KADİR HAS UNIVERSITY
SCHOOL OF GRADUATE STUDIES
PROGRAM OF COMPUTER ENGINEERING

CAPTURING THE DATA SIMILARITY AMONG ORGANIZATIONS OF SAME NATURE

WAQAR ISHAQ

DOCTOR OF PHILOSOPHY THESIS

İSTANBUL, JUNE, 2021

CAPTURING THE DATA SIMILARITY AMONG ORGANIZATIONS OF SAME NATURE

WAQAR ISHAQ

Ph.D. THESIS

Submitted to the School of Graduate Studies of
Kadir Has University in partial fulfillment of the requirements for the degree of
Ph.D in Computer Engineering

İSTANBUL, JUNE, 2021

DECLARATION OF RESEARCH ETHICS /
METHODS OF DISSEMINATION

I, WAQAR ISHAQ, hereby declare that;

- this Ph.D. thesis is my own original work and that due references have been appropriately provided on all supporting literature and resources;
- this Ph.D. thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- I have followed *Kadir Has University Academic Ethics Principles prepared in accordance with The Council of Higher Education's Ethical Conduct Principles.*

In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations.

Furthermore, both printed and electronic copies of my work will be kept in Kadir Has Information Center under the following condition as indicated below (SELECT ONLY ONE, DELETE THE OTHER TWO):

☐ The full content of my thesis will be accessible from everywhere by all means.

WAQAR ISHAQ
17.06.2021

KADİR HAS UNIVERSITY

SCHOOL OF GRADUATE STUDIES

ACCEPTANCE AND APPROVAL

This work entitled CAPTURING THE DATA SIMILARITY AMONG ORGANIZATIONS OF SAME NATURE prepared by WAQAR ISHAQ has been judged to be successful at the defense exam on 17.06.2021 and accepted by our jury as Ph.D. thesis.

APPROVED BY:

Assoc. Prof. Habib ŞENOL (Advisor) .....
Kadir Has University

Assoc. Prof. Tamer DAĞ .....
Kadir Has University

Asst. Prof. Ayşe Bahar DELİBAŞ .....
Kadir Has University

Prof. Songül VARLI .....
Yıldız Technical University

Prof. Cafer ÇALIŞKAN .....
Antalya Bilim University

I certify that the above signatures belong to the faculty members named above.

.....

Prof. Mehmet Timur AYDEMİR

Director of School of Graduate Studies

DATE OF APPROVAL: 17.06.2021

TABLE OF CONTENTS

CAPTURING THE DATA SIMILARITY AMONG ORGANIZATIONS OF SAME NATURE

ABSTRACT

The vertical collaborative clustering aims to unravel the hidden structure of data (similarity) among different sites, which will help data owners to make a smart decision without sharing actual data. For example, various hospitals located in different regions want to investigate the structure of common disease among people of different populations to identify latent causes without sharing actual data with other hospitals. Similarly, a chain of regional educational institutions wants to evaluate their students' performance belonging to different regions based on common latent constructs. The available methods used for finding hidden structures are complicated and biased to perform collaboration in measuring similarity among multiple sites. In this dissertation, the author proposed two approaches of vertical collaborative clustering, namely (1) Vertical Collaborative Clustering Model (2) Vertical Collaborative Clustering based on Bit-Plane Slicing, with superior accuracy over the state of the art approaches.

The Vertical Collaborative Clustering Model ($VCCM$) manages the collaboration among multiple data sites using Self-Organizing Map (SOM). It includes standard procedure and tuning of the exchanged information in specific proportionality to augment the learning process of the clustering via collaboration. Moreover, the VCCM unravels hidden information without compromising the data confidentiality. The aim of the model is to set an ideal environment for the collaboration process among multiple sites. The VCCM is evaluated by purity measurement, using four datasets (Iris, Geyser, Cancer and Waveform). The findings of this study show the significance of the VCCM by comparing the collaborative results with the local results using purity measurement. The VCCM unlocks possible reasons determining impact of collaboration based on related and unrelated patterns. The results demonstrate that the proposed VCCM improves local learning by collaboration and

also helps the data owner to make better decisions on the clustering. Additionally, the results obtained have better accuracy than the existing approaches.
The proposed Vertical Collaborative Clustering based on Bit-Plane Slicing (VCC-BPS) is simple and unique approach with improved accuracy, manages collaboration among various data sites. The VCC-BPS transforms data from input space to code space, capturing maximum similarity locally and collaboratively at a particular bit plane. The findings of this study highlight the significance of those particular bits which fit the model in correctly classifying clusters locally and collaboratively. Thenceforth, the data owner appraises local and collaborative results to reach a better decision. The VCC-BPS is validated by Geyser, Skin and Iris datasets and its results are compared with the composite dataset. It is found that the VCC-BPS outperforms existing solutions with improved accuracy in term of purity and Davies-Bouldin index to manage collaboration among different data sites. It also performs data compression by representing a large number of observations with a small number of data symbols.

Keywords: Collaborative clustering, Collaboration, Vertical collaborative clustering, Cluster combination, Purity measurement, Similarity measurement.

TEZİN TÜRKÇE ADI

ÖZET

Dikey işbirlikçi kümeleme, farklı siteler arasındaki gizli veri yapısını (benzerliği) ortaya çıkarmayı amaçlayarak, veri sahiplerinin gerçek verileri paylaşmadan akıllıca bir karar vermelerine yardımcı olacaktır. Örneğin, farklı bölgelerde bulunan çeşitli hastaneler, gerçek verileri diğer hastanelerle paylaşmadan gizli nedenleri belirlemek için farklı popülasyonlardan insanlar arasındaki ortak hastalık yapısını araştırmak ister. Benzer şekilde, bir bölgesel eğitim kurumları zinciri, öğrencilerinin farklı bölgelere ait performanslarını ortak örtük yapılara göre değerlendirmek ister. Gizli yapıları bulmak için kullanılan mevcut yöntemler karmaşıktır ve birden çok site arasındaki benzerliği ölçmede işbirliği yapmak için önyargılıdır. Bu tezde, yazar iki dikey işbirlikçi kümeleme yaklaşımı önerdi, yani (1) Dikey İşbirlikçi Kümeleme Modeli (2) Bit Düzlemi Dilimlemeye dayalı Dikey İşbirliğine Dayalı Kümeleme, son teknoloji yaklaşımlarına göre üstün doğrulukla.

Dikey İşbirlikçi Kümeleme Modeli (VCCM), Kendi Kendini Düzenleyen Harita (SOM) kullanarak birden çok veri sitesi arasındaki işbirliğini yönetir. İşbirliği yoluyla kümelemenin öğrenme sürecini artırmak için, belirli orantılı olarak değiş tokuş edilen bilginin standart prosedürü ve ayarlanmasını içerir. Dahası, VCCM gizli bilgileri veri gizliliğinden ödün vermeden çözer. Modelin amacı, birden çok site arasında işbirliği süreci için ideal bir ortam oluşturmaktır. VCCM, dört veri seti (Iris, Geyser, Cancer ve Waveform) kullanılarak saflık ölçümüyle değerlendirilir. Bu çalışmanın bulguları, işbirlikçi sonuçları saflık ölçümünü kullanarak yerel sonuçlarla karşılaştırarak VCCM'nin önemini göstermektedir. VCCM, ilişkili ve ilgisiz modellere dayalı olarak işbirliğinin etkisini belirleyen olası nedenleri ortaya çıkarır. Sonuçlar, önerilen VCCM, nin işbirliği yoluyla yerel öğrenmeyi geliştirdiğini ve ayrıca veri sahibinin kümeleme konusunda daha iyi kararlar almasına yardımcı olduğunu göstermektedir. Ek olarak, elde edilen sonuçlar mevcut yaklaşımlardan daha iyi doğruluğa sahiptir.

Bit Düzlemi Dilimlemeye (VCC-BPS) dayalı önerilen Dikey İşbirliğine Dayalı Kümeleme,

gelişmiş doğrulukla basit ve benzersiz bir yaklaşımdır ve çeşitli veri siteleri arasındaki işbirliğini yönetir. VCC-BPS, verileri giriş alanından kod alanına dönüştürerek, belirli bir bit düzleminde yerel olarak ve işbirliği içinde maksimum benzerliği yakalar. Bu çalışmanın bulguları, modele uyan belirli bitlerin, sınıf etiketlerini yerel olarak ve işbirliği içinde doğru bir şekilde sınıflandırmadaki önemini vurgulamaktadır. Bundan sonra, veri sahibi daha iyi bir karara varmak için yerel ve işbirliğine dayalı sonuçları değerlendirir. VCC-BPS, Gayzer, Skin ve Iris veri kümeleri tarafından doğrulanır ve sonuçları bileşik veri kümesiyle karşılaştırılır. VCC-BPS'nin, farklı veri siteleri arasındaki işbirliğini yönetmek için saflık ve Davies-Bouldin indeksi açısından iyileştirilmiş doğrulukla mevcut çözümlerden daha iyi performans gösterdiği bulunmuştur. Ayrıca, çok sayıda gözlemi az sayıda veri sembolü ile temsil ederek veri sıkıştırması gerçekleştirir.

Anahtar Sözcükler: İşbirlikçi kümeleme, İşbirliği, Dikey işbirlikçi kümeleme, Küme kombinasyonu, Saflık ölçümü, Benzerlik ölçümü.

ACKNOWLEDGEMENTS

All praise be to Allah, the Lord of the universe. Every thing belongs to Allah and to Him, it shall return.

First and foremost, the author thanks the thesis advisor Assoc. Prof. Habib ŞENOL for his support and cooperation to manage the final contribution of my PhD defense. The author would like to sincerely thank Dr. Eliya BÜYÜKKAYA for her support and guidance throughout this study. The author is thankful to the University and Computer Engineering Department in particular for the financial support extended to him during the study.

The author is grateful in particular to Prof. Feza KERESTECİOĞLU to arm him with skills to think creatively and outside the box to find solutions to the problems came across at different stages of this work. The author would also like to thank members of the research committee for their support and suggestions during this study.

The author is indebted of respectful gratitude to his father (***Dada***), brother (***Bhaijan***) and sisters (***Honey, Nazgul & Kitty with Bhabee***) for their affection and prayers during the entire research work. Special appreciations are due to the author's spouse (***Aasia Waqar***), daughters (***Maryum & Arva***) and sons (***Muhammad Hussain Ishaq & Muhammad Haider Ishaq***) for their patience, understanding, and love has been the source of encouragement to him.

This thesis is dedicated to our beloved ***Prophet Muhammad*** peace be upon ***him***, ***his companions***, author's parents (***Dada & Ammee***), brother (***Muhammad Ishfaq Khan***), sisters, wife and kids with my maternal and paternal nephews and nieces. They all have been source of spiritual, mental and physical support during the challenges of entire research work.

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

C_v	Code vector
C_M^i	Code map of the i^{th} data site
P^i	Local purity of the i^{th} data site
$\overline{P}$	Collaborative purity
GP	Global purity
DB	Local Davies Bouldin index
$\overline{DB}$	Collaborative DB index
S_i and S_j	Local dispersions of the i^{th} and j^{th} clusters
d(i,j)	Local inter-cluster distance between the i^{th} and j^{th} clusters
D(i^A,j^B)	Centroid to centroid distance between i^{th} and j^{th} clusters of data site A and B respectively
HCC	Horizental Collaborative Clustering
VCC	Vertical Collaborative Clustering
SOM	Self-Organizing Mapping
GTM	Generative Topographic Mapping
BPS	Bit Plane Slicing
VCCM	Vertical Collaborative Clustering Model
VCC-BPS	Vertical Collaborative Clustering based on Bit Plane Slicing
VCC-SOM	Vertical Collaborative Clustering using Self-Organizing Mapping
VCC-GTM	Vertical Collaborative Clustering using Generative Topographic Mapping

1. INTRODUCTION

1.1 Preface

Data have become valuable asset for the data owners such as companies, institutes, organizations etc. to make smart decision. This requires data behavior understanding to unlock the hidden information. Having a dataset A, with number of observations $n = \{n_1, n_2....n_n\}$ that are measured by number of features $X = \{x_1, x_2....x_m\}$, provokes an intensive task of finding answers for questions that can contribute to the data owner benefits. In other words, how can we learn from the behavior of this dataset? Ultimately, to learn something, first, you must have a goal. If the goal is to predict an output value y for given dataset A, that can be done via decision procedure known as $h : X \rightarrow Y$ where $Y = \{y_1, y_2....y_n\}$. In this case, A turns into a training dataset that helps to train a mathematical algorithm model "SA" where its output is to predict a value y. Straightforward, substitute an observation n_j measured value vector $X = \{x_1, x_2....x_m\}$ to model "SA" the output is a prediction of y_j value. This goal is known as a supervised learning approach. This approach can be easily evaluated by using the evaluation set or cross-validation technique to predict an output value of the model [1]. On the contrary, if the goal is not to predict an output value of y, but to disclose possible hidden structure in dataset A, then such an approach would be known as unsupervised learning. The mathematical algorithm of this approach can find the type of underlying structure that the user has established either directly or indirectly in their approach. Sometimes, besides, the approach also provides some level of significance of the discovered structure.

Clustering as a type of unsupervised learning approach, segregates observations into groups, called clusters, which may be mutually exclusive or overlap, relying on the

technique used. The observations within a cluster are more similar to each other than the observations from another cluster. The similarity measure is of outstanding importance to define clusters that can be disclosed in the data. Different types of distances have been introduced in the literature with respect to the problem and context of the study [3].

1.2 Research Focus

Supposing that the owner of the dataset A has the goal of using an unsupervised learning approach. For the sake of argument, let's assume that the owner has essentially adhered to some criteria before adopting the approach, criteria such as defining the type of cluster will be looked for, organizing search space, validation methods (all of which will be introduced in the literature section). As a result of applying the approach, Figure 1.1 illustrates the clustering result of the dataset, represented in two dimensions. This result of clustering has been reached using $\{x_1, x_2, ..., x_m\}$ of features on $\{n_1, n_2, n_3, ..., n_n\}$ of observations where m and n represent number of features and observations respectively for dataset A.

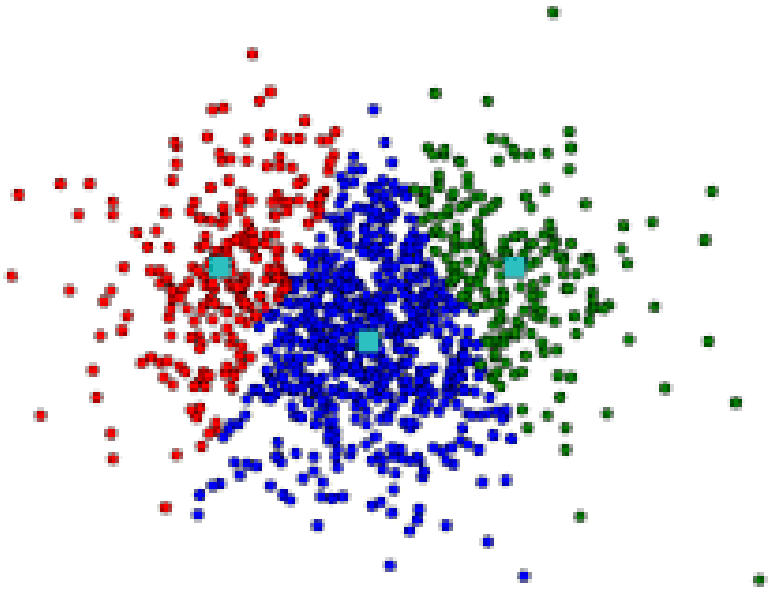


Figure 1.1 An example of clustering

The question is with same number of features $\{x_1, x_2, ..., x_m\}$, however added a

greater number of observations $\{n_1, n_2, n_3, ..., n_{n+1}, n_{n+2}, n_{n+n}\}$, will the result of clustering remain the same? Will it be any better or worse? Another scenario, if the features have changed to a greater number, with the same observations, will the result again remain the same? Or even if both features and observations have changed, the same questions apply. Learning the behavior of 100 observations can help to discover a pattern. However, if the number grows to 1000 or more, positively the pattern would be more intelligent. To this end, a concept of "combining the clustering" is introduced. According to "combining the clustering" approach, the same clustering method can be applied over two or more datasets, then results are shared and merged to associate clusters of one site with other site(s) to identify similarity. This requires the selection of suitable clustering method(s), an adjustment in parameter values, number of features and observations, etc., to obtain unbiased outputs. This approach adds parallelization, scalability and robustness to the desired solution [4, 27]. One of combining the clustering inspiration ideas is known as collaborative clustering.

Let us assume that the owners of dataset A and B, having the same features, apply the same clustering algorithm and obtain local results R_A and R_B respectively. Now, is there any way that both owners can exchange information about the two results R_A and R_B? If that is possible, then both owners will evaluate the final clustering result obtained from other site(s) in addition to their local data. The benefit here is augmenting the learning process of clustering the local data through external clustering information from other site(s). Technically speaking, different approaches are introduced to implement the idea, one of which is known as *horizontal* collaborative clustering (HCC) and the other is known as *vertical* collaborative clustering (VCC). In HCC approach, different datasets have same observations with different features, while in VCC approach, different datasets contain same features with different observations collaborate the clustering results [2, 28].
Different researchers worked on both of these approaches to explore hidden information and measure similarity among various independent datasets. This study focuses on vertical collaborative clustering by considering two or more independent

organizations having data of same nature (feature space). For example, different hospitals located in various regions want to investigate the common disease among people of different populations. Notably, these organizations, companies or hospitals expressed in same feature space, are not allowed to share the actual data due to data confidentiality. Moreover, pooling of different organizational data with same feature space into single tall dataset (combined dataset with a large number of observations) is not feasible for data analysis. The reasons are bandwidth restriction and performance issues such as latency and large memory computations etc. The vertical collaborative clustering using Self-Organizing Mapping (SOM) [20, 22] and Generative Topographic Mapping (GTM) [10, 16, 29, 34] are existing approaches to apprehend data information among different data sites but have certain limitations which are mentioned as under:

1. SOM is sensitive to learning rate and neighborhood function in generating results which affects similarity measurement [23]. In SOM, all results rely on size of the map and a collaborative matrix which consists of collaborative coefficients, determines strength of each collaborative link, and degrade results if not set correctly [20, 32, 33]. Moreover, it lacks simplicity in calculating coefficients, which affects accuracy and performance. For further reading on SOM (see e.g.[15, 17])
2. GTM is non-linear approach of unsupervised learning and more precise than linear approaches but has higher run time complexity than linear approaches [16, 32]. GTM uses likelihood function for fast convergence and better tuning of topographic map parameters, which may not guarantee global convergence for all algorithms. Moreover, fast convergence does not ensure results of good quality [14, 31].
3. Accuracy of existing solutions is not verified by comparison of local and collaborative results with global result for which datasets are pooled (all datasets are combined). Additionally, test data results are not mentioned to evaluate model generalization.

This study aims to use unsupervised learning approach to associate clusters similarity of one data site with that of other independent site(s) in distributed environment by sharing data results, not the actual data. In this dissertation, two approaches are proposed to implement the concept of the vertical collaborative clustering (VCC). These approaches are namely, the *Vertical Collaborative Clustering Model* ($VCCM$) and *Vertical Collaborative Clustering based on Bit-Plane Slicing* ($VCC-BPS$) to overcome the above-mentioned limitations.

The Vertical Collaborative Clustering Model ($VCCM$) manages the collaboration among multiple data sites using Self-Organizing Map (SOM). It includes standard procedure and tuning of the exchanged information in specific proportionality to augment the learning process of the clustering via collaboration. Moreover, the VCCM unravels hidden information without compromising the data confidentiality. The aim of the model is to set an ideal environment for the collaboration process among multiple sites. The findings of the VCCM outputs show its significance by comparing the collaborative results with the local results using purity measurement. The VCCM unlocks possible reasons determining impact of collaboration based on related and unrelated patterns. The results demonstrate that the proposed VCCM improves local learning by collaboration and also helps the data owner to make better decisions on the clustering. Additionally, the results obtained have better accuracy than the existing approaches.

The another proposed approach, the Vertical Collaborative Clustering based on Bit-Plane Slicing (VCC-BPS) is simple, accurate and compresses data, managing collaboration among different data sites. It performs clustering, data reduction and visualization simultaneously. The VCC-BPS consists of two phases i.e. local and collaborative phase to find a bit plane at which model fits the data to identify maximum similarity locally and collaboratively. The working principle of this approach is to transform input data to code (discrete or latent) space using bit plane slicing approach, where the model fits the data with maximum similarity at particular bit plane. This approach consists of the two phases as follows:

1. *Local Phase*: The object of the local phase is to look for that specific bit plane

at which observations are grouped based on maximum similarity within the local dataset. It is an iterative process, searching for a bit plane at which similar observations are represented by particular code map where model fits the data to capture maximum similarity locally.

2. *Collaborative Phase*: In a collaborative phase, the basic requirements of the VCC are fulfilled to capture similar behavior among participating data sites. This is achieved by exchanging the local result table developed at each local site with that of other participating site(s). And then both the local and external results (result received from other site(s)) are merged with respect to certain rules to identify maximum similarity. The rules in detail are mentioned in section 4.2, ensure symmetry in behavior among participating sites.
   Finally, the data owner decides whether collaboration brings any new insight to uncover hidden information (similarity). The local and collaborative results are evaluated by purity and David-Bouldin index.

The author contributions in this dissertation are:

1. Interaction is developed between two or more data sources having same feature space to reveal similarities among participating sites without compromising data confidentiality. The proposed approaches do clustering, data reduction and visualization simultaneously.
2. The VCCM works as a coordinator to organize and manage the collaboration process between local and outdoor site(s) using standard procedure to enhance local learning. Moreover, the proposed model provides suitable environment for collaboration.
3. The VCC-BPS is simple novel approach, can be used as a tool by different organizations to make smart decisions without compromising data confidentiality. It performs compression to represent a large number of observations by small data codes.

1.3 Overview of the Study

The overview of the dissertation is organized as follows. **Chapter** 1 presents importance, problem statement and aim of the study with motivation. **Chapter** 2 includes comprehensive review of the literature, describing topics and theories related to the clustering, collaborative clustering and its types with different approaches. **Chapter** 3 elaborates the proposed vertical collaborative clustering model (VCCM) to analyze the data. **Chapter** 4 explains the proposed vertical collaborative clustering based on bit-plane slicing (VCC-BPS) approach to examine and interpret the data among different sites. **Chapter** 5 consists of the main results obtained from the proposed study. **Chapter** 6 discusses the analytical findings with final implications. Finally, **chapter** 7 mentions the summary of the theoretical contributions and also discuss the limitations with potential avenues for the future work.

2. LITERATURE REVIEW

This chapter explains clustering, collaborative clustering with its requirements, types, importance, existing approaches and bit plane slicing.

2.1 Clustering

The goal of clustering as a type of unsupervised learning, is to group clusters of observations that can be mutually exclusive or overlapped. The similarity is an important factor to decide the observations that are grouped together. Two approaches are mostly known for clustering [5]:

1. *Generative* approach is often based on statistical model, where the objective is to determine parameters that maximize how well the model fits the data.
2. *Discriminative* approach mostly depends on optimization criteria and similarity measurements to group the data.

Let us consider a buzzword known as ill-defined problem [6, 7]. This problem is considered from the idea that mathematically, the similarity is not a transitive relation while belonging to the same cluster. In other words, different methods may give inconsistent clustering outputs for the same data. Moreover, proper heuristics be employed to manage the computational cost. The following points need to be considered before adopting clustering approaches:

1. The first thing is to define the type of clusters being looked for, which relies on the context and our goal. The reason is that the same set of observations can be clustered in different ways, depending on type of distance used [8].

For example, measuring a distance between observations in the input space or between an observation and a cluster, may lead to a different clustering model [3].

2. The learning process of clustering is affected by the organization of the search space which is based on the number of features, their degree of dependence, the type of normalization, etc. [2] discusses how the escalation of dimensionality increases the volume of the space exponentially.
3. The last matter is considering the validation step of the clustering model. Since there is no post validation method to compare true classes with the classes discovered by the algorithm for test data, clustering output evaluation is a delicate task. A few statistical procedures have been introduced to test the importance of the clustering result. They are based on measuring deviations of some statistical quantities but have certain limitations which create hindrances in getting the true findings [21].

2.2 Collaborative Clustering and the Vertical Type

The clustering algorithms use two types of information during their computations:

1. Information about observation membership.
2. Information about internal parameters, such as the number of clusters anticipated, the coordinates of observations, and so on.

If we consider these two kinds of information developed from each local dataset, then the question is: Is it possible to exchange this information with another site that has a similar structured dataset? The answer is "yes". The concept, of doing so, is known as the collaborative clustering [9]. The goal is that the local clustering process can benefit from the work done by the other collaborator. In other words, collaborative clustering helps the local algorithm to escape from local minima (i.e. by only operating over a local dataset) by discovering better solutions (i.e. by exchanging information with another site that has a similar dataset). The validity

is measured by the assumption that useful information be shared between the local sites. Important benefits of collaboration occur due to [2]:

1. Operating on local data in addition to information from other sites can help the algorithm to enhance the learning process.
2. The algorithm can escape local optima by using external information to get better solutions.
3. The local bias can be managed by using external information. However, this information can also be subject to other types of bias.

The core of this approach is accomplished by the exchange of information. Here information can be about the local data, or current hypothesized local clustering, or the value of one algorithm's parameters. In other words, what can be shared between experts is information about data (e.g. features found useful, distances used, etc.) or information on the observation itself, such as the characteristic of an observation measured by a fixed feature vector. Important to mention is the performance measurement in clustering. It is hard to introduce an answer to such a question or in other words, no perfect answer can be reached. Therefore, there is no specific way of measuring the absolute quality of partitioning the data points. However, as mentioned above in the validity line that the assumption always lays on as the useful information is shared between the local sites. Nonetheless, some measurements are still a pioneer metric to measure the performance or the validity of the cluster. For example, [11] introduces a technique by defining the similarity between input clusters based on the graph structure. Notable, that the way the cluster is viewed can be a good matter of measuring the performance or the validity of the cluster. But still, as the goal is to look for a common structure among different datasets, it is no longer possible to make direct comparisons at the level of the observation since they are different. Only descriptions of the clusters found by the local algorithm can be exchanged, and a consensus measure must be defined at this level [12].

Following the above paragraph, it is important to discuss an important question that

is how to control the collaboration phase? There are different approaches introduced in this domain, here are some of them [13]:

1. Synchronous or asynchronous operations: The former occurs when each local clustering process has its own goal and exchanges information only in the search of its local goal. The latter one is generally needed when the result depends on all local achievements.
2. Iterative or one-time process: The former occurs when the computations performed by each local algorithm can consider partial solutions shared by other algorithm sites and is therefore iterative. In a one-time process, all algorithms compute their local solution, after which a master algorithm combines them and outputs the final solution.
3. Local or global control: The former works with an asynchronous control strategy, while the latter is linked with the computation of a final combined overall solution [14].

However, regardless of the method of controlling, termination condition is of main concern in collaborative approaches. Where it is required that a clustering algorithm stops when a condition is met, even though the solution obtained might not be meeting the global optimization criterion [14].
To conclude this section, we will introduce the two most common types of implementing collaborative clustering. Noteworthy, other types are there, however, the focus of this paper is to discuss one type in particular. The two types known in collaborative clustering are [10]:

1. Horizontal collaborative clustering, the idea of it as the name may suggest, same observations are expressed in different feature space as shown in Figure 2.1 (a). In other words, let dataset A with set of observations $\{n_1, n_2....n_n\}$, operates over the feature space $\{x_1, x_2....x_m\}$. Another dataset with the same set of observations can be investigating again, however in different feature spaces such as $\{z_1, z_2....z_m\}$ as shown in the Figure 2.1 (b). For example, same

patient visits kidney and heart disease hospitals as shown in the Figure 2.1 (b). This forms independent datasets of same patients in two or more hospitals. For further detail on the HCC, see e.g. [26, 30, 31].

2. Vertical collaborative clustering describes datasets in the same feature space but with different observations. In other words, let dataset A with set of observations $\{n_1, n_2....n_n\}$, operates over the feature space $\{x_1, x_2....x_m\}$. Another dataset B with different set of observations $\{o_1, o_2....o_n\}$, however, operates in the same feature space $\{x_1, x_2....x_m\}$ as shown in the Figure 2.2 (a). For example, different patients having common disease (kidney), visits various hospitals located in different regions. These Kidney disease hospitals have same features to diagnose common disease among people of different regions as shown in Figure 2.2 (b). This forms independent datasets, consisting of various observations expressed in same feature space.

The proposed work considers the last type of collaborative clustering which is the vertical collaborative clustering (VCC), has the following basic requirements [22]:

1. Type and number of features must be the same among data sites.
2. Share local findings with other sites, such that collaborative results obtained at each site are as if obtained from the pooled dataset (all datasets are combined).

The benefits of such an approach are as follow [22]:

1. Reduces time and space complexity.
2. Keeps data confidentiality.
3. Enhances scalability.

2.3 Self-Organizing and Generative Topographic Mapping

Self-Organizing Map (SOM) is linear approach of unsupervised learning, performs clustering by mapping high dimensional data into two-dimensional map. It displays clustering and visualization, as well (Algorithm 1) [16, 17]. In SOM, the dataset

Figure 2.1 Horizontal Collaborative Clustering Description: (a) Various independent datasets having same set of observations, expressed in different feature space in the distributed environment. (b) Dataset A and B are of Kidney and Heart disease hospitals, where same patient having both diseases visit them (features of Kidney disease data are different from that of Heart).

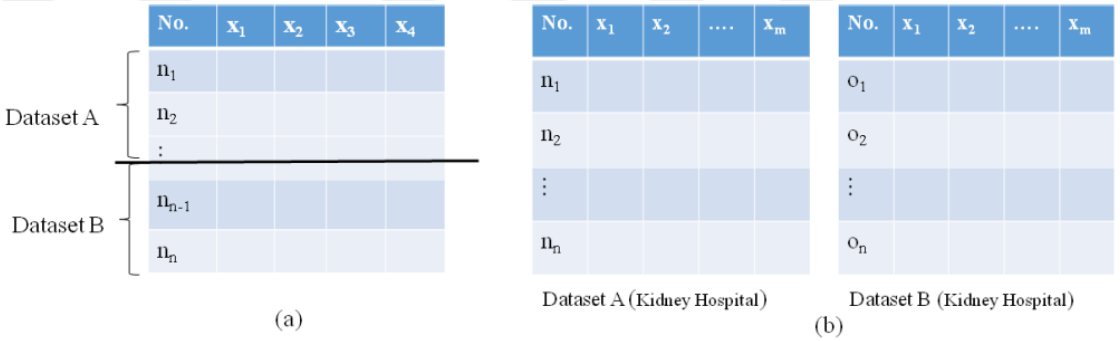


Figure 2.2 Vertical Collaborative Clustering Description: (a) Various independent datasets having same feature space with different observations in the distributed environment. (b) Both datasets A and B are Kidney disease hospitals. Different patients visit Kidney disease hospitals having same features (data of same nature).

is fed to a randomly initialized map, where the weight vector of the nodes in the map are gradually adjusted towards observations in the dataset. The SOM uses two stages i.e. competition and cooperation stage to associate similar observations from input space to same node in the discrete (latent) space. In the competition stage, the node in map whose weight vector is the most similar to the input observation vector is selected as the *best matching unit* (BMU) for that observation (Line 5). In cooperation stage, the weight vectors of the BMU and nodes close to the BMU in the map are adjusted with respect to the input observation vector (Line 9). The influence of one node over other nodes depends upon the degree of closeness. We use Gaussian distribution to capture such relation among neighbor nodes due to its

monotonic decaying property (Line 8) [6, 15]. The weight of an activated node is updated using the following formula:

$$w_j(n+1) = w_j(n) + \alpha(n) \times \theta_{j,i}(n) \times (x(n) - w_j(n)) \tag{2.1}$$

Where n refers to the index of current iteration, i is the index of the BMU for the current observation $x(n)$, j is the index of activated node, w_j is the weight vector of activated node, $\theta_{j,i}(n)$ is the neighbourhood function describing the distance between the BMU i and the activated node j at iteration n, and $\alpha(n)$ is a learning-rate parameter. The change in weight of nodes decrease with time and distance from the BMU via $\theta_{j,i}(n)$ and $\alpha(n)$.

In [33], hybrid collaborative clustering approach which is a combination of vertical and horizontal collaborative clustering, uses the Self Organizing Map (SOM) algorithm to find common structure by exchange of information. [35] explains collaborative classification among different information sources (data sites) with same features using SOM to reveal common structure of distributed data. Collaborative filtering makes use of available preference to predict unknown preferences based on clustering similarity measurement [36]. Alternatively, Generative topographic mapping (GTM) is probabilistic model using expectation maximization as an alternate to SOM based on following limitations of SOM:

- Neighborhood preservation is not guaranteed.
- Convergence of prototypes are not guaranteed as well.
- There is no theory about parameter initialization.

GTM is non-linear approach of unsupervised learning and more precise than linear approaches but has higher run time complexity than linear approaches [23]. Recently [10, 34], proposed the probabilistic approaches of the collaborative learning using generative topographic mapping (GTM) based on principle of vertical collaborative clustering to exchange the information for tunning the topographic maps parameters. [32] introduces nonlinear classification approach to interpolate missing data and performs nonlinear mapping between data and latent space using Gener-

ative Topographic Map (GTM). GTM uses likelihood function for fast convergence and better tuning of the topographic map parameters, which may not guarantee global convergence for all algorithms. Moreover, fast convergence does not ensure results of good quality [22]. Additionally, the expectation maximization algorithm relies on type of data distribution to be known.

Contrary to GTM, our concern is to develop a technique that ensures unbiased environment among multiple data sites. For this purpose, VCCM use SOM with a concern that all datasets must be initialized with same weight, grid size and its orientation, number of iterations and size of data sets. Such symmetry in initialization among multiple data sites ensures preservation in neighborhood. This neighborhood function is based on radius which in turn is function of iteration. Moreover, convergence of prototypes can be guaranteed by symmetry in number of iterations using heuristic approach. Based on above assumptions, an unbiased environment can be developed to manage collaborative process among multiple data sites.

2.4 Bit Plane Slicing

The Bit Plane Slicing (BPS) is an image compression technique that divides a pixel of 8 bits image into 8-bit planes. Bit plane ranges from least significant bit (LSB) represented as bit-level 0 to most significant bit (MSB) marked as bit-level 7. The least and most significant bit plane contains all low and high order bits in the byte respectively. Change in low order bits of LSB does not change value much because they lack high contrast, while the change in high order bits of MSB signifies the change in data. Therefore, the most significant bit contains the majority of significant data and forms an image approximately similar to the original 8-bit image. This highlights the relative importance of specific bits in the image to reduce the image size. Based on such a strong characteristic of BPS, an 8-bit image containing a large amount of data is compressed into an image of small size with high similarity [24, 25]. The pictorial representation of bit plane slicing is shown in Fig 2.3 for an image composed of pixels, where each pixel occupies 8-bits memory and is represented by eight single-bit planes. The equation (2.2) is used to form k^{th} bit

plane with respect to k^{th} bit selected from all pixels. [38]:

$$BitPlane_k = Remainder\{\frac{1}{2}floor[\frac{1}{2^k}Image]\} \tag{2.2}$$

Where the value of k varies from 0 to 7. Suppose a gray scale image contains a pixel of intensity value 220. To find appropriate value for fourth bit plane, equation (2.2) will return 1.

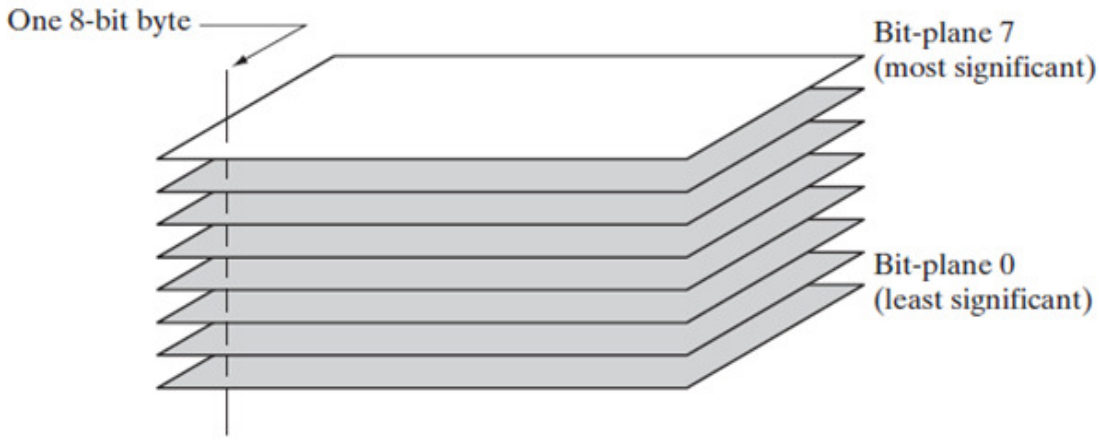


Figure 2.3 Bit Plane Slicing Description [37]

3. VERTICAL COLLABORATIVE CLUSTERING MODEL

This chapter mentions the functionality, architecture and technical detail of the proposed approach *Vertical Collaborative Clustering Model* (VCCM).

3.1 VCCM Functionality

The VCCM functionality is to operate as a coordinator among different sites which are interested to share the clustering results. The proposed VCCM will enable a local data owner (e.g. business companies, government agencies and institutions, etc.) to justify whether the collaborating information from outdoor site(s) would unravel more hidden structure in implementing unsupervised clustering technique on the local data. As a result, that hopefully would lead to a better decision. The VCCM architecture is illustrated in the Figure 3.1 to organize and manage collaboration process between the local and outdoor site(s). Following are the VCCM steps to enhance local learning:

1. Clustering process is run over local data using same initialization parameters such as weight (W_o), map size etc. to produce local SOM map at each data site.
2. Local clustering process is re-run over the local data (say A) using the external shared result say W_B. The output of this step is termed as collaborative SOM map.
3. SOM map may be shared among more than two sites one after another, forms chain mapping (W_{BC}) as shown in the Figure 3.1.
4. K-means approach is applied over the local and collaborative SOM maps to

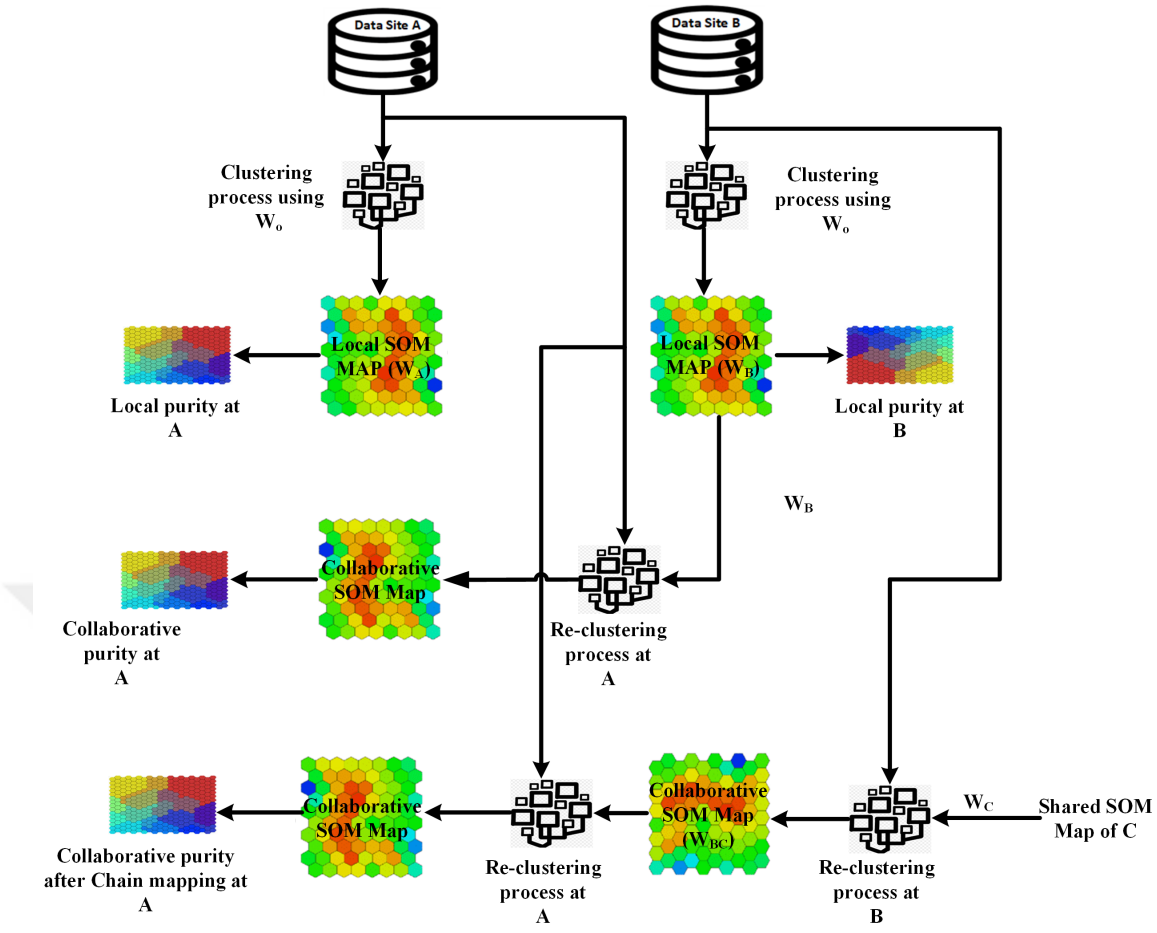


Figure 3.1 Description of the VCCM Architecture

extract clusters.

5. The clusters over SOM map are evaluated by the purity index.

3.2 VCCM Technicality

Technically in the $VCCM$, the collaboration between the data sites is based on mapping of the similar observations to same nodes using Self-Organizing Map (SOM). Since this learning requires unbiased environment to manage collaboration among sites, we propose local clustering be initialized with parameters of same values. Such initialization ensures preservation in the neighborhood and prototype convergence. The collaborative re-clustering phase (re-clustering process) processes local data using shared map of other data site(s). This shared map represents mapping information from continuous observation space (input space) to discrete map space

for respective data sites. Moreover, if more than two data sites are involved in collaboration, the map shared among more than two sites, initialized with same parameters forms chain mapping to represent similar observations in discrete map space as shown in the Figure 3.1. The output of re-clustering process gives final map to capture the local behavior with respect to the shared information of the outdoor site(s). In other words, this information exchanged expands local learning to discover hidden structure in local data by adding clustering information from other sites, which can lead to a better local clustering result. The proposed VCCM algorithm consists of three phases to manage collaboration between data sites, which are local clustering, collaborative re-clustering and evaluation. Algorithm 2 elaborates our VCCM proposition for a given local data site A, while other sites B, C, etc. participate in collaboration with site A.

3.2.1 Local Clustering

In local clustering phase, to ensure unbiased environment for the collaboration, data sites initialize parameters with same values, such as map size, learning rate, neighborhood parameter, size of data set and number of iterations etc., prior to the SOM clustering algorithm. The SOM algorithm 1 is then applied over the local data set to compute the clustering results, namely the local map W_A^{loc} (Line 1 of Algorithm 2).

3.2.2 Collaborative Re-clustering

To establish collaboration, the VCCM exchanges local clustering map information among data sites (Lines 2 and 3 of Algorithm 2). The VCCM then accommodates the proportionality of collaboration among data sites via collaborative map W' based on coefficient per site (σ_i) and respective exchanged clustering map W_i^{loc} (Lines 4 and 5 of Algorithm 2). In Line 6 of Algorithm 2, a new SOM is reconstructed from operation on the local dataset X_A in addition to the collaborative map W', to get

Algorithm 1 Self-organizing map (SOM)

Input: Dataset X, Map W

$x(t)$: observation vector

w_j : weight vector of activated node j

i : index of the best matching unit (BMU) with respect to $x(t)$

$\theta_{j,i}$: neighbourhood function describing the distance between BMU i and j

α_0 : value of learning-rate parameter α at the initiation of the SOM

r_0 : value of radius parameter r at the initiation of the SOM

Output: Adjusted map W

for $n = 1$ to λ **do** ▷ iterations
  $\alpha \leftarrow \alpha_0 \times \frac{\lambda - n}{\lambda}$ ▷ learning-rate parameter
  $r \leftarrow e^{n \times \frac{log(r_0)}{\lambda}}$ ▷ radius parameter
  for $t = 1$ to $|X|$ **do** ▷ observations
    $w_i \leftarrow \arg\min_j ||x(t) - w_j||, j \in W$ ▷ finding the BMU i
    for $j = 1$ to $|W|$ **do** ▷ map nodes
      $d^2_{j,i} \leftarrow ||r_j - r_i||^2$ ▷ r_j: position of node j, r_i: position of BMU i
      $\theta_{j,i} \leftarrow e^{-\frac{d^2_{j,i}}{r^2}}$ ▷ neighbourhood between i and j
      $w_j \leftarrow w_j + \alpha \times \theta_{j,i} \times (x(t) - w_j)$ ▷ adjusting node weight
    end for
  end for
end for

the collaborative map W_A^{col} by modifying equation (3.1) as following:

$$w'_j(n+1) = w'_j(n) + \alpha(n) \times \theta_{j,i}(n) \times (x(t) - w'_j(n)) \tag{3.1}$$

Where $w'_j(n)$ is collaborative mapping weight. The term $x(t) - w'_j(n)$ determines the impact of collaboration among datasets in equation (3.1). Larger the difference means that local data observation and collaborative map node have different patterns, and hence the collaboration captures new information. On the other hand, the small difference means that local data observation and collaborative map node follow similar patterns. To put it simply, the i^{th} and j^{th} observations of data sites

Algorithm 2 Vertical Collaborative Clustering Model (VCCM)

Input: Dataset X_A, Initialized map W_0

Output: Adjusted map W_A

1: $W_A^{loc} \leftarrow SOM(X_A, W_0)$ ▷ call SOM with initialized parameters

2: send W_A^{loc} to datasites $B, C, ...$

3: $W_B^{loc}, W_C^{loc}, ... \leftarrow$ get map from datasites $B, C, ...$

4: $\sigma_B, \sigma_C, ... \leftarrow$ determine coefficients such that $\sum_{i \in W} \sigma_i = 1$ ▷ $W = \{W_B^{loc}, W_C^{loc}, ...\}$

5: $W'_{BC} \leftarrow \sum_{i \in W} \sigma_i \times W_i^{loc}$

6: $W_A^{col} \leftarrow SOM(X_A, W'_{BC})$

7: $W_A \leftarrow$ Evaluate W_A^{loc} and W_A^{col} based on purity

A and B respectively, have smaller distance from common node on local and shared map. This means both observations are highly similar and belong to same cluster. Therefore, collaboration discloses similar patterns among different data sites via new SOM map constructed from operation on local dataset + information from outdoor site(s).

3.2.3 Evaluation

To evaluate the local map W_A^{loc} and the collaborative map W_A^{col} results, we calculate the purity of each map based on clustering of the map and label of observations (Line 7 of Algorithm 2). The purity measures the percentage of observations belonging to majority of class labels in the given cluster [10], as follows:

$$Purity = \frac{1}{|X|} \sum_{k \in C} \max_{l \in L} \mid c_k^l \mid \tag{3.2}$$

Where X refers to observations, C refers to clusters, L refers to labels and $|c_k^l|$ is the number of observations with label l in cluster k. Based on the comparison between the local and collaborative purity results, the data owner makes decision whether the collaboration unlocks the hidden pattern among different data sites.

3.3 Summary

In this chapter, the Vertical Collaborative Clustering Model ($VCCM$) is proposed to manage the collaboration among multiple data sites using Self-Organizing Map (SOM). It includes standard procedure and tuning of the exchanged information in specific proportionality to augment the learning process of the clustering via collaboration. Moreover, the VCCM unravels hidden information without compromising the data confidentiality. The aim of the model is to set an ideal environment for the collaboration process among multiple sites.

4. VERTICAL COLLABORATIVE CLUSTERING BASED ON BIT-PLANE SLICING

The prime reason for proposing vertical collaborative clustering using bit plane slicing, in addition to all benefits of using collaborative clustering, will enable a local data owner (e.g. business owners, government and private institutions, individuals, etc.) to find hidden structure in the process of implementing clustering techniques. This aim is logically explained since local data owner has the local capacity within the size of his/her data. However, enlarging the narratives that help to find hidden structure (in term of similarity among data sites without sharing data) by adding other information about clustering results from different sites, which happen to have same feature space with different observations, and can lead to better local clustering results by collaboration. In other words, clusters of one data site are associated with that of other site (s) with respect to certain criteria, identifying similarities where participating sites have same feature space. For example, various hospitals located in different regions want to investigate the structure of common disease among people of different populations, identifying latent causes without sharing actual data with other hospitals. Similarly, a chain of regional educational institutes wants to evaluate their students' performance belonging to different regions based on common latent constructs.

The proposed approach is termed as Vertical Collaborative Clustering based on Bit Plane Slicing (VCC-BPS), which performs collaboration among data sites by associating the local clustering outputs of one data site with that of other data site(s) to capture similarity. The working principle of this approach is to transform the input data to the code space using bit plane slicing approach, where the model fits the data with maximum similarity as shown in Fig 4.1. In other words, mapping of similar

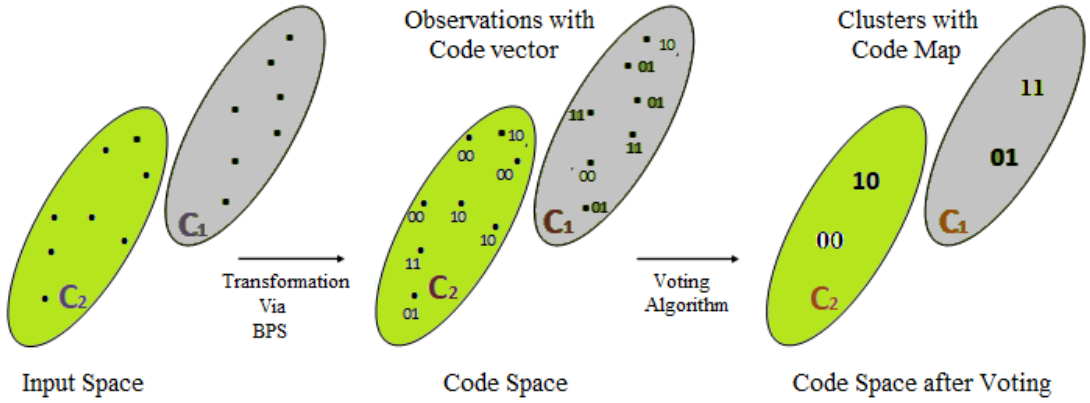


Figure 4.1 Transformation from Input space to Code space

data inputs to a particular code map is done by searching for adequate common bit plane among sites where the model fits data with maximum similarity. The novelty of this approach is to capture not only similarity in local behavior but it also qualifies for collaboration to apprehend similarity among different data sites concerning common code space. This learning demands an unbiased environment where data of the same nature at different sites performs vertical collaborative clustering based on the following assumptions:

- Number of features and their type be the same (Requirement of VCC).
- Type of clustering method must be the same at all sites to avoid the influence of one clustering method over the other [27].
- Binary form after the decimal point is considered for computation.
- Bit plane consists of a single bit per feature to generate code.
- Number of clusters be the same at all sites to deal with inconsistent output during collaboration [27].

The vertical collaborative clustering using bit plane slicing consists of two phases i.e. local and collaborative phase to manage collaboration among data sites. The block diagram of the proposed approach is shown in Figure 4.2.

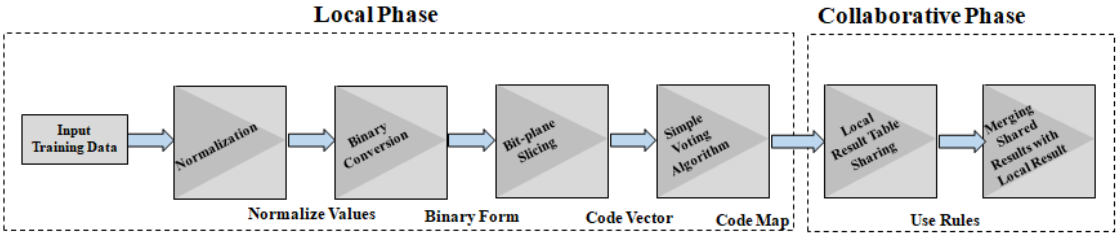


Figure 4.2 Proposed Approach Block Diagram

4.1 Local Phase

According to the local phase, the unlabeled dataset is first normalized to form common analytical plate form and then K-means clustering approach is used to cluster the given data. The normalized value of each observation for given feature space is converted into binary form. Then, BPS approach is used to transform binary input data into code space where code vectors are developed and associated with corresponding cluster for each observation. Finally, simple voting algorithm is applied over the training data in code space to find cluster for the code vector called code map, based on its most frequent occurrence for selected bit plane. It captures data behavior where similar observations are represented by particular code map associated with same cluster. In this phase, large volume of local data is compressed and represented as code map. Following are the different steps involved in the local phase:

1. Conversion to binary form and bit plane generation: In this step, dataset with features X_1 and X_2 is first normalized using following min-max normalization equation [40]:

$$N = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4.1}$$

Where x is observation value, min and max denotes minimum and maximum value. The normalized values lie between 0 and 1. The normalized valued are then clustered using the K-Means clustering approach as shown in column 7 (Cluster predicted) of Tablee 4.1. Then the next is to convert the normalized values into binary form as shown in column 4 (Binary form) of Table 4.1. Moreover, the observation value of each feature consists of 8 bits and represents

8 bit-plane. Bit plane is defined as a set of bits corresponding to the same bit position among all observations in a data array (feature vector) shown in column 5 (Bit Plane 4 and 3) of Table 4.1.

2. Code vector generation: We are inspired by the Bit Plane Slicing approach [24, 25] which compresses image with high resemblance to the original one by considering the most significant bits. This study exploits such a strong characteristic of BPS when used with vertical collaborative clustering, highlights the relative importance of specific bits whether they are most or least significant bits or combination of both least and most significant bits in data, capturing maximum similarity. The purpose of using BPS is to transform binary input obtained from step 1 of the local phase into code space for a particular bit plane. Then code vector (C_V) is formed for each observation by concatenating bit plane (feature vector or column vector) of one feature with that of other feature in the local data as shown in column 6 (Code Vector) of Table 4.1. The generated code vector column in Table 4.1 for feature X1 and X2 are with respect to 4 and 3 bit planes which belongs to 5^{th} and 4^{th} bit position in the byte of respective features. The size of the code vector depends on the number of features. For example, number of features are two in a dataset, then number of bits per code(b) are two, assuming 1 bit per feature. Moreover, number of code vectors are $C_v = 2^b = 2^2 = 4$ (00,01,10,11). Similarly, in case of four features, $C_v = 2^4 = 16$. A code vector is a compressed form of actual data for a given observation at a particular bit plane.
3. Voting Algorithm: This step aims to correctly map the code vectors obtained from step 2 of the local phase with the respective cluster predicted (column 7). It is found in step 2 that certain observations have the same code vector mapping to different clusters, thus forming the dual nature of the code vector. Notably, the same code vector must not belong to more than one cluster. Such dual nature is shown in column 6 (Code Vector) versus column 7 (Cluster Predicted) of Table 4.1. To solve such dual behavior, simple voting algorithm is used to group observations with same code vector, associated with most frequent cluster predicted (column 7) i.e. clusters in majority with respect to

Table 4.1 Code Map with Bit Plane Clusters

Obs.	Features Values		Normalized Values		Binary form		Bit Plan 4 and 3		Code Vector	Cluster Predicted	Code Map	Bit Plane Cluster
	X1	X2	X1	X2	X1	X2	X1	X2				
1	1.933	49	0.0951	0.1132	00011000	00011100	1	1	11	C-1	11	C-1
2	4.35	74	0.7857	0.5849	11001001	10010101	0	0	00	C-2	00	C-2
3	4.933	88	0.9523	0.8491	11110011	11011001	1	1	11	C-2	11	C-1
4	1.867	53	0.0763	0.1887	00010011	00110000	1	0	10	C-1	10	C-2
5	2.883	55	0.3666	0.2264	01011101	00111001	1	1	11	C-1	11	C-1
6	4.8	94	0.9143	0.9623	11101010	11110110	0	0	00	C-2	00	C-2
7	4.65	90	0.8714	0.8868	11011111	11100011	1	0	10	C-2	10	C-2
8	4	70	0.6857	0.5094	10101111	10000010	0	0	00	C-2	00	C-2
9	1.7	59	0.0286	0.3019	00000111	01001101	0	1	01	C-1	01	C-1
10	2.483	62	0.2523	0.3585	01000000	01011011	0	1	01	C-1	01	C-1
11	4.5	84	0.8286	0.7736	11010100	11000110	1	0	10	C-2	10	C-2
12	4.367	82	0.7906	0.7358	11001010	10111100	0	1	01	C-2	01	C-1
13	4.567	84	0.8477	0.7736	11011001	11000110	1	0	10	C-2	10	C-2
14	1.817	59	0.0620	0.3019	00001111	01001101	0	1	01	C-1	01	C-1
15	2.133	67	0.1523	0.4528	00100110	01110011	0	0	00	C-1	00	C-2

code vector. Such clusters developed after simple voting algorithm are called bit plane clusters and corresponding code vector is known as code map. In other words, the simple voting algorithm helps to train the model at particular bit plane where code vectors are associated with the most frequent clusters as shown in column 8 (Code Map) and 9 (Bit Plane Cluster) of Table 4.1 respectively. Such mapping via simple voting algorithm correctly groups similar observations to large extent with the least inaccuracy.

The local phase of the proposed approach is an iterative approach to look for those bit planes in the local dataset at which observations are grouped based on maximum similarity in code space as shown in Fig 4.1. Here, a search is made to determine a bit plane at which there is large contrast among observations to correctly group (cluster) them with least inaccuracy. For example, three observations with code vector 00 are clustered as cluster-2 (C-2) and one observation as cluster-1 (C-1) as shown in Fig 4.1, after transformation from input to code space using BPS. It is found in Table 4.1 that code vector 00 represents three observations $\{2, 6, 8\}$ belonging to C-2 and one observation

{15} belonging to C-1, thus reveals its dual behavior when bit plane is (4,3). Now in such scenario, using simple voting approach at particular bit plane (4,3) of feature X_1 and X_2 respectively, cluster C-2 dominates in match to C-1 at code vector 00, therefore all observations i.e. $\{2, 6, 8, 15\}$ in local dataset corresponding to code vector 00 are updated as cluster C-2 (cluster in the majority called bit plane cluster). Moreover, observation {15} whose actual cluster is C-1 for code vector 00, is misclassified by the proposed approach. The same analogy is applied to other code vectors as shown in Table 4.1. The detail description of the local phase in the context of the simple voting algorithm is explained in Figure 4.3. It is important to mention that the same approach can be applied to the datasets with more than two clusters.

In this phase, data is compressed to code map with most frequent clusters for selected bit planes. This forms the most important attribute of the proposed approach capturing not only similarity in local behavior but also qualifies for collaboration to apprehend similarity among different data sites for the same shared code space.

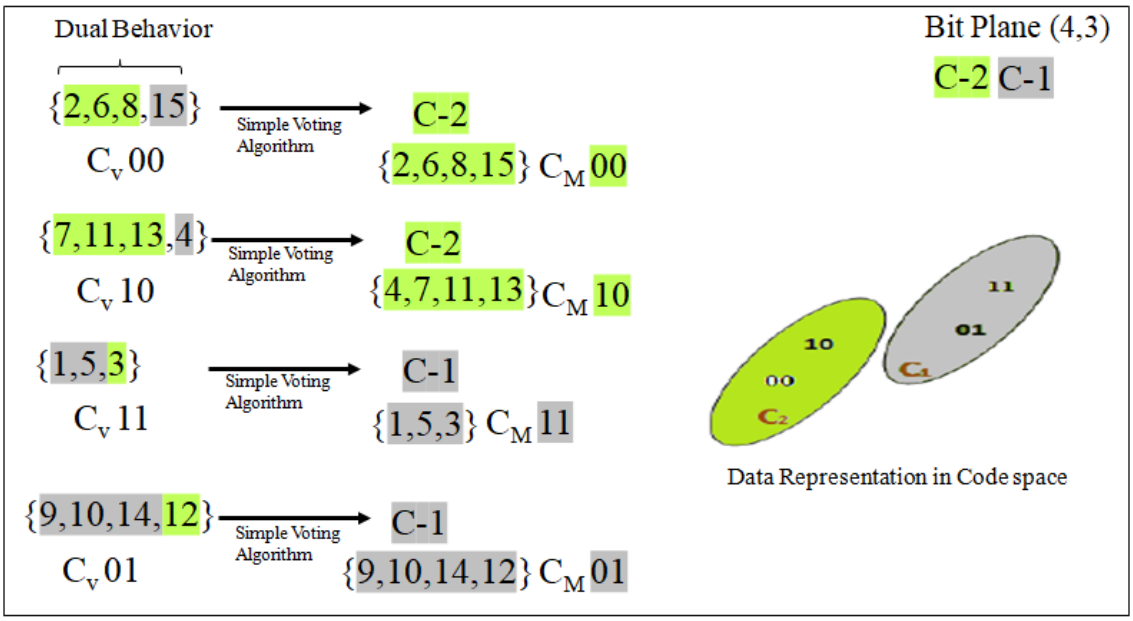


Figure 4.3 The Description of Voting Algorithm Step in the Local Phase

4.2 Collaborative Phase

This phase aims to fulfill the basic requirement of vertical collaborative clustering, which is to share the local findings with other sites, such that collaborative results obtained at each site are as if obtained from the pooled dataset [22]. This challenging task is addressed by the proposed approach using certain rules to identify similarities among participating sites. The rules are as follow:

1. The same code map must represent the same cluster among all participating sites at a particular bit plane. For example, code map 00 maps to the cluster C-2 at A with respect to particular bit plane, then the same code map must represent the same cluster for the same bit plane at B as shown in Figure. 4.4.
2. There must be a common bit plane during the collaboration phase.
3. Only those local bit plane combinations are considered for collaboration that give local purity greater than 70% as threshold level.
4. More than one code map may represent the same cluster locally. For example, at site A, code map 11 for an observation (x_1), maps to cluster C-1. Similarly, code map 01 for other observation (x_2), maps to cluster C-1 at A. It means both code maps fall in the same group (cluster) as C-1 as shown in Figure. 4.5.

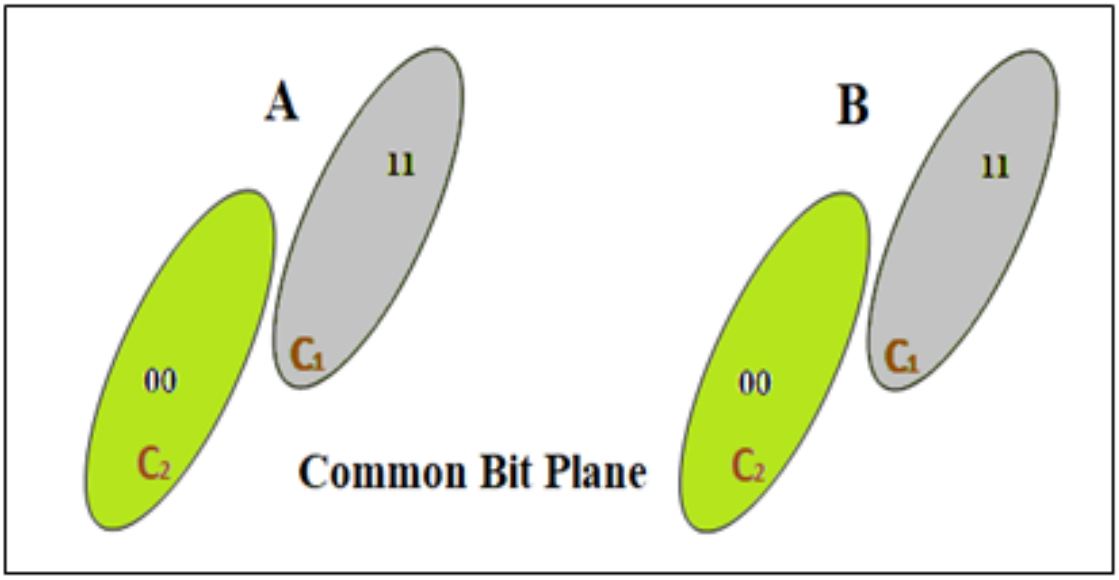


Figure 4.4 The Visual Description of Rule 1 in the Collaborative Phase

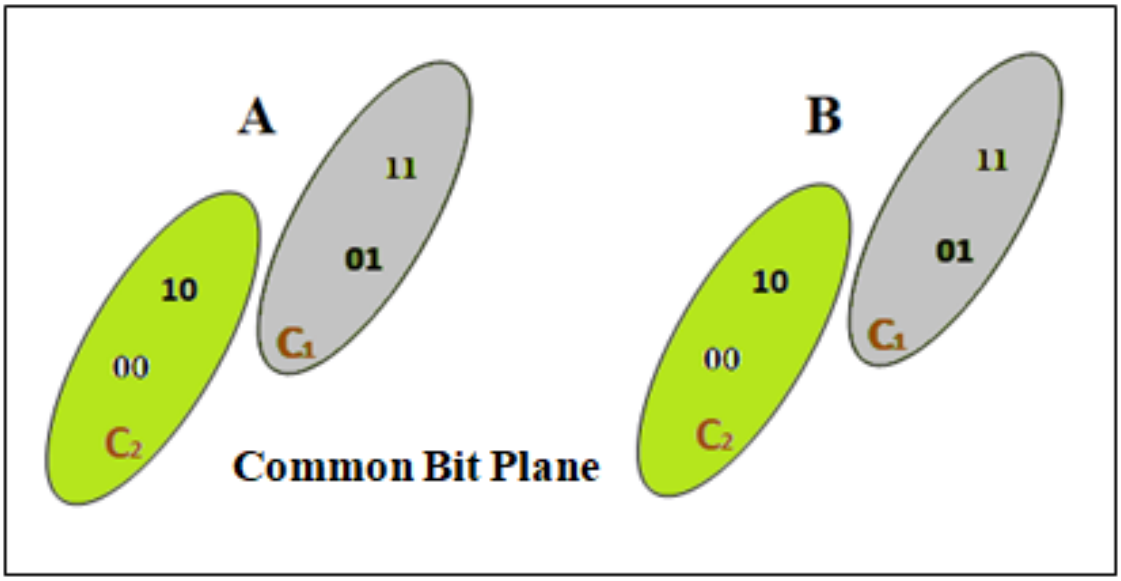


Figure 4.5 The Visual Description of Rule 4 in the Collaborative Phase

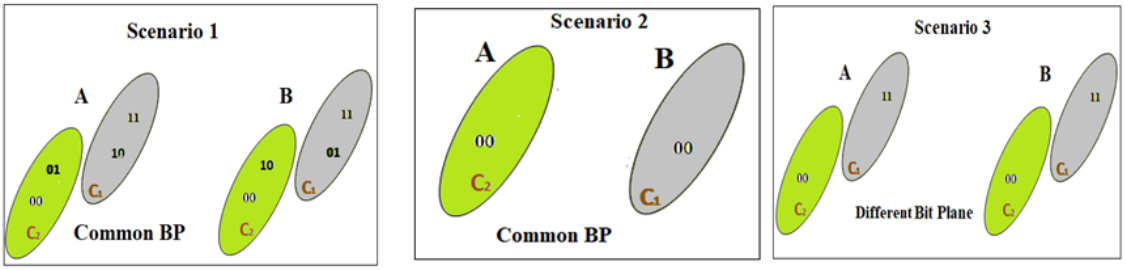


Figure 4.6 Description of Different Scenarios that do not qualify for Collaboration

It is noticeable that those bit plane combinations where local results do not obey the above-mentioned rules, do not qualify for collaboration due to mismatch in behavior among participating sites. For example, observations with code map 00 maps to the cluster C-2 at A and the same code map represents observations with different cluster (C-1) at B with respect to common bit plane. Then such bit plane combination in the light of first rule do not qualify for collaboration due to mismatch in behavior among the participating sites as shown in scenario 2 of Figure 4.6. Likewise, if observations with code maps 00 and 01 maps to the cluster C-2 at site A and the code maps 00 and 10 represent observations with cluster C-2 at site B with respect to common bit plane, then such bit plane combination is not considered for collaboration in light of the fourth rule as shown in scenario 1 of Figure 4.6. Here the code maps at A differ from that of B to capture similar behavior. Similarly, scenario 3 in Figure 4.6 shows that data behavior at site A and B is similar but bit plane combination at A

is different from that at B, do not qualify for collaboration under rule 2.
In collaborative phase, the participating sites share their local results called the local result table. It consists of code vector with predicted cluster and code map with bit plane cluster (code vector associated with the cluster in the majority) for different bit plane combinations. When one data site local behavior matches with that of another data site in light of the above mentioned rules, then the results are merged to compute the collaborative purity. The collaborative purity is measured as the mean of local purities by merging shared results with local results at common bit plane combination such that code map(s) of local site exactly matches with that of other data site(s). These rules ensure symmetry i.e. code map of one data site is exactly similar to another site with respect to common bit plane combinations. Such symmetry gives collaborative purity as the mean of all local purities with respect to common bit plane combinations among the participating sites where code map(s) at one site is similar to that at another site. Likewise, the collaborative DB index is measured by sharing local data clusters centroid and their variances among the participating sites under same rules.

4.3 Summary

This chapter presents the proposed approach, the vertical collaborative clustering using a bit plane slicing (VCC-BPS), as a simple and unique approach with improved accuracy. It manages collaboration among various data sites by transforming data from input space to code space. This results in capturing maximum similarity locally and collaboratively at a particular bit plane. The findings of this study highlight the significance of those particular bits which fit the model to correctly cluster the data locally and collaboratively. Thenceforth, the data owner appraises local and collaborative results to reach a better decision.

5. RESULTS EVALUATION

This chapter consists of the main results obtained from the proposed study. It includes the result evaluation details for the VCCM and VCC-BPS in the sections 5.1 and 5.2 respectively.

5.1 VCCM Experimental Details

This section introduces the datasets used, evaluation metrics and experimental results.

5.1.1 VCCM Datasets

To evaluate the Vertical Collaborative Clustering Model (VCCM), we applied our algorithm on four multivariate datasets with features of real values, which are Geyser [18], Iris, Breast Cancer Wisconsin (Cancer) and Waveform datasets [19] (Appendix A.1). To set distributed environment for the vertical collaborative clustering, datasets are randomly divided into two data sites with same features, named as A and B except Waveform dataset which is divided into four sites and named as A, B, C and D as mentioned in the Table 5.1.

Table 5.1 VCCM Dataset Description

Dataset	**# of Observations**	**# of Features**	**# of Classes**
Iris	75 × 2 sites	4	3
Geyser	136 × 2 sites	2	2
Cancer	284 × 2 sites	30	2
Waveform	1250 × 4 sites	40	3

5.1.2 Evaluation Metrics and Experimental Results

The phases of VCCM, which are local clustering, collaborative re-clustering and evaluation, are processed at each data sites. The local and collaborative purity index are measured using equation (3.2) to evaluate local and collaborative maps. In the local clustering, both data sites apply the SOM algorithm 1 on their local data with a given initialized map. Data is normalized prior to SOM algorithm. We use 5×5 map for Iris, Geyser and Cancer data sites and 10×10 for Waveform data sites. To ensure unbiased environment, data sites are initialized with parameters of the same values, including the map.

In Collaborative re-clustering phase, the local clustering map results are first exchanged between both sites. At each data site, the SOM algorithm 2 is then re-run over their local data with respect to the shared collaborative map. It is important to note that during the VCCM process (1) no real data is exchanged and therefore, data confidentiality is maintained, (2) the local site is fed through map representing the data of other sites, (3) chain mapping is done if there are more than two participating sites and (4) local data site discloses hidden pattern by exploiting map from other sites.

In Evaluation, each data site compares the local and collaborative map results based on purity measurement. This facilitates the data owner to make decision whether collaboration brings any new insight to uncover hidden information. It is important to mention that since SOM does not perform direct clustering and is coupled with K-means approach to extract clusters [10]. The experimental results of the VCCM are mentioned in the Table.5.2.

5.2 VCC-BPS Experimental Details

This section mentions the datasets used, reason of their selection, evaluation metrics and experimental results.

Table 5.2 Experimental Results of VCCM

Dataset	**Site**	**Local Purity**	**Collaborative Purity**	σ_A	σ_B	σ_C	σ_D	**Impact**
Iris	A_{iris}	80.00	86.67	-	1	n/a		↗
	B_{iris}	80.00	89.33	1	-			↗
Geyser	A_{geyser}	93.38	93.38	-	1			-
	B_{geyser}	96.32	96.32	1	-			-
Cancer	A_{cancer}	91.54	92.25	-	1			↗
	B_{cancer}	92.95	93.30	1	-			↗
Waveform	A_{wave}	59.28	59.28	-	1	-	-	-
			57.76	-	-	-	1	↘
			61.92	-	0.33	0.67	-	↗
			61.92	-	0.25	0.50	0.25	↗
	B_{wave}	55.68	55.68	1	-	-	-	-
	C_{wave}	57.92	57.92	-	-	-	1	-
	D_{wave}	56.08	56.08	-	-	1	-	-
			57.04	1	-	-	-	↗
			59.12	0.40	0.60	-	-	↗
			59.12	0.25	0.25	0.50	-	↗

5.2.1 VCC-BPS Datasets

To evaluate the proposed approach (VCC-BPS), three multivariate datasets i.e. Geyser [18], Skin segmentation (Skin) [19] and Iris [39] are used with features of real values (Appendix A.1). Skin dataset is tall dataset, consists of large number of observations with three features. To avoid computational complexity, these datasets of low dimensionality are chosen to explain and implement this novel approach simply and clearly. For example, a dataset with two features has 8-bit planes per feature and thus has (8^2) 64-bit plane combinations for given feature space. Similarly, in case of five features, search space for finding optimal solution consists of (8^5) 32,768-bit plane combinations. This shows how much the search space explodes with the increase in number of features as shown in Fig 5.1. Therefore, datasets with small feature space are selected to reduce the search space to find a suitable bit plane. Such selection does not mean that proposed approach has limited application to perform clustering and visualization simultaneously.
It is important to mention that as all features in the data are not necessary to give desirable solutions. Analysis with large features consumes large memory space and computational power. The domain experts' use feature reduction approaches to remove redundant features. This can lead to subset of features that preserves relevant structure of the data in particular domain, producing desirable solutions. Such process of data reduction be applied in pre-processing step prior to the proposed approach. The combination of both approaches will perform clustering, data reduction and visualization simultaneously to produce clustering results for given data.

To prepare the datasets for vertical collaborative clustering, having same features in a distributed environment (i.e. fulfilling the first requirement of VCC), the dataset is randomly divided into two data sites with same features, which are named as dataset A and B as shown in the Table 5.3. The datasets mentioned in the Table 5.3 are subjected to the K-means approach for clustering the observations into two groups (i.e. C-1 and C-2) in case of Geyser and Skin data, while three groups in case of Iris

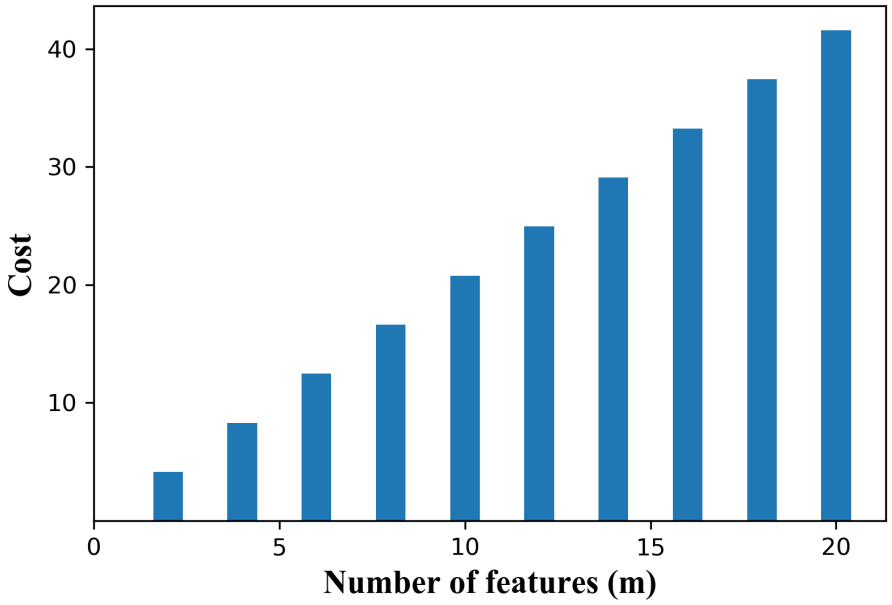


Figure 5.1 Computational Cost of Measuring the Purity

data before BPS. The local and collaborative phases of the proposed approach are processed at each data site and then evaluated by purity and Davies-Bouldin index.

Table 5.3 VCC-BPS Dataset Description

Dataset	# of Observations	# of Features	# of Clusters
Geyser [18]	136 × 2 sites	2	2
Skin [19]	122528 × 2 sites	3	2
Iris [39]	75 × 2 sites	4	3

5.2.2 Evaluation Metrics

To evaluate the local results, the local purity is calculated for given observations based on their respective cluster predicted and bit plane cluster at a particular bit plane using equation (5.1).

$$P^i = \frac{1}{|n|} \sum_{k \in C} \max_{l \in L} \mid c_k^l \mid \tag{5.1}$$

Where P^i denotes local purity of the i^{th} data site, n and C refers to number of the observations and clusters respectively. L denotes the labels (bit plane clusters) and $|c_k^l|$ describes the number of observations with label l in cluster k. The purity is the average proportion of the majority label in each cluster [22, 28].

The equation (5.2) is used to compute collaborative purity under certain rules (refer to section 4.2) to merge respective results. The proposed collaborative purity equation (5.2) is used to associate the clusters of one data site with that of other data site such that code map of one data site exactly matches with that of other site at common bit plane.

$$\overline{P} = \underset{C_M^i \sim C_M^j}{Avg} \; (\mathrm{P}^i, \mathrm{P}^j)^{BP} \tag{5.2}$$

Where $\overline{P}$ is collaborative purity, P^i and P^j denote local purities of the i^{th} and j^{th} data sites with respect to common bit plane combinations BP such that code map(s) at i^{th} data site (C_M^i) must be similar to that at j^{th} site. In addition to local and collaborative purity, the global purity is also used to evaluate the accuracy of collaborative outcome. For measuring the global purity, datasets are pooled and then purity is measured over the combined dataset clustering final map. The global purity with the local and collaborative purity is visually explained in Fig 5.2.

In addition to the purity as external index, Davies-Bouldin index (DB) is used as internal quality index to assess the compactness and separation of the local clustering results [10] using equation (5.3).

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \frac{S_i + S_j}{d(i,j)} \tag{5.3}$$

Where S_i and S_j are local dispersion of i^{th} and j^{th} clusters, $d(i,j)$ is centroid to centroid (inter-cluster) distance for K number of given clusters of local dataset and DB refers to local Davies-Bouldin index. Local dispersion S_i and their corresponding inter-cluster distance, i.e. d(i,j) can be computed using equation (5.4) and (5.5).

$$S_i = \frac{1}{T_i} \sum_{l=1}^{T_i} \| \; x_l - \mu_i \; \|^2 \tag{5.4}$$

$$d(i,j) = \| \; \mu_j - \mu_i \; \|^2 \tag{5.5}$$

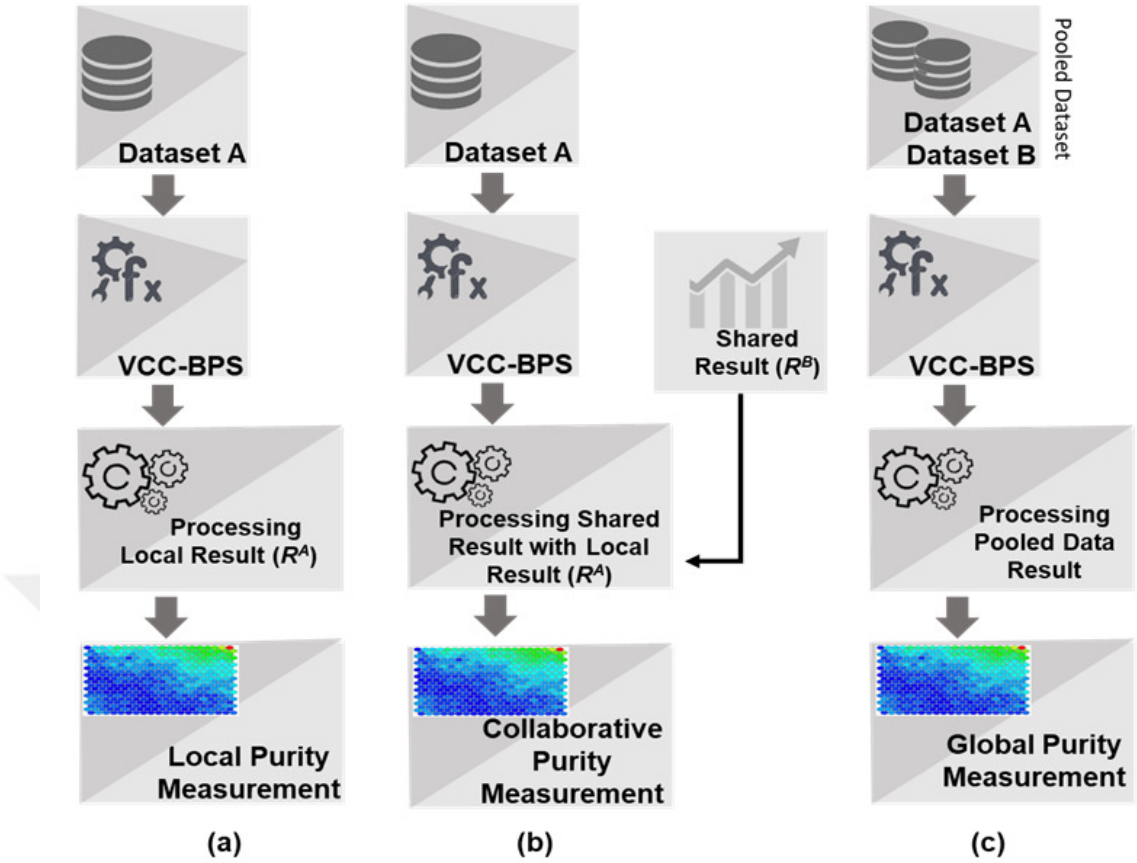


Figure 5.2 The Local, Collaborative and Global purity measurements: (a) The local purity is computed at data site A without collaboration.(b) The collaborative purity is computed at A with respect to the result shared from site B, enhancing the learning while data confidentiality is maintained.(c) The global purity is measured with respect to the pooled dataset where the data confidentiality is compromised. This is done to check whether the collaborative results similarity are close to global similarity result.

Where x_l is an input of observation of the given dataset, associated with i^{th} cluster of size T_i, having centroid μ_i. Moreover, μ_i and μ_j refers to the centroid of the i^{th} and j^{th} cluster of same dataset (local dataset). The two clusters are considered similar, if they have large dispersion relative to their distance. Lower value of local DB indicates a cluster of better quality. Equation (5.6) is used to associate the clusters of dataset A with that of B to measure collaborative DB index:

$$\overline{DB} = \frac{1}{K} \sum_{i,j=1}^{K} \max_{i,j \in K} \frac{S_i^A + S_j^B}{D(i^A, j^B)} \tag{5.6}$$

Where $\overline{DB}$ is collaborative DB index, $D(i^A, j^B)$ is the centroid to centroid distance between i^{th} and j^{th} cluster of dataset A and B respectively. Likewise, S_i^A and S_j^B are

dispersion of i^{th} and j^{th} clusters of dataset A and B respectively. It is noticeable that low local DB value means observations within clusters are compact and clusters are well separated, whereas high collaborative DB value for dataset A and B means both have similarity in behavior and vice versa. In other words, equation (5.3) reveals that local DB value is small when inter-cluster distance ($d(i,j)$) is large. Likewise, equation (5.6) shows that collaborative DB value is large when centroid to centroid distance (D(i^A, j^B)) between clusters of A and B is small i.e. i^{th} cluster of A is similar to j^{th} cluster of B.

5.2.3 Experimental Results-VCC-BPS

In this section, the local and collaborative results are evaluated by purity and DB index. It also presents the comparison of proposed approach (VCC-BPS) with existing approaches VCC-SOM [20] and VCC-GTM [10].

In local phase of VCC-BPS, the normalized training datasets (A and B) of Geyser are clustered using K-means approach. Then normalized values are converted into binary form and then subjected to BPS generating code vector associated with clusters, followed by simple voting algorithm to find code map with clusters in the majority at particular bit planes as shown in Table 5.4. The Table 5.4 consists of Geyser Data Local Result Table for site A and B, which explains that dataset A has 46 and 68 observations in majority, belonging to cluster C-2 and C-1, respectively, using simple voting algorithm with respect to bit plane (6,7). Similarly, the dataset B has 38 and 77 observations in majority, belonging to cluster C-2 and C-1, respectively at bit plane (6,7). The code maps 00 and 10 participate to associate observations with cluster C-2 and C-1 at A and B, respectively as shown in local and collaborative code map diagram column of Table 5.8. Code maps 01 and 11 do not participate in capturing similarity locally at A and B when BP (6,7). The local purity is measured using equation (5.1) for Geyser data at A and B, consisting of 120 training observations each as follows: Local purity at A = P^A =(46+68)/120 = 0.95 and P^B = (38+77)/120 = 0.958 such that code maps are 00 and 10 at both sites with BP (6,7). The detail about other bit plane combinations for Geyser data

are shown in the discussion column of Table 5.4. The detail about Iris data having three classes, are mentioned in Table 5.5 with respect to only single bit plane combination (5,7,1,2) to avoid large computational local result table. The same analogy is applied to Skin datasets to generate local result table.

In collaborative phase of VCC-BPS, data site A and B of Geyser share their local result table (R^A and R^B) and then collaborative purity is measured with respect to common bit plane as shown in Table 5.8. The collaborative purity for Geyser data at site A and B is measured using equation (5.2) as follows: $\overline{P} = \underset{C_M^A \sim C_M^B}{Avg}(P^A, P^B)^{BP} = \underset{00,10}{Avg}((46+68)/120, (38+77)/120)^{(6,7)} = 0.954$. The Table 5.8 consists of the local, collaborative, global purity and DB indexes at site A and B with respect to particular bit plane for Geyser datasets. The same analogy is applied to Skin and Iris data consisting of 2 and 3 clusters with detail mentioned in discussion column of Tables 5.9 and 5.10 respectively.

The Davies-Bouldin index is used to evaluate the results of our proposed approach for Geyser, Skin and Iris data locally and collaboratively at A and B. The local and collaborative DB index values are computed using equations (5.3) and equation (5.6) respectively. The details about computing local and collaborative DB for Iris data are mentioned in Table 5.6 and Table 5.7. The same analogy is applied to Geyser and Skin data to measure their respective local and collaborative DB values as mentioned in Table 5.8 and Table 5.9. To check the generalization of the proposed approach, test data is passed through the model and accuracy is determined for different bit planes as shown in Table 5.8, 5.9 and 5.10.

The existing approaches which are VCC using SOM [20] and GTM [10] are implemented and tested over Geyser, Skin and Iris datasets for comparison with the proposed approach as shown in Table 6.1. The results of VCC-SOM and VCC-GTM approaches are topographic maps, representing compressed form of original dataset for given number of clusters mentioned in Table 5.3. Since SOM and GTM do not perform direct clustering, but are coupled with K-means approach and EM algorithm respectively over final map to extract clusters [10]. Then purity and Davies-Bouldin index are measured over final map using equations (5.1) and (5.3) respectively [10, 29, 34]. The size of the map is 5×5 for existing approaches to

Table 5.4 Geyser Dataset Local Result Table for A and B (R^A, R^B) using Purity Index

Bit Plane (BP)	Dataset (DS)	Cluster	Code Vector				Code map				Discussion
			00	01	10	11	00	01	10	11	
(6,7)	DS-A	C-1	4	0	68	0	0	0	68	0	With BP (6,7) of features 1 and 2 respectively, the majority of observations at A and B belong to cluster C-2 are 46 and 38, are represented by code map 00, whereas misclassified are 4 and 3 respectively. Likewise, 68 and 77 observations at A and B belong to cluster C-1,are represented by code map 10, whereas misclassified are 2 each respectively.
		C-2	46	0	2	0	46	0	0	0	
	DS-B	C-1	3	0	77	0	0	0	77	0	
		C-2	38	0	2	0	38	0	0	0	
(5,7)	DS-A	C-1	24	0	48	0	0	0	48	0	With BP (5,7), code map 00 receives 47 and 37 votes of majority to represent cluster C-2 at A and B, whereas misclassified are 24 and 31 respectively. Similarly, code map 10 receives 48 and 49 votes of majority to represent cluster C-1 at A and B, whereas 1 and 3 are misclassified respectively.
		C-2	47	0	1	0	47	0	0	0	
	DS-B	C-1	31	0	49	0	0	0	49	0	
		C2	37	0	3	0	37	0	0	0	
(5,6)	DS-A	C-1	1	33	1	41	0	33	0	41	With BP (5,6), code map 00 receives 30 and 27 votes of majority to represent cluster C-2 at A and B, whereas misclassified are 1 and 1 respectively. Likewise, code map 01 receives 33 and 35 votes of majority to represent cluster C-1 at A and, B respectively. Similarly, code map 10 receives 14 and 13 votes of majority to represent cluster C-2 at A and B respectively, whereas 1 and 2 are misclassified. Likewise, code map 11 has 41 and 42 votes of majority to represent C-1 at A and B respectively.
		C-2	30	0	14	0	30	0	14	0	
	DS-B	C-1	1	35	2	42	0	35	0	42	
		C-2	27	0	13	0	27	0	13	0	
(4,6)	DS-A	C-1	1	32	2	37	0	32	0	37	With BP (4,6), code map 00 receives 31 and 29 votes of majority to represent cluster C-2 at A and B respectively, whereas 1 is misclassified at A. Likewise, code map 01 has 32 and 41 votes of majority representing C-1 at A and B respectively. Similarly, code map 10 receives 17 and 11 votes to represent C-2 at A and B, whereas 2 and 3 are misclassified respectively. Likewise, code map 11 has 37 and 36 votes of majority representing C-1 at A and B respectively.
		C-2	31	0	17	0	31	0	17	0	
	DS-B	C-1	0	41	3	36	0	41	0	36	
		C-2	29	0	11	0	29	0	11	0	
(3,6)	DS-A	C-1	3	34	0	35	0	34	0	35	With BP (3,6), code map 00 receives 23 and 22 votes of majority to represent cluster C-2 at A and B respectively, whereas 3 and 1 are misclassified. Likewise, code map 01 has 34 and 35 votes of majority representing C-1 at A and B respectively. Similarly, code map 10 receives 25 and 18 votes to represent cluster C-2 at A and B respectively, whereas 2 are misclassified at B. Likewise, code map 11 has 35 and 42 votes of majority representing C-1 at A and B respectively.
		C-2	23	0	25	0	23	0	25	0	
	DS-B	C-1	1	35	2	42	0	35	0	42	
		C-2	22	0	18	0	22	0	18	0	

Table 5.5 Iris Dataset Local Result Table for A and B (R^A, R^B) using Purity Index

Bit Plan	Data Set	Cluster	Code Vector								Code Map								Discussion
			0000	0001	0010	0011	1000	1001	1010	1011	0000	0001	0010	0011	1000	1001	1010	1011	
(5,7,1,2)	A	C-1	13	0	0	0	7	0	0	0	13	0	0	0	7	0	0	0	At BP (5,7,1,2), code map 0000 receives 13 and 9 votes of majority to represent cluster C-1 at A and B respectively, whereas 1 observation is misclassified at A. Likewise, code map 0001 has 2 and 4 votes of majority representing C-3, whereas misclassified are 1 each at A and B respectively. Similarly, code map 0010 has 8 and 5 votes in majority to represent cluster C-2, whereas misclassified observations are 4 and 2 at A and B respectively. Code map 0011 has 8 votes of majority each to represent cluster C-3, whereas misclassified observations are 3 and 2 at A and B respectively. Code map 1000 has 7 and 12 votes of majority to represent cluster C-1 at A and B respectively. Code map 1001 has 3 and 7 votes of majority to represent cluster C-2 at A and B respectively, whereas misclassified observations are 4 at B. Code Map 1010 has 7 and 4 majority of votes to represent cluster C-2, whereas 3 and 2 are misclassified observations at A and B respectively. Code map 1011 has 6 and 4 votes of majority to represent cluster C-3 at A and B respectively, whereas 2 are misclassified observations at B. The code maps not mentioned, are not involved in measuring similarity locally and collaboratively.
		C-2	1	1	8	3	0	3	7	0	0	0	8	0	0	3	7	0	
		C-3	0	2	4	8	0	0	3	6	0	2	0	8	0	0	0	6	
	B	C-1	9	0	0	0	12	0	0	0	9	0	0	0	12	0	0	0	
		C-2	0	1	5	2	0	7	4	2	0	0	5	0	0	7	4	0	
		C-3	0	4	2	8	0	4	2	4	0	4	0	8	0	0	0	4	

Table 5.6 Iris Dataset Local Result Table for A and B (R^A, R^B) using Local Davies Bouldin Index

Data site	Bit Plane	Cluster	Cluster Centroid				Local Dispersion (S_i)	Local cluster to cluster distance d(i,j)			$\frac{S_i+S_j}{d(i,j)}$			Local DB
								Cluster						
			X1	X2	X3	X4		1	2	3	1	2	3	
A	(7,3,1,2)	1	0.276	0.61	0.142	0.123	0.275	0.000	0.716	1.087	0	0.690	0.446	0.839
		2	0.46	0.304	0.599	0.543	0.219	0.716	0.000	0.470	0.690	0	0.913	
		3	0.696	0.460	0.820	0.846	0.210	1.087	0.470	0.000	0.446	0.913	0	
B		1	0.198	0.567	0.110	0.096	0.242	0	0.757	1.151	0	0.637	0.410	0.877
		2	0.419	0.284	0.584	0.564	0.240	0.757	0	0.471	0.637	0	0.997	
		3	0.733	0.385	0.815	0.809	0.230	1.151	0.471	0	0.410	0.997	0	
A	(7,4,5,2)	1	0.18	0.5960	0.075	0.051	0.136	0	0.674	1.147	0	0.673	0.324	0.956
		2	0.4370	0.3750	0.496	0.454	0.318	0.674	0	0.505	0.673	0	1.097	
		3	0.667	0.438	0.788	0.79	0.236	1.147	0.505	0	0.324	1.097	0	
B		1	0.144	0.56	0.073	0.062	0.176	0	0.681	1.140	0	0.678	0.410	0.982
		2	0.364	0.322	0.5	0.482	0.286	0.681	0	0.509	0.678	0	1.133	
		3	0.697	0.378	0.768	0.753	0.291	1.140	0.509	0	0.410	1.133	0	
A	(7,6,1,2)	1	0.2220	0.6330	0.072	0.059	0.163	0	0.788	1.140	0	0.464	0.350	0.837
		2	0.4730	0.3080	0.576	0.505	0.203	0.788	0	0.429	0.464	0	1.023	
		3	0.645	0.432	0.791	0.81	0.236	1.140	0.429	0	0.350	1.023	0	
B		1	0.155	0.563	0.074	0.063	0.173	0	0.770	1.161	0	0.565	0.351	0.900
		2	0.398	0.296	0.556	0.543	0.262	0.770	0	0.466	0.565	0	1.068	
		3	0.723	0.395	0.785	0.764	0.235	1.161	0.466	0	0.351	1.068	0	
A	(6,2,1,2)	1	0.2220	0.6330	0.072	0.059	0.163	0	0.813	1.134	0	0.466	0.345	0.838
		2	0.4360	0.2920	0.585	0.545	0.216	0.813	0	0.434	0.466	0	1.024	
		3	0.69	0.454	0.79	0.78	0.228	1.134	0.434	0	0.345	1.024	0	
B		1	0.161	0.579	0.074	0.063	0.117	0	0.784	1.122	0	0.418	0.306	0.837
		2	0.393	0.275	0.565	0.54	0.211	0.784	0	0.417	0.418	0	1.047	
		3	0.679	0.37	0.761	0.752	0.226	1.122	0.417	0	0.306	1.047	0	
A	(5,7,1,2)	1	0.2130	0.6230	0.074	0.06	0.156	0	0.775	1.114	0	0.465	0.350	0.818
		2	0.4200	0.3040	0.561	0.527	0.204	0.775	0	0.441	0.465	0	0.994	
		3	0.687	0.449	0.777	0.762	0.234	1.114	0.441	0	0.350	0.994	0	
B		1	0.161	0.579	0.074	0.063	0.184	0	0.764	1.105	0	0.509	0.399	0.841
		2	0.352	0.245	0.543	0.528	0.205	0.764	0	0.459	0.509	0	1.006	
		3	0.675	0.38	0.754	0.737	0.257	1.105	0.459	0	0.399	1.006	0	

Table 5.7 Collaborative Davies Bouldin Measurement for Iris Data

Bit Plane	Cluster		S_i^A	S_j^B	D(i^A,j^B)			$\frac{S_i^A+S_j^B}{D(i^A,j^B)}$			$\overline{DB}$
	A	B			Cluster						
	i	j			1	2	3	1	2	3	
(7,3,1,2)	1	1	0.275	0.242	0.098	0.719	1.088	5.253	0.717	0.464	6.272
	2	2	0.219	0.240	0.759	0.052	0.446	0.607	8.758	1.008	
	3	3	0.210	0.230	1.152	0.493	0.092	0.393	0.913	4.804	
(7,4,5,2)	1	1	0.136	0.176	0.052	0.689	1.135	5.986	0.612	0.376	5.231
	2	2	0.318	0.286	0.673	0.198	0.481	0.734	3.051	1.267	
	3	3	0.236	0.291	1.153	0.532	0.079	0.357	0.981	6.656	
(7,6,1,2)	1	1	0.163	0.173	0.097	0.783	1.146	3.464	0.543	0.347	4.533
	2	2	0.203	0.262	0.783	0.087	0.425	0.480	5.329	1.030	
	3	3	0.236	0.235	1.153	0.454	0.098	0.355	1.097	4.806	
(6,2,1,2)	1	1	0.163	0.117	0.082	0.795	0.795	3.432	0.471	0.489	5.568
	2	2	0.216	0.211	0.807	0.051	0.373	0.413	8.434	1.186	
	3	3	0.228	0.226	1.150	0.478	0.094	0.300	0.918	4.839	
(5,7,1,2)	1	1	0.156	0.184	0.068	0.775	1.092	4.987	0.466	0.378	5.250
	2	2	0.204	0.205	0.771	0.092	0.390	0.503	4.455	1.182	
	3	3	0.234	0.257	1.130	0.513	0.078	0.370	0.855	6.308	

Table 5.8 Geyser Purity Measurement and Code Map Description

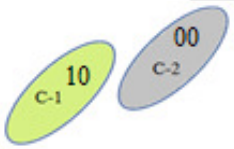
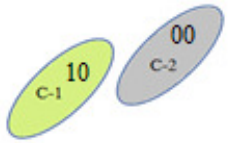
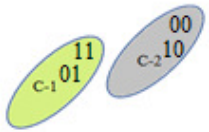
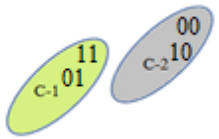
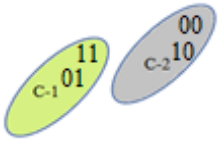
Local Result	Bit Plane	Local Purity	Collaborative Purity	Global Purity (GP)	Local DB (DB)	$(\overline{DB})$	Test Data Accuracy	Local & Collaborative Code Map Diagram	Discussion
R^A	(6,7)	0.95	0.9542	0.9583	0.411	6.93	0.9062		Code map 00 and 10 represents all observations which are clustered as C-2 and C-1 respectively using voting algo. at bit plane (6,7) . Code map 00 and 10 captures similarity at A and B by measuring local purity 95% and 95.8% respectively at this BP. The collaborative purity based on local result table shared is 95.42% for common bit plane (6,7) between A and B. Local purity at A is less than collaborative purity mean local learning at A has expanded and reveal significance of collaboration in capturing hidden information. On other side, local learning of B dropped by collaboration. GP is 95.83%. Code map 11 and 01 do not participate in capturing behavior. $\overline{DB}$ after collaboration is 6.93, whereas local DBindex for A and B are 0.411 and 0.355 respectively. Test data accuracy is 90.62% at this BP.
R^B		0.958			0.355				
R^A	(5,7)	0.7917	0.7542	0.8133	0.728	8.150	0.7187		Code map 00 and 10 represents all observations which are clustered as C-2 and C-1 respectively using voting algo. at bit plane (5,7) . Code map 00 and 10 captures similarity at A and B by measuring local purity 79.17% and 71.7% respectively at this BP. The collaborative purity based on local result table shared is 75.42% for common bit plane (5,7) between A and B. Local purity at B is less than collaborative purity, means local learning at B has expanded and reveal significance of collaboration in capturing hidden information. On other side, local learning of A dropped by collaboration. GP is 81.33 %. $\overline{DB}$ after collaboration is 8.15, whereas local DB index for A and B are 0.728 and 0.955 respectively. Test data accuracy is 71.87 %.
R^B		0.717			0.955				
R^A	(5,6)	0.9833	0.9792	0.9875	0.389	7.144	0.9062		Code maps 00,10 and 11,10 represent all observations which are clustered as C-2 and C-1 respectively using voting algorithm at bit plane (5,6) . These code maps capture similarity at A and B by measuring local purity 98.33% and 97.5% respectively at this BP. The collaborative purity based on local result table shared is 97.92% for common bit plane (5,6) between A and B. Local purity at B is less than collaborative purity, means local learning at A has expanded and reveal significance of collaboration in capturing hidden information. On other side, local learning of A dropped by collaboration. GP is 98.75 %. $\overline{DB}$ after collaboration is 7.144, whereas local DB for A and B are 0.389 and 0.423 respectively. Test data accuracy is 90.62 %.
R^B		0.975			0.423				
R^A	(4,6)	0.975	0.975	0.9875	0.452	7.715	0.9375		Code maps 00,10 and 11,01 represent observations which are clustered as C-2 and C-1 respectively using voting algorithm at bit plane (4,6) . These code maps capture similarity at A and B by measuring local purity 97.5% and 97.5% respectively at this BP. The collaborative purity based on local result table shared is 97.5% for common bit plane (4,6) between A and B. Local purity at B is less than collaborative purity, means local learning at A has expanded and reveal significance of collaboration in capturing hidden information. On other side, local learning of A has dropped by collaboration. GP is 98.75 %. $\overline{DB}$ after collaboration is 7.715, whereas local DB for A and B are 0.452 and 0.449 respectively. Test data accuracy is 93.75 %.
R^B		0.975			0.449				
R^A	(3,6)	0.975	0.975	0.9875	0.439	7.189	0.875		Code maps 00,10 and 11,01 represent observations which are clustered as C-2 and C-1 respectively using voting algorithm at bit plane (3,6) . These code maps capture similarity at A and B by measuring local purity 97.5% and 97.5% respectively at this BP. The collaborative purity based on local result table shared is 97.5% for common bit plane (3,6) between A and B. This shows that both sites have similar information when bit plane is (3,6). GP is 98.75%. $\overline{DB}$ after collaboration is 7.189, whereas local DB for A and B are 0.439 and 0.442 respectively. Test data accuracy is 87.5%.
R^B		0.975			0.442				

Table 5.9 Skin Purity Measurement and Code Map Description

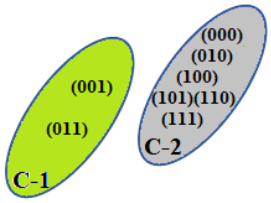
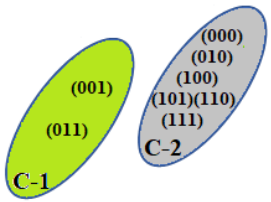
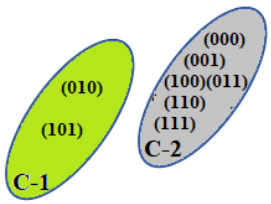
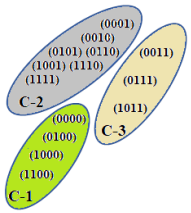
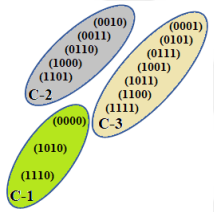
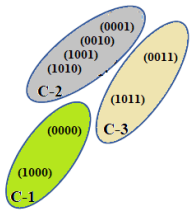
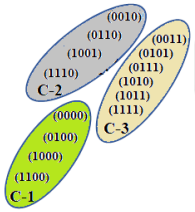
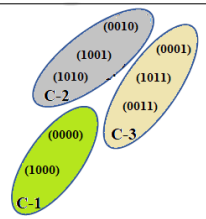
Local Result	Bit Plane	Local Purity	Collaborative Purity	Global Purity (GP)	Local DB (DB)	$(\overline{DB})$	Test Data Accuracy	Local & Collaborative Code Map Diagram	Discussion
R^A	(7,7,7)	0.7552	0.7383	0.7633	0.815	0.865	0.6277		Code maps (001) and (011) denote all observations at A and B clustered as C-1 at bit plane (7,7,7). Same analogy applies to other code maps clustered as C-2. Local purity at B is less than collaborative purity , reveals significance of collaboration to expands its learning via sharing. On other side, learning at A dropped by collaboration. GP is 76.33 %. $\overline{DB}$ after collaboration is 0.865, whereas, local DB for A and B are 0.815 and 0.914 respectively. Test data accuracy is 62.77 %.
R^B		0.7214			0.914				
R^A	(7,6,7)	0.7423	0.7570	0.7810	0.953	0.916	0.6516		Code maps (001) and (011) denote all observations at A and B clustered as C-1 at bit plane (7,6,7). Same analogy applies to other code maps clustered as C-2. Local purity at A is less than collaborative purity , reveals significance of collaboration to expand its learning via sharing. On other side, learning at B dropped by collaboration. GP is 78.10%. $\overline{DB}$ after collaboration is 0.916, whereas local DB for A and B are 0.953 and 0.863 respectively. Test data accuracy is 65.16 %.
R^B		0.7716			0.863				
R^A	(5,5,7)	0.7233	0.7274	0.7525	0.923	0.997	0.5956		Code maps (010) and (101) denote all observations at A and B clustered as C-1 at bit plane (5,5,7). Same analogy applies to other code maps clustered as C-2. Local purity at A is slightly less than collaborative purity, reveals minor expansion in learning via sharing. On other side, learning at B has slightly dropped by collaboration. This shows that both sites have almost similar information at bit plane is (5,5,7). GP is 75.25 %. $\overline{DB}$ after collaboration is 0.997, whereas local DB for A and B are 0.923 and 0.861 respectively. Test data accuracy is 59.56 %.
R^B		0.7315			0.861				

Table 5.10 Iris Purity Measurement and Code Map Description

Local Result	Bit Plane	Local Purity	Collaborative Purity	Global Purity (GP)	Local DB (DB)	($\overline{DB}$)	Test Data Accuracy	Local & Collaborative Code Map Diagram	Discussion
R^A	(7,3,1,2)	0.8636	0.8561	0.8867	0.839	6.272	0.8333		Code maps (0011),(0111) and (1011) denote all observations at A and B clustered as C-1 at bit plane (7,3,1,2). Same analogy applies to other code maps clustered as C-2 and C-3. Local purity at B is less than collaborative purity , reveals significance of collaboration to expands its learning via sharing. On other side, learning at A dropped by collaboration. GP is 88.67 %. $\overline{DB}$ after collaboration is 6.272, whereas local DB for A and B are 0.839 and 0.877 respectively. Test data accuracy is 83.33 %. Code maps (1010) and(1101) do not participate in capturing similarity.
R^B		0.8485			0.877				
R^A	(7,4,5,2)	0.8182	0.8030	0.8333	0.956	5.231	0.7222		Code maps (0000),(1010) and (1110) represents all obs. at A and B clustered as C-1 at bit plane (7,4,5,2). Same analogy applies to other code maps clustered as C-2 and C-3. Local purity at B is less than collaborative purity , reveals significance of collaboration to expands its learning via sharing. On other side, learning at A has fallen by collaboration. GP is 83.33 %.$\overline{DB}$ is 5.231, whereas, local DB at A and B is 0.956 and 0.982 respectively. Test data accuracy is 72.22 %. Code map (0100) does not participate in capturing similarity at A and B.
R^B		0.7879			0.982				
R^A	(7,6,1,2)	0.8182	0.8106	0.84	0.837	4.533	0.7778		Code maps (0000) and (1000) denote all observations at A and B clustered as C-1 at bit plane (7,6,1,2). Same analogy applies to other code maps clustered as C-2 & C-3. Local purity at B is less than collaborative purity which reveals significance of collaboration in expanding its learning. On other side, local learning of A dropped by collaboration. GP is 84 %. $\overline{DB}$ after collaboration is 4.533, whereas, local DB at A and B is 0.837 and 0.900 respectively. Test accuracy is 77.78 %.
R^B		0.803			0.900				
R^A	(6,2,1,2)	0.8636	0.8636	0.8933	0.838	5.568	0.8889		Code maps (0010),(0110),(1001) and (1110) denotes all obs. at A and B clustered as C-2 at bit plane (6,2,1,2). Same analogy applies to other code maps clustered as C-1 and C-3. Local and collaborative purity at A and B are same which means data at A and B are highly similar. GP is 89.33%. $\overline{DB}$ is 5.568, whereas, local DB at A and B are 0.838 and 0.837 respectively. Test accuracy is 88.89%. Code map(0001) and (1101) do not participate in capturing similarity at A and B.
R^B		0.8636			0.837				
R^A	(5,7,1,2)	0.8182	0.8106	0.86	0.818	5.25	0.8333		Code maps (0100),(0101),(0110),(0111),(1100),(1101),(1110) and (1111) do not participate in capturing similarity locally and collaboratively at A and B at bit plane (5,7,1,2). Code maps mentioned in the figure at bit plane (5,7,1,2), represents all observations at A and B with their respective clusters. GP is 86%. $\overline{DB}$ is 5.25, whereas, local DB at A and B are 0.818 and 0.841 respectively.%. Test accuracy is 83.33 %.
R^B		0.8030			0.841				

capture similar behavior among participating data sites. The Table 6.1 mentions local and collaborative results (i.e. collaboration of A with B and vice versa) for existing approaches using purity and DB index.

5.3 Summary

In this chapter, the proposed approaches are evaluated by various datasets using different evaluation metrics. The over all findings of this study shows the significance of the proposed approach by comparing the local, collaborative and global results.

6. DISCUSSION

In this chapter, the analytical findings with final implications are discussed. Section 6.1 includes the discussion for the VCCM approach, whereas section 6.2 mentions discussion for the VCC-BPS.

6.1 VCCM Discussion

The experimental results for all data sites are shown in Table 5.2, explained as follows:

- At site A_{iris} and B_{iris}, collaborative purities are higher than the respective local purities at A_{iris} and B_{iris}, which reveals the significance of collaboration to capture hidden information. Moreover, impact of collaboration is high at B_{iris} than A_{iris} (i.e $89.33\% > 86.67\%$).
- In case of Geyser, local and collaborative purity remains same at both A_{geyser} and B_{geyser}. This means that both sites have similar information, therefore VCCM does not reveal unique hidden pattern.
- In case of Cancer, collaboration is effective for both A_{cancer} and B_{cancer} to reveal hidden information. The reason of slight improvement is that the majority of patterns between both sites are similar.
- In case of Waveform data, four data sites participate in collaboration:
  - When only site B_{wave} shares its local map with site A_{wave}, purity before and after collaboration at A_{wave} remains same, meaning that both sites have similar information.
  - When site A_{wave} only uses the local map of site D_{wave}, purity at A_{wave} decreases after collaboration, which is due to unrelated patterns at D_{wave}.

Therefore, A_{wave} may avoid collaboration with D_{wave}.

- When site B_{wave} and C_{wave} collaborate with A_{wave} with coefficients 0.33 and 0.67 respectively, this expands learning from 59.28% to 61.92%.
- However, adding site D to the previous scenario with coefficients 0.25, 0.50 and 0.25 for B,C and D, respectively have no effect to enhance collaborative purity beyond 61.92%. This reveals that D_{wave} participates as redundant in process of collaboration.
- The same analogy is applied to other sites acting as local site.

In [20], if local purity at one data site is low while high at other data site, then data site with low purity will benefit from collaboration while other does not. According to our findings, in VCCM, (1) if observation patterns among sites are similar, then purity will remain unchanged, (2) if patterns are unrelated, then purity will fall, (3) if patterns are unique among sites then purity will increase. Contrary to [20], we decide the impact of collaboration via related and unrelated patterns instead of the magnitude of local purity, providing better collaboration to disclose hidden information.

6.2 VCC-BPS Discussion

In our proposed work, the vertical collaborative clustering using bit plane slicing approach is studied and applied over all eight-bit planes per feature of Geyser, Skin and Iris datasets. Since there are 2, 3 and 4 features in Geyser, Skin and Iris data, therefore, the numbers of bit plane combinations are 64 (8^2), 512 (8^3) and 4096 (8^4) respectively. The findings reveal BP (5,6) and (7,6,7) as the most significant bits combinations for Geyser and Skin data, whereas BP (6,2,1,2) as both most and least significant bits combinations for Iris data, capturing similarity. These bit plane combinations have such important bits which capture contrast between the given clusters as the least or most significant bits or both to correctly group the observations at particular bit plane with the maximum similarity. They have high purity with good compactness (DB) value in comparison to existing approaches. It is not necessary

that the collaborative purity will always be the mean of the local purities for all methods. For instance, our proposed approach (VCC-BPS) returns exact mean of local purities but existing approaches return collaborative purity not exactly equal to the mean of the local purities. The reason behind such symmetric and asymmetric collaborative purity is that the existing approaches have topographic map of fix size having nodes to represent similar observations. These nodes may be surplus and do not participate to represent data or have the least data observations at one site in match to another site. This deteriorates the final map results once K-mean algorithm is applied [27]. As a result, the collaborative purity corresponding to the existing approaches is asymmetric. Therefore, the collaborative purity measured at site A is different from that at B. Our proposed approach is very effective to deal with such problem by considering only those code maps which participate to capture similarity locally and collaboratively, and discard other code maps which do not participate. This forms symmetry, giving collaborative purity as the mean of the local purities.

This study shows that 120 training observations of the Geyser data at site A and B each, are compressed into 4 code maps (2 bits per code map) locally and collaboratively at bit plane (5,6). In case of Iris data, the contrast among three given clusters is captured and 66 training observations at site A and B are compressed into 14 code maps (4 bits per code map) and 2 code maps (0001 and 1101) do not participate in data compression locally and collaboratively at bit plane (6,2,1,2). Likewise, using same approach for Skin dataset, contrast between two given clusters is captured and more than 98000 training observations at site A and B each, are compressed into 8 code maps locally and collaboratively at bit plane (7,6,7).

The proposed collaborative purity reflects the similarity among the participating sites as high if the difference between the local and collaborative purity is low and vice versa. Moreover, if the local purity is less than the collaborative purity, means local learning enhanced by collaboration and accordingly collaborative DB increases based on equation (5.6). Such increase in collaborative DB confirms similarity between respective clusters of different data sites, whereas the low local DB shows the quality clustering within the local data. The Table 6.1 shows the out-performance

of the proposed approach in comparison to the existing approaches with quality clustering in terms of increased purity and collaborative DB with high test data accuracy. Notably, the bit plane at which an optimal solution is obtained, varies from dataset to dataset. Moreover, if the dataset with large number of features is used then the accuracy will not be compromised but computational cost will increase. Additionally, the collaborative purity results are closer to the global purity, verifies accuracy of our proposed approach. It also reveals that the proposed approach is successful to capture distributed hidden behavior which is similar to that of pooled dataset. The performance comparison based on evaluation metrics for the existing and proposed approaches using different data, is graphically shown in the Figure 6.1.

Table 6.1 Comparison of Existing and Proposed Work

Methods		Data site	Purity		Davies Bouldin Index	
			Local	Collaborative	DB	$\overline{DB}$
Existing	VCC-SOM [20]	A_{Geyser}	93.38	94.85	0.546	0.531
		B_{Geyser}	96.32	95.48	0.533	0.554
		A_{Skin}	73.16	71.89	0.865	0.901
		B_{Skin}	70.62	72.13	0.881	0.876
		A_{Iris}	80	80	0.702	0.702
		B_{Iris}	80	82.45	0.702	0.678
	VCC-GTM [10]	A_{Geyser}	93.4	94.64	0.547	0.536
		B_{Geyser}	95.88	94.23	0.541	0.567
		A_{Skin}	74.64	72.77	0.872	0.875
		B_{Skin}	70.91	73.12	0.88	0.866
		A_{Iris}	84.3	85.17	0.712	0.701
		B_{Iris}	86.04	84.29	0.668	0.691
Proposed	VCC-BPS at BP (5,6)	A_{Geyser}	98.33	97.92	0.389	7.144
		B_{Geyser}	97.5		0.423	
	VCC-BPS at BP (7,6,7)	A_{Skin}	74.23	75.70	0.953	0.916
		B_{Skin}	77.16		0.863	
	VCC-BPS at BP (6,2,1,2)	A_{Iris}	86.36	86.36	0.838	5.568
		B_{Iris}	86.36		0.837	

Figure 6.1 Graphical comparison between existing and proposed approaches for Geyser, Skin and Iris data using purity and DB indices.

7. CONCLUSION AND FUTURE WORK

This chapter mentions the summary of the theoretical contributions. It also discusses the limitations with potential avenues for the future work. Section 7.1 mentions the conclusion and future work for the VCCM. Likewise, section 7.2 includes conclusion and potential avenue for the future work of VCC-BPS.

7.1 VCCM Conclusion

This work proposes Vertical Collaborative Clustering Model (VCCM) as a unique model to manage the collaborative clustering process among different sites using a vertical approach and self-organizing map (SOM). In principle, VCC is a process where two or more data owners work together to reveal hidden structure in the local dataset with respect to the knowledge shared by outdoor data sites. However, the challenge is how to ensure that collaboration can bring improvement locally. And it is also much more challenging to control the process without any sort of a standard procedure while implementing. Instinctively, the collaboration is valuable provided the final local clustering have higher quality than if there had been no exchange of information between the local processes. Therefore, extreme attention must be taken to ensure that the collaborative process can improve the performance of each local clustering algorithm, and the controlling approach must be carefully crafted. Thus, VCCM sets an ideal environment (i.e. a standard procedure) for the collaboration by implementing methodical steps at both sites, which are interested to collaborate, before delivering the final collaborative results to the data owner. As a result, that would reduce the risk of misjudgment by the data owner, after implementing the collaborative clustering in a biased environment. In addition, it ensures that the collaboration brings improvement locally. VCCM ensures unbiased envi-

ronment for the collaboration by proposing same initialization parameters among all participating sites. Additionally, the VCCM improves clustering by exchanging local clustering results without compromising data confidentiality and accommodates collaboration by tuning collaborative map in a specific proportionality to disclose hidden patterns. The results demonstrate that the proposed VCCM improves local learning by collaboration and also helps the data owner to make better decisions on the clustering.

However, in the proposed VCCM, the findings show that the map consisting of the nodes of fixed size represents similar data. These nodes may be empty or have least observations at one site in comparison to other which deteriorates the clustering results. The possible extension of the proposed work would be to deal with such maps which participate in collaboration.

7.2 VCC-BPS Conclusion

The vertical collaborative clustering based on bit plane slicing manages collaboration among different sites. In this novel approach, an adequate common bit plane is determined among participating data sites, at which model fits the data with maximum similarity to unlock hidden patterns. Investigation shows that there is at least one-bit plane which captures relative important information commonly shared among different data sites. Notably, the bit planes, which contribute the most to represent relative important information, vary from dataset to dataset. The comparison of the proposed with the existing approaches reveals that VCC-BPS outperforms by having superior accuracy in term of high purity with improved DB and compress a large number of observations into smaller code space. The proposed collaborative results are close to that of pooled data output which verifies its accuracy. Additionally, it develops interaction between two or more data sources having same feature space to reveal similarities among datasets without compromising data confidentiality. The proposed approach does clustering, data reduction (compress large number of observations to small code map) and visualization simultaneously.

However, the proposed approach has a vast search space finding bit planes with

an adequate solution for a dataset with large feature space. This requires further investigation to add an extra computational layer such as using a data compression technique before simple voting algorithm to unravel the most informative bit plane and reduce the computational cost of measuring similarity both locally and collaboratively. Additionally, the probabilistic approach could also be used as an alternate solution to reduce the size of search space in finding the optimal bit plane, capturing maximum similarity. Moreover, we plan to develop correlation between local and collaborative evaluation metrics to validate the clustering outputs.

APPENDIX A:

The general profile of different datasets used in this study are as follow:

A.1 Datasets

- *Skin Segmentation Dataset*: This dataset consists of 245057 observations with 3 features i.e. B,G,R. Moreover, it has 2 classes, labeled as skin and non-skin.
- *Wisconsin Diagnostic Breast Cancer (WDBC) Dataset*: This dataset has 569 observations with 32 features. Each observation is labeled as benign or malignant .
- *Iris Dataset*: This is multi-variate dataset, consists of 4 features with 150 observations with 3 clusters.
- *Geyser Dataset*: It is multi-variate dataset of real values. It consists of 272 observations with 2 features.
- *Waveform Dataset*: This dataset consists of 40 features with 5000 observations, grouped into 3 clusters.

REFERENCES

1. Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008, July). An empirical evaluation of supervised learning in high dimensions. In Proceedings of the 25th international conference on Machine learning (pp. 96-103). ACM.
2. Cornuejols, A., Wemmert, C., Gancarski, P. & Bennani, Y. (2018), Collaborative clustering: Why, when, what and how, Information Fusion, 39, pp.81-95.
3. Celebi, M. E. (Ed.). (2014). Partitional clustering algorithms. Springer.
4. Fred, A. L., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. IEEE transactions on pattern analysis and machine intelligence, 27(6), 835-850.
5. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., & West, M. (2007). Generative or discriminative? getting the best of both worlds. Bayesian statistics, 8(3), 3-24.
6. Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
7. Ackerman, M., Ben-David, S., & Loker, D. (2010). Towards property-based classification of clustering paradigms. In Advances in Neural Information Processing Systems (pp. 10-18).
8. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics
9. Forestier, G., Gançarski, P., & Wemmert, C. (2010). Collaborative clustering with background knowledge. Data & Knowledge Engineering, 69(2), 211-228.
10. Sublime, J., Grozavu, N., Cabanes, G., Bennani, Y., & Cornuéjols, A. (2015). From horizontal to vertical collaborative clustering using generative topographic maps. International journal of hybrid intelligent systems, 12(4), 245-256.
11. Lancichinetti, A., & Fortunato, S. (2012). Consensus clustering in complex networks. Scientific reports, 2, 336.
12. Fagnani, F., Fosson, S. M., & Ravazzi, C. (2014). Consensus-like algorithms for estimation of Gaussian mixtures over large scale networks. Mathematical Models and Methods in Applied Sciences, 24(02), 381-404.

13. Gionis, A., Mannila, H., & Tsaparas, P. (2007). Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 4.
14. Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Learning in graphical models (pp. 355-368). Springer, Dordrecht.
15. Kohonen, T., Schroeder, M. R., Huang, T. S., & Maps, S. O. (2001). Springer-Verlag New York. Inc., Secaucus, NJ, 43(2).
16. Bishop, C. M., Svensén, M., & Williams, C. K. (1998). GTM: The generative topographic mapping. Neural computation, 10(1), 215-234.
17. Zehraoui, F., & Bennani, Y. (2005). New self-organizing maps for multivariate sequences processing. International Journal of Computational Intelligence and Applications, 5(04), 439-456.
18. Azzalini, A. & Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. Applied Statistics, 39, 357-365. Available: https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/faithful.html
19. A. Frank & A. Asuncion,(2010) UCI machine learning repository, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation
20. N. Grozavu, M. Ghassany, & Y. Bennani,(2011). Learning confidence exchange in collaborative clustering in Neural Networks (IJCNN), The 2011 International Joint Conference on, 5 (1), 872-879.
21. Bock, H.-H. (1985): On some significance tests in cluster analysis. Journal of Classification 2, 77-108
22. Ghassany, M., Grozavu, N. & Bennani, Y. (2013), Collaborative multi-view clustering, in 'The 2013 International Joint Conference on Neural Networks (IJCNN)', pp. 1-8.
23. W. Natita, W. Wiboonsak, & S. Dusadee,(2016),Appropriate Learning Rate and Neighborhood Function of Self-organizing Map (SOM) for Specific Humidity Pattern Classification over Southern Thailand, in 'International Journal of Modeling and Optimization'.
24. J. Kim, M. Sullivan, E. Choukse & M. Erez,(2016), Bit-Plane Compression: Transforming Data for Better Compression in Many-Core Architectures,

ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Seoul, pp. 329-340.

25. Podlasov, Alexey. (2006). Lossless image compression via bit-plane separation and multilayer context tree modeling. J. Electronic Imaging.
26. J. Sublime, D. Maurel, N. Grozavu, B. Matei and Y. Bennani, "Optimizing exchange confidence during collaborative clustering," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1-8.
27. W. Ishaq and E. Buyukkaya, "Dark patches in clustering," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 806-811.
28. P. Rastin, G. Cabanes, N. Grozavu and Y. Bennani, "Collaborative Clustering: How to Select the Optimal Collaborators?," 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, 2015, pp. 787-794.
29. J. Sublime, B. Matei, N. Grozavu, Y. Bennani and A. Cornu, "Entropy Based Probabilistic Collaborative Clustering", Pattern Recognition, vol. 72, pp. 144-157, 2017.
30. N. Grozavu, G. Cabanes and Y. Bennani, "Diversity analysis in collaborative clustering," 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, 2014, pp. 1754-1761.
31. J. Sublime, B. Matei and P. Murena, "Analysis of the influence of diversity in collaborative and multi-view clustering," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 4126-4133.
32. Bed Jatin and Toshniwal Durga, "SFA-GTM: Seismic Facies Analysis Based on Generative Topographic Map and RBF",2018.
33. A. Filali, C. Jlassi and N. Arous, "A Hybrid Collaborative Clustering Using Self-Organizing Map," 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), Hammamet, 2017, pp. 709-716.
34. J. Sublime, N. Grozavu, Y. Bennani and A. Cornu, "Vertical Collaborative Clustering using Generative Topographic Maps",In IEEE 7th International Conference on Soft Computing and Pattern Recognition, SocPaR 2015.
35. Ghassany Mohamad , Grozavu Nistor, Bennani Younes, "Collaborative cluster-

ing using prototype-based techniques" 2012. International Journal of Computational Intelligence and Applications.

36. Falih Issam, Grozavu Nistor, Kanawati Rushed , Bennani Younes, "Topological multi-view clustering for collaborative filtering" 2018, Procedia Computer Science. 144. 306-312.
37. Rafael C. Ganzalez, Richard E. Woods "Digital Image Processing", Second edition, Pearson Education, ISBN: 81-7808-629-8.
38. Hassan K. Albahadily, V. Yu. Tsviatkou & V.K. Kanapelka, "Gray Scale Image Compression using Bit Plane Slicing and Developed RLE Algorithms", 2017, International Journal of Advanced Research in Computer and Communication Engineering.
39. A. Frank & A. Asuncion, (2010) UCI machine learning repository, [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Iris
40. Data Mining , J. Han-M. Kamber, Morgan-Kaufman, Academic Press, 2001, ISBN: 1-55860-901-6