comes back to **CPA**.

This is explained by the fact that in the last step of **CPA** algorithm, we force the assignment of the instances to the new centroids protected from the views which explain the agreement between the labels predict from the consensus clustering and the true one. While on the **CNR** we cluster a new representation of the data.It should be noted that We choose this index to compare our approaches with the classical method of ensemble clustering.

However, as long as we are comparing between true and predict labels, this index doesn't emphasize the unsupervised clustering.To further evaluate the performance, we compute a measurement score by following [71]:

$$Score(M_i) = \sum_j \frac{R_n(M_i, D_j)}{\max_i R_n(M_i, D_j)} \tag{3.5}$$

where $R_n(M_i, D_j)$ indicates the R_n value of M_i method on the D_j data sets. This score gives an overall evaluation on all the data sets, which shows our approaches outperforms the other methods substantially in most cases.

3.5 Summary and discussion

We have proposed a new way to explore the advantages and usefulness of multi-view-clustering. We have reviewed the multi-view-clustering within the framework of optimal transport theory by proposing two new clustering algorithms that allow to combine the structures discovered by several views, in the form of a consensus. One algorithm based on projections and another uses the partitions found by different views to create new representations of the data. Experiments on seven real-world data sets showed the effectiveness of the two proposed algorithms compared with both a single-view approach and another classic state-of-the-art technique.

The results obtained show that optimal transport can provide for this machine learning task a formal framework and algorithmic flexibility that marks an improvement in performance over the existing system. On the other hand, optimal transport is becoming increasingly popular in the field of machine learning, with several applications to data science under different learning paradigms. In large dimensions, we are often confronted with the intrinsic instability of optimal transport with regard to input dimensions.

Indeed, the complexity of approximating Wasserstein's distances can grow exponentially in size, making it difficult to estimate these distances accurately. A multi-view approach can help to mitigate this phenomenon by breaking down the overall problem into subproblems, each representing a view. Future work includes the investigation of learning capabilities of new data representations through our proposed approaches. We also want to see how to use the deep learning paradigm to enrich our multi-view approaches.

4 Collaborative clustering through Optimal Transport

4.1 Introduction

Data clustering is one of the main interests in unsupervised Machine Learning research [35]. A large number of clustering algorithms have been proposed in the literature [61], divided into different families based on the cost function to optimize [35, 67].

Clustering task is known to be difficult and suffer from several issues. Most of the problems come from the fact that unsupervised algorithms work with very little information about the expected result [65]. Therefore, the choice of the cost function to optimize, the algorithm to use and the values of the parameters require a lot of expertise to obtain the desired output [20]. In addition, modern data-sets are often very large (both in size and dimension) and distributed into several sites [66], which limit the efficiency of most classical clustering algorithms.

In an attempt to solve these issues, the scientific community has suggested several ways of combining the results of different algorithms [64]. Several approaches have been proposed in that direction, based on the idea of several algorithms working on the data, either with each algorithm optimizing a different cost function or working with different values of the parameters on the same data-set, or with each algorithm working on a subset of the data, usually trying to optimize the same cost functions.

These approaches can be classified into two main categories: In Ensemble Learning approaches, several algorithms are trained on the data and the set of results are merged into a global consensus [60]. In Collaborative Clustering several models are trained simultaneously on the data-set, usually each algorithm working on a sub-set of the data, and exchange information during the learning process [22]. In this study we focus on the later approach.

Generally speaking, the problem of Collaborative Clustering can be defined as follows: Given a finite number of disjoint data sites, collaborative clustering is a scheme of collective development and reconciliation of fundamental cluster structures across these sites [41]. The general framework for collaborative clustering is based on two principal steps:

Local step: Each algorithm will train on the data it has access to and produce a clustering result, e.g. a model of the local data subset.

Collaborative step: The algorithms share their output in order to confirm or improve their models, with the goal of finding better clustering solutions.

In this chapter, we propose to study the unsupervised collaboration framework through the optimal transport theory, thus benefiting from this mathematical formalism to analyze and describe the process of collaboration between the different algorithms. In this case, the collaboration, that consists of exchanges of information between algorithms, will be modeled in the form of bi-directional or even multi-directional transports.

The rest of the Chapter is organized as follow. Section 4.2 develops the state of the art about prototype based collaborative methods. In Section 4.3 we introduce the novel framework of collaborative clustering using optimal transport theory. In Section 4.4 we provide an experimental validation and discuss the quality of the proposed approach comparing to classical methods. Finally, in Section 4.5 a conclusion and some perspective work are given.

4.2 Prototype based Collaboration

Collaborative Fuzzy C-Means

The first Collaborative clustering was introduced by [41] under the name "Collaborative Fuzzy Clustering” (CoFC). This approach was based on extended version of Fuzzy C-means adapted to distributed data. The algorithm was based on two steps, the first step aims to find c clusters for each collaborator where each objects is assigned to some cluster with a certain degree membership stored in matrix S. The second step consists to exchange the information stored in the matrix S or the prototypes of each cluster. The algorithm of Fuzzy C-Means is trained again for each collaborator taking into account the shared information.

Several studies had been done to develop several algorithms and approaches on this framework, such as CoEM in [6], CoFKM [11] and collaborative EM-like algorithm (EM for Expectation–Maximization) based on Markov Random Fields [31]. All this approaches follow the same principle as Collaborative Fuzzy C-Means.

However, These algorithms display similar limitations: they require the same number of cluster in each site, the same same model trained in each site, and the algorithm can only happen between instances of the same algorithm.

Collaborative Clustering through SOM

The collaborative clustering was also developed based on Self-Organization Maps (SOM) [26, 27] by adapting the original objective function to distributed data. The main idea was to add a term inspired by the classical SOM neighborhood function to the original SOM objective function, where this term aims to compare neighborhoods of each prototype in each sites. This neighborhood term is adaptable to either horizontal or vertical collaboration.

The same principle can also be adapted to the Generative Topographic Maps (GTM) [23] with a modification in the M-step of the EM algorithm. The modification consists to add collaborative term inspired from the penalized likelihood estimation [25].

Although these approach have presented positive results, and the number of cluster does not mutter either in collaborative SOM or collaborative GTM, but in both approaches, the maps of each collaborator must have the number of neurons and topologically close to each each other. This condition is very restraining since the collaboration is based on the information that exists in the neurons. Moreover, the collaboration based on SOM and GTM requires an input collaborative confidence parameter, which defines the importance of distant collaborators, this parameter is updated during the collaboration based on the quality of each collaborator. However, since we are working in unsupervised learning case, choosing and learning this parameter is very tricky which can affect strongly the final results.

SAMARAH method

The SAMARAH algorithm was proposed in [22, 58], where the main idea is focused only on horizontal collaboration with the advantage that it does not require a smoothness function or the same number of clusters or prototypes. The idea behind this approach is to improve quality of each model of clustering trained on the same data sets by creating the collaboration between them, to reduce their diversity in order find a consensus model easier with a higher quality.

Precisely, the SAMARH method requiers tree steps:

- Local step: generation of local clustering results by performing some clustering algorithm such as (K-mean, EM..).
- Collaborative step : refinement of the models through the collaboration using o confusion matrix that maps all the models and detects the conflict between different partitions, and then solves the conflict by splitting the clusters, or merging them in single cluster or delete the weak one.
- Consensus step : aggregation of the improved results to form unified model of all the data sets.

To sum up, we can that the SAMARAH method has many strengths and covers all the steps of learning also doesn't require the same algorithm for all the collaborators. However, this method has a weakness side : it doesn't covers the vertical collaboration, the principle of solving the conflict based on pairwise criterion can make the process volatile.

Study of diversity

Recent works have been done to develop the collaborative clustering and make it more flexible [36, 50], ensuring a collaboration between different algorithms without fixing a unique number of clusters for all of the collaborators. The advantage of this approach is that different families of clustering algorithms can exchange information in a collaborative framework.

One of the most strong challenges in collaborative clustering is measure of diversity between collaborators. The measurement of diversity can resolve many limitations of collaborative clustering like the order of the collaboration which is build from the choice of distant collaborators. This choice must take into account the quality and diversity to be pertinent. In [42], the authors develop a new heuristic criterion to select the optimal collaborator. They showed that the diversity between collaborators could be an important impact on collaboration. recently a study of the influence of diversity on the collaboration was done based on the entropy [49] and showed trade-off between the gain quality and diversity between the collaborators.

In the next section we detail the proposed the novel frame work of collaborative clustering based on optimal transport, where we aim to resolve this many limitation in collaborative clustering.

4.3 Proposed approach

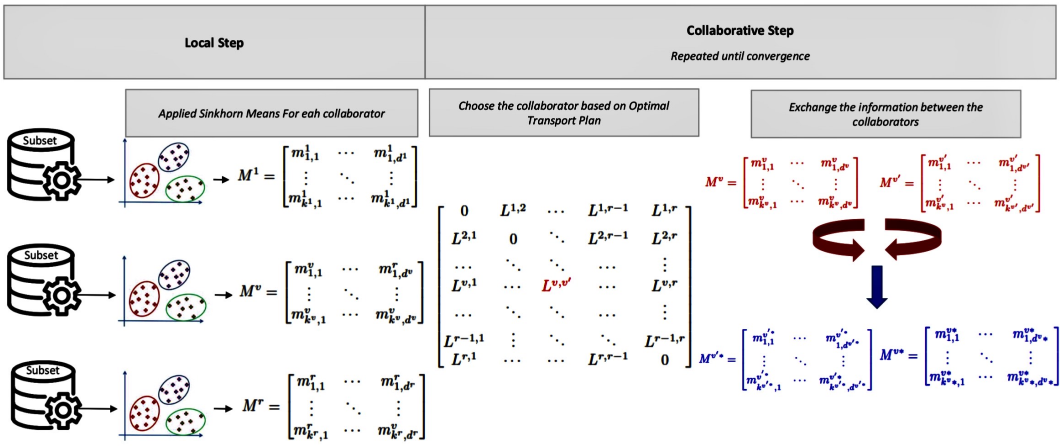

Local Step
Collaborative Step
Repeated until convergence
Applied Sinkhorn Means For eah collaborator
Choose the collaborator based on Optimal Transport Plan
Exchange the information between the collaborators
Subset
Subset
Subset

FIGURE (4.1) The mechanism of the proposed approach where the local step consist to find the centroids for each collaborator, the collaborative step consist to find the best collaborator by computing a transport plan between the local distribution and the final step consist to influence the local distribution by the distant distribution prototype.

4.3.1 Motivation

With the development of hardware technology, a huge amount of data represented in different views and different structures have been generated in real word applications. This kind of data is considered as a new challenge to develop the existing clustering algorithms, designed for single view data, to be more adaptable to multi-view data.

One of the most difficult challenges in collaborative learning is how to choose the right collaborator to collaborate with, which construct the order of the collaboration, not only to increase local quality of each model, but also to ensure the convergence and to avoid over-fitting (Figure 4.1).

Classical collaborative algorithms are based on two steps. The first one consists to cluster the data locally, the second consists to send and receive information between the local models. Despite the quantity of work on this framework, it still requires many restrictions to ensure the convergence: usually each algorithm must work on the same representation space and must compute the same number of clusters. These restrictions limit the flexibility of collaborative clustering approaches for the analysis of real data.

On the other hand, Optimal Transport theory has shown very significant results, especially in transfer learning [13] and for comparison of distributions. Based on this idea, our intuition is to model collaborative learning as a bi-directional knowledge transfer and improve the optimization of the cost function based on the comparison of the distributions of local subset, in order to weight the mutual confidence of the collaborators and use a transport plan to transfer the information between them.

In the next section we detail the proposed approach based on Optimal Transport theory, either in vertical or horizontal collaboration.

4.3.2 Collaborative Learning algorithms

The main goal of the proposed approach is to improve the quality and the stability collaboration and guarantee the convergence without over-fitting. In collaborative clustering, we distinct two principal approaches: vertical and horizontal collaboration.

In the vertical collaboration, the collaborators learn from different instances represented in the same space, while in horizontal collaboration the collaborators work on the same instances in different representation space.

In general, different frameworks must be used for vertical and horizontal collaboration. Here we propose a unified framework adapted to both approaches.

Local step

Let consider r collaborators, where the data of each collaborator v, $X^v = \left\{x_i^v\right\}_{i=1}^{n^v}$ with $x_i^v \in \mathbb{R}^{d^v}$ and d^v the dimension of the space representation of v, corresponds to a distribution $\mu^v = \frac{1}{n^v}\sum_{i=1}^{n^v}\delta_{x_i^v}$.

We seek in the local step to find the centroids $M^v = \{m_1^v, .., m_{k^v}^v\}$, corresponding to a distribution ν^v, that represents the local clusters of each collaborator v, such as to minimizes the Optimal Transport plan $L^v = \left\{l_{ij}^v\right\}_{i,j=1}^{i,j=n^v,k^v}$ between the the local data X^v and the centroids M^v.

To achieve this, we will solve the following minimization problem (3.3), where the first minimization of L^v consists to find the Optimal Transport plan between the data and the centroids, and the second minimization M^v aims to update the distribution of the centroids so that the transport plan is optimal between the data and the centroids.

$$\underset{L^v\in\Pi(\mu^v,\nu^v),M^v}{\text{argmin}} \quad < L^v, C(X^v, M^v) >_F -\frac{1}{\lambda}H(L^v) \tag{4.1}$$

Subject to $\sum_{j=1}^{k^v} l_{ij}^v = \frac{1}{n^v}$ and $\sum_{i=1}^{n^v} l_{ij}^v = \frac{1}{k^v}$ and $C : X^v \times M^v \to \mathbb{R}_+$ s.t $C_{ij} = c(x_i^v, m_i^v) = ||x_i^v - m_j^v||^2$ the euclidean distance between sample x_i^v and the centroids m_i^v.

It should be noticed that resolving (4.1) is equivalent to a Lloyd's problem which is the Expectation Minimization algorithm when $d = 1$ and $p = 2$ without any constraints on the weights. This is why to resolve this problem we alternate between computing the Sinkhorn matrix L^v to assign instances the closest cluster and updating the centroids to decrease the transportation cost.

Algorithm 4: Sinkhorn-Means local algorithm

```
Input  : $X^v = \{x_i^v\}_{i=1}^{n^v}$: data of collaborator $v$ with distribution $\mu^v$
         $k_v$ : number of local clusters
         $\lambda$ : entropic constant
Output : The OT matrix $L^v$ and the centroids $M^v$

Initialize the centroids $M^v = \left\{m_j^v\right\}_{j=1}^{k^v}$ randomly
Compute the associated distribution $\nu^v = \frac{1}{k^v}\sum_{j=1}^{k^v}\delta_{m_j^v}$
repeat
    Compute the OT matrix $L^v = \left\{l_{ij}^v\right\}_{i,j=1}^{i,j=n^v,k^v}$ :
        $$(L^v)^* = \underset{L^v\in\Pi(\mu^v,\nu^v)}{\text{argmin}} < L^v, C(X^v, M^v) >_F - \frac{1}{\lambda}H(L^v)$$
    Update the centroids $M^v = \left\{m_j^v\right\}_{j=1}^{k^v}$:
        $$m_j^v = \sum_i l_{ij}^* x_i^v \quad 1 \leq j \leq k^v$$
until convergence;
return $(L^v)^*$ and $M^v$
```

Algorithm 4 details the computation of the local objective function (4.1), proceeding similarly to k-means but with the advantage of using the Wasserstein distance. This allows to get soft assignment of the data, in contrary to k-means, which means that the components of the assignment matrix $l_{ij} \in [0, \frac{1}{n}]$. Besides, the penalty term based on the entropy regularization guarantees a solution with higher entropy which increases the stability of the algorithm and ensures a uniform assignation of the instances.

Global step

The global step aims to compute the collaboration between the models where each collaborator can update its local clustering based on information exchanged with the other collaborators, until stabilization of clusters with improved quality. In the proposed approach, the collaboration step could be seen as two simultaneous phases.

The first phase aims to create an interaction plan based on Sinkhorn matrix distance which compares the local distribution of each collaborator to the others. The idea behind this phase is to allow each model to select the best collaborator to exchange information with, in other words the algorithm will also learns the best order of the collaborations in each iteration. The heuristic work in [42] proved that a collaboration with a model proposing a very different data distribution decreases the local quality, while a collaboration between very similar models is ineffective. Thus, the most beneficial collaboration is the one with models of median diversity. Hence, after the construction of the transport plan using the Sinkhorn algorithm, which compares the local structures, the proposed algorithm learns to choose for each model the collaborator with the median distribution similarity.

The second phase consists to exchange information between collaborators to improve local quality of each model. More precisely, we are looking to transport the prototypes to influence the location of the local prototypes; in order to get a higher local quality of each collaborator.

Considering the same notation above, we seek to minimize the following objective function:

$$\underset{L^v,M^v}{\operatorname{argmin}}\{< L^v, C(X^v, M^v) >_F -\frac{1}{\lambda}H(L^v) + \sum_{v'=1.v'\neq v}^{r} \alpha_{v',v}(< L^{v,v'}, C(M^v, M^{v'}) >_F \\ -\frac{1}{\lambda}H(L^{v,v'}))\} \tag{4.2}$$

Where the first term deals with the local clustering and the remaining is the collaboration term and represents the influence on the local centroids' distribution by the distant centroid's distributions. $\alpha_{v',v}$ are non-negative coefficients proportional to the diversity between the collaborators and the difference of local quality, and $L^{v,v'}$ is the Optimal Transport plan between the centroids of the v^{th} and v'^{th} collaborators.

Algorithm 5 explains the computation steps of the proposed approach and shows how it learns to select the best collaborator to learn from at each iteration, based on Sinkhorn comparisons between the distributions, and how it alternates between influencing the local centroids based on the confidence coefficient relative to the chosen collaborator and its local centroids distribution and update of the centroids relative to local instances in order to improve the clusters' quality.

It should be pointed out that in each iteration, each collaborator chooses successively the collaborators to exchange information with, based on the Sinkhorn matrix distance. More accurately, in each iteration, each model exchanges information with the collaborator having the median similarity between the two modelled distributions, computed with the Wasserstein metric. If this exchange increases the quality of the model (here we use the Davies-Bouldin index [17]), the centroids of the model are updated. Otherwise, the selected collaborator is removed from the list of possible collaborators and the process is repeated with the remaining collaborators, until the quality of the clusters stops increasing.

```
Algorithm 5: Collaborative clustering (Co-OT)
Input  : $\{X^v\}_{v=1}^r$: the $r$ collaborators' data with distributions $\{\mu^v\}_{v=1}^r$
         $\{k^v\}_{v=1}^r$ : the numbers of clusters
         $\lambda$ : the entropic constant
         $\{\alpha_{v,v'}\}_{v,v'=1}^{v,v'=r}$ : the confidence coefficient matrix
Output : The partition matrix $\{(L^v)^*\}_{v=1}^r$ and the centroids $\{M^v\}_{v=1}^r$

Initialize the centroids $M^v = \left\{m_j^v\right\}_{j=1}^{k^v}$ randomly
Compute the associated distribution $\nu^v = \frac{1}{k^v}\sum_{j=1}^{k^v}\delta_{m_j^v}, \forall v \in \{1,...,r\}$
repeat
    for $v = 1,...,r$ do
        Update the centroids $M^v$ and the partition matrix $(L^v)^*$ using a local
          algorithm (e.g. Sinkhorn-Means, SOM, GTM, $k$-Means...)
        Update the centroids distribution $\nu^v = \frac{1}{k^v}\sum_{j=1}^{k^v}\delta_{m_j^v}$
        for $v' = 1,...,r$ and $v \neq v'$ do
            Compute the OT matrix $(L^{v,v'})^* = \{l_{jj'}\}_{j,j'=1}^{j,j'=k^v,k^{v'}}$ between the centroids
              of collaborators $v$ and $v'$:
                $$(L^{v,v'})^* = \underset{L^{v,v'}\in\Pi(\nu^v,\nu^{v'})}{\text{argmin}} < L^{v,v'}, C(M^v, M^{v'}) >_F - \frac{1}{\lambda}H(L^{v,v'})$$
        Chose the median collaborator:
                $$v^* = median_{v'}\left\{(L^{v,v'})^*\right\}_{v'=1}^r, v' \neq v$$
        Update the local centroids based on the collaborator's information, if the
          internal quality is increased (see below):
                $$m_j^v = \alpha_{v,v*}\sum_{j'} l_{jj'}^{v,v*} m_{j'}^{v*} \quad 1 \leq j \leq k^v$$
until convergence;
return $\{(L^v)^*\}_{v=1}^r$ and the centroids $\{M^v\}_{v=1}^r$
```

It must be highlighted that the proposed algorithm can be adapted to both horizontal a vertical collaboration, since the inputs of the algorithm requires the distributions that represent the local structure of each collaborator, where it can be either sharing the same space but different samples (vertical collaboration), which formally means that $X^v = \{x_i^v\}_{i=1}^{n^v}$ such that $x_i^v \in \mathbb{R}^d$ or built from different spaces but share the same instances (horizontal collaboration), which means $X^v = \left\{x_i^v\right\}_{i=1}^n$ such that $x_i^v \in \mathbb{R}^{d_v}$.

Another important advantage of the proposed algorithm is its adaptability with all the prototype-based algorithms, more precisely instead of using the Sinkhorn-Means as a local algorithm to get the centroids, we can use other prototype models like k-means, SOM, EM, etc. The proposed algorithm has therefore the capability to work with hybrid models. This will be detailed in Section 4.4.2.

4.4 Experiments

4.4.1 Setting

Data-sets

We consider the following data-sets provided by the UCI Machine Learning Repository [19], described in Table 4.1. Each data-set is split between several collaborators.

TABLE (4.1) Some characteristics of the experimental real-world data-sets

Data-sets	#instances	#Attributes	#Classes
Glass	214	10	7
Spambase	4601	57	6
Waveform-noise	5000	40	3
Wdbc	569	33	2
Wine	178	13	3

Data-set splitting

In order to test experimentally the proposed algorithm, we first proceeded with a data pre-processing in order to create the local subsets.

For vertical collaboration, we aim create samples from the original data, which means different instances represented with same characteristics. We split the data horizontally into 10 random subsets X^v, each subset is represented by the distribution μ^v we train the algorithm 1 to get the local centroids partitions ν^v, and then applied the collaborative algorithm 5 between the subsets in order to increase their local quality. To do so, we split the data as showed in figure 4.2 where the data base is rated into v samples that share the same features.

For horizontal collaboration, the main idea is to split each chosen data set to 10 subsets, (see figure 4.3) that share the same instances but represented with different features

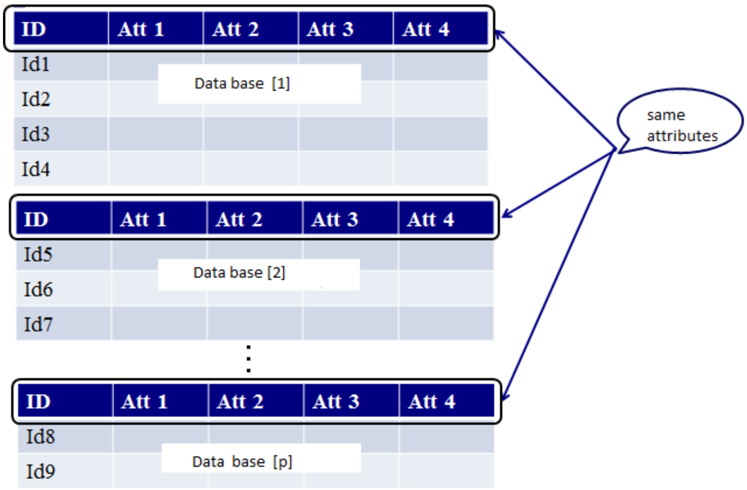


FIGURE (4.2) Splitting of the Vertical collaboration

in each subset, selected randomly with replacement. Considering the notation above, each subset X^v will be represented by the distribution μ^v that will be considered as the input of algorithm 1 to get the distribution of the local centroids ν^v. Algorithm 5 is then applied to influence the location of the local centroids by the centroids of the distant learners without having access to their local data.

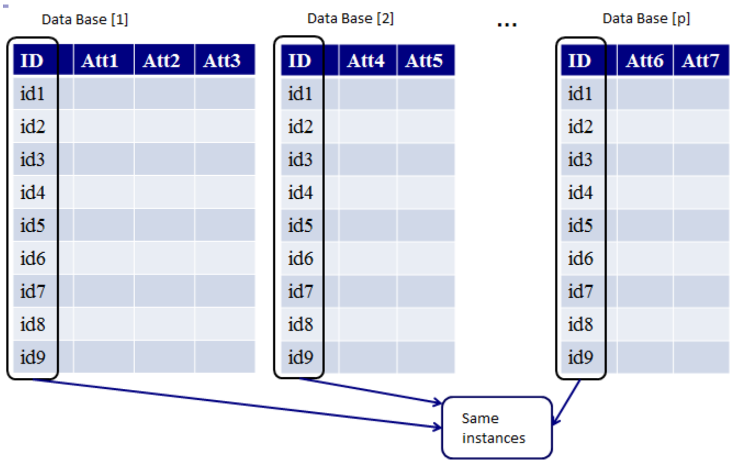


FIGURE (4.3) Splitting of the Horizontal collaboration

Quality measures

The proposed approach was evaluated with two internal quality indexes: Davies-Bouldin (DB) and Silhouette indexes, as well as an external criterion: Adjusted Rand Index (ARI). DB-Davies Bouldin index [17] is defined in the previous Chapter 3.4.

The silhouette index [43], is based on the measurement of the difference between the average of the distance between the instance x_i and the instances belonging to the same cluster a_i and the average distance between the instance x_i and the instances belonging to other clusters b_i, the closer the silhouette value is to 1 means that the instances are assigned to the right cluster.

$$S = \frac{1}{K} \sum \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{4.3}$$

Moreover, since the data-sets we proposed in the experiments provide available labels, we choose to add an external quality index the Adjusted Rand Index (ARI) [47].

The Adjusted Rand Index [61] defined as follow 4.4:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} / \binom{n}{2}}{\frac{1}{2}(\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}) - \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} / \binom{n}{2}} \tag{4.4}$$

Where $n_{ij} =\mid C_i \cap Y_j \mid$ and C_i is the ith cluster and Y_j is jth real class provides from the real label of the data-sets, and a_i is the number of instances belonging to the same cluster with the same class while b_j is the number of instances belonging to different cluster with different class.

The ARI index measures the agreement between two partitions, one provided from the proposed algorithm and the second one provided from the labeled data-sets. The values of ARI are between 0 and 1 and the quality is better when the value of ARI is close to 1.

We therefore applied algorithm 1 on local data, then the coefficient matrix α is computed based on a diversity index between the collaborators [42]. This coefficient is used to control the importance of the terms of the collaboration. Algorithm 5 in trained 20 times in order to estimate the mean quality of the collaboration and a 95% confidence interval for the 20 experiments. The experimental results of horizontal collaboration were compared with SOM-collaborative [23]. Both approaches were trained on the same subsets and on the same local model, a 3×5 map, with the parameters suggested by the authors of the algorithm [23]. The last part of the experiments results consists to compare the proposed algorithm with the collaborative algorithms proposed in the state of the art, where the algorithms is trained on only two collaborators, we followed the same split as mentioned in [23] to compare the gain quality brought from the collaboration, based on DB Davis-Bouldin index.

Computation tools

A nice feature of the wasserstein distance is that their computation is vectored, which means the computation of a n distances, whether from one histogram to many, or many to many, can be carried out simultaneously using elementary linear algebra operations. To do so we use the PyTorch version of Sinkhorn-means, on GPGPU's. Moreover, the data collaborators were parallelized in order to compute local algorithm at the same time. For the experiment results, we used Alienware area-51m with GeForce RTX 2080/PCIe/SSE2 / NVIDIA Corporation graphic card.

4.4.2 Results and discussion

In this section we evaluate the approach on several data-sets for both vertical and horizontal clustering, based on different quality indexes, either internal or external one.

We also compare the proposed algorithm with state-of-the-art approaches of collaborative clustering based on prototypes exchanges: Self-Organizing Maps collaboration (Co-SOM) and Generative-Topographic Maps collaboration (Co-GTM).

Vertical Collaboration case

To evaluate the proposed approach in a vertical collaboration case, we computed the algorithm 5 on several sub-sets that share the same features but have different size and complexity.

TABLE (4.2) Values of the different quality indexes before and after a vertical collaboration for each collaborator built from the Spambase data set.

Models	DB		Silhouette		ARI	
	Before	After	Before	After	Before	After
collab1	0.681	0.561	0.524	0.539	0.169	0.165
collab2	0.769	0.540	0.532	0.625	0.118	0.119
collab3	0.751	0.653	0.530	0.548	0.127	0.130
collab4	0.714	0.576	0.538	0.574	0.168	0.169
collab5	0.653	0.673	0.539	0.535	0.149	0.148
collab6	0.714	0.569	0.552	0.556	0.156	0.160
collab7	0.705	0.589	0.536	0.563	0.174	0.154
collab8	0.720	0.576	0.544	0.590	0.163	0.165
collab9	0.717	0.711	0.503	0.577	0.149	0.178
collab10	0.665	0.605	0.495	0.561	0.159	0.192

TABLE (4.3) Average values ($\pm CI_{95\%}$) of the different quality indexes before and after the vertical collaboration for each data set over 20 executions.

Indexes		**DB**		**Silhouette**		**ARI**	
Data-sets		**before**	**after**	**before**	**after**	**before**	**after**
Glass	Average	0.984	0.689	0.369	0.471	0.223	0.244
	$\pm CI_{95\%}$	±0.09	±0.17	±0.04	±0.06	±0.05	±0.07
Spambase	Average	0.711	0.603	0.529	0.567	0.153	0.158
	$\pm CI_{95\%}$	±0.02	±0.03	±0.01	±0.01	±0.01	±0.01
Waveform-noise	Average	2.819	2.768	0.078	0.080	0.285	0.291
	$\pm CI_{95\%}$	±0.05	±0.05	±0.002	±0.002	±0.01	±0.01
WDBC	Average	0.675	0.629	0.448	0.513	0.290	0.374
	$\pm CI_{95\%}$	±0.02	±0.05	±0.02	±0.02	±0.03	±0.06
Wine	Average	0.525	0.496	0.568	0.574	0.306	0.308
	$\pm CI_{95\%}$	±0.02	±0.02	±0.02	±0.02	±0.03	±0.02

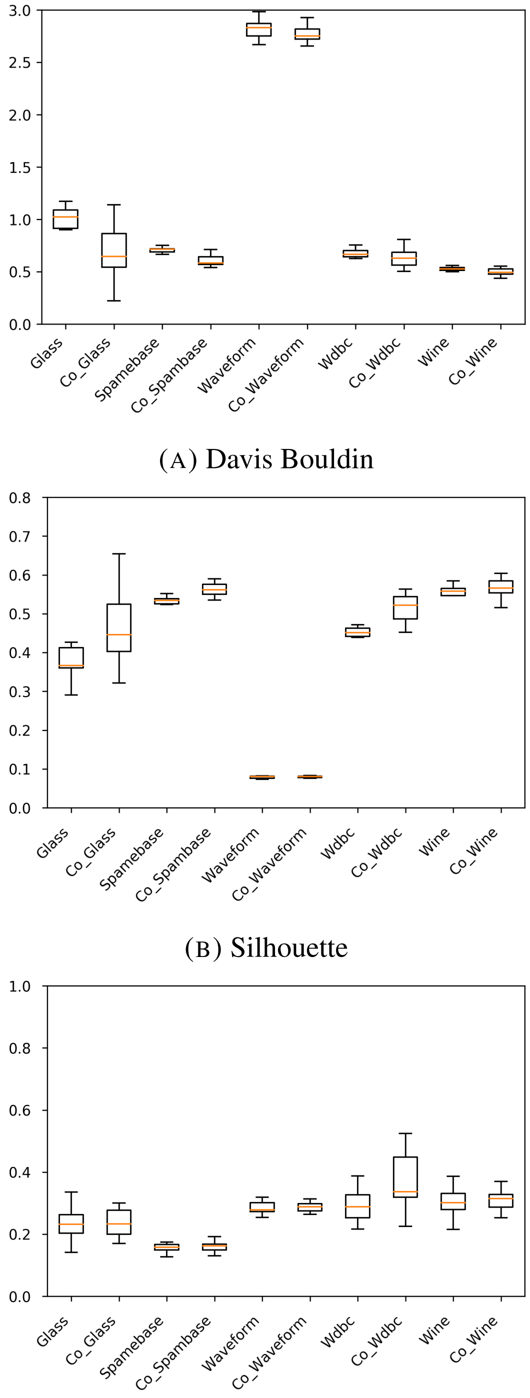
As one can see, the proposed approach shows in general an acceptable capacity at improving the DB index of the clustering before and after a vertical collaboration table 4.3. This is not surprising, considering that the proposed algorithm evaluates the gain of quality based on this index. The DB index is computed at each iteration in order to learn whether or not the collaborator can benefit from this collaboration. To make sure of the validity of the algorithm, we used the silhouette internal index. As shown in Table 4.3, the value of Silhouette increases after collaboration, which is a confirmation that the proposed approach increases the quality of each collaborator. However, the quality gain resulting from the collaboration is not always very high for some data-sets. This is due to the structure of the database and its horizontal splitting.

If the data is very sparse (notably Spambase), we can observe that the collaboration increases more the quality than non-sparse data (for example Waveform data-set).

Table 4.3 shows the results achieved on this index and highlighted the performance of our algorithm and confirm that the quality of each collaborator increases after the collaboration.

As one can see, the results are generally positive but the difference between the values either in internal indexes ($Silhouette$ and DB) or in external index ARI before and after collaboration is not very impressive, this is explained by the horizontal splitting which gives small subsets that practically have the same structure, which means that the collaboration can be seen as a bidirectional exchange of information between subsets of the same given database.

As we will see later on, this is not the case in horizontal collaboration, in which the impact of the collaboration is more important since the data are represented with different features for each collaborator. In addition, we chose one data set (due to page limitation) to detail the effect of the proposed algorithm on each collaborator. Table 4.2 shows the values of different quality indexes of each collaborator built from Spambase data set, and confirm that the quality does increase the quality of most collaborators in the process.



(A) Davis Bouldin

(B) Silhouette

(C) ARI

FIGURE (4.4) Sensitivity Box-Whiskers plots for the vertical collaboration case

Sensitivity Box-Whiskers plots (figure 4.4) are drawn for the 20 experiments scores for each dataset before and after collaboration process. They enable us to study the distributional characteristics of scores as well as the level of the scores. To begin with, scores are sorted over the 20 tests. Then four equal sized groups are made from the ordered scores. That is, 25% of all scores are placed in each group. The lines dividing the groups are called quartiles, and the groups are referred to as quartile groups. Usually, we label these groups 1 to 4 starting at the bottom. The median (middle quartile) marks the mid-point of the scores and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less. The middle “box” represents the middle 50% of scores for the group. The range of scores from lower to upper quartile is referred to as the inter-quartile range. The middle 50% of scores fall within the inter-quartile range.

As can be seen from these graphs, the overall performance behavior shows a clear improvement as a result of the collaboration process. For example, for the DB Index, we can see a decrease in index values for all databases due to the contribution of collaboration. For the other two quality indices, we rather observe an increase in the values of the indices showing an improvement in the qualities of the solutions found.

Horizontal collaboration case

In this section we validate the effectiveness of the proposed approach on different date-sets for horizontal collaboration, where each collaborator represents the instances with different features (in a different representation space) see figure 4.3.

We show how the exchange of information between the collaborators can improve the local results of each collaborator. Moreover, we show that the gain of quality is much important comparing to classical collaboration (SOM and GTM collaboration)

Besides Davis-Bouldin index 3.4, which is trained in the algorithm, we validated the proposed approach with silhouette 4.3, the Adjusted Rand Index 4.4.

Table 4.5 shows that the collaboration step in the proposed approach increases the local quality of the models in regard to internal indexes DB and $Silhouette$, in a horizontal collaboration framework, for different data-sets. Similarly, the ARI index values show that the clusters computed by the models are closer to the expected output after the collaborations (table 4.5). One can notice that horizontal collaboration, between models that do not share the same representation space, is much more beneficial compared to vertical collaboration, where the models are computed in different spaces. This is due to the fact that in the vertical framework, the random splitting of the data-sets produce sub-sets of different instances represented in the same space (i.e., same features) with similar distributions due to the random process of the split. Therefore, each local model should be quite similar to the others and few exploitable information is exchanged in the collaborative step. This could be confirmed by the comparison between the index values of Spambase data set in vertical collaboration (table 4.2) and the horizontal collaboration (table 4.4) where the difference between the score of the indexes is much more important for each collaborator comparing to vertical collaboration.

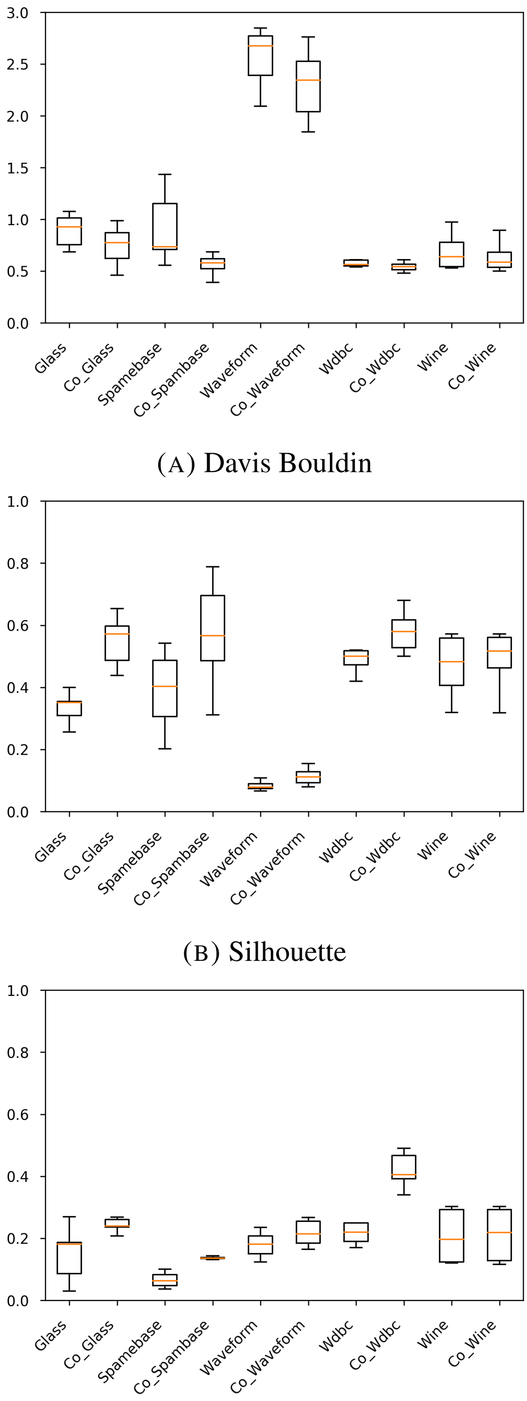
TABLE (4.4) Values of the different quality indexes before and after the horizontal collaboration for each collaborator built from the Spambase data set.

Models	DB		Silhouette		ARI	
	Before	After	Before	After	Before	After
collab1	0,583	0.565	0.415	0.532	0.045	0.137
collab2	0.751	0.690	0.392	0.452	0.086	0.136
collab3	0.555	0.495	0.543	0.788	0.043	0.135
collab4	1.436	0.578	0.315	0.631	0.073	0.118
collab5	0.714	0.459	0.507	0.717	0.057	0.136
collab6	1.067	0.706	0.287	0.578	0.058	0.139
collab7	1.183	1.099	0.304	0.312	0.157	0.144
collab8	0.722	0.511	0.505	0.470	0.101	0.143
collab9	0.707	0.503	0.435	0.555	0.036	0.136
collab10	1.370	0.418	0.202	0.755	0.069	0.132

TABLE (4.5) Average values ($\pm CI_{95\%}$) of the different quality indexes before and after the horizontal collaboration for each data set over 20 executions.

Indexes		**DB**		**Silhouette**		**ARI**	
Data-sets		**before**	**after**	**before**	**after**	**before**	**after**
Glass	Average	1.028	0.608	0.335	0.552	0.155	0.237
	$\pm CI_{95\%}$	±0.23	±0.18	±0.02	±0.04	±0.04	±0.01
Spambase	Average	0.903	0.481	0.390	0.579	0.072	0.135
	$\pm CI_{95\%}$	±0.20	±0.12	±0.06	±0.09	±0.02	±0.004
Waveform-noise	Average	2.578	2.310	0.078	0.108	0.179	0.218
	$\pm CI_{95\%}$	±0.15	±0.20	±0.008	±0.01	±0.02	±0.02
WDBC	Average	0.601	0.550	0.483	0.566	0.219	0.439
	$\pm CI_{95\%}$	±0.17	±0.09	±0.02	±0.05	±0.05	±0.13
Wine	Average	0.688	0.643	0.470	0.490	0.206	0.212
	$\pm CI_{95\%}$	±0.10	±0.09	±0.06	±0.05	±0.05	±0.05

Sensitivity Box-Whiskers plots (figure 4.5) represents a synthesis of the scores into five crucial pieces of information identifiable at a glance: position measurement, dispersion, asymmetry and length of Whiskers. The position measurement is characterized by the dividing line on the median (as well as the middle of the box). Dispersion is defined by the length of the Box-Whiskers (as well as the distance between the ends of the Whiskers and the gap). Asymmetry is defined as the deviation of the median line from the center of the Box-Whiskers from the length of the box (as well as by the length of the upper Whiskers from the length of the lower Whiskers, and by the number of scores on each side). The length of the Whiskers is the distance between the ends of the Whiskers in relation to the length of the Box-Whiskers (and



(A) Davis Bouldin

(B) Silhouette

(C) ARI

FIGURE (4.5) Sensitivity Box-Whiskers plots for the horizontal collaboration case

the number of scores specifically marked). These graphs show the same overall performance behavior observed in the case of vertical collaboration. They show a clear improvement as a result of the collaboration process. This improvement is observed for all quality indices used.

Comparison with other collaborative approaches

In this section, the proposed collaborative algorithm 5 is based on Sinkhorn-Means (Sin-Mean) local algorithms as described in algorithm 1 (this framework is thereafter called Co-Sin-OT) and we illustrate the adaptability of the proposed collaborative approach by alternatively using Self-Organizing Maps (SOM) as local algorithms (Co-SOM-OT). Both are compared to popular state-of-the-art collaborative algorithms based on Self-Organized-Maps (Co-SOM) [27] and Generative-Topographic-Maps (Co-GTM) [23]. We focus here on the horizontal collaboration case as in [27] and [23]. Indeed, horizontal collaboration is usually more useful and applicable to real problems comparing to vertical collaboration, it is also more difficult. In the first part of the experiments, we test the quality of the collaboration for 10 collaborators. As the Co-GTM algorithm is designed for only two collaborators, it is not included in the comparisons. In the second part, only two collaborators are trained and the Co-GTM algorithm is included in the protocol.

TABLE (4.6) Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the *Silhouette* index for the Glass data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions.

	Sinkhorn-based			SOM-based				
	Sin-Means	Co-Sin-OT	Gain	SOM	Co-SOM	Gain	Co-SOM-OT	Gain
collab1	0.353	0.582	0.229	0.088	0.240	0.152	0.240	0.152
collab2	0.294	0.461	0.167	-0.009	-0.009	0.000	0.131	0.140
collab3	0.351	0.653	0.303	-0.036	-0.036	0.000	0.156	0.192
collab4	0.400	0.569	0.169	0.395	0.395	0.000	0.340	-0.055
collab5	0.256	0.439	0.182	-0.008	0.320	0.329	0.320	0.329
collab6	0.339	0.602	0.263	0.070	0.070	0.000	0.102	0.032
collab7	0.299	0.565	0.266	0.222	0.257	0.034	0.253	0.031
collab8	0.358	0.444	0.086	0.410	0.410	0.000	0.439	0.029
collab9	0.351	0.635	0.284	0.073	0.183	0.110	0.223	0.150
collab10	0.355	0.574	0.220	-0.053	-0.051	0.001	0.003	0.056
Average	0.335	0.552	0.216	0.115	0.177	0.063	0.221	0.106
$\pm CI_{95\%}$	±0.02	±0.04	±0.12	±0.10	±0.10	±0.06	±0.07	±0.06

The first set of experiments are thus restricted to Co-SOM, Co-SOM-OT and Co-Sin-OT in order to be able to work with several collaborators. All collaborative approaches are applied on the same subsets. In SOM-based approaches, each local collaborator starts with the same 5×3 SOM. The approaches are compared using the Silhouette index. As shown in Tables 4.6 to 4.10, the results obtained with the proposed approach are globally better for this index. One can note that, for some collaborators, the quality of the collaboration leads to very similar results in both cases, despite very different quality before collaboration. The OT-based approach (Co-SOM-OT) provides a much more stable quality improvement over the set of collaborators. In addition, the use of Sinkhorn-Means as the local algorithm (Co-Sin-OT) provide the best results comparing a SOM-Based local clustering (Co-SOM and Co-SOM-OT). This can be explained by the fact that the mechanism of the SOM-based collaborative algorithms is constrained by the neighborhood's functions.

Moreover, it was built for a collaboration between two collaborators, then extended to allows multiple collaborations, unlike the proposed approach where each learner exchange information with all of the others at each step of the collaboration.

TABLE (4.7) Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the *Silhouette* index for the Spambase data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions.

	Sinkhorn-based			SOM-based				
	Sin-Means	Co-Sin-OT	Gain	SOM	Co-SOM	Gain	Co-SOM-OT	Gain
collab1	0,415	0,532	0,118	0,224	0,483	0,260	0,346	0,122
collab2	0,392	0,452	0,060	0,038	0,080	0,042	0,124	0,086
collab3	0,543	0,788	0,246	-0,137	-0,137	0,000	0,005	0,142
collab4	0,315	0,631	0,316	-0,308	-0,091	0,216	-0,103	0,205
collab5	0,507	0,717	0,211	-0,101	-0,028	0,073	0,052	0,153
collab6	0,287	0,578	0,291	-0,039	0,036	0,075	0,153	0,192
collab7	0,304	0,312	0,008	-0,035	-0,035	0,000	0,087	0,122
collab8	0,505	0,470	-0,035	0,314	0,524	0,210	0,511	0,197
collab9	0,435	0,555	0,119	-0,260	-0,069	0,191	0,023	0,283
collab10	0,202	0,755	0,553	0,041	0,041	0,000	0,114	0,073
Average	0.390	0.579	0,188	-0.026	0.035	0.106	0,131	0,158
$\pm CI_{95\%}$	±0.06	±0.09	±0.10	±0.12	±0.12	±0.06	±0.11	±0.03

TABLE (4.8) Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the silhouette index for the Waveform-noise data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions.

	Sinkhorn-based			SOM-based				
	Sin-Means	Co-Sin-OT	Gain	SOM	Co-SOM	Gain	Co-SOM-OT	Gain
collab1	0,108	0,155	0,047	0,025	0,030	0,006	0,064	0,039
collab2	0,074	0,097	0,023	0,036	0,036	0,000	0,062	0,026
collab3	0,075	0,091	0,016	0,070	0,070	0,000	0,069	-0,001
collab4	0,067	0,088	0,021	0,043	0,047	0,003	0,064	0,021
collab5	0,081	0,127	0,046	0,054	0,058	0,004	0,069	0,015
collab6	0,067	0,079	0,012	0,063	0,063	0,000	0,067	0,004
collab7	0,093	0,128	0,036	0,026	0,026	0,000	0,063	0,037
collab8	0,095	0,140	0,045	0,040	0,044	0,004	0,067	0,027
collab9	0,081	0,120	0,039	0,031	0,038	0,007	0,065	0,034
collab10	0,075	0,104	0,029	0,025	0,032	0,007	0,063	0,038
Average	0.078	0.108	0,0313	0.043	0.045	0.003	0,065	0,024
$\pm CI_{95\%}$	±0.008	±0.01	±0.008	±0.01	±0.009	±0.01	±0.05	±0.01

In the second set of experiments, we compare the proposed approach to classical collaborative algorithms based on Self-Organized-Maps (Co-SOM) [27] and Generative-Topographic-Maps (Co-GTM) [23]. The three approaches are compared using DB index, as in [23, 27]. As shown in Table 4.11, the results obtained with the proposed approach (Co-Sin-OT), in comparison to the state-of-the-art, is generally better than the classical approaches. The lowest qualities are expressed by the older approach, SOM-based collaborative clustering (Co-SOM), followed by the GTM-based approach (Co-GTM). Unlike Co-SOM and Co-GTM, the proposed approach aims to find a local optimum for each collaborator. More precisely, at the end of the local training, each collaborator exchange information based on a stopping criterion that ends the collaboration with each collaborator as soon as the quality of the collaboration starts decreasing, which is not the case in the other approaches. Furthermore, Table 4.11 compares the quality gain brought by the collaboration from each approach. The proposed approach increases the quality of each collaborator on all of the data-sets, which implies a positive gain quality. On the contrary, in SOM-based collaboration the gain can be negative for some data-sets. Finally, in order to evaluate the general performance of the approaches, we define the following score measurement:

$$Score(M_i) = \sum_j \frac{G(M_i, D_j)}{\max_i G(M_i, D_j)} \tag{4.5}$$

Where G indicates the gain quality of each approach M_i of each data-sets D_j. This score gives an overall vision of the best approach on all the data-sets. As shown in table 4.11, the best score belongs to the proposed collaboration based on Optimal Transport theory, followed by the GTM collaborative approach (Co-GTM) and the SOM-based collaboration (Co-SOM). These results highlight the the performance of the proposed algorithm, due to the strong theoretical back-ground of Optimal Transport theory.

TABLE (4.9) Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the silhouette index for the WDBC data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions.

	Sinkhorn-based			SOM-based				
	Sin-Means	Co-Sin-OT	Gain	SOM	Co-SOM	Gain	Co-SOM-OT	Gain
collab1	0,470	0,632	0,162	0,233	0,233	0,000	0,244	0,011
collab2	0,514	0,620	0,106	0,185	0,208	0,024	0,211	0,026
collab3	0,485	0,614	0,129	0,246	0,303	0,057	0,401	0,155
collab4	0,521	0,608	0,087	0,015	0,074	0,059	0,126	0,111
collab5	0,422	0,499	0,077	0,102	0,182	0,080	0,341	0,239
collab6	0,490	0,549	0,059	0,240	0,278	0,038	0,334	0,094
collab7	0,516	0,685	0,168	0,298	0,337	0,039	0,445	0,147
collab8	0,388	0,392	0,004	0,125	0,125	0,000	0,323	0,198
collab9	0,518	0,516	-0,003	0,202	0,213	0,011	0,367	0,165
collab10	0,513	0,548	0,035	0,110	0,123	0,013	0,236	0,126
Average	0.483	0.566	0,082	0.175	0.204	0.032	0,303	0,127
$\pm CI_{95\%}$	±0.02	±0.05	±0.03	±0.05	±0.05	±0.01	±0.06	±0.04

TABLE (4.10) Comparison of SOM-based and Sinkhorn-based collaborative approaches, using the silhouette index for the Wine data set. The average values ($\pm CI_{95\%}$) is computed over 20 executions.

	Sinkhorn-based			SOM-based				
	Sin-Means	Co-Sin-OT	Gain	SOM	Co-SOM	Gain	Co-SOM-OT	Gain
collab1	0.560	0.562	0.002	0.3921	0.4033	0.011	0.446	0.054
collab2	0.559	0.560	0.001	0.4288	0.4288	0.000	0.431	0.002
collab3	0.572	0.573	0.001	0.4945	0.4945	0.000	0.542	0.048
collab4	0.320	0.318	-0.001	0.1223	0.1255	0.003	0.224	0.102
collab5	0.446	0.499	0.054	0.1444	0.2116	0.067	0.221	0.077
collab6	0.394	0.450	0.056	0.1558	0.1768	0.021	0.201	0.045
collab7	0.329	0.351	0.022	0.1107	0.1107	0.000	0.257	0.146
collab8	0.520	0.534	0.014	0.1243	0.2210	0.097	0.348	0.224
collab9	0.444	0.498	0.054	0.0715	0.2286	0.157	0.356	0.284
collab10	0.559	0.560	0.002	0.3923	0.3923	0.000	0.395	0.003
Average	0.470	0.490	0.020	0.244	0.279	0.036	0.342	0.098
$\pm CI_{95\%}$	±0.06	±0.05	±0.01	±0.10	±0.09	±0.03	±0.07	±0.06

TABLE (4.11) Comparison of DB index Between SOM, GTM and OT based approaches on different data-sets for two collaborators

Approaches		**Co-SOM**			**Co-GTM**			**Co-Sin-OT**		
Data-sets		**before**	**after**	**gain**	**before**	**after**	**gain**	**before**	**after**	**gain**
Glass	Collab1	1.010	0.985	0.002	0.740	0.970	0.000	1.109	0.774	0.253
	Collab2	0.902	0.924		1.280	1.050		0.908	0.731	
Spambase	Collab1	2.924	2.436	-0.077	1.120	1.060	1.014	1.467	1.055	0.187
	Collab2	0.960	1.748		0.870	0.900		0.775	0.766	
Waveform-noise	Collab1	6.488	6.488	0.027	1.140	1.310	0.378	3.592	2.579	0.256
	Collab2	7.269	6.898		3.750	1.310		3.732	2.868	
Wdbc	Collab1	0.640	0.641	0.001	0.970	0.920	0.010	0.755	0.612	0.219
	Collab2	0.651	0.649		0.870	0.900		0.928	0.701	
Score				**-1.740**			**1.377**			**3.581**

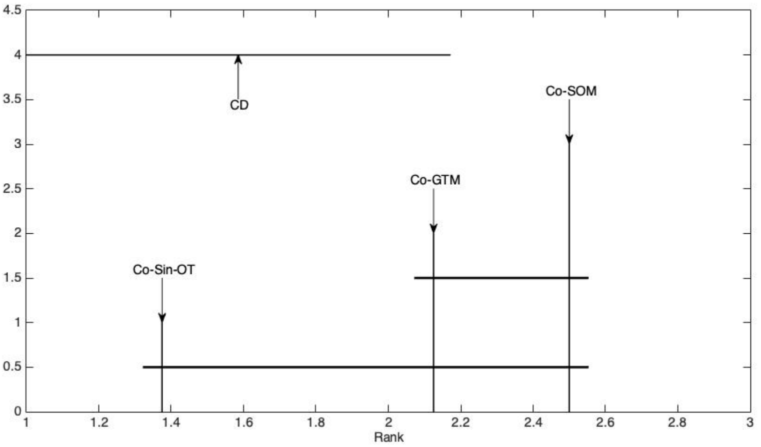


FIGURE (4.6) Friedman and Nemenyi test for comparing multiple approaches over multiple data sets: Approaches are ordered from left (the best) to right (the worst)

In order to assess the performance of our approaches, we use the Friedman test and Nemenyi test recommended in [18]. The Friedman test is conducted to test the null-hypothesis that all approaches are equivalent in respect of accuracy. If the null hypothesis is rejected, then the Nemenyi test will be performed. In addition, if the average ranks of two approaches differ by at least the critical difference (CD), then it can be concluded that their performances are significantly different. In the Friedman test, we set the significant level $\alpha = 0.05$. The figure 4.6 shows a critical diagram representing a projection of average ranks approaches on enumerated axis. The approaches are ordered from left (the best) to right (the worst) and a thick line which connects the approaches were the average ranks not significantly different (for the level of 5% significance). As shown in figure 4.6, Co-Sin-OT achieves significant improvement over the other proposed techniques (Co-GTM and Co-SOM) since during collaboration phase it is stable and the process stops the collaboration for some learners when their local quality stars to decrease, which prevents common issue of collaborative approaches.

Compared to the most cited approaches of the state of the art, the positive impact of using the collaborative learning based on this theory is:

- The proposed algorithm is based on a strong and well defended theory that becomes increasingly popular in the field of machine learning.
- Its strength is highlighted by experimental validation on both for artificial and real data-sets.
- The stopping criteria that we proposed based on the measure of the gain quality brought after each collaboration, because it guarantee the convergence once the gain quality tends towards zero.
- The choice of the distant collaborator which is very important and allows to give an optimal order of the collaboration. In the proposed algorithm we solved this paradigm based on the Optimal Transport Matrix plan, that aims to compare

all distribution of the centroids in each site. In this way each collaborator will be enable to choose the best one.

- The proposed algorithm stops the negative collaboration, based on the measure of the gain quality between each collaboration and updates the centroids if the gain quality is positive. Otherwise, it moves to the other distant collaborator.

Finally, the proposed approach ensures the adaptability of working with different local models, this is lead us to introduce some managerial applications of our work, like the management system learning where the collaborative learning could offer in interaction between learners to make them work cooperatively rather than competitively and helps to create sub-networks of collaboration where the diversity is decreased, and manage the conflict learning by using a one-to-one collaboration. Besides the exchange information using the proposed algorithm preserve the privacy of each collaborator, and ensures the control of the shared information with each collaborator, and filters the received information to avoid affecting the real structure of local data. Thus, all the collaborators can explore the distributed data that could containing some mutual information while keeping the control on received and transmitted information.

However, the proposed algorithm still suffers from some limitations, in particular considering the same dimension in every site, and also the curse of height dimensionality that Optimal Transport still suffers from, which leads us to increase the penalty coefficient of the regularization in order to avoid the over-fitting.

4.5 Conclusion

In this paper, we proposed a new framework of collaborative learning inspired by Optimal Transport theory, where the collaborators aim to increase their local quality based on the information exchanged from other learners. We explained the motivation and the intuition behind our approach and we proposed a new algorithm of collaborative clustering based on the Wasserstein distance. The proposed approach allows to exchange information between collaborators either in vertical or horizontal collaboration. The results are stable and the process stops the collaboration for some learners when their local quality stars to decrease, which prevents common issue of collaborative approaches.

The approach proposed in this paper is the first step into a new family of algorithms for the collaborative leaning task. We plan to develop further collaborative clustering algorithms based on Gromov-Wasserstein distance that ensure the comparison between the distribution coming from heterogeneous spaces, in order to make the collaborative algorithms more flexible, and to improve the quality and the stability of the collaborations.

There are several perspectives to this work. On the short term we are working to improve the approach in order to learn the confidence coefficient at each iteration, according to the diversity and the quality of the collaborators. This could be based on comparisons between sub-sets' distributions using the Wasserstein distance. This would lead us to another extension where the interaction between collaborators could be modeled as graph in a Wasserstein space, which would allow the construction of a theoretical proof of convergence.

5 Subspace guided collaborative clustering based on optimal transport

5.1 Introduction

Clustering in of the most exploratory framework in machine learning, where the main idea is to cluster the data on different clusters according to same similarity criteria, where the objects that belong to the same cluster are more similar than the ones how belong to other cluster. In literature, several approaches have been proposed based on the idea of maximizing the similarity intra-cluster and dissimilarity inter-cluster. With the development of hardware technology, a huge amount of data represented in different views and different structures have been generated in real word applications. This kind of data considered as a new challenge to develop the existing clustering algorithms designed for a single view data to be more adaptable to multi-view data. Several approaches have been proposed in the context of Multi-View Clustering, where the main idea is to train different models originating from different views to form a global model called the consensus clustering [64] the leads to a global minimization, or to aggregate different information between views in order to find a local minimum for each learner, i.e. learn from other learners [22]. However, one of the most challenging problem in the Multi-view clustering is how to make models provided form each view comparable. In other word, the diversity between views is very high which can lead to lose the sense of learning between sites or the views, especially when the characteristics of each view are very different. Furthermore,some of the views may be of a high dimensionality which leads to a high computational complexity and could decreased the quality of the models of each view. Feature selection is one of the most popular approach to address this type of problem, and can both simplify the calculation and make the models more accurate to comparison. On the other hand, Collaborative learning is on of the new area of research in data analysis that aims to exchange different information between models in order to enhance the performances of learning models. The collaborative learning categorized up to now in unsupervised learning. The main interest of this discipline is to help local clustering algorithms to make decision about how to improve there local cluster by using different information provided from other algorithms trained on different sites that could either have the same instances represented by different characteristics, or different samples that sharing the same characteristics. This kind of algorithms have proved a very useful to learn from distant representation of data spread in multiples sites while preserving the privacy of each sites.

Feature selection methods have proved a significant results in Multi-views consensus clustering [72]. For this reason we had the intuition to perform this method on collaborative clustering to guide the diversity between local models, and also to get more accurate local models. We propose in this worker a new algorithm of collaborative learning guided by feature selection, where the main idea is to improve the learning from the collaborative by selecting the most discriminative feature according to evolution of the considered collaborator. The rest of the chapter will be organized as follow: in the next section will explain the motivation behind the proposed approach, and will be followed by section where we details the proposed approach. In section.. we will investigate and discuss experimentally the quality of our approach, and finally we will conclude by a summary about the proposed approach and some perspective work in this framework.

5.2 Feature selection

Due to the significant growth of data in recent year, the learning performance has become more and more affected. This issues gave birth to a new framework in data mining, where the main idea is to select the best feature from the original data that maximally improves the quality of the learning algorithms. This issue had attract many researchers to perform many methods of feature selection where the main idea is to select small subset from the original set of feature basing on some predefined criterion. However most of these approaches were proposed in supervised classification where they require the information about the class labels to determine whether a feature relevant or not [29], while in unsupervised learning it is not possible to measure the relevancy of the features when the class labels are unknown. Yet some methods have been proposed to extend the feature selection into unsupervised learning through the use of performance information such as structure of the cluster, the quality of the prototypes, and even the visualisation of the data.

One of the most well-known of these techniques is the dimensionality reduction where the basic idea is to remove all the redundant and irrelevant features that degrade the performance of learning. The most popular approaches in this field which had proved their utility are The Principal Component Analysis (PCA) [59], the Singular Value Decomposition (SVD) [24], and Linear Discriminant Analysis (LDA) [56]. However, even those approaches can achieve the advantages of feature selection, they still considered as feature extraction ant not feature selection, where the main difference is feature selection the originality of feature are kept, and my be needed in the intractability of the models while in feature extraction, we lose this information about the importance of the data in the original data.

Generally speaking, we categorized the feature selection approaches as follow (see figure 5.1:

- Filter approaches [33]: aim to filter the feature based the characteristics of the data instead of the learning algorithm. Basically it is considered as prepossessing of the data based on the correlation between the features and some evaluation score depending on the type of the data set.

- Wrapper approaches [38]: contrary to filter approaches, wrappers need a learning algorithms to measure the importance of the features in term of performance. Mainly, we wrappers methods select subset of feature basing on some research strategies, and the evaluated through a learning algorithm. However, in term of complexity, wrapper methods are considered more intensive comparing to filter methods.

- Hybrid approaches [46]: aim to combine between the wrapper and filter methods to get the advantages of both methods. Generally there two ways of combination, the first one is to use two level of feature selection, where the first one will select the best feature based on some data criterion, and then a wrapper method will be used to evaluate the selected feature subsets. The second way of combination is to make an overlapping between the filter and wrapper methods alliteratively. It must be mentioned that this kind of approaches tend to be efficient.

- Embedded approaches [57]: Contrary to hybrid approaches, the Embedded approaches tend to find a trade-off between wrapper and filter approaches. The principle idea if the make the two method work in a cooperation with learning algorithm, without using it several time. Yet the wrapper approach still better and give better result comparing to this method.

To summary feature selection is a very significant process that aims to performance of learning model through a selection of the most relevant features in term of complexity, quality and interpretability. However the choice of the methods of feature selection is very important and depend of the nature of the problem.

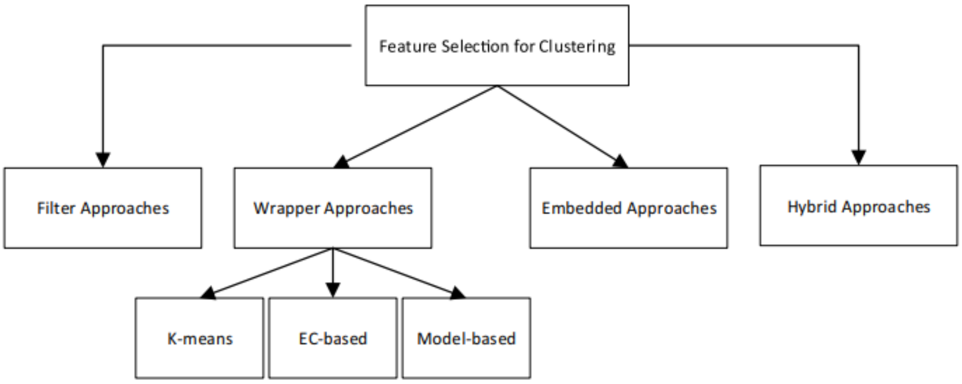


FIGURE (5.1) Feature selection approaches for clustering

In the next section we will explicit the main idea of the proposed approach and the motivation behind introducing feature selection into collaborative clustering

5.3 Proposed approach

5.3.1 Motivation

Recent work have been done on the impact of diversity between the collaborators and the gain quality brought from each collaboration [42]. The main idea of this of this work was to analyse the evolution of the quality during the collaboration, this work proved that the collaborators should not be very diverse otherwise will decrease the quality or will completely lose there own structure.

From this point, we had the intuition to guide this diversity by using feature selection technique. The idea behind is to make the collaboration strengthened by selecting not only the most discriminating feature for each site but also the features that will increase the quality of the distant collaborator. This selection will be done alliteratively for each selected collaborator. This way we will control diversity during the collaboration. Consequently the the choice of the collaborator will be more accurate and the collaboration will be more controlled.

In what follows we will explain the proposed algorithm and detail the mechanism of the method .

5.3.2 Proposed algorithm

We remind that this work aims to improve the collaborative algorithm based on optimal transport which had already proved higher performance comparing the collaborative prototype based methods (see chapter 3). Let consider r collaborators, where the data of each collaborator v, $X^v = \left\{x_i^v\right\}_{i=1}^{n^v}$ with $x_i^v \in \mathbb{R}^{d^v}$ and d^v the dimension of the space representation of v, and corresponds to a distribution $\mu^v = \frac{1}{n^v}\sum \delta_{x_i^v}$.

Starting with the first step of the proposed approach which is selecting the features. The main idea of this step is to choose the best feature for the collaboration so we can reduce the complexity and the diversity between collaborator. Thus, in this step we will first rank each feature according to its discrimination, to do so, we will use the improved F-score 5.1 which measures the discrimination among several samples [62].

considering the d^{th} vector, $d = 1, ..., d^v$ $f_{i,d}, i = 1, \ldots, n$, and l samples of the data-set, the improved $F-$score is defined as follow :

$$F_d = \frac{\sum_{j=1}^{l}(\overline{f_d}^{(j)} - \overline{f_d})^2}{\sum_{j=1}^{l}\frac{1}{n_j-1}\sum_{k=1}^{n_j}(f_{i,d}^{(j)} - \overline{f_d}^{(j)})^2} \tag{5.1}$$

Where n_j is the size of the j^{th} sample, $\overline{f_d}, \overline{f_d}^{(j)}$ are the average of the d^{th} feature of the whole data-set, and the j^{th} sample respectively. Then numerator indicates the discrimination between each sample, and the denominator indicates the one within each of sample.The larger the improved $F-$score is, the feature is more likely to be discriminative The idea behind ranking each feature before the threshold step is to avoid the biased selection because some feature my have some overlaps. Therefor, considering only threshold selection my affect the performance of the proposed algorithm. given for example of the following Scenario where there are two features f_1 and f_2, using both of them will increase the quality by 5%, while using just f_1 would increase the quality by 4%, and 5% for f_2. The percentage threshold is 3% imagine that the the algorithm selects first f_1, in this case it will select f_2 because the quality will be increased only by 1%. Consequently, the quality will be less than its actual potential. For this purpose we chose the rank each feature sort then in a descending order according to their discrimnation rank.

Algorithm 6 details the mechanism of this hybrid technique of feature selection that combines between ranking the importance of the feature using F-score 5.1, and forward selection basing on the variation of the clustering quality index $Q_{S_d^v}$ by measuring it using the Threshold p^v.

It should be mentioned that the feature selection algorithm depends also on the quality of the collaborator that needs information. More precisely when the algorithm chose a collaborator that he will exchange the information with him, the selection of the feature will take in consideration if the exchange is beneficial or not for the distant collaborator, this will be detailed in algorithm 7.

Considering the same notation, in the collaboration step, we seek to minimize the following objective function:

$$\underset{L^v,M^v}{\operatorname{argmin}}\{< L^v, C(X^v, M^v) >_F - \frac{1}{\lambda}(H(L^v) + \sum_{v'=1.v'\neq v}^{r} \alpha_{v',v}(< L^{v,v'}, C(M^v, M^{v'}) >_F \\ - \frac{1}{\lambda} H(L^{v,v'}))\} \tag{5.2}$$

Where, $M^v = \{m_1^v, .., m_{k^v}^v\}$, corresponding to a distribution ν^v, that represents the local clusters of each collaborator v, $L^v = \left\{ l_{ij}^v \right\}_{i,j=1}^{i,j=n^v,k^v}$ is optimal transport plan between the the local data X^v and the centroids M^v, and and $L^{v,v'}$ is the optimal transport plan between the centroids of the v^{th} and v'^{th} collaborators.

The first term refers to the first step on the collaboration where it aims to find the distribution of the local centroids. The second term of the objective function refers to second step which consists to exchange to the information between the collaborators, in order to increase the quality, this exchange based on influencing the local distribution of the centoirds by the distant distribution. $\alpha_{v',v'}$ are non-negative coefficients proportional to a mutual information which measure the diversity between the collaborators, called "confidence coefficients".

The last term of the of the objective function is a combination of a penalty term based on entropy regularization locally between instances and local centroids, and between centroids distant.

Algorithm 7 details the computation of proposed approach. The idea behind is to make the collaboration more blinded en the the exchange of the information is more punctual and targeted by introducing the forward feature selection for each collaborators based on its local quality of clustering as a threshold, and then chose the median collaborator based on the comparison of similarity between the distribution of local centroids to the distant collaborators. Following [42] we choose at each iteration the median one to exchange with, in this case if the quality is increased, the centroids are updated, otherwise, we give another shot to the chosen collaborator to execute another shot of forward selection basing on the quality index of the collaborator. In this case either the quality will be increased or this collaborator will be removed from the list and the process will be repeated until the quality of the collaborators stops increasing. It must highlighted that proposed algorithm is adaptable with all the prototype based algorithms, more precisely instead of using the Sinkhorn-Means as a local algorithm

Algorithm 6: Threshold-Forward Feature selection

```
Input  : $F^v = \{f^v_d\}^{d^v}_{d=1}$: Set of the features of collaborator $v$
         $p^v$ : the parameter of selected percentage
Output : The the set of the selected feature $S^v_{best}$
Initialize the set of the best selected feature$S^v_{best} = S^v_0 = \emptyset$
The quality index of clustering $Q^v_{best} = Q^v_0$
repeat
    Calculate the rank value for each feature $f^v_d$,   $1 < d < d_v$
    Sort features in descending order according to their rank.
    $S^v_d = S^v_{d-1} \cup f^v_d$
    Perform the clustering algorithm.
    if $| Q_{S^v_{d+1}} - Q_{S^v_d} |> p^v$ then
    $Q^v_{best} = Q_{S^v_{d+1}}$
    $S^v_{best} = S^v_{d+1}$
until convergence;
return $S^v_{best}$
```

to get the centroids, we can use other prototype models like k-means, SOM, EM, etc. The proposed algorithm has therefore the capability to work with hybrid models. However, in our case we chose to use Sinkhorn-means as local model to keep all the mechanism based on Optimal transport.

5.4 Experiments Results

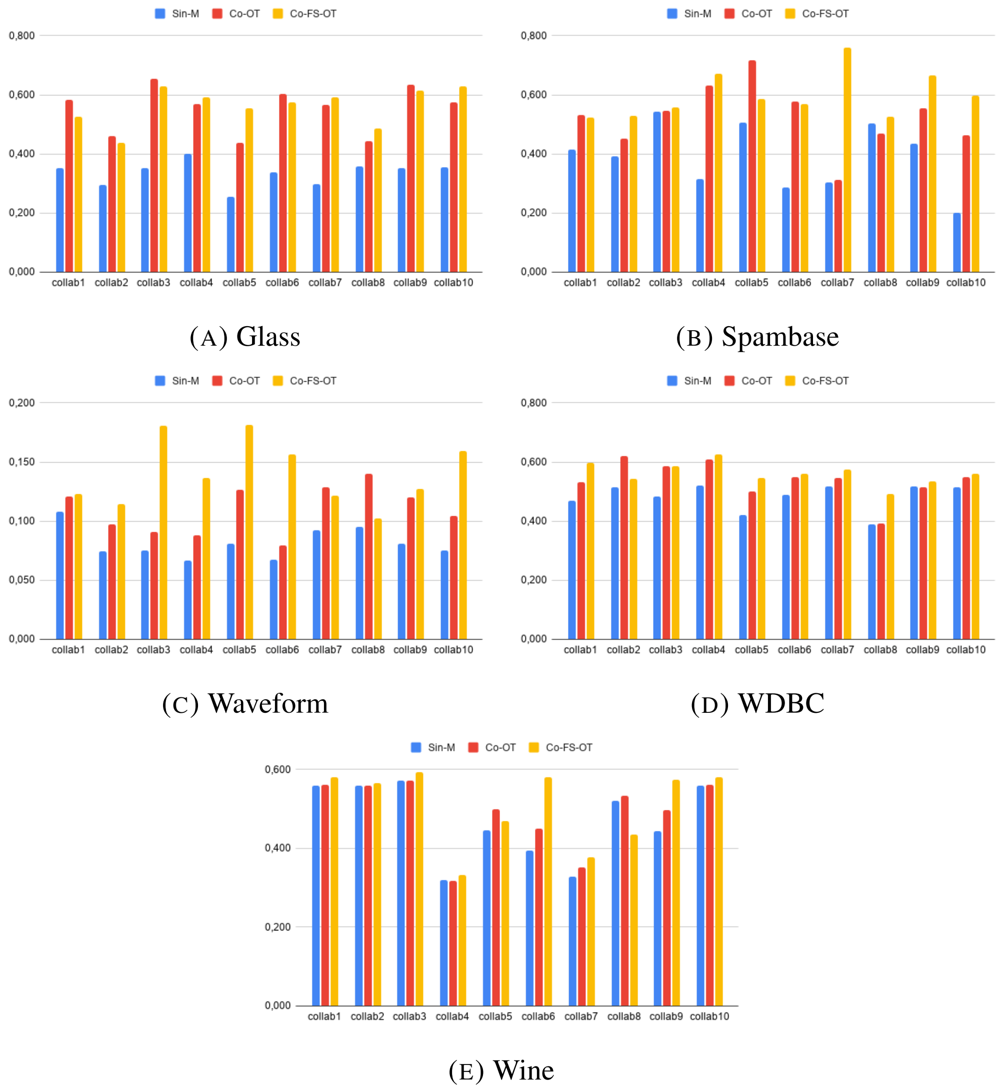
5.4.1 Protocol experiments

In order to evaluate the proposed approach experimentally, we start with a pre-processing of the data, where we create the collaborators in the form of the local subsets that share the same instances but represented with different characteristics. We split the data horizontally into 10 collaborators (subets), each collaborators X^v that will be considered as the inputs of algorithm 7. The first step consists to applied forward selection followed by Sinkhorn-means [14] in order to get the centroids. The, the coefficient is computed based on diversity between the distributions of each collaborator. Algorithm 7 in trained 20 times in order to estimate the mean quality of the collaboration and a 95% confidence interval for the 20 experiments. The comparison will be done between a collaboration with optimal transport only, and by using the feature selection during the training. Both approaches will be tested on the same subsets and the same local model(Sinkhorn-Means). due to the limit of pages we chose to use only two index of comparison, an internal and external one that will be defined in the next section.

5.4.2 Experiments Results

In this section we evaluate the proposed approach using several data sets, described in table 4.1. We chose for this evaluation two type of quality index,an internal one which is $Silhouette$ index [43], and an external one since original data is labeled which the Adjusted Rand Index(ARI) [61]. It must be mentioned that during the training of the algorithm we used another unsupervised quality index to control the collaboration and feature selection which is Davis-Bouldin [17]. The reason why we chose to add another unsupervised index instead of using only DB index is that the collaboration depends on the variation of this index, which means it iterates while DB is decreasing and converges when the index stops changing.

The $Silhouette$ index is based on measurement of the difference between the average of distances between the instances belonging to the same cluster and the average distance between to other clusters. The closer the $silhouette$ value is to 1, means that the instances belongs the right cluster. Table 5.1 chose the result of this index comparing the quality of the clusters using Sinkhorn-Means locally, the collaboration based on Optimal Transport interactions (Co-OT) and the collaboration guided by the feature selection (Co-FS-OT). As one can see, over 20 executions the average value does prove that that forward feature selection improve the quality of the cluster on



(A) Glass

(B) Spambase

(C) Waveform

(D) WDBC

(E) Wine

FIGURE (5.2) Histogram that compares *Silhouette* index values for 10 collaborators for each data sets.

different data sets. Moreover, figure 5.2 shows the details of each collaborator and proves that each collaborator succeeds in improving its local quality by giving the possibility to feature selection shots during the collaboration.

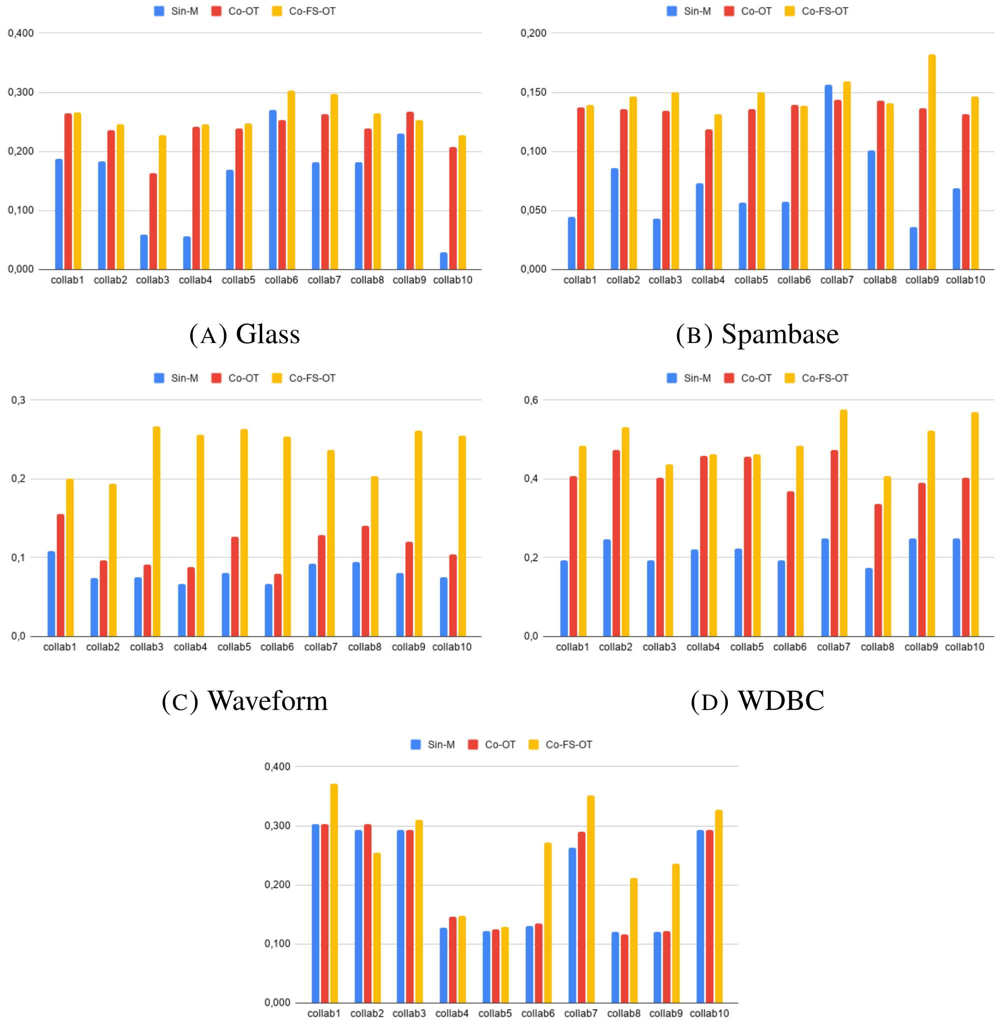
TABLE (5.1) Average values ($\pm CI_{95\%}$) of Silhouette index for each data set over 20 executions.

Models		Sin-Means	Co-OT	Gain	Co-FS-OT	Gain
Glass	Average	0.336	0.553	0.217	0.563	0.227
	$\pm CI_{95\%}$	±0.025	±0.04	±0.04	±0.03	±0.03
Spambase	Average	0.390	0.526	0.135	0.599	0.208
	$\pm CI_{95\%}$	±0.06	±0.06	±0.07	±0.04	±0.01
Waveform noise	Average	0.082	0.110	0.028	0.140	0.059
	$\pm CI_{95\%}$	±0.01	±0.01	±0.01	±0.01	±0.02
WDBC	Average	0.484	0.540	0.056	0.562	0.078
	$\pm CI_{95\%}$	±0.02	±0.04	±0.02	±0.01	±0.02
Wine	Average	0.470	0.491	0.020	0.509	0.038
	$\pm CI_{95\%}$	±0.06	±0.05	±0.01	±0.06	±0.04

TABLE (5.2) Average values ($\pm CI_{95\%}$) of ARI index for each data set over 20 executions.

Models		Sin-Means	Co-OT	Gain	Co-FS-OT	Gain
Glass	Average	0.155	0.237	0.120	0.257	0.103
	$\pm CI_{95\%}$	±0.04	±0.02	±0.05	±0.01	±0.03
Spambase	Average	0.072	0.136	0.063	0.148	0.076
	$\pm CI_{95\%}$	±0.02	±0.01	±0.02	±0.01	±0.02
Waveform	Average	0.082	0.113	0.031	0.239	0.158
noise	$\pm CI_{95\%}$	±0.01	±0.02	±0.01	±0.02	±0.02
WDBC	Average	0.219	0.417	0.198	0.494	0.274
	$\pm CI_{95\%}$	±0.02	±0.02	±0.02	±0.03	±0.02
Wine	Average	0.207	0.213	0.006	0.261	0.054
	$\pm CI_{95\%}$	±0.05	±0.05	±0.006	±0.11	±0.12

Further more, since the data that we used provide labels we use ARI index which measures the agreement between two partitions, one provided from the origin labels and the second one provided from the proposed approach. The value of the ARI is between 0 and 1, the more closer the ARI value to 1, the better agreement is, consequently better cluster. Table 5.2 shows result of the average of this index over 20 executions of the 10 collaborators, the results in figure 5.3 confirm that guidance of the collaboration by the feature selection increases the agreement between the two partitions which means that we are getting close the true label of each instance. This proves that even that we are working on unsupervised clustering, the proposed approach its quality to cluster the data correctly comparing to its labels. It must mentioned that although up to our comparisons, the collaboration based on optimal transport (Co-OT) is the best prototype based method, the Co-FS-OT has improved



(A) Glass

(B) Spambase

(C) Waveform

(D) WDBC

(E) Wine

FIGURE (5.3) Histogram that compares Adjusted Rand Index (ARI) values for 10 collaborators for each data sets.

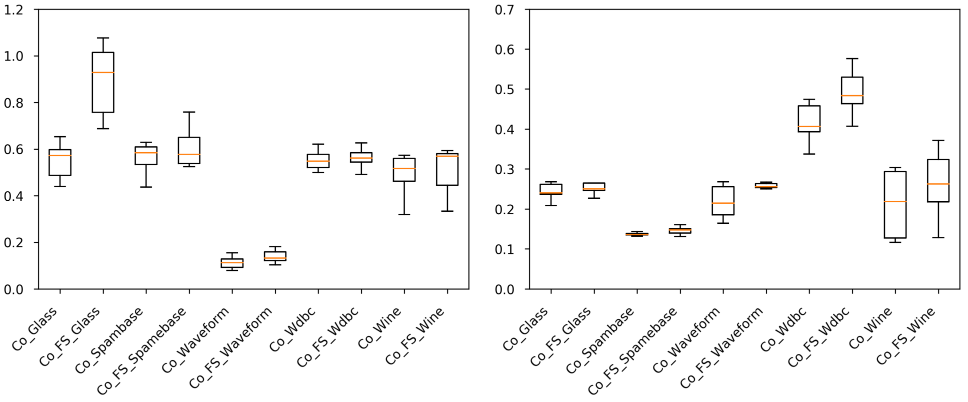
that feature selection algorithm chose the best representation of each collaborator to learn from. In this way we gave the opportunity to the collaborator to control the distant clustering, while preserving privacy and anonymity.

Sensitivity Box-Whiskers plots 5.4 are drawn for the 20 experiments scores for each dataset for the collabration Co-OT and Co-FS-OT, They enable us to study the distributional characteristics of scores as well as the level of the scores. To begin with, scores are sorted over the 20 tests. Then four equal sized groups are made from the ordered scores. That is, 25% of all scores are placed in each group. The lines

Algorithm 7: Collaborative clustering (Co-FS-OT)

```
Input  : $\{X^v\}_{v=1}^r$: the $r$ collaborators' data with distributions $\{\mu^v\}_{v=1}^r$
         $\{k^v\}_{v=1}^r$ : the numbers of clusters
         $\lambda$ : the entropic constant
         $\{\alpha_{v,v'}\}_{v,v'=1}^{v,v'=r}$ : the confidence coefficient matrix
Output : The partition matrix $\{(L^v)^*\}_{v=1}^r$ and the centroids $\{M^v\}_{v=1}^r$
Initialize the centroids $M^v = \left\{m_j^v\right\}_{j=1}^{k^v}$ randomly
Compute the associated distribution $\nu^v = \frac{1}{k^v}\sum_{j=1}^{k^v}\delta_{m_j^v}, \forall v \in \{1,...,r\}$
repeat
    for $v = 1,...,r$ do
        Update the centroids $M^v$ and the partition matrix $(L^v)^*$ using a local
          algorithm (e.g. Sinkhorn-Means, SOM, GTM, $k$-Means...)
        Update the centroids distribution $\nu^v = \frac{1}{k^v}\sum_{j=1}^{k^v}\delta_{m_j^v}$
        for $v' = 1,...,r$ and $v \neq v'$ do
            Compute Feature Selection Algorithm 6 on $X^{v'}$
            Compute the OT matrix $(L^{v,v'})^* = \{l_{jj'}\}_{j,j'=1}^{j,j'=k^v,k^{v'}}$ between the centroids
              of collaborators $v$ and $v'$:
              $$(L^{v,v'})^* = \underset{L^{v,v'}\in\Pi(\nu^v,\nu^{v'})}{\mathrm{argmin}} < L^{v,v'}, C(M^v, M^{v'}) >_F - \frac{1}{\lambda}H(L^{v,v'})$$
        Chose the median collaborator :
          $$v^* = median_{v'}\left\{(L^{v,v'})^*\right\}_{v'=1}^r, v' \neq v$$
        if Quality increased then
            Update the local centroids based on the collaborator $X^{v*}$:
              $$m_j^v = \alpha_{v,v*}\sum_{j'} l_{jj'}^{v,v*} m_{j'}^{v*} \quad 1 \leq j \leq k^v$$
        else
            Compute Feature Selection Algorithm 6 on the median collaborator
              $X^{v*}$
            Update the local centroids if the quality is increased.
until convergence;
return $\{(L^v)^*\}_{v=1}^r$ and the centroids $\{M^v\}_{v=1}^r$
```

dividing the groups are called quartiles, and the groups are referred to as quartile groups. Usually we label these groups 1 to 4 starting at the bottom. The median



(A) Silhouette

(B) ARI

FIGURE (5.4) Sensitivity Box plost CO-OT and CO-FS-OT comparaison

(middle quartile) marks the mid-point of the scores and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less. The middle "box" represents the middle 50% of scores for the group. The range of scores from lower to upper quartile is referred to as the inter-quartile range. The middle 50% of scores fall within the inter-quartile range. As can be seen from these graphs, the overall performance behaviour shows a clear improvement using the feature selection. Moreover, we remark that for ARI box the diversity between the collaborators is reduced, which mean that the CO-FS-OT improve the distribution to be nearer to the real distribution of the data-set.

5.5 Summary and discussion

To summarise, in this chapter we improved a recent prototype-based collaborative approach by introducing the forward feature selection who contributed in the control of the collaboration and enforce interactions between the collaborators in each iteration while preserving the privacy of each learners. Comparing to the collaborative approach based on optimal transport, the proposed algorithm had given a positive results and improved the quality of the exchange of the information, and its strength is highlighted by good experimental results both for artificial and real data-sets.

For the perspective work, we still working on the improvement of the notion of the confidence between the collaborators, we believe that this notion has a strong relation with the diversity between the collaborators and their local quality. Thus we want to build a system that is capable to measure this diversity and quality at the same time, which would lead to create better interaction between learners and reduce the training time of the algorithm and guarantee better convergence.

Conclusion and perspectives

In this thesis, the research outlined concerns the development of Multi-Models clustering approaches to learn from distributed. We work on two principal framework in Multi-Model clustering,

5.6 Conclusion

The first one is the Multi-view clustering where we aim to learn a global optimum model from the views. We proposed two approaches based on optimal transport theory. The first approach (PCA) consists to find a consensus model from local models, by projecting the learned distributions on the global space. The second approach (CNR) aims to learn a new consensus distribution from the local representation of the distributions. In this framework we tackled the motivation behind the proposed approach, we also showed how the transition between the Multi-view clustering to a unified consensus could be modeled though Optimal transport. Finally we proposed extensive experiments on several data sets to prove the utility of our proposed approaches, comparing to a single view clustering and classical method.

The second framework, we tackled in Multi-Models clustering, is the collaborative clustering where the main idea is to exchange the knowledge between different collaborators in order to improve their local models. We introduce in this framework, a new approach based on optimal transport theory (Co-OT) that aims to improve the mechanism of the collaboration, and how to transport the information between collaborators with the minimum cost possible. Furthermore, we proposed an objective function of collaboration based on the Wasserstein distance proposed a solution of how to choose the right collaborators through the comparisons of the local distribution of the centroids and the study of the diversity between the collaborators. Moreover, we prove the adaptability of the proposed approach to different prototype based model like (SOM, K-means...). Besides, we validated the proposed approach using several unsupervised quality index, and also we added an external index since the data has the real label. Finally we compared our contribution to the classical methods in prototype based collaboration.

In the last part of in this thesis, we presents novel model of collaborative learning guided by the feature selection. the main idea of this approach was to create more interaction between the collaborator in order to make the collaboration more beneficial for each collaborator. We proposed a hybrid feature selection algorithm aims to rank the feature according to its importance in the data set for each collaborator, after that build a trade-off between the threshold selection based on the local gain quality, and distant gain quality of the collaborator where we knowledge was transferred, while preserving the privacy of each collaborator. On the other hand, the proposed method has reduced the diversity between the collaborators which tended to give positive results comparing to classical methods. The Collaborative clustering was developed within the framework of optimal transport the theory. Extensive experiments on multiple data-sets to evaluate the proposed approaches and demonstrated its utility in distributed data.

5.7 Future perspectives

Possible future perspectives of this thesis are many. the comparison between the behaviour of the collaborators during the collaboration with and without using the feature selection confirms the the diversity between the collaborators is very important. While our proposed approaches has proved their utility to get better quality of clustering, we believe that we barely tackled the impact of diversity between the collaborators. Although we proposed an heuristic methods that aims the compare the diversity between distributions. we believe that we can build theoretical work that proves the quality of the collaboration has a strong link to diversity. More precisely, it seems that the confidence between collaborators could be analysed through the behaviour of the gain quality and the diversity of each collaborators. Moreover, we can even a function the predict the order of the collaboration based only on initial local quality and diversity.

Another possible prospective to this thesis is to combine the multi-view consensus and the collaborative learning to build to give birth to federated learning which is very similar to collaborative learning. In this context we believe that we the consensus clustering can be seen as central server that controls the collaboration between the views during different steps, and guide exchange between the collaborators will be guided by the central server while preserving the privacy of each collaborator, which that they will be no one to one exchange and the identity of the collaborator will be anonymous, and the exchange will be available only with consensus view.

A very interested extension to this thesis, is the modeling of the collaborative learning as networks, where the each collaborator can be build a local graph clustering and the exchange between the collaborators can be represented by edges of a global graph. Moreover, we believe that these edges will be weighted basing on the quality and diversity between the collaborators. Furthermore, this idea could be developed in the optimal transport framework by introducing the Gromov WAsserstein distance which allows to compare two distributions coming from completely different spaces. This will also resolve limitation of using the same dimension in different sub spaces.

A UIC Data sets

In this appendix, we describe a few data sets taken from the UCI repository that have been used in the experimental parts of this Thesis.

Dermatology data set

This database contains 366 instances described by 34 attributes, 33 of which are linear valued and one of them is nominal.

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris which gives 6 calsses. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages.

Glass data set

Glass Identification data set was generated to help in criminological investigation. At the scene of the crime, the glass left can be used as evidence, but only if it is correctly identified. This data set contains 214 instances, 10 numeric attributes and class name. Each instance has one of 7 possible classes.

Sateimage data set

The sample database was generated taking a small section (82 rows and 100 columns) from the original data. The binary values were converted to their present ASCII form by Ashwin Srinivasan. The classification for each pixel was performed on the basis of an actual site visit by Ms. Karen Hall, when working for Professor John A. Richards, at the Centre for Remote Sensing at the University of New South Wales, Australia. Conversion to 3x3 neighbourhoods and splitting into test and training sets was done by Alistair Sutherland. The database consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the classification associated with the central pixel in each neighbourhood. The aim is to predict this classification, given the multi-spectral values. In the sample database, the class of a pixel is coded as a number.

Escherichia coli data set (EColi)

This data set contains 336 instances describing cells measures. The original data set contains 7 numerical attributes (we removed the first attribute containing the sequence name). The goal of this data set is to predict the localization site of proteins by employing some measures about the cells. There are 4 main site locations that can be divided into 8 hierarchical classes.

Iris data set (Iris)

This data set has 150 instances of iris flowers described by 4 integer attributes. The flowers can be classified in 3 categories: Iris Setosa, Iris Versicolour and Iris Virginica. Class structures are "well behaved" and the class instances are balanced (50/50/50).

Pen digits data set

We create a digit database by collecting 250 samples from 44 writers. The samples written by 30 writers are used for training, cross-validation and writer dependent testing, and the digits written by the other 14 are used for writer independent testing. This database is also available in the UNIPEN format.

Spam base data set (Spam Base)

The Spam Base data set contains 4601 observations described by 57 attributes and a label column: Spam or not Spam (1 or 0).

Waveform data set (Waveform)

This data set consists of 5000 instances divided into 3 classes. The original base included 40 variables, 19 are all noise attributes with mean 0 and variance 1. Each class is generated from a combination of 2 of 3 "base" waves.

Wisconsin Diagnostic Breast Cancer (WDBC)

This data has 569 instances with 32 variables (ID, diagnosis, 30 real-valued input variables). Each data observation is labeled as benign (357) or malignant (212).

Variables are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Wine data set (Wine)

This data set contains 178 instances of Italian wines from three different cultivars. All wines are described by 13 numerical attributes and the classes to be found are the 3 cultivars of origin. Class structures are “well behaved” in this data set, but the class instances are unbalanced (59/71/48).

B Main publications

International Journals

Ben Bouazza F., Bennani Y., El Hamri M., Cabnaes G., Matei B., Touzani A ., (2019) Multi-view clustering through optimal transport, Australian Journal of Intelligent Information Processing Systems.

Ben Bouazza F., Bennani Y., Cabnaes G., Touzani A ., (2020) Unsupervised collaborative learning based on Optimal Transport theory, Journal of Intelligent Systems.

International Conferences

Ben Bouazza F., Bennani Y., ., Cabnaes G., Touzani A ., (2020) Collaborative clustering through optimal transport, (ICANN'20) International Conference on Artificial Neural Networks, Bratislava, Slovakia.

Ben Bouazza F., Bennani Y., Cabnaes G., Touzani A ., (2020) Subspace guided collaborative clustering based on optimal transport,(SoCPaR'20) International Conference on Soft Computing and Pattern Recognition. (Submitted)

Bibliography

[1] Isabelle Abraham, Romain Abraham, Maïtine Bergounioux, and Guillaume Carlier. Tomographic reconstruction from a few views: a multi-marginal optimal transport approach. *Applied Mathematics & Optimization*, 75(1):55–73, 2017.

[2] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[3] Hanan G Ayad and Mohamed S Kamel. On voting-based consensus of cluster ensembles. *Pattern Recognition*, 43(5):1943–1953, 2010.

[4] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[5] Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *ICDM*, volume 4, pages 19–26, 2004.

[6] Steffen Bickel and Tobias Scheffer. Estimation of mixture models using co-em. In *ECML*, pages 35–46. Springer, 2005.

[7] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100. ACM, 1998.

[8] Ulf Brefeld and Tobias Scheffer. Co-em support vector learning. In *ICML*, pages 16–24. ACM, 2004.

[9] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–594, 2015.

[10] Wei Cheng, Xiang Zhang, Zhishan Guo, Yubao Wu, Patrick F Sullivan, and Wei Wang. Flexible and robust co-regularized multi-domain graph clustering. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 320–328, 2013.

[11] Guillaume Cleuziou, Matthieu Exbrayat, Lionel Martin, and Jacques-Henri Sublemontier. Cofkm: A centralized method for multiple-view clustering. In *2009 Ninth IEEE International Conference on Data Mining*, pages 752–757. IEEE, 2009.

[12] Roberto Cominetti, José A Soto, and José Vaisman. On the rate of convergence of krasnosel'ski-mann iterations and their connection with sums of bernoullis. *Israel Journal of Mathematics*, 199(2):757–772, 2014.

[13] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *ECML*, pages 274–289. Springer, 2014.

[14] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[15] Marco Cuturi and David Avis. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.

[16] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *ICML*, pages 685–693, 2014.

[17] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 1(2):224–227, 1979.

[18] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

[19] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.

[20] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings 1995*, pages 194–202. Elsevier, 1995.

[21] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[22] Germain Forestier, Cédric Wemmert, and Pierre Gançarski. Collaborative multi-strategical classification for object-oriented image analysis. In *Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications in conjunction with IbPRIA*, pages 80–90, 2007.

[23] Mohamad Ghassany, Nistor Grozavu, and Younes Bennani. Collaborative clustering using prototype-based techniques. *International Journal of Computational Intelligence and Applications*, 11(03):1250017, 2012.

[24] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Linear Algebra*, pages 134–151. Springer, 1971.

[25] Peter J. Green. On Use of the EM Algorithm for Penalized Likelihood Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):443–452, 1990.

[26] Nistor Grozavu. *Collaborative Unsupervised Learning and Cluster Characterization*. PhD thesis, The Paris 13 University, 2009.

[27] Nistor Grozavu and Younes Bennani. Topological collaborative clustering. *Australian Journal of Intelligent Information Processing Systems*, 12(3), 2010.

[28] Quanquan Gu and Jie Zhou. Learning the shared subspace for multi-task clustering and transductive transfer classification. In *2009 Ninth IEEE International Conference on Data Mining*, pages 159–168. IEEE, 2009.

[29] Emrah Hancer, Bing Xue, and Mengjie Zhang. A survey on feature selection approaches for clustering. *Artificial Intelligence Review*, pages 1–27, 2020.

[30] Nhat Ho, Xuan Long Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via wasserstein means. In *ICML*, pages 1501–1509, 2017.

[31] Tianming Hu, Ying Yu, Jinzhi Xiong, and Sam Yuan Sung. Maximum likelihood combination of multiple clusterings. *Pattern Recognition Letters*, 27(13):1457–1464, 2006.

[32] Zhanxuan Hu, Feiping Nie, Wei Chang, Shuzheng Hao, Rong Wang, and Xuelong Li. Multi-view spectral clustering via sparse graph learning. *Neurocomputing*, 384:1–10, 2020.

[33] Pierre-Emmanuel Jouve and Nicolas Nicoloyannis. A filter feature selection method for clustering. In *International Symposium on Methodologies for Intelligent Systems*, pages 583–593. Springer, 2005.

[34] Leonid V Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4):1381–1382, 2006.

[35] Sotiris Kotsiantis and Panayiotis Pintelas. Recent advances in clustering: A brief survey. *WSEAS Transactions on Information Science and Applications*, 1(1):73–81, 2004.

[36] Antoine Lachaud, Nistor Grozavu, Basarab Matei, and Younès Bennani. Collaborative clustering between different topological partitions. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4111–4117. IEEE, 2017.

[37] Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Lei Wang, Chang Tang, Jianping Yin, Dinggang Shen, Huaimin Wang, and Wen Gao. Late fusion incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2410–2423, 2018.

[38] Majdi Mafarja and Seyedali Mirjalili. Whale optimization approaches for wrapper feature selection. *Applied Soft Computing*, 62:441–453, 2018.

[39] BG Mirkin. Additive clustering and qualitative factor analysis methods for similarity matrices. *Journal of Classification*, 4(1):7–31, 1987.

[40] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

[41] Witold Pedrycz. Collaborative fuzzy clustering. *Pattern Recognition Letters*, 23(14):1675–1686, 2002.

[42] Parisa Rastin, Guénaël Cabanes, Nistor Grozavu, and Younes Bennani. Collaborative clustering: How to select the optimal collaborators? In *2015 IEEE Symposium Series on Computational Intelligence*, pages 787–794. IEEE, 2015.

[43] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[44] Erwin Schrödinger. *Über die umkehrung der naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u . . . , 1931.

[45] K Schwarzschild. Sitzungsberichte preuss. *Akad. Wiss*, 424, 1916.

[46] Saúl Solorio-Fernández, J Ariel Carrasco-Ochoa, and José Fco Martínez-Trinidad. A new hybrid filter–wrapper feature selection method for clustering based on ranking. *Neurocomputing*, 214:866–880, 2016.

[47] Douglas Steinley. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3):386, 2004.

[48] Alexander Strehl, Joydeep Ghosh, and Claire Cardie. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[49] Jérémie Sublime, Guénaël Cabanes, and Basarab Matei. Study on the influence of diversity and quality in entropy based collaborative clustering. *Entropy*, 21(10):951, 2019.

[50] Jérémie Sublime, Basarab Matei, Guénaël Cabanes, Nistor Grozavu, Younès Bennani, and Antoine Cornuéjols. Entropy based probabilistic collaborative clustering. *Pattern Recognition*, 72:144–157, 2017.

[51] Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. From ensemble clustering to multi-view clustering. In *IJCAI*, 2017.

[52] Anusua Trivedi, Piyush Rai, Hal Daumé III, and Scott L DuVall. Multiview clustering with incomplete views. In *NIPS workshop*, volume 224, 2010.

[53] Grigorios Tzortzis and Aristidis Likas. Kernel-based weighted multi-view clustering. In *2012 IEEE 12th international conference on data mining*, pages 675–684. IEEE, 2012.

[54] Sandro Vega-Pons and José Ruiz-Shulcloper. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372, 2011.

[55] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[56] Huan Wang, Shuicheng Yan, Dong Xu, Xiaoou Tang, and Thomas Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *2007 IEEE Conference on*

Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007.

[57] Suhang Wang, Jiliang Tang, and Huan Liu. Embedded unsupervised feature selection. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[58] Cédric Wemmert. *Classification hybride distribuée par collaboration de méthodes non supervisées*. PhD thesis, Strasbourg 1, 2000.

[59] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[60] Junjie Wu, Hongfu Liu, Hui Xiong, Jie Cao, and Jian Chen. K-means-based consensus clustering: A unified view. *IEEE transactions on knowledge and data engineering*, 27(1):155–169, 2014.

[61] Junjie Wu, Hui Xiong, and Jian Chen. Adapting the right measures for k-means clustering. In *SIGKDD*, pages 877–886. ACM, 2009.

[62] Juanying Xie and Chunxia Wang. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications*, 38(5):5809–5815, 2011.

[63] Xijiong Xie and Shiliang Sun. Multi-view clustering ensembles. In *International Conference on Machine Learning and Cybernetics*, volume 1, pages 51–56, 2013.

[64] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[65] Miin-Shen Yang and Kuo-Lung Wu. Unsupervised possibilistic clustering. *Pattern Recognition*, 39(1):5–21, 2006.

[66] Yan Yang and Hao Wang. Multi-view clustering: A survey. *Big Data Mining and Analytics*, 1(2):83–107, 2018.

[67] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, 1995.

[68] Kun Zhan, Feiping Nie, Jing Wang, and Yi Yang. Multiview consensus graph clustering. *IEEE Transactions on Image Processing*, 28(3):1261–1270, 2018.

[69] Jianwen Zhang and Changshui Zhang. Multitask bregman clustering. *Neurocomputing*, 74(10):1720–1734, 2011.

[70] Xiaotong Zhang, Xianchao Zhang, Han Liu, and Xinyue Liu. Multi-task multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3324–3338, 2016.

[71] Handong Zhao and Yun Fu. Dual-regularized multi-view outlier detection. In *IJCAI*, 2015.

[72] Pengfei Zhu, Wencheng Zhu, Qinghua Hu, Changqing Zhang, and Wangmeng Zuo. Subspace clustering guided unsupervised feature selection. *Pattern Recognition*, 66:364–374, 2017.