VISUAL ANALYTICS FOR PARTITION COMPARISON AND EVALUATION

BY

JIANGPING ZHOU

B.E. ZHEJIANG UNIVERSITY (1986)
M.E. SICHUAN UNIVERSITY (1990)
M.S. FITCHBURG STATE COLLEGE (2001)

A DISSERTATION SUBMITTED TO THE FACULTY OF THE COMPUTER SCIENCE DEPARTMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF SCIENCE IN COMPUTER SCIENCE
UNIVERSITY OF MASSACHUSETTS LOWELL
2007

Signature of Author: ______________________________ Date: __________________

Signature of Dissertation / Director:
Dr. Georges Grinstein

Signature of Other Dissertation Committee Members:
Dr. Haim Levkowitz
Dr. Kenneth Marx
VISUAL ANALYTICS FOR PARTITION COMPARISON AND EVALUATION

BY

JIAPING ZHOU

ABSTRACT OF A DISSERTATION SUBMITTED TO THE FACULTY OF THE COMPUTER SCIENCE DEPARTMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF SCIENCE IN COMPUTER SCIENCE UNIVERSITY OF MASSACHUSETTS LOWELL 2007

Dissertation Supervisor:  Georges Grinstein, Ph.D.
Professor, Department of Computer Science
ABSTRACT

Due to the rapid increase of data size and complexity in the world of information technology, cluster analysis with the assistance of visualization has become a major methodology for data analysis and exploration. This research focuses on visualization for mutual comparison and evaluation of multiple partitions, known as cluster stability analysis. In this context, a partition is constructed from the decomposition of a dataset into a family of disjoint groups. Partitions may refer to flat clustering results, categorical dimensions, binned numerical dimensions, predetermined class labeling dimensions, or prior knowledge in mutually exclusive format (one data item is associated with one and only one outcome). Cluster stability analysis is a kind of comparative cluster analysis, and the primary step in identifying “optimal” cluster structure, building cluster ensemble, or conducting cluster validation. Visualization for mutual comparison and evaluation of multiple partitions is a new research area in which there have been very few publications.

In this research, we define and extend a cluster stability metric system to record stability, group stability, and partition similarity measures. We describe a visualization tool CComViz (Cluster Comparison Visualization) that performs mutual comparison and evaluation of multiple partitions of the same dataset. CComViz utilizes a novel algorithm for the layout of record and dimension order in a display. We use it to visualize data stability, data flow, density distribution and hierarchy, data correlation at the record, the group, and the dimension levels within a single graphical display. We show how CComViz can, using brushing and linking, also help in the identification of interesting records and patterns in exploratory data analysis.
ACKNOWLEDGEMENTS

My long journey of study towards a doctoral degree has finally approached its completion. I am so grateful to so many people for their support in so many ways.

First and foremost, I would like to express my deep gratitude to my advisor, Dr. Georges Grinstein for his guidance and mentorship. His knowledge, passion and dedication have always enlightened me not only on my work and research but also on my personal life. I owe a special debt to Dr. Gary Livingstone, who helped me get my doctoral study started. I am also indebted to other two members of my thesis committee, Dr. Kenneth Marx and Dr Haim Levkowitz for their valuable suggestions to my dissertation research.

I am fortunate to have the generosity of the IVPR members. Dr. Chih-Hung Chiang and Dr. Howie Goodell are such persons who are always there when you need them. Dr. Alexander Gee gave me inspiration to add some analysis techniques to my CComViz tool. Mr. Brian Drohan provided expert opinions to the case study of my thesis for the MGH data. Mr. Shawn Konecni helped editing the manuscript. My thanks also go to many friends who have offered me support in many forms.

My deepest gratitude belongs to my loving wife Lin, daughter Shuyue, and other members of our families, for their ever-present support, whole-hearted understanding, and unfailing patience through the years.
# TABLE OF CONTENTS

ABSTRACT.......................................................................................................................... ii  
ACKNOWLEDGEMENTS....................................................................................................... iii  
TABLE OF CONTENTS......................................................................................................... iv  
LIST OF TABLES.................................................................................................................. vi  
LIST OF FIGURES............................................................................................................... vii  

1. INTRODUCTION ........................................................................................................ 1  
   1.1 Problem Statement and Motivation ................................................................. 2  
   1.2 Research Goals and Tasks ............................................................................. 8  
   1.3 Contributions................................................................................................. 10  

2. RELATED WORK ........................................................................................................ 12  
   2.1 Cluster Analysis............................................................................................... 12  
   2.2 Clustering Quality Measures and Cluster Stability Measures ...................... 17  
   2.3 Cluster Comparison and Visualization ........................................................... 23  
   2.4 Parallel Coordinates and Extension ............................................................... 35  

3 PARTITION STABILITY METRIC........................................................................... 47  
   3.1 Label Correlation Coefficient between Groups ............................................. 48  
   3.2 Label Correspondent Relationship ............................................................... 49  
      3.2.1 Polygamous............................................................................................. 50  
      3.2.2 Asymmetric........................................................................................... 51  
      3.2.3 Intransitive........................................................................................... 52  
      3.2.4 No-correspondence............................................................................... 53  
   3.3 Dimensional Record Stability ................................................................. 53  
   3.4 Record Stability ......................................................................................... 54  
   3.5 Group Stability under Threshold .............................................................. 55  
   3.6 Group Stability in Average ........................................................................ 56  
   3.7 Relative Partition Similarity ....................................................................... 56  
   3.8 Partition Similarity ..................................................................................... 57  

4. CCOMVIZ VISUALIZATION ................................................................................. 58  
   4.1 Why Record and Dimension Display Order Rearrangement .................. 60  
   4.2 Dimension Display Order Rearrangement ............................................... 63  
      4.2.1 Sorting Approach............................................................................... 63  
      4.2.2 Sequential Selection Approach......................................................... 64  
   4.3 Record Display Order Rearrangement ..................................................... 65  
   4.4 Other Modes for Record Display Order Rearrangement ......................... 73  
   4.5 Record Rendering Layer .......................................................................... 73  
   4.6 Change of Hot Dimension ....................................................................... 75  
   4.7 Boolean Operation for Record Selection ............................................... 75
4.8 Handling of Dimension with Continuous Value ........................................ 77
4.9 Addition of Record Stability Dimension .................................................. 77
4.10 Computational Complexity ....................................................................... 78

5. DISCUSSION OF STABILITY .................................................................... 79

6 CASE STUDIES ............................................................................................ 89
6.1 Salamander Gene Expression Data ............................................................ 90
6.2 MGH Breast Cancer Data ......................................................................... 96

7 CONCLUSION ............................................................................................... 104

8 FUTURE WORK ........................................................................................... 108
8.1 Label Correspondent Relationship ............................................................ 108
8.2 Interpretation and Validation of Stability .................................................. 108
8.3 CComViz Layout ....................................................................................... 109
8.4 User Study ................................................................................................ 109

LITERATURE CITED ..................................................................................... 110
LIST OF TABLES

Table 3-1        Synthetic Dataset ................................................................. 48
Table 3-2        Label Correlation Coefficient Matrix for the Synthetic Dataset .... 49
Table 3-3        DRS for Each Partition............................................................. 54
Table 3-4        RS Values........................................................................... 55
Table 3-5        GST Values with $T = 0.9$.......................................................... 56
Table 3-6        GSA Values........................................................................... 56
Table 3-7        RPS Matrix........................................................................... 57
Table 3-8        PSM Values........................................................................... 57
Table 5-1        CDM matrix for the Synthetic Dataset....................................... 81
Table 5-2        Comparison of RS and CDM measures...................................... 82
Table 6-1        XM clustering result and decompositions.................................. 98
Table 6-2        Contingence table................................................................... 99
Table 6-3        Statistic calculation................................................................. 100
Table 6-4        Variable ranges of entire CXM5 cluster and its stable subset.... 102
LIST OF FIGURES

Fig. 2.1-1  PCA plot on a gene expression data [86].......................................................... 16
Fig. 2.1-2  Radviz with properly ordered dimensions [56]............................................. 16
Fig. 2.1-3  Color-filled Dendrogram [19]...................................................................... 17
Fig. 2.1-4  Probability Model[19]................................................................................ 17
Fig. 2.3-1  Histogram of correlation similarities on a synthetic dataset with a mixture of four Gaussians distribution sub-datasets [12]..................................... 24
Fig. 2.3-2  Cumulative distribution on a synthetic dataset with a mixture of four Gaussians distribution sub-datasets [12].................................................. 24
Fig. 2.3-3  Change in cluster size due to different clustering methods as the number of clusters is equal to 36 [24]................................................................. 26
Fig. 2.3-4  Hierarchical tree using WADP distance measure between clustering results of different methods [24]................................................................. 27
Fig. 2.3-5  Silhouette Plots [Mathworks].............................................................. 27
Fig. 2.3-6  Adjusted FOMs plot on a synthetic dataset for three clustering methods and a baseline [126]................................................................. 28
Fig. 2.3-7  F-measure performances of five clustering methods [68]............................... 29
Fig. 2.3-8  Weighted bipartite graph with crossing minimization [119].................................. 32
Fig. 2.3-9  Expression Profiler for comparison of a hierarchical clustering and a k-means clustering on a real life dataset [119]........................................... 33
Fig. 2.3-10 Cluster Comparison visualization [19]............................................................ 33
Fig. 2.3-11 Stable Matrix for Iris dataset with 8 runs on k-mean and the number of clusters as 8 for each run [29]................................................................. 34
Fig. 2.4-1  Parallel coordinates polylines vs. scatter plot projections for three data items with 3 dimensions [13]............................................................. 35
Fig. 2.4-2  Severe occlusion and clutter in regular parallel coordinates for a dataset containing 5 dimensions and 3848 data items [124]........................................ 36
Fig. 2.4-3  Parallel coordinate plot with saturation controls on the same dataset as in Fig. 2.4-2 [124]................................................................. 37
Fig. 2.4-4  Parallel coordinate plot with filtering-out of sparse areas on the same dataset as in Fig. 2.4-2 [124]................................................................. 37
Fig. 2.4-5  Hierarchical parallel coordinates plot (implemented in XmdvTool) on Iris dataset [121]................................................................. 39
Fig. 2.4-6  Parallel coordinates for visualizing the fuzzy rules of Iris dataset [13]................................................................. 39
Fig. 2.4-7  The Detailed and abstract representations of the same data [98].......................... 40
Fig. 2.4-8  Attribute value distribution in two classes over all dimensions [2].................. 41
Fig. 2.4-9  The outcomes with different transfer function for the same data [83].............. 42
Fig. 2.4-10  Color-encoded bars for displaying distribution [111]................................... 43
Fig. 2.4-11 Parallel histogram [101]................................................................. 44
Fig. 2.4-12 Parallel boxplots [105]................................................................................. 44
Fig. 2.4-13  Ellipse plots overlapping with parallel coordinates [2]............................. 44
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 6-3</td>
<td>Patterns and expression levels of stable records</td>
</tr>
<tr>
<td>Fig. 6-4</td>
<td>Patterns and expression levels of unstable records</td>
</tr>
<tr>
<td>Fig. 6-5</td>
<td>CComViz projections of stable records ($rs &gt; 0.6$) (upper) and unstable records ($rs \leq 0.25$) (lower)</td>
</tr>
<tr>
<td>Fig. 6-6</td>
<td>MDS projection of $C_{XM5}$ cluster</td>
</tr>
<tr>
<td>Fig. 6-7</td>
<td>Parallel coordinates projection of $C_{XM5}$ cluster</td>
</tr>
</tbody>
</table>
1 INTRODUCTION

In recent decades, the speed of data acquisition in molecular biology, public health studies, and drug development largely exceeded our ability to process data efficiently as ever-increasing data size and complexity placed extraordinary challenges in data analysis and exploration. Therefore, the development of more effective and more efficient data analysis techniques has become one of the most important tasks in most scientific research and industrial fields. By using interactive approaches, information visualization provides individuals with improved insight and better understanding of data structures, patterns, anomalies, and other interesting features occurring in databases. Visualization has now become essential in data analysis and exploration. “The analyst user during the highly iterative model building process can quickly obtain insights from the visualization that suggest the adequacy of the solution and what further experiments to conduct. Alternatively, the business user can examine and query the final clustering solution using the visualization.” [33].

In reality, no data is isolated and no single computational learning method is uniformly recommended. Comparison of information from different sources and computing results is a routine and nontrivial task. Without external assistance, this task is sometimes considered as tedious and painful. A framework or workspace featuring information visualization, systematic comparison, evaluation and validation across many sources obviously provides the needed assistance.
1.1 Problem Statement and Motivation

Cluster analysis is one of the most commonly used data analysis and exploration methods. It identifies clusters from a collection of data items. The data items in each of the clusters are more “similar” than those from others [60]. Depending on the task or goal, the notion of similarity can be measured in many different ways, which could be distance-based, density-based, correlation-based or domain-specific in terms of prior knowledge of the problem. The clustering method differentiates homogeneous and heterogeneous data items to provide insight and a better understanding of the problem, which is necessary for decision-making. Widely available clustering algorithms differ in efficiency, effectiveness, and robustness regarding datasets with different features, structures or sizes. Data clustering results are often sensitive to clustering algorithms and the parameters used. The diversity of datasets and clustering algorithms makes “no single established method to estimate the significance of an observed degree of relationship obtained by cluster prediction techniques” [15]. Hence, in reality we do not have a universal clustering algorithm which works well on all datasets. We do not even have a general guideline to select the right clustering algorithms with the right parameter settings before the applied datasets become well known [35, 44]. From computational point of view, many basic clustering problems, such as the k-median problem [96], the set cover problem [75], and the maximum coverage problem [37], are recognized as NP-hard [55]. On this basis, the identification of optimal clustering result is also considered NP-hard [14, 91]. The current solution to this problem is based on heuristics and experiments with repeated clustering, comparison, evaluation, and validation in terms of similarity.
assumptions and various robust techniques. Although the challenge for determining experimental scopes, particularly the clustering methods and the numbers of clusters still remains, domain experts and data analysts have in-depth knowledge.

Comparative cluster analysis deals with cluster evaluation, optimization, and validation through comparison of multiple partitions. In this context, we adopt the term “partition” from mathematical terminology as “the decomposition of a set into a family of disjoint sets” [http://www.thefreedictionary.com/partition]. Based on this mathematical definition, partition in this context is defined as follows.

**Definition**

Let $X$ be a dataset with $n$ data items. Set $P = \{g_1, g_2, \ldots, g_p\}$ is defined as a partition of $X$, in which the subsets $g_j$, $j = 1, 2, \ldots, p$, are disjoint and exhaustive groups, such that $g_x \cap g_y = \emptyset$, $x \neq y$; $|P| = |g_1| + |g_2| + \ldots + |g_p| = n$.

By this definition, the categorical dimension of a dataset is a typical partition because its categories are naturally disjoint. A flat clustering result, namely the cluster assignment, is comprised of disjoint groups, and every data item is assigned one and only one group label. So, the flat clustering result is a partition. Besides the flat clustering result, clustering results with other structures can also be treated as partitions after appropriate conversion (Cluster structure will be reviewed in next chapter). Other examples of partitions include binned numerical dimensions, predetermined class labeling dimensions, and prior knowledge in mutually exclusive format (One data item is associated with one and only one outcome) [57].

13
In most literature, “clustering” refers to a flat clustering result. Many studies about comparative cluster analysis were with regards to the flat clustering result only. With respect to the original description, the terms “clustering”, “clustering result”, “partition”, and “partitioning result” are interchangeable in this dissertation depending on context; so are the terms “cluster” and “partitioning group” or simply “group”.

Comparative cluster analysis often addresses the following questions:

• How well do the obtained clustering results fit the data?
• How well are the obtained clustering results in agreement with prior knowledge of the problem?
• Is the clustering result obtained by a method better than other methods?
• How sensitive are the clustering results to clustering methods and parameters used?
  What are their relationships?
• How consistent, or stable, are the data memberships over multiple partitions?
• Where are the outliers recognized by different criteria reflected in partitions?
• Are the clustering results sensitive to random noise added to input datasets?

The process of a complete comparative cluster analysis can be divided into three stages. The first stage is partition comparison and evaluation; the second is clustering selection or cluster ensemble; and the last one is cluster validation. This research focuses on the first stage. The results of partition comparison and evaluation can be used not only to identify interesting data features, but also to provide help and guidance for clustering
selection, cluster ensemble, and cluster validation. There are several reasons why partition comparison and evaluation are important.

- “The clustering quality relies on the use of criteria that attempt to capture those aspects that humans perceive as the properties of a good clustering solution.” [70]. Clustering criteria characterize clustering algorithms and clustering evaluation methods. Defining and choosing appropriate criteria are usually not easy and controversial since the diverse criteria are sometimes partially complementary, even conflicting, or may favor different types of solutions. The choice of criteria for evaluating clustering results is fully empirical. Comparative analysis is a natural and practical approach to better decision-making.

- Clustering algorithms may behave differently depending on chosen data features (subsets of data dimensions) and parameters used, and hence result in varying partitions [64, 65]. Choosing a partition to be used for the next step in the process is a difficult decision. It is wise to bring various clustering results into a framework for comparison and evaluation. Afterwards, a partition can be chosen which optimizes certain quality measures, or a new one can be created using a consensus of clustering results. The first approach to partition selection is known as model selection, while the second is called cluster ensemble [70]. Since cluster ensemble combines multiple clustering results in a voting scheme, it often leads to a better and robust result [113, 50]. However, this technique works only when majority of participant clustering results have good quality and maintain a certain degree of consistency. Early partition
comparison and evaluation to screen qualified clustering results is a practical approach that guarantees cluster ensemble quality.

- The observation of diverse clustering results and their membership changes over various clustering methods and configurations helps reveal hidden information for better understanding of data from various perspectives. When conceptually different clustering methods produce similar outcomes, scientists become more confident about the clustering experiments and discovered cluster structure [68]. The information obtained from partition comparison and evaluation helps elucidate relationships between clustering methods, parameter settings, and data features. This insight will become domain knowledge, and is likely very valuable in study of the domain problem.

- When prior knowledge or predetermined classification is available, it is important to verify the agreement between it and the obtained clustering results. Only highly corroborated results give scientists confidence in the accuracy of the clustering results [126]. When prior knowledge is absent, cluster comparison and quantitative assessment often serve as an important reference for various tasks, such as identifying clustering paradigm that is most suitable for a domain problem, or significant data features (dimensions) through discriminant analysis [73].

- Predetermined classifications often contain misclassified or misdiagnosed instances, called class noise [133]. Without paying attention to these instances, supervised learning and validation relying on these instances will cause misleading results.
Comparative cluster analysis is an practical approach to detecting these instances and keeping data analysis and exploration in track.

- The distinction between outliers and misclassified records can be small [110]. It is not reliable to judge outliers based on only a few clustering results or predetermined classifications. The conclusion should be made through the comparison of various partitions.

- Cluster cross-validation provides mathematical confidence to clustering reliability and robustness, and is often required in cluster analysis for a large dataset. The procedure of cluster cross-validation involves one or more data-driven statistical tests. These tests are considered to be very expensive [81, 93, 59]. When many clustering results need validation, it is wise to perform their comparison first, and then select or build the most ‘attractive’ one before cross-validation.

- Several cluster stability analysis techniques, such as perturbation-based cluster analysis [85] and cluster analysis on resampling data [52, 87, 57], aim at inferring the number of clusters in terms of the consistency of multiple runs of the same clustering methods under different conditions. The measurements of consistency totally rely on the comparison of many experimental results.

To systematically perform partition comparison and evaluation for large and complex data, we need to have a convenient framework with relevant metrics and analysis tools. This framework should facilitate not only the comparison and evaluation of large number of partitions, but also data representation, manipulation, and user
interaction as well. Although several text, table, and chart based frameworks for comparative cluster analysis have been proposed [4,113], they do not satisfy a necessary standard of intuitiveness, convenience, efficiency, and effectiveness as a whole for large and complex datasets. Visualization is an ideal solution that makes it possible to build such a framework and various tools. In addition to the visual representation of data, visualization provides the means to bring powerful human perception into the data exploration process, which helps to tune and steer future exploration.

Partition comparison and evaluation can benefit much from better visualization because human visual perception has evolved to enable superior image comparison. For a set of clustering results, a successful visualization can more quickly and intuitively tell how well their clusters are defined, what differences and relationships they have, and the degree of membership of data items when compared to computational methods. Data patterns, trends, anomalies, outliers, relationships, and parameter settings can be more easily discerned when visualization is applied to partition comparison and evaluation. Visualization for partition comparison and evaluation is also a basic component of interactive frameworks featuring information consolidation, evaluation and validation. To our best knowledge, few studies have been done before in these areas.

1.2 Research Goals and Tasks

The goal of this research is to develop a partition stability metric system and a visualization tool facilitating the analysis of data stability, similarity, density distribution, hierarchies and correlation at record, group and dimension levels based on comparison of
multiple partitions. Using these techniques at the record level, stability and membership migration rate of individual records across multiple partitions can be measured and observed. At the group level, label correspondent relationships, density distributions and hierarchies are visualized. And at the dimension level, partitions are evaluated by different criteria. To accomplish this goal, five major research tasks are involved. They are discussed from Chapter 2 to Chapter 6 respectively. Below is a brief description for each chapter.

Chapter 2: Related Work

Review related work which is part of the basis of this research that provides inspiration for our work. Through the review, several limitations of previous work are recognized.

Chapter 3: Partition Stability Metric System

Extend a perturbation-based cluster stability metric system to the partition stability metric system. This system contains six measures, including Dimensional Record Stability (DRS), Record Stability (RS), Group Stability under Threshold (GST), Group Stability in Average (GSA), Relative Partition Similarity (RPS), and Partition Similarity (PSM). In addition, three fundamental properties and one feature related to this metric system are explored.

Chapter 4: CComViz Visualization
Develop a visualization tool, namely CComViz (Cluster Comparison Visualization), for partition comparison and evaluation. A novel methodology for designing a layout algorithm is presented to support this tool. This algorithm is used to rearrange record and dimension display order in CComViz for achieving visual aesthetics and crossing minimization. A number of interactive functions are designed to increase this tool’s usability.

**Chapter 5: Discussion of Stability**

Discuss the record stability (RS) measure defined in the partition stability metric by comparing the RS measure with another stability measure (CDM), and accessing geometrical properties of stable records in CComViz, parallel coordinate, MDS (Multidimensional Scaling) plot, scatter plot under the UVP data and tool linking environment.

**Chapter 6: Case Studies**

Apply CComViz, record stability measure and other data analysis techniques to Salamander gene expression data and MGH breast cancer risk factors data for exploring data stability feature, and identifying interesting records and data patterns.

**1.3 Contributions**

This research responds to one of the most demanding requests from the exploratory data analysis community by presenting a suite of techniques for mutual comparison and evaluation of multiple partitions. Our major contributions include:
1. Define partition comparison and evaluation to address mutual comparison and evaluation of multiple variables that are either homogeneous or heterogeneous, including categorical dimensions, numerical attributes, clustering results, classifications and prior knowledge.

2. Extend a perturbation-based cluster stability metric system to the partition stability metric system. Three fundamental properties and one feature related to this metric system are recognized.

3. Develop CComViz visualization tool which contains a number of interactive functions for partition comparison and evaluation within a single graphic display.

4. Present a novel methodology and a novel layout algorithm used to rearrange record and dimension display order in CComViz for achieving visual aesthetics and crossing reduction, as well as visual metaphor for stable and unstable records.
2 RELATED WORK

2.1 Cluster Analysis

Cluster analysis is one of the most commonly used quantitative data analysis and exploration methods in almost every scientific and industrial field [41]. As a major step for cluster analysis, data clustering groups homogenous data in certain way in terms of homogeneity or similarity measure. A data item is a record. Homogenous records are called intra-class members, and heterogeneous records are inter-class members. Based on the ways about how data are clustered and how homogeneity is defined, various clustering algorithms and techniques have been developed [82, 78, 40].

There are three basic ways to cluster data, which lead to three types of cluster structures. Flat clustering assigns each record a cluster identification label in mutually exclusive format (one record is associated with one and only one identification label), such that records with same identification labels share more similar characteristics than others [81]. Hierarchical clustering gives each record a nested sequence representing a path in a hierarchical tree, such that records are organized into hierarchies without inconsistent links (the link whose distance is significantly larger than the average of the nearby link distances on both sides), and the records beneath a node share similar characteristics [6]. Probabilistic clustering, or called mixture model by some researchers, [22, 44] gives each record a vector of probabilities, each probability corresponds to a
cluster category, called model state, and indicates possibility of the record remaining in
the cluster. The conversion between the three types of clustering structures is possible.
For example, a hierarchical clustering result can be converted to a flat structure by cutting
off hierarchical tree to a specific depth. For probabilistic clustering result, the conversion
can be done by assigning the state name with highest probability for a data item as cluster
label for this data item. Every type of clustering can be generated through either
unsupervised or supervised learning depending on whether predetermined classification
or prior knowledge is used. Detailed surveys about a large number of clustering methods
can be found in publications [81, 82, 86].

Comparative cluster analysis is an important part of cluster analysis. It deals with
cluster comparison and evaluation, cluster ensemble, and cluster validation. Since real
data cluster structure can only be accessed by running various clustering algorithms to
reveal different aspects of data features, cluster comparison and evaluation is an effective
way to capture intrinsic data characteristics.

Cluster ensemble is a clustering technique using combination, consensus, fusion,
or weighted voting to re-construct cluster structure relied on comparison of multiple
clustering results without directly accessing the original data features. This technique
provides a means to consolidate results from a portfolio of individual clustering results.
The effectiveness of this approach lies under the fact that highly agreed clusters, in
expectation, often translate to prominent data features. “Combining multiple
classification or regression models often provides superior results compared to using a
single, well-tuned model” [113]. Although the study of cluster ensemble started in as
early as mid-80s [9, 97], it wasn’t popular until recent years. Some successful cluster ensemble techniques include re-sampling scheme-based consensus [94], voting scheme-based consensus [47, 48, 49, 50], and multi-objective clustering (MOCK) [70].

Cluster validation includes cluster validity index analysis and cluster cross-validation. Cluster validity index analysis identifies the best clustering result of which a validity index value is optimal among a set of clustering results. Various cluster validity indices that measure clustering quality are defined [61, 4]. Their applications for cluster validation can be found in publications [3, 4, 88, 12, 24, 73, 119]. Cluster cross-validation involves a certain amount of statistical data-driven re-sampling, clustering experiment, testing and evaluation. It gives mathematical confidence about how the obtained clustering result is reliable and robust with less concern about overfitting. Overfitting is a common machine learning problem with which a small change in training data, such as data size or re-sampling, will cause significant impact on the outcome, thus individual outcomes overfit training data [28]. In cluster analysis, overfitting causes significant changes of cluster assignments and correlations regarding different data re-sampling. For cluster cross-validation, the original dataset is sampled into multiple folders, and clustering, testing, and evaluation are conducted on different folders to avoid overfitting. According to different statistical tests used, varieties of cluster validation techniques have been suggested, for examples, Monte Carlo test [81, 96, 59], the bootstrap hypothesis test [38], and Jackknife style statistical analysis (leave-one-out) [126, 30].
Visual cluster analysis has gained much popularity in recent years due to its ability to represent data with visual distinction between data clusters. Since visualization harnesses human visual perception to evaluate clusters through various interactions, it helps retrieve information and gain insights into supporting and complementing statistical summaries. There are three application scenarios for visual cluster analysis. The first scenario is a straightforward projection of data in graphic space without dimensional transformation. Typical visualizations used in this scenario include scatterplot, heatmap and parallel coordinates. The second scenario involves dimensional topology-preserving transformation, data projection is therefore an artifact for displaying data geometrical property and other important data features. PCA (Principal Component Analysis) plot [86] (Fig 2.1-1), MDS (Multidimensional Scaling) plot [129] and RadViz [76, 56] (Fig 2.1-2) are some successful tools in this scenario. The last scenario emphasizes cluster structure rather than data geometrical property. The tools in this scenario include Dendrogram (Fig 2.1-3), Silhouette plot [130] and Probability Model [19] (Fig 2.1-4) etc.
**Fig. 2.1-1** PCA plot on a gene expression data [86].

**Fig. 2.1-2**: Radviz with properly ordered dimensions [56]
2.2 Clustering Quality Measures and Cluster Stability Measures

Clustering quality measures utilize key features of data domain and serve as the basis of data clustering and cluster validation. In different literatures they are called cluster validity indices or clustering criteria [21]. Theodoridis and Koutroumbas (1999) [114] divide these measures into three types: internal, external and relative according to the use of information source. All measures are equally important to partition evaluation and validation. Internal measure uses internal data information or intrinsic property. The forms of internal data information are associated with mathematical data model employed. They reflect data’s statistical and geometrical properties, such as variance, density,
cluster shape, connectivity, intra-cluster compactness, and inter-cluster separation etc. [41, 71, 68]. By internal clustering criteria, a good quality clustering minimizes intra-cluster distances for achieving compactness, and maximizes inter-cluster distance for separation.

External clustering criteria utilize external or given data information based upon domain knowledge. Various diagnostic results, functional and validated classifications are considered as external information. For example, in molecular biology the external information applied to gene expression data may be regulatory regions [74,132], enzymatic classifications [88], metabolic pathways and protein structure [89], or functional classifications of the genes [126], such as structured and shared vocabulary which allows the annotation of gene products across different model organisms to be involved in the cluster validation [134, 92,103]. External information is usually expressed as labeling dimensions in mutually exclusive format [125, http://en.wikipedia.org/wiki/Mutually_exclusive]. External criteria statistically assess the degree of agreement between one partition and a labeling dimension.

Relative clustering criteria are based on comparisons of partitions generated by the same algorithm with different parameter settings, or applied to different re-sampling subset of the same dataset. Partitioning consistency and cluster stability are outcome of partition comparisons. External and relative clustering quality measures are defined to assess these qualities.

Partitioning consistency or cluster stability is a clustering quality measure often used to validate partitions and the number of clusters in comparative cluster analysis. Since it is an outcome of partition comparisons, some external or relative clustering
criteria can be used to assess this quality. Lange et al. (2003) [87] introduced a set of stability measures to assess the ability of a clustering result on one sampling subset to predict the clustering of another re-sampling subset from the same data source. Roth et al. (2002) [105] defined stability measures as the data self-consistence independent of external information. Their measurements are based on imitating independent sample datasets by way of re-sampling. Ben-Hur et al. (2002) [12] and Giurcaneanu et al. (2003) [57] suggested using similarity measures on intersected subset of two independent sample datasets to measure cluster stability. The similarity is measured based on cluster assignments generated by the same clustering algorithm with same parameters on two sample datasets respectively.

Cvek (2004) [29] proposed a cluster stability measure, called CDM (C-Dimension Measure). CDM measure is a 2D symmetric matrix in which both dimensions are record indices in the same order. The cell of CDM matrix is a vote number among participant clusterings, called c-dimensions in CDM, for recognizing the pair of records staying in the same cluster. CDS measure is intended to capture groups of stable records which are clustered together in majority of clusterings without considering their population and relationship between groups. Fred and Jain (2005) [50] did a similar research to Cvek’s and named the normalized CDM matrix as co-association matrix by dividing every cell of CDS matrix by the number of participant clusterings.

Katz and Grinstein (2006) [85] defined a stability metric system, which consists of a set of stability measures at the record, cluster, and global levels, for perturbation-based cluster analysis. Perturbation-based cluster analysis studies the response of various
clustering experiments to random noise added to input data in order to find out the most stable clustering results through a large quantity of experiments. The calculation of these measures requires the setting of number of clusters to be same in every perturbation run, which equal to the number of clusters of the original (unperturbed) clustering. The correspondent cluster pairs between unperturbed and perturbed clustering results need to be identified before the calculation.

The correspondent cluster pair is two clusters in two separate clustering results, which correspond to each other in affiliate relationship. The records falling into both clusters of a correspondent cluster pair are considered to maintain the same membership. If a record belongs to only one of correspondent cluster pair, this record is considered to migrate. Records with high migration rate across many clustering results are considered as unstable records. The identification of correspondent pair is called label correspondence problem [94, 113, 118]. Because of diversity of clustering algorithms and metric systems used, label correspondent relationship often goes beyond one-to-one, and might be one-to-many, many-to-one, or even many-to-many. When data stability is measured at the record and group levels, the label correspondence problem has to be solved beforehand.

Two approaches for identifying correspondent cluster pair are suggested by Katz and Grinstein. The first approach uses normalized distance between two clusters. The normalized Euclidean distance between clusters $c_1$ and $c_2$ is defined as:

$$\text{distance}(c_1, c_2) = \sqrt{\frac{|c_1| + |c_2| - 2 \cdot |c_1 \cap c_2|}{|c_1| + |c_2|}}$$
Such that two clusters with minimum normalized distance is a correspondent cluster pair.

The second approach uses correlation coefficient of two clusters. The correlation coefficient between clusters $c_1$ and $c_2$ is by

$$\text{correlation}(c_1, c_2) = \frac{|c_1 \cap c_2|}{\sqrt{|c_1| \times |c_2|}}$$

Such that two clusters with maximum correlation coefficient is a correspondent cluster pair.

Once the cluster correspondent relationship for a dataset is recognized, data stability at the record, cluster and global levels can be measured.

The stability at record level is the estimate of maximum probability for a record to stay in a certain cluster.

Let $J$ be the number of clusters of unperturbed clustering, $P$ be the total number of perturbation runs, $R_{k,j}$ be the number of perturbation runs in which the cluster that the record with index $k$ belongs to and unperturbed cluster with index $j$ are a correspondent pair. The record stability is defined as:

$$S(k) = \frac{\max (R_{k,j})}{P}$$

To eliminate the impact of number of clusters, the record stability index is then defined.

$$I(k) = \left( S(k) - \frac{1}{J} \right) \times \frac{1}{J - 1}$$

Two types of cluster stability are defined: cluster stability frequency and cluster stability score.
Let C be a given confidence level, N(m, C) be the number of records in unperturbed cluster with index m, whose record stability values exceed threshold 1 − C, and T(m) be the total number of records in unperturbed cluster with index m. The cluster stability frequency is defined as:

$$SP(m, C) = \frac{N(m, C)}{T(m)}$$

And the cluster stability score is defined as:

$$SA(m) = \frac{1}{T(m)} \sum_{k \in \text{cluster}(m)} S(k)$$

The global stability index is the average of cluster stability over all clusters and all perturbation cycles.

$$S(J) = \frac{1}{P} \sum_{p=1}^{P} \left( \frac{1}{J} \sum_{m=1}^{J} S(m, J, p) \right)$$

where P is the numbers of perturbation runs. S(m, J, p) is SP(m, C) or SA(m) at the p-th perturbation run.

Through the review of stability measures, we notice that various stability definitions from different authors are conceptually different. Among them the stability metric system especially causes our interest because it measures data stability at the record, cluster and global levels.

2.3 Cluster Comparison and Visualization
In the first chapter, we list a number of reasons about the importance of cluster comparison. In this section, we review various cluster comparison techniques incorporated with visualization.

Estimation or evaluation of the number of clusters is a major application of cluster comparison. Ben-Hur et al. (2002) [12] proposed a cluster stability-based method to validate the number of clusters. This technique groups clustering results by the number of clusters, and employs a cosine similarity metric introduced by Fowlkes and Mallows (1983) [43] to compute the correlations between pairs of clustering results for each group. To make the validation more reliable, a number of clustering results with different number of clusters on dataset re-samplings generate a distribution of stability. This distribution is then used to determine the appropriate number of clusters based on the concentration of group correlation (similarity) spectrum. For example, the histogram of correlation similarities (Fig. 2.3-1) indicates $k = 2$ is an appropriate number of clusters due to the narrowest similarity spectrum. Fig. 2.3-2 plots cumulative distribution of similarity values for different numbers of clusters, and provides information for the analysis of phase change beneath the distributions, for example, bigger gap between two settings of the numbers of clusters indicates phase transition from stable clustering to less stable. The proposed technique works for both flat clustering algorithms for finding appropriate number of clusters and hierarchical clustering algorithm for finding appropriate depth at which the most stable clustering occurs. Since this technique re-
Fig. 2.3-1 Histogram of correlation similarities on a synthetic dataset with a mixture of four Gaussians distribution sub-datasets [12].

Fig. 2.3-2 Cumulative distribution on a synthetic dataset with a mixture of four Gaussians distribution sub-datasets [12]
samples dataset for clustering, it is applicable to various datasets with different distributions.

Azuaje (2001) [5] suggested a voting strategy to estimate the correct number of clusters by embedding cluster internal validity indices into SOM algorithm. Milligan (1985) [93], Goldberg (1989)[59], Tibshirani et al.(2000)[116], and Bandyopadhyay and Maulik (2001) [7] developed other techniques to determine and validate the number of clusters using various internal validity indices, as well as their weighted combinations. Many of similar techniques can be found in publications [88, 24, 73].

With regard to an embryonic stem cell gene expression dataset, Chen et al. (2002) [24] demonstrated the application of various internal and external validity indices to evaluate the performance of several clustering algorithms that are often used by molecular biologists, such as hierarchical clustering, k-means, PAM, and SOM. In the application, clustering robustness, cluster size and clustering consistencies are particularly analyzed using WADP (Weighted Average Discrepant Pairs) index, which measures robustness of clustering against perturbed data input [15]. Fig. 2.3-3 shows the change of cluster size due to different clustering algorithms. Fig. 2.3-4 is a tree representation of clustering similarity among these algorithms based on WADP distance (average of WADP(A, B) and WADP(B, A), where A and B are two clustering results), namely average discrepancy rate of gene pairs. The height in tree represents the distance between two merging nodes. This tree is interpreted by the authors: “k-means was similar to PAM, while average linkage and SOM r1 tended to produce clusters not overlapping with those of other methods. However, note that even the distance between k-means and
PAM was larger than 0.45, which meant more than 45% of gene pairs in one clustering result were separated by the other method.” One of significant contributions of this study is that it brought biological insights into computational analysis by giving biological interpretation to computational clusters and examining the distribution of functionally classified genes, therefore, the gene functions can be better known and classified according to their biological and geometrical homologies. Through demonstration on particular dataset, this study proposed a general guideline for selecting proper clustering methods which may help extract meaningful biological information.

Fig. 2.3-3 Change in cluster size due to different clustering methods as the number of clusters is equal to 36 [24]
Fig. 2.3-4 Hierarchical tree using WADP distance measure between clustering results of different methods. The height in tree represents the distance between two merging nodes [24].

Silhouette plot [www.mathworks.com] is intuitive and convenient for cluster comparison and evaluation. Silhouette index [106] is an internal clustering criterion and widely used for validating cluster quality in terms of data coherence and separation features. Fig. 2.3-5 shows three silhouette plots on the same synthetic dataset. From these plots, we can compare three clustering results, each has different number of clusters, and conclude the clustering result with 4 clusters is the best because more records in the 4 clusters have high silhouette index values.

(a) The second cluster contains many records with low silhouette values
(b) Many records in all four cluster have high silhouette values
(c) The first and the fourth clusters have fewer records with high silhouette values

Fig. 2.3-5 Silhouette Plots [Mathworks]
In order to compare and evaluate clustering results from different clustering methods over a range of numbers of clusters, Yeung et al. (2001) [126] developed a quantitative data-driven framework based on Figure of Merit (FOM) and adjusted FOM quality measures. The plots of adjusted FOM versus numbers of clusters for different clustering methods provide convenience and intuition for comparing performances of clustering methods on applied datasets and estimating the “right” numbers of clusters. For instance, from the adjusted FOMs plot (Fig. 2.3-6) on a synthetic dataset for three clustering methods and a baseline reference, we can clearly see the clustering results with 5-10 clusters from CAST and single-link clustering algorithms are better than those from k-means. Note that smaller adjusted FOM value indicates better prediction power.

![Adjusted FOMs plot on a synthetic dataset for three clustering methods and a baseline.](image.png)
Handl et al (2005) [68] reviewed a number of validity index-based cluster validation techniques applied to post-genomic data. For the review, various plots of performances versus the numbers of clusters are generated. The performances are measured by several internal and external clustering criteria. Fig. 2.3-7 shows one of these performance plots on a synthetic dataset containing two elongated classes (see bottom right) with normal distribution, where F-measure [120] is a consistence measure between a clustering result and a class labeling dimension. Apparently, the performances of the five included clustering methods in terms of F-measure are quite different. According to the performance plots, the authors claimed that validity index-based validation depending on individual internal or external measures suffer from biases with respect to the number of cluster, which make it hard to identify appropriate numbers of clusters and clustering methods. Therefore, as conclusion, alternative complementary and multi-objective validations are recommended for achieving more robust results.

Fig. 2.3-7  F-measure performances of five clustering methods [68]
Bolshakova and Azuaje (2003) [17] proposed another validity index analysis framework, Machaon-CVE (Clustering and Validation Environment), to support biomedical knowledge discovery. Machaon-CVE provides an integrated environment for clustering generation and clustering quality assessment using cluster validity indices. Within this framework, multiple clustering results from different clustering methods and a range of numbers of clusters, as well as their validity index values are represented in text trees on which the best number of clusters is highlighted.

He et al. (2004) [73] conducted a systematic study of quantitative cluster comparison using both internal and external criteria in order to give insightful illustrations of various factors that affect clustering quality. In this study, the authors defined three levels of external criteria – cluster entropy, class entropy and overall entropy, and applied them to evaluate the clustering results obtained from k-means, SOM, ART-C (Adaptive Resonance Theory under Constraints) clustering methods under varying number of clusters. The three levels of entropies are calculated based on a predetermined class labeling dimension, and able to identify the “optimal” number of clusters and rank clustering results.

Expression Profiler [119, 89] is an official, open-source clustering comparison and visualization program developed at EBI (European Bioinformatics Institute) [www.ebi.ac.uk/expressionprofiler]. In this program, a new method for comparing various clustering results in either flat or hierarchical structure is presented. For the comparison of two flat clustering results, a bipartite graph is used to express their relationships. The weight of edge between two clusters belonging to two individual
clusterings is proportional to the number of overlapped records in both clusters. The display order of cluster nodes in each clustering is properly rearranged by using gravity-center algorithm, so that the number of crossing edges is minimized.

Fig. 2.3-8 demonstrates the concept about how two clustering results are laid out to minimize edge crossing and merge “superclusters”. Fig. 2.3-9 displays the comparison of a hierarchical clustering result and a k-means clustering result on a real life dataset in Expression Profiler.

SeqExpress [www.seqexpress.com], developed by Boyle (2004) [19], is an integrated analysis and desktop visualization program for gene expression experiments. In this program, cluster comparison and evaluation can be conducted through Cluster Comparison visualizations. Like Expression Profiler, Cluster Comparison tool uses bipartite graph to represent two partitions to be compared. The two partitions could both be clustering results. Or one is a clustering result, and another is a set of biologically relevant categories, e.g. cell cycles, disease states, or regulatory networks information etc. Fig. 2.3-10 is an example showing the comparison of a set of clusters against a prediction of yeast gene regulatory modules.

Cvek (2004) [29] proposed a visualization technique, called SM (Stable Matrix) visualization, to perform cluster stability analysis. Stable Matrix is actually the graphical representation of a symmetric 2D matrix in heatmap or color mosaic visualization. This matrix is a properly ordered CDM (C-Dimension Measure) matrix [29] or co-association matrix [50]. Both dimensions of SM matrix are record indices in the same order and with size $n \times n$, where $n$ is the total number of records. The cell value of this matrix is
calculated by SM \((i, j) = \frac{n_{ij}}{N}\); where \(N\) is the total number of clustering results involved in the computation, and \(n_{ij}\) is the number of clustering results which recognize the pair of records \((i, j)\) staying in the same cluster. The record index order in SM is arranged by using CDSM (C-Dimension Stability Measure) algorithm. This algorithm is developed based on the concept of pCluster model [135] by which records are grouped in terms of their memberships among a set of clustering results. Since both CDM and CDSM techniques are based on the comparison of multiple clustering results in voting fashion, SM visualization is a good technique to perform visual cluster stability analysis and help building-up clustering ensemble [50]. As shown in Fig. 2.3-11 for Iris dataset, the record stability status (yellow as high and blue as low) and outliers located in vertical or horizontal overlapping regions are well observed.

**Fig. 2.3-8** (a) Weighted bipartite graph representing two clusterings A and B of the same dataset. (b) Redraw of the graph after minimizing edge crossing (c) Identification of “superclusters” [119]
**Fig. 2.3-9**  Screenshot of Expression Profiler for comparison of a hierarchical clustering and a k-means clustering on a real life dataset [119]

**Fig. 2.3-10**  Cluster Comparison visualization [19]
Fig. 2.3-11  Stable Matrix for Iris dataset with 8 runs on k-mean and the number of clusters as 8 for each run [29]
2.4 Parallel Coordinates and Extensions

Parallel coordinates [80] is a graphical data representation, in which variable coordinates, the axes starting from minimum to maximum, are laid out in parallel, and data items are represented by polylines connecting coordinate points of them on all axes. Fig. 2.4-1 illustrates parallel coordinate projection in contrast to scatterplot.

With many preferential features, parallel coordinates plot has been widely recognized as one of the best visualizations for high-dimensional data, and adopted in various applications for data representation, analysis and exploration. These features include

- Capability for large dimensionality (no limit for dimensionality in some visualization tools in which dimensional zooming function is available, e.g. the UVP parallel coordinates [56])
• No dimensional transformation and regression ("non-projective mapping between N-Dimensional and 2-Dimensional sets" [79]), so that the visual data representation is intuitive and easy to interpret. This feature also leads to high computational efficiency of parallel coordinates.

• Single display rather than multiple to represent whole data space [20].

• Both numerical and categorical dimensions are handled as well. No concerns about the mapping of categorical data to numerical representations [10].

• Elegant duality properties with ordinary Cartesian coordinates make data mapping unique, so that the data projections directly reflects data’s statistical features [124].

• Intuitive indication of intra-cluster compactness, inter-cluster separation and data correlation in whole space and subspace simultaneously.

![Image of severe occlusion and clutter in regular parallel coordinates](image)

**Fig. 2.4-2** Severe occlusion and clutter in regular parallel coordinates for a dataset containing 5 dimensions and 3848 data items [124]
Although regular parallel coordinate plot has no problem with any dimensionality, it loses its effectiveness as the dataset size (length) becomes large due to a large amount of overlapping data and crossing polylines, namely occlusion and clutter problem. With this problem, hundreds and thousands of data items can hardly be distinctive since the polylines merge into a blob (see Fig. 2.4.2). In the past two decades, a variety of techniques were developed to address this problem.

Wegman and Miller (122, 123) took the ASH (Averaged Shifted Histograms) [108] ideas to visualize line densities within parallel coordinates. The ASH technique attempts to properly choose bins, and represents these bins with density estimates when computing the frequency histograms. Wegman and Luo (1996) [124] presented two strategies to extend parallel coordinate density plot in order to ease the identification of major clusters. The first strategy is to adjust color hue, saturation and brightness values in...
terms of line density, so that the vivid projection image reveals the presence of clusters (see Fig. 2.4.3). And the second strategy is by filtering out sparse areas while emphasizing dense areas where clusters are assumed existing (see Fig. 2.4-4). In similar strategies, Artero, Oliveira and Levkowitz (2004) [3] proposed a unique and efficient algorithm to filter out low frequency or low density data and draw frequency plot and density plot within parallel coordinates. The proposed algorithm is based on image processing techniques such as grayscale manipulation and thresholding, and works in integer arithmetic on discrete raster representation of parallel coordinates for achieving computational efficiency.

Fua et al. (1999) [53] presented a multi-resolution view of data in a novel extension to parallel coordinates, known as hierarchical parallel coordinates, to convey aggregation information from hierarchical clustering results. As shown in Fig. 2.4-5, this technique shows populations and extents of clusters with variable-width opacity bands. These bands gradually fade from opaque/dense centers to transparent/sparse edges that visually encode cluster information. The thickness of the band across each axis section represents the extents of the cluster in that dimension. The color saturation and brightness of each cluster band indicate proximity in hierarchy. And color hue differentiates the clusters. Within hierarchical parallel coordinates implemented in XmdvTool [121], users can easily see, explore and compare data structures at desired level of details to user-interested data regions through a number of interactive functions for navigation and filtering, such as drilldown/rollup, dimension zooming, structure-based brushing, extent scaling, and dynamic masking. A similar technique to hierarchical parallel coordinates
was developed by Berthold et al. (2003) [13] for visualizing fuzzy clustering results (see Fig. 2.4-6). This technique transparently projects color-encoded bands representing the ranges of classes. The mixture of band colors reveals data clustering structure and association between classes and record attribute values. This visualization provides a better interpretation to overlay of fuzzy rules.

![Hierarchical parallel coordinates plot](image1)

**Fig. 2.4-5** Hierarchical parallel coordinates plot (implemented in XmdvTool) on Iris dataset [121]

![Parallel coordinates plot](image2)

**Fig. 2.4-6** Parallel coordinates for visualizing the fuzzy rules of Iris dataset [13]
As an example of implementing level-of-detail design principal in visualization [23], Novotny (2004) [98] introduced an approach to represent data using different striped textures, as shown in Fig. 2.4-7, to help users distinguish clusters at the user-controlled abstraction level. At different abstraction level, information about record cluster membership, cluster density, distribution and hierarchy is retrieved, and reflected in striped textures colored by encoded information values.

![Figure 2.4-7](image)

Fig. 2.4-7  The Detailed and abstract representations of the same data [98]

To achieve clutter reduction, Andrienko et al. (2004) [2] presented a modified parallel coordinates suitable for cluster comparison and exploration. The proposed technique divides the value range of each variable axis of parallel coordinates into equal
frequency intervals in response to user-desired number of subintervals, and uses “striped envelopes” to represent distributions of the characteristics in classes and the entire dataset. In Fig. 2.4-8, the colored strips give impression about the change of class distribution over all dimensions. In the modified parallel coordinates, the objective classes can be either data predefined, or user-defined in various ways. The class (cluster) comparison can take place by switching the color-by class or launching multiple plots, each on different color-by class.

![Fig. 2.4-8 Attribute value distribution in two classes over all dimensions [2]](image)

Most recently, Johansson et al. (2005) [83] developed an additional extension of parallel coordinates to highlight clustering structure from varying aspects of cluster characteristics. This technique uses high-precision textures to represent clusters. The mathematical expressions of textures encode clustering structure information. Different transfer functions operate on high-precision textures to get various views of them, such that some hidden and interesting data features, such as outliers and trends, can possibly
be revealed. The transfer function maps high-precision textures to different views with the change of opacity values (Fig. 2.4-9). In addition to system pre-defined transfer functions, user-defined transfer functions can also be created and manipulated for supporting exploration needs.

(a) Linear transfer function is applied to provide overview of the data.

(b) Logarithmic transfer function is applied to highlight a cluster with low density.

(c) Square root transfer function is applied; outliers are identified from high-density regions.

**Fig. 2.4-9** The outcomes with different transfer function for the same data [83]
To expand expression extent of parallel coordinates, some techniques add additional graphics on variable axes to display more information. For examples, color-encoded bars [111] (Fig. 2.4-10) and histograms [101] (Fig. 2.4-11) are added along axes to represent distributions of attribute values. Theus (2002) [115] put box plot to each axis to display various statistics about underlying dimension (Fig. 2.4-12). Andrienko et al. (2004) [2] presented ellipse graphics overlapping on viable axes. This “ellipse plots” is an alternative to the “striped envelope plot” (see Fig. 2.4-13(a). The ellipse represents the subinterval; its horizontal diameter is proportional to the subinterval length, and the vertical diameter is proportional to the size of the subsets represented by the ellipse. 2.4-13(b) demonstrates the overlapping effect of parallel coordinates and ellipse plots.

![Fig. 2.4-10. Color-encoded bars for displaying distribution [111]](image)
Fig. 2.4-11. Parallel histogram [101]

Fig. 2.4-12. Parallel boxplots [115]

Fig. 2.4-13. Ellipse plots overlapping with parallel coordinates [2]
Parallel Sets [10, 11] is a new visualization tool adopting axes layout from parallel coordinates and designed for displaying categorical data (Fig. 2.4-14). In this visualization, every categorical dimension is represented as an independent set of frequency-scaled boxes. And data streams representing subsets of categorical subset are drawn across all dimensions. In Parallel Sets, the display orders of frequency-scaled boxes and dimensions can be manually and interactively manipulated for displaying different projection patterns. Parallel Sets presents a new visual metaphor mapping categorical variables to visual entities. This metaphor clearly reveals categorical data frequency and distributions, which can be hardly visualized in regular parallel coordinates. Parallel Sets deals with numerical dimensions by binning and uses histogram boxes to replace frequency-scaled boxes (Fig. 2.4.14b). Parallel Sets is a highly interactive visualization tool, a suite of interactive data reorganization functions are available.

Using the same data representation as Parallel Sets, Interactive Sankey Diagrams (Fig. 2.4-15) is a successful visualization to show quantitative information about the flow of energy [102]. In Fig. 2.4-15, a node on the bottom is selected and the distributions from different branches of the flow graph are highlighted. This visualization tool was designed for usability and visual aesthetics. The issues about node and edge layout, edge drawing order were recognized as surprisingly difficult problem and addressed by presenting grouping and level of detail, as well as sorting approaches to reduce occlusion and crossing.
(a) Three categorical dimensions. Categories are grouped to facilitate the density hierarchies.

(b) Three numerical dimensions and one categorical dimension.

**Fig. 2.4-14** Parallel Sets visualization [10]

**Fig. 2.4-15** Interactive Sankey Diagrams [102]
3 PARTITION STABILITY METRIC

When conducting partition comparison and evaluation, it is important to have objective measures to evaluate partitioning quality in quantitative manner [73]. We previously reviewed a cluster stability metric system used for perturbation-based cluster analysis proposed by Katz and Grinstein (2006) [85]. This system contains stability measures bound to label correspondent relationships at record, cluster and global levels. With these measures, data stability is measured based on iterative comparisons between clustering results of perturbed dataset and the clustering result of the original (unperturbed) dataset. In this research, we extended this system to partition similarity and stability measures, and used them for partition comparison and evaluation. We therefore measured stability based on mutual comparisons among all participant partitions – every partition is compared with others in both directions (A comparing to B and B comparing to A differ in meanings). Our extended metric system works under an assumption similar to the one made by Fred and Jain (2005) [50] for consolidating multiple clustering results into a single and better partition: “patterns belonging to a natural cluster are very likely to be co-located in the same cluster in different data partitions”.

In addition to the partition notation described in the first chapter, we adopted notations below to describe the partition stability metric system.

N: A set of partitions involved in partition comparison;
C: Dimension or partition;

c: Group, a disjoint subset of C.

In order to illustrate computational processes in this dissertation, a synthetic dataset is created, where \( C_1, C_2, C_3 \) are partitions. The partitioning groups for each partition are as follows.

\[
C_1 = \{ c_{11}: \{0, 2, 3, 4, 6, 7\}, c_{12}: \{1, 5, 8, 11\}, c_{13}: \{9, 10, 12, 13, 14, 15\}\};
\]

\[
C_2 = \{ c_{21}: \{2, 6, 7, 8, 11, 14, 15\}, c_{22}: \{0, 1, 3, 4, 5, 9, 10, 12, 13\}\};
\]

\[
C_3 = \{ c_{31}: \{0, 3, 9, 15\}, c_{32}: \{1, 8, 10, 13, 14\}, c_{33}: \{2, 4, 5, 6, 7, 11, 12\}\}.
\]

<table>
<thead>
<tr>
<th>Record ID</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

3.1 Label Correlation Coefficient between Groups

Label correlation coefficient between group \( c_a \) and \( c_b \) of two partitions \( A \) and \( B \) is defined as:
$$\text{corCl}(c_a, c_b) = \frac{|c_a \cap c_b|}{\sqrt{|c_a||c_b|}}$$  \hspace{1cm} (3-1)$$

where $|x|$ represents the size of set $x$. Note that the label correlation coefficient is symmetrical, i.e. $\text{corCl}(c_a, c_b) = \text{corCl}(c_b, c_a)$

Regarding the synthetic dataset, label correlation coefficients between groups of $C_1$, $C_2$ and $C_3$ are displayed in Table 3-2.

**Table 3-2**  
Label Correlation Coefficient Matrix for the Synthetic Dataset

<table>
<thead>
<tr>
<th>Group</th>
<th>$c_{11}$</th>
<th>$c_{12}$</th>
<th>$c_{13}$</th>
<th>$c_{21}$</th>
<th>$c_{22}$</th>
<th>$c_{31}$</th>
<th>$c_{32}$</th>
<th>$c_{33}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_{11}$</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>0.46</td>
<td>0.41</td>
<td>0.41</td>
<td>0</td>
<td>0.62</td>
</tr>
<tr>
<td>$c_{12}$</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>0.38</td>
<td>0.33</td>
<td>0</td>
<td>0.45</td>
<td>0.38</td>
</tr>
<tr>
<td>$c_{13}$</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>0.31</td>
<td>0.54</td>
<td>0.41</td>
<td>0.55</td>
<td>0.15</td>
</tr>
<tr>
<td>$c_{21}$</td>
<td>0.46</td>
<td>0.38</td>
<td>0.31</td>
<td>/</td>
<td>/</td>
<td>0.19</td>
<td>0.34</td>
<td>0.57</td>
</tr>
<tr>
<td>$c_{22}$</td>
<td>0.41</td>
<td>0.33</td>
<td>0.54</td>
<td>/</td>
<td>/</td>
<td>0.5</td>
<td>0.45</td>
<td>0.38</td>
</tr>
<tr>
<td>$c_{31}$</td>
<td>0.41</td>
<td>0</td>
<td>0.41</td>
<td>0.19</td>
<td>0.5</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>$c_{32}$</td>
<td>0</td>
<td>0.45</td>
<td>0.55</td>
<td>0.34</td>
<td>0.45</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>$c_{33}$</td>
<td>0.62</td>
<td>0.38</td>
<td>0.15</td>
<td>0.57</td>
<td>0.38</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>

* Red font represents the maximum value in the row of each block

### 3.2 Label Correspondent Relationship

Label correspondent relationship is the result of solving label correspondence problem [94, 95]. Label correspondence problem is about recognizing the most similar pairs of groups between two partitions. Any cluster similarity or distance measures can be used to identify these pairs, for example, normalized Euclidean distance [85], label correlation coefficient [85] and F-measure [120]. We adopted label correlation coefficient in this dissertation.
Let \( c_a \) and \( c_b \) be two groups of partition \( C_u \) and \( C_v \) respectively. \( C_u \) is called comparison partition, and \( C_v \) is called reference partition. \( c_b \) is defined as the correspondent group of \( c_a \) if pairwise label correlation coefficient

\[
\text{corCl} (c_a, c_b) = \max (\text{corCl} (c_a, c_i)), \; c_a \in C_u; \; c_i \in C_v, \; i = 1, 2, ..., |C_v|
\]  

(3-2)

The relation is expressed as \( c_a \propto c_b \).

For the synthetic dataset, the label correspondent relationship can be built based on Table 3-2. The diagram below (Fig. 3-1) shows the relationship, where arrows represent \( \propto \) relations and double ends arrows indicate that the relations are bi-directional.

![Diagram](image)

**Fig. 3-1** The label correspondent relationship for the synthetic dataset

Based on (3-2), the relation \( \propto \) has three important properties and one feature.

### 3.2.1 Polygamous

One group in comparison partition may have more than one correspondent group in reference partition, i.e. many-to-one. And more than one group in comparison partition may also correspond to the same group in reference partition, i.e. one-to-many.
The polygamy of the relation occurs in three situations:

Let $c_b$ in $C_v$ be a correspondent group of $c_a$ in $C_u$.

(a) $c_b$ is not unique with respect to $c_a$ because more than one group in $C_v$ may hold or be equal to max $(\text{corCl} (c_a, c_i))$, $c_i \in C_v$, $i = 1, 2, \ldots, |C_v|$. For instance, $\text{corCl} (c_{31}, c_{11}) = \text{corCl} (c_{31}, c_{13})$ referring to Table 3-2.

(b) $c_a$ is not unique with respect to $c_b$ because more than one group in $C_u$ may hold or be equal to max $(\text{corCl} (c_j, c_b))$, $c_j \in C_u$, $j = 1, 2, \ldots, |C_u|$.

(c) If the number of groups in $C_u$ is more than that in $C_v$, at least one group in $C_u$ will pair with at least two groups in $C_v$ based on the Pigeonhole principle [136].

3.2.2 Asymmetric

$c_a \propto c_b$ does not imply $c_b \propto c_a$, and vice versa.

The example below (Fig 3-2) illustrates this property. In this figure, the numbers inside blocks represent the numbers of overlapping records.

![Fig.3-2 Illustration of the asymmetric property.]

\[
\text{corCl} (c_{11}, c_{21}) = \frac{3}{\sqrt{(2+3) \times (3+4)}} \approx 0.51;
\]
\[ \text{corCl} (c_{11}, c_{22}) = \frac{2}{\sqrt{(2+3) \cdot (2+4)}} \approx 0.36; \]
\[ \text{corCl} (c_{12}, c_{21}) = \frac{4}{\sqrt{(4+4) \cdot (3+4)}} \approx 0.53; \]
\[ \text{corCl} (c_{12}, c_{22}) = \frac{4}{\sqrt{(4+4) \cdot (2+4)}} \approx 0.58; \]

In terms of above label correlation coefficients, \( \text{corCl} (c_{11}, c_{21}) > \text{corCl} (c_{11}, c_{22}) \)

\[ \text{s.t. } c_{11} \propto c_{21}; \text{corCl} (c_{21}, c_{12}) = \text{corCl} (c_{12}, c_{21}) > \text{corCl} (c_{21}, c_{11}) = \text{corCl} (c_{11}, c_{21}) \text{ s.t. } c_{21} \propto c_{12}. \text{ Therefore, } c_{11} \propto c_{21} \text{ doesn’t imply } c_{21} \propto c_{11} \]

3.2.3 **Intransitive**

\( c_a \propto c_b \) and \( c_b \propto c_c \) does not imply \( c_a \propto c_c \).

The example below (Fig 3-3) illustrates this property. In this figure, the numbers inside blocks represent the numbers of overlapping records

![Fig.3-3 Illustration of the intransitive property](image)

\[ \text{corCl} (c_{11}, c_{21}) = \frac{6}{\sqrt{(2+4+3) \cdot (2+4)}} \approx 0.82; \]
\[ \text{corCl} (c_{11}, c_{22}) = \frac{3}{\sqrt{(2+4+3) \cdot (3+2)}} \approx 0.44; \]
\[ \text{corCl} (c_{11}, c_{23}) = 0; \]
\[ \text{corCl} (c_{21}, c_{31}) = \frac{4}{\sqrt{(2+4) \cdot 4}} \approx 0.82; \]
\[ \text{corCl} (c_{21}, c_{32}) = \frac{2}{\sqrt{(2+4) \cdot (2+3+2)}} \approx 0.31; \]
\[ \text{corCl} (c_{21}, c_{33}) = 0; \]
\[ \text{corCl} (c_{11}, c_{31}) = \frac{4}{\sqrt{(2+4+3) \cdot 4}} \approx 0.66; \]
\[ \text{corCl} (c_{11}, c_{32}) = \frac{5}{\sqrt{(2+4+3) \cdot (2+3+2)}} \approx 0.75; \]
\[ \text{corCl} (c_{11}, c_{33}) = 0; \]
In terms of above label correlation coefficients, $c_{11} \propto c_{21}$, $c_{21} \propto c_{31}$, and $c_{11} \propto c_{32}$.

Obviously, $c_{11} \propto c_{31}$ is not implied.

### 3.2.4 No-correspondence

No-correspondence is a feature of the label correspondence relation. According to polygamous property of the label correspondence relation, some groups in comparison partition do not correspond to any of the groups in reference partition, even though comparison partition has fewer or equal number of groups in reference partition. These groups are called non-correspondent groups.

### 3.3. Dimensional Record Stability

Dimensional record stability (DRS) is the probability of one record falling into its correspondent groups over $N$.

Let $C$ be a partition in $N$, $c$ a group of $C$, $k$ the index of a record which belongs to $c$, $N_k$ the number of partitions in $N - C$ in which the record $k$ falls into the correspondent group of $c$, $N_k \leq |N| - 1$.

The dimensional record stability for record $k$ with regard to $C$ is

$$\text{DRS}(k, C) = \frac{N_k}{|N| - 1} \quad (3-3)$$

The value of DRS is in the range $[0, 1]$. The value zero occurs when a record does not fall into either correspondent group of its own.
Regarding the synthetic dataset, dimensional record stability for $C_1$, $C_2$ and $C_3$ is calculated and shown in Table 3-3.

3.4 Record Stability

Record stability (RS) is the average of dimensional record stabilities in $N$.

$$\text{RS}(k) = \frac{1}{|N|} \sum_{i=1}^{|N|} \text{DRS}(k, C_i)$$  \hspace{1cm} (3-4)

Based on the definition above, record stability is interpreted as the estimate of probability of a record staying stable in $N$. Its value is in the range $[0, 1]$. The record stability for a record is zero when no more than one partition considers this record as maintaining the same membership.

Regarding the synthetic dataset, record stability is calculated and shown in Table 3-4.

<table>
<thead>
<tr>
<th>Record ID</th>
<th>DRS($C_1$)</th>
<th>DRS($C_2$)</th>
<th>DRS($C_3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>12</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>15</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Group Stability under Threshold

Group stability under threshold (GST) is the percentage of the number of records whose dimensional record stability values exceed a threshold value over the total record number in a partitioning group.

Let $C$ be a partition, $C \in \mathbb{N}$, $c$ a group of $C$, $T$ a threshold of dimensional record stability value, $0 \leq T \leq 1$, and $N_T$ the number of records in $c$ whose dimensional record stability values exceed $T$.

$$GST\ (T,\ c) = \frac{N_T}{|c|}$$

GST value is between 0 and 1.

Regarding the synthetic dataset, GST value with $T = 0.9$ is calculated for each group of $C_1$, $C_2$ and $C_3$, as shown in Table 3-5.

<table>
<thead>
<tr>
<th>Record ID</th>
<th>RS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>0.33</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0.33</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.83</td>
</tr>
<tr>
<td>11</td>
<td>0.5</td>
</tr>
<tr>
<td>12</td>
<td>0.33</td>
</tr>
<tr>
<td>13</td>
<td>0.83</td>
</tr>
<tr>
<td>14</td>
<td>0.33</td>
</tr>
<tr>
<td>15</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table 3-4 RS Values
3.6 Group Stability in Average

Group stability in average (GSA) is the average dimensional record stability values for a partitioning group.

Let $C$ be a partition, $C \in \mathbb{N}$, $c$ a group of $C$,

$$GSA(c) = \frac{\sum_{k \in c} DRS(k, C)}{|c|}$$  

(3-6)

GSA value is between 0 and 1.

Regarding the synthetic dataset, GSA value is calculated for each group of $C_1$, $C_2$ and $C_3$, as shown in Table 3-6.

<table>
<thead>
<tr>
<th>Group</th>
<th>GST (0.9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_{11}$</td>
<td>0.33</td>
</tr>
<tr>
<td>$c_{12}$</td>
<td>0</td>
</tr>
<tr>
<td>$c_{13}$</td>
<td>0.5</td>
</tr>
<tr>
<td>$c_{21}$</td>
<td>0.14</td>
</tr>
<tr>
<td>$c_{22}$</td>
<td>0.22</td>
</tr>
<tr>
<td>$c_{31}$</td>
<td>0.75</td>
</tr>
<tr>
<td>$c_{32}$</td>
<td>0.4</td>
</tr>
<tr>
<td>$c_{33}$</td>
<td>0.14</td>
</tr>
</tbody>
</table>

3.7 Relative Partition Similarity

Relative partition similarity (RPS) is the average correlation coefficients for a pair of partitions based on the label correspondent relationship.
Let $C_u$ be reference partition, $C_v$ comparison partition, and $G(C_u)$ the number of groups in $C_u$.

$$\text{RPS} (C_u, C_v) = \frac{1}{G(C_u)} \sum_{c_a \in C_u, c_b \in C_v} \text{corCl} (c_a, c_b)$$  \hspace{1cm} (3-7)$$

Note that this similarity measure is asymmetric, i.e. $\text{RPS} (C_u, C_v)$ may not be equal to $\text{RPS} (C_v, C_u)$ because of the asymmetric property of the label correspondence relation.

Regarding the synthetic dataset, RPS value is calculated for each pair of partitions and shown in Table 3-7.

<table>
<thead>
<tr>
<th>Table 3-7</th>
<th>RPS Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition</td>
<td>$C_1$</td>
</tr>
<tr>
<td>$C_1$</td>
<td>1</td>
</tr>
<tr>
<td>$C_2$</td>
<td>0.5</td>
</tr>
<tr>
<td>$C_3$</td>
<td>0.53</td>
</tr>
</tbody>
</table>

### 3.8 Partition Similarity

Partition similarity (PSM) for a partition is the average RPS of this partition relative to the rest in $N$.

$$\text{PSM} (C) = \frac{1}{|N| - 1} \sum_{i=1}^{[N]} \text{RPS} (C, C_i), \hspace{1cm} C \neq C_i$$  \hspace{1cm} (3-8)$$

Regarding the synthetic dataset, PSM value is calculated for each partition and shown in Table 3-8.

<table>
<thead>
<tr>
<th>Table 3-8</th>
<th>PSM Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition</td>
<td>PSM</td>
</tr>
<tr>
<td>$C_1$</td>
<td>0.5</td>
</tr>
<tr>
<td>$C_2$</td>
<td>0.52</td>
</tr>
<tr>
<td>$C_3$</td>
<td>0.52</td>
</tr>
</tbody>
</table>
As we reviewed previously, parallel coordinates and its various extensions are effective for visualizing data pattern and clusters. It also has great potential to play important roles in comparative cluster analysis thanks to its intuitiveness and feature showing parallel axes. When data variables and partitions are projected in parallel coordinates, it allows us conveniently exploring data at record, cluster and dimension levels within a single graphic display. The single display feature is especially valuable to comparative cluster analysis since it helps saving screen space and enhances difference recognition [1, 2]. However, due to inherent occlusion and clutter problem in parallel coordinates, many data characteristics become invisible.

In order to seek a visualization tool more suitable for comparative cluster analysis, we developed CComViz (Cluster Comparison Visualization) visualization. CComViz has the same data representation model as Parallel Sets [10, 11] and Interactive Sankey Diagrams [102]. Fig. 4-1 uses several records from Car dataset to depict how data are projected in CComViz. In CComViz, one of the most important visual properties is called hot dimension. The hot dimension, like Maker dimension in Fig. 4-1, is a selected dimension by which record lines are colored. Any active dimension can be selected as the hot dimension. Choosing hot dimension provides a convenient way to visualize data flows, density distributions and hierarchies starting from different dimensions.
Like parallel coordinates, CComViz lays dimension axes out in parallel. But these axes are not coordinates. Every axis is an expression of individual record display order and represented as a set of continuous density bars. Each bar corresponds to a partitioning group with its length scaled according to its density in dimension. In CComViz, data items are also represented as polylines that link their points on every axis. The point position of each data item on axis is determined by its rank in the record display order. Since each data item has a unique rank on axis, no more than one data item is projected to the same point on each axis (without considering the rounding of projection position to screen pixel). Therefore, no more than one data item is projected to the same polyline. This feature benefits in particular categorical dimensions in contrast to regular parallel coordinates, in which data items with the same category value are projected to the same point on axis.

CComViz is designed to perform partition comparison and evaluation. It contains a sophisticated layout algorithm to rearrange record and dimension display order in terms
of a label correspondent relationship, selection of the hot dimension, and partition
stability measures. The mechanism behind this algorithm is based on consideration and
utilization of user perception modeling, partition stability measurement, and interactive
functions involved in the course of data exploration and pattern recognition. The
methodology used in the algorithm design makes CComViz an ideal tool to visualize data
stability, data flow, density distribution and hierarchy, as well as data correlation across a
number of partitions within a single graphic display.

Unlike parallel coordinates, CComViz focuses on achieving visual analytics
rather than geometrical data mapping. The observation of geometrical data properties in
CComViz may be disabled, meaning that the position of record projection has no direct
connection to its value. Although this may inadvertently impact on numerical dimensions,
by means of data and tool linking mechanism, as in the UVP [56], this disadvantage can
be compensated through using other visualization tools. As a matter of fact, parallel
coordinates and CComViz share many common characteristics and interfaces. Using both
visualization tools in an interactive and linked environment can generate a strong and
unified power for comparative cluster analysis.

4.1 Why Record and Dimension Display Order Rearrangement

Good data representation with minimum clutter and occlusion presents challenge
to every visualization tool. The feature, that no more than one record is projected to the
same point (without consideration of rounding projection point to screen pixel) on each
axis, avoids record overlap in CComViz. But clutter due to crossing remains concern. Fig.
4-1 illustrates that it will make it difficult to observe data pattern without proper record and dimension display order rearrangement, even for only several hundreds of records.

From visualization point of view, visually representing data with less clutter is not enough. Data representation should amplify human cognition [23]. Now that CComViz is intended to display data stability, density distribution and hierarchy across multiple partitions, the record and dimension display order needs to enhance readability of these data features so cognition can be amplified.

Information visualization is fundamentally dependent upon the properties of human perception. According to visualization study [23], grouping, sorting, and positioning are among the most effective ways to support human perceptual inference, enhance pattern detection, as well as reduce information search time and memory usage. Another influential factor in human perceptual inference is related to semantics. As CComViz is designed to distinguish stable and unstable records, we expect their visual appearances being different. Relying on semantic associations, to make stable records look smooth and unstable records fluctuate can help with human perceptual inference. In CComViz, in order to achieve the match between data representation model and human perception model, these concepts are extensively used to develop a layout algorithm for record and dimension display order rearrangement.

The methodology used in developing the algorithm is based on considerations and expectations illustrated in Fig. 4-2. This figure is a CComViz projection of Iris dataset. It includes nine clustering results generated by Kmean, hierarchical, and SOM clustering algorithms with the number of clusters at 3, 4 and 5. In this figure, Classes dimension is
selected as the hot dimension. The color of record line, therefore, indicates type of Iris flower. Each dimension axis has an independent color legend for its own partitioning groups.

**Fig. 4-2** Considerations and expectations in developing the algorithm for record and dimension display order rearrangement

As many considerations and expectations are involved, it is difficult to satisfy all of them. Our methodology takes an approach to reaching a reasonable compromise.

In CComViz, the process for rearranging record and dimension display order is in a pipeline sequence. The computation of record display order rearrangement is based on dimension display order. Any change in adding, removing or reordering dimensions will cause a change in the record display order.
4.2 Dimension Display Order Rearrangement

For partition comparison and evaluation, it is especially important to identify distinctions and variations between similar partitions. The similarity is defined by different metrics, such as clustering criteria and partitioning consistency. Since similar partitions have a high degree of partitioning consistency, there should be fewer crossing lines if similar partitions are adjacent and the display orders of partitioning groups along their axes maintain a correspondent relationship (To be discussed later). Therefore, from visual aesthetics point of view, putting similar partitions together assists the observation of outliers or odd records while from human visual perception point of view, it facilitates comparison and perceptual inference.

As the layout of dimension axes in CComViz is a linear sequence, there are two approaches to dimension display order rearrangement, which in CComViz is an optional function. By default, the dimension display order is arranged by users to achieve flexibility.

4.2.1 Sorting Approach

Cluster evaluation through sorting by criteria is common in comparative cluster analysis. CComViz provides a natural way to visualize sorted clustering results. In general, any kind of clustering criteria or clustering quality measures can be used for sorting. Due to their varying characteristics, the comparison results can be either consistent or inconsistent. For example, data stability pattern may not be observed better through sorting by internal criteria than by external or relative criteria, since internal
criteria are not intended to measure data stability. However, it is still worthwhile to project CComViz in different sorting orders to explore partitioning features and verify partitioning reliability from various perspectives.

As one of CComViz projection goals is to minimize crossing, the stability-based criteria, such as our proposed PSM, are ideal for that purpose. Regarding the synthetic dataset displayed in Table 3-1, in terms of sorting by PSM (Table 3-7), the dimension display order is \(<C_2, C_3, C_1>\).

4.2.2 Sequential Selection Approach

Sequential selection is a process for rearranging the dimension display order by using relative partition similarity measure. Once the first dimension which user thinks the best or the most interesting in terms of certain criteria is given, a dimension that is most similar to current dimension among all unselected candidate dimensions is chosen as the immediate successive dimension. Our proposed RPS is one of this kind of similarity measures.

In the synthetic dataset, when \(C_2\), which has higher PSM value over others, is chosen as the first dimension, the dimension display order by sequential selection using RPS (Table 3-6) remains to be \(<C_2, C_3, C_1>\).

The dimension display order rearrangement is capable of improving not only visual aesthetics but also comparison result. As we previously discussed, the stability-based partition comparison and evaluation work in voting fashion. When too many inferior partitions or divergent partitions are involved in comparison, the results will not
be reliable. By visually checking partition quality or similarity, and throwing inferior or divergent partitions away from the sorted sequence, the results of next run on the trimmed series of partitions will be improved. In summary, a proper dimension display order makes users focus more on high quality or interesting partitions when many partitions are involved in comparison and evaluation. It also leads to progressive refinement of comparison results.

4.3 Record Display Order Rearrangement

To some extent, record display order rearrangement in CComViz is similar to the Bipartite Graph Crossing Minimization (BGCM) problem [131]. A BGCM algorithm, called gravity-center algorithm, has been successfully used in Expression Profiler visualization tool [119] to reduce crossing when two clustering results are compared. However, a distinction exists: BGCM algorithms deal with one pair of partitions, whereas record display order rearrangement in CComViz deals with unlimited number of partitions. Our methodology does not directly treat record display order rearrangement as a BGCM problem; instead it implicitly achieves crossing minimization. As discussed previously, reducing crossing is not the only goal of CComViz visualization. Other goals include facilitating observation of data stability, density distribution and hierarchy, as well as other interesting relationships. To achieve these goals, we take the grouping and reordering approach to developing a layout algorithm that enforces the formation of record projections in envelope (band) shapes with fewer crossings [53, 2, 10, 11].
This algorithm contains five steps and involves operations of four levels of record grouping and one reordering of partitioning groups. These operations require information from the label correspondent relationship which should be done in data pre-processing.

We now use the synthetic dataset to illustrate the steps of the algorithm. For this dataset, the label correspondent relationship has been built and is shown in Fig. 3-1. Fig. 4-3 is the data projection before rearranging the record display order. In this figure, the dimension display order \(<C_2, C_3, C_1>\) follows PSM (Partition Similarity) sorting; \(C_2\) is selected as the hot dimension; and the number inside each block represents record index.

At the beginning, records in every dimension are arranged in their original order.

![Fig 4-3 Projection with the original record display order. Records are colored by \(C_2\)](image)

The first step is to group records within each dimension, as shown in Fig. 4-4. Notice that the record display order of each group needs to keep sorted by record indices.

This rule also applies to sub-groups in the rest of steps.
The second step is to reorder partitioning groups within each dimension, sequentially starting from the first dimension (leftmost) to the last dimension (rightmost).

In Fig. 4-5, the display order of partitioning groups of the first dimension follows the sorting order by their GSA (Group Stability in Average) values. The top position corresponds to the group that has the highest GSA value. As in the synthetic dataset, $\text{GSA}(c_{21}) = 0.5 > \text{GSA}(c_{22}) = 0.44$ (Table 3-5), the group display order of $C_2$ is $\langle c_{21}, c_{22} \rangle$ (from top to bottom). And now starting from the second dimension to the last dimension, the group display order of every dimension is arranged by choosing correspondent pairs versus the immediate previous dimension in terms of the label correspondent relationship.

The matching processes follow the group display order of the immediate previous dimension starting from the top in greedy fashion. Extra groups, if exist, after the
matching are appended to the end and their display order still follows the sorting order by their GSA values.

As in the synthetic dataset, since \( c_{21} \propto c_{33} \), \( c_{22} \propto c_{31} \), for \( C_{3} \) \( c_{33} \) should correspond to \( c_{21} \) and \( c_{31} \) should correspond to \( c_{22} \). \( c_{32} \) is the only extra group left and it is put at the bottom. So, the group display order of \( C_{3} \) is \( < c_{33}, c_{31}, c_{32} > \) versus \( < c_{21}, c_{22} > \).

For situations where many-to-one pairs occur, the following rules are applied.

**Precedence Rule:** Given \( c_{ai} \propto c_{b} \), \( c_{aj} \propto c_{b} \), where \( c_{ai}, c_{aj} \in C_{a} \), \( \text{corCl}(c_{ai} \propto c_{b}) < \text{corCl}(c_{aj} \propto c_{b}) \), \( c_{aj} \) has higher precedence to pair with \( c_{b} \).

**Replacement Rule:** When \( c_{ai} \propto c_{b} \) but another group in \( C_{a} \) has higher precedence to pair with \( c_{b}, c_{b} \) will not pair with \( c_{ai} \). Instead, another available group in \( C_{b} \) whose label

---

**Fig 4-5** Step 2: Reordering of partitioning groups, sequentially starting from the first dimension to the last dimension in terms of the label correspondent relationship.
correlation coefficient value with \( c_{ai} \) is the next highest will be chosen as a replacement to
\( c_b \) to pair with \( c_{ai} \).

As in the synthetic dataset, from the relations: \( c_{31} \propto c_{11}, c_{31} \propto c_{13}, c_{32} \propto c_{13}, c_{33} \propto c_{11} \), we see that either \( c_{11} \) or \( c_{13} \) may pair with \( c_{31} \). But by checking \( \text{corCl}(c_{31}, c_{11}) = 0.41 < \text{corCl}(c_{33}, c_{11}) = 0.62 \), and \( \text{corCl}(c_{31}, c_{13}) = 0.41 < \text{corCl}(c_{32} \propto c_{13}) = 0.55 \), \( c_{11} \) should pair with \( c_{33} \) and \( c_{13} \) should pair with \( c_{32} \). Only is \( c_{12} \) left as replacement to pair with \( c_{31} \). Finally, the group display order of \( C_1 \) is \( <c_{11}, c_{12}, c_{13}> \) versus \( <c_{33}, c_{31}, c_{32}> \).

The third step is to perform a second level grouping based on the hot dimension. The display order of the sub-groups after the grouping follows the group display order of the hot dimension. For example, after this operation, \( c_{33} \) is split into two groups \( \{2, 6, 7, 11\} \) and \( \{4, 5, 12\} \) (Fig. 4-6). The first group belongs to \( c_{21} \) and the second group belongs to \( c_{22} \). The display order of sub-groups in \( c_{33} \) is \( <\{2, 6, 7, 11\}, \{4, 5, 12\}> \) (upper to lower), which follows the group display order of the hot dimension \( <c_{21}, c_{22}> \).
The last two steps are third and fourth level grouping based on the immediate next dimension and the immediate previous dimension respectively. After grouping, the display orders of sub-groups follow the group display orders of the immediate next dimension and the immediate previous dimension respectively. For example, after the third level grouping, a sub-group of $C_3$ and its display order are $\langle \{4\}, \{5\}, \{12\} \rangle$, which follows the group display order of $C_1$: $\langle c_{11}, c_{12}, c_{13} \rangle$; after the fourth level grouping, a sub-group of $C_1$ and its display order are $\langle \{12\}, \{9\}, \{12, 13\} \rangle$, which follows the group display order of $C_3$: $\langle c_{33}, c_{31}, c_{32} \rangle$. Fig. 4-7 and Fig. 4-8 show the results after completing these two steps.
**Fig 4-7** Step 4: Third level grouping based on the immediate next dimension

**Fig 4-8** Step 5: Fourth level grouping based on the immediate previous dimension
With these steps, the layout algorithm behaves like an undo-sorting from completely sorting orders where there is no crossing at all to rearranged orders where crossing reduction and visual aesthetics are achieved. In contrast to Fig. 4-3, Fig. 4-9, although a few of crossings, more clearly reveals information which is hidden before. For example, in Fig. 4-9, the complicated hierarchies related to records 4, 5, 12, 0, 3, 9, 10 and 13 are clearly displayed; Smooth record lines for records 2, 6, 7 10 and 13 well reflect their high stability. Although the record display order in Fig. 4-9 does not minimize crossing, for example, the sub-order <12, 9, 15, 14> instead of <15, 14, 12, 9> in C_1 may reduce a small number of crossings, it is not as important as grouping 12, 9, 10 and 13 together to reveal density distribution regarding C_2 in c_{13}. It is these effects and features that we attempt to achieve using our novel methodology.
4.4 Other Modes for Record Display Order Rearrangement

The layout algorithm for the record display order rearrangement is applied only in CComViz automatic mode. Besides this mode, several other modes are also available for exploration needs. With these modes, users may freeze a previous display order during continuous play or rearrange the record display order for each dimension separately. Fig. 4-10 is an example that demonstrates CComViz’s diverse approaches to exploring data.

Fig. 4-10 contains CComViz projections of Iris dataset with a selection of dimensions, including the original Classes dimension and several clustering results. In Fig. 4-10 (a) and (b), the record display order of every group in each dimension is individually sorted by silhouette width. This mode allows users easily investigating how record silhouette values change over clustering results in contrast to automatic mode, as in Fig. 4-10 (c), where displaying record stability is in the first priority. Since the dimension display order is sorted by dimension silhouette index, silhouette plots in Fig. 4-10 (a) present good visual aesthetics for the comparison of clustering quality.

4.5 Record Rendering Layer

For most real-world datasets, record occlusion in CComViz can not be avoided even with the help of proper dimension and record display order rearrangement. Under certain circumstances, transparent drawing provides helps for small datasets. But for large and complex datasets, visual aesthetics in transparent drawing can not be achieved. In this situation, the control of record rendering layers offers an ideal and flexible solution. With this control, multiple record rendering layers are present in the order of
Fig 4-10  CComViz projections of Iris dataset in Silhouette and automatic modes

(a) Silhouette plots. Record lines are rendered in transparent layer

(b) Silhouette plots are removed but the record display order remains unchanged

(c) Record display order is rearranged in automatic mode

Color By Classes:
- Iris-setosa
- Iris-virginica
- Iris-versicolor

Dimension Names:
- Classes: Original category
- KM-3: Kmeans clustering with 3 clusters
- KM-4: Kmeans clustering with 4 clusters
- SOM1x3: SOM clustering with grid 1 x 3
- SOM2x2: SOM clustering with grid 2 x 2

Fig 4-11  Illustration of record rendering layers. The hot group is Iris-versicolor on upper projection and Iris-virginica on lower projection

Dimension Names:
- Classes: Original category
- KM-3: Kmeans clustering with 3 clusters
- KM-4: Kmeans clustering with 4 clusters
- KM-5: Kmeans clustering with 5 clusters
- SOM1x3: SOM clustering with grid 1 x 3
- SOM2x2: SOM clustering with grid 2 x 2
- SOM1x5: SOM clustering with grid 1 x 5
interestingness and the most user-interested layer is rendered on the front. In CComViz, this control is implemented through the selection of hot group. A hot group is a partitioning group of the hot dimension. The records belonging to the hot group are brought to the front rendering layer. The selection of a hot group is made by clicking the mouse on a density bar which represents the hot group. With a sequence of selections, the order of layers is historically stored. As a result, the projection image looks like a stack of multiple record group layers.

Fig. 4-11 illustrates this effect using Iris dataset. In this figure, the dimension display order is rearranged by using sequential selection by RPS.

4.6 Change of Hot Dimension

The operation of changing the hot dimension is in the same way as changing record rendering layer by clicking mouse on the designated dimension axis. Since record display order rearrangement relies on the selection of the hot dimension, the rearrangement is recomputed once the hot dimension is changed. As an effective interactive function, changing the hot dimension during data exploration provides an easy way to investigate data flows, density hierarchies and distributions starting from different dimensions. Fig. 4-12 displays data flow and density hierarchies by selecting different hot dimension.

4.7 Boolean Operation for Record Selection

Boolean operations AND, OR, NAND, and NOR are implemented to make combined record selections in CComViz. They are useful to identify interesting records,
such as stable/unstable records, outliers, anomalies, etc. As shown in Fig. 4-12, the record selection is made by AND operation on selected subsets in KM-5, SOM-1x5, and KM-4 dimensions.

**Fig 4-12** Data flows and hierarchies in different selection of the hot dimension.

**Fig 4-13** Illustration of AND Boolean operation on blocks A, B and C for record selection. The visual property configuration in this figure is the same as in Fig. 4-11
4.8 Handling of Dimension with Continuous Value

As we discussed at the beginning, dimensions with continuous values can be treated as partitions after binning. In CComViz, binning operations can be performed in terms of user specified bin number prior to adding a dimension to the active dimension list. Fig. 4-14 shows the four attribute dimensions and Classes category of Iris dataset in CComViz. The bin number for each attribute dimension is equally 5. From this single projection, we can easily see attribute hierarchies regarding class type.

![Fig 4-14 CComViz projection of Iris dataset. The numbers of bins for attribute dimensions are equal to 5.](image)

4.9 Addition of Record Stability Dimension

Record stability (RS) is a data characteristic measure at record level. A vector of RS measurements covering all records creates a new dimension. In CComViz, this dimension is calculated in terms of active dimensions and can be appended to the original dataset. Under the UVP linking environment, this dimension can be added to any visualization tools for further exploration and analysis. Fig. 4-15 includes the RS
dimension and a set of clustering results based on which the RS dimension is calculated.

From this figure, we can clearly see that all of setosa flowers are stable and most of unstable records occur in versicolor flowers.

![Graph showing dimension and clustering results](image)

**Fig 4-15** RS dimension and a set of partitions based on which the RS dimension is calculated

### 4.10 Computational Complexity

The major cost of the layout algorithms for the dimension and record display order rearrangement comes from set union, set intersection, and set difference operations. None of them is excessively costly. The overall computational time complexity for this algorithm is in $O(M \times C^2)$, where $M$ is the number of record, and $C$ is the number of active dimensions.
5  DISCUSSION OF STABILITY

The definitions of stability by different authors differ in meaning. But one thing is in common: stable records within clusters share consistent characteristics. By our definition, stability is a measure about the degree of agreement that multiple partitions recognize data items compliant with a label correspondent relationship. Stable records are a set of dense groups which are mostly compliant with a label correspondent relationship, whereas unstable records are a set of sparse groups which are rarely compliant with a label correspondent relationship. A record’s compliance with a label correspondent relationship means the compliance of the record’s memberships in partitions with the label correspondent relationship. For example, assume a record’s memberships in three partitions are \( c_{11}, c_{21} \) and \( c_{31} \), if all relations \( c_{11} \propto c_{21}, c_{21} \propto c_{11}, c_{11} \propto c_{31}, c_{31} \propto c_{11}, c_{21} \propto c_{31} \) and \( c_{31} \propto c_{21} \) hold in the label correspondent relationship, this record has the strongest compliance with the label correspondent relationship and is, therefore, the most stable record. This chapter attempts to give intuitive interpretation to the definition for better understanding of the stability.

The idea of visual illustration demonstrated in Fig. 5-1 was originally proposed by Skupin [112] for evaluating clustering solutions. We adopt it for interpreting the stability. Suppose the data items in the synthetic dataset (Table 3-1) are projected in 2D space via certain transformation (Fig. 5-1). When putting the three partitioning results
overlapping, we get 12 disjoint subsets which are intersections of partitioning groups. From this figure, we may notice that the stable records, including 2, 6, 7, 10 and 13, in terms of their RS values are within dense subsets, and their membership $c_{11}$, $c_{21}$ and $c_{31}$ for 2, 6, 7, as well as $c_{12}$, $c_{22}$ and $c_{32}$ for 10, 13 strongly comply with the label correspondent relationship. This in turn suggests an interpretation: stable records are within dense regions that are intersections of partitioning groups mostly compliant with a label correspondent relationship, whereas unstable records are within sparse regions that

**Fig 5-1** Overlapping of three partitions of the synthetic dataset
are intersections of partitioning groups rarely compliant with a label correspondent relationship. Because the number of data items in the synthetic dataset is small, Fig. 5-1 is not perfect to interpret the stability. A better illustration can be seen later when this idea is applied to Iris dataset.

Although the partition stability metric is extended from the perturbation-based cluster stability metric, it has a new usage, which is more similar to that of CDM measure [29]. So, for interpretation of the stability, it makes more sense to compare results with the same dataset between CDM measure and RS measure.

Table 5-1 is the CDM matrix for the synthetic dataset (Table 3-1). Before the comparison, one problem has to be solved. Since Table 5-1 is a matrix and the RS dimension (Table 3-4) is a vector, we need to find a way to make them comparable.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5-1 CDM matrix for the Synthetic Dataset
Referring to Table 5-1, the numbers of each row represent a survey of similarity of the correspondent record among whole record population. According to surveys, if a record shows even similarity to whole population, this record is unstable because it does not have tendency. If a record shows strong similarity to part of records and no similarity to others, this record is stable because it insists on coherence or consistency. This variance can be well captured by standard deviation. Therefore, we may induce that high standard deviation of a row indicates the correspondent record has strong tendency to be grouped together with part of records in the row. In other words, the record with high standard deviation of its row is considered as stable. So, verifying the correlation between the normalized standard deviation dimension of the CDM matrix and the RD dimension is a way to make connection between the two measures, which help interpretation of the stability. Table 5-2 contains the NSD (Normalized Standard Deviation) and RS values for each record.

<table>
<thead>
<tr>
<th>Record ID</th>
<th>NSD</th>
<th>RS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.487</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>0.373</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.487</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>0.597</td>
<td>0.33</td>
</tr>
<tr>
<td>5</td>
<td>0.131</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0.373</td>
<td>0.33</td>
</tr>
<tr>
<td>9</td>
<td>0.703</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.805</td>
<td>0.83</td>
</tr>
<tr>
<td>11</td>
<td>0.597</td>
<td>0.5</td>
</tr>
<tr>
<td>12</td>
<td>0.131</td>
<td>0.33</td>
</tr>
<tr>
<td>13</td>
<td>0.805</td>
<td>0.83</td>
</tr>
<tr>
<td>14</td>
<td>0.373</td>
<td>0.33</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0.33</td>
</tr>
</tbody>
</table>
With Pearson's correlation coefficient $r = 0.89$ and Spearman Rank correlation $r_r = 0.89$ [statistic online calculator: www.wessa.net], the NSD dimension and the RS dimension are considered as significantly correlated. This result encourages us to assert that there exists correlation between these two measures (Proof is beyond this research scope). Another method to interpret stability is by way of using different visualization tools under a linking environment to check data projection. This method can also be used to verify visual aesthetics of CComViz.

Fig. 5-2 – Fig. 5-6 are data projections of full records, stable records and unstable records in Iris dataset in four visualization tools. RS value is computed in terms of $KM-3$, $KM-4$, $KM-5$, $SOM1x3$, $SOM2x2$, $SOM1x5$ clustering results and Classes dimension. The stable and unstable records are defined by RS thresholds. The thresholds in this particular case are 0.8 for stable records and 0.5 for unstable records. These records are selected via the RS dimension within parallel coordinates in Fig. 5-2. This figure clearly shows all data attribute values within a single graphic display.

Fig. 5-3 is an MDS plot with hand-drawing boarders added. This figure visually reveals the relationship between records and various clustering results. The projection of stable records confirms that stable records are within dense subsets that are intersections of partitioning groups.

In most clustering algorithms, clusters are generated based on distances between a cluster central point and records. The central area around the central point is often dense. As stable records are within dense subsets whereas unstable record are within sparse subset, we can infer that stable records of a partitioning group stay close to a cluster
Fig. 5-2 Parallel Coordinates for Iris dataset
**Fig. 5-3** MDS Plot for Iris dataset with selected stable records and cluster boarders

**Fig. 5-4** Multidimensional Scaling Plot for Iris dataset
Color By Classes:
- Iris-setosa
- Iris-virginica
- Iris-versicolor

Fig. 5-5  Scatter Plot for Iris dataset
Fig. 5-6  CComViz for Iris dataset

Fig. 5-7  Correlation between CDM measure and RS measure
central area of the partitioning group, whereas unstable records remain in remote areas.

Fig. 5-2, Fig. 5-3, Fig 5-4, and Fig 5-5 confirm this inference. Relying on this inference, we anticipate that stable records reflect main characteristics of partitioning groups whereas unstable records remain questionable with uncertain features.

Fig. 5-6 shows that stable records move smoothly across multiple partitions in CComViz as opposed to unstable records. This metaphor allows us to quickly estimate record’s stability status without computing its record stability value.

To verify the correlation between the NSD dimension and the RS dimension, a scatter plot is drawn (Fig. 5-7). This figure shows a strong correlation between these two measures. Their Pearson's correlation coefficient and Spearman Rank correlation are 0.78 and 0.91 respectively [statistic online calculator: www.wessa.net]. The plot as shown in Fig. 5-7 presents an interesting abstraction to represent data. Its features deserve further investigation
6 CASE STUDIES

Case studies of this research involve the analysis and exploration of two typical data types: microarray gene expression data and clinical patient management data. Microarray gene expression data usually contain a large number of genes and a small number of experimental conditions (e.g. sampling times, cell or tissue lines etc.). The mission of cluster analysis for microarray data is to infer causal relationships between genes associated with conditions [104]. In terms of objects to be clustered, there are two types of cluster analysis: gene cluster analysis and array (experimental condition) cluster analysis [39]. The application of gene cluster analysis to time series gene expression data is an extensive research area in analytic biology. A variety of analysis approaches have been proposed [8, 25, 18, 84].

Clinical patient management data are considered as one of the most difficult types of data for cluster analysis. They often contain both numerical dimensions (e.g. clinical test data, evaluation scores, etc.) and categorical dimensions (e.g. patient profiles, medical and treatment histories, diagnostic results, etc.). Having large quantity of uncertainties, such as noisy, missing and unavailable values, is one of the most remarkable features of these data.

In the case studies, we use Salamander gene expression data and MGH (Massachusetts General Hospital) breast cancer risk factor data to demonstrate the usefulness of record stability measure and CComViz in conjunction with other UVP visualization tools used to identify interesting records and data patterns.
6.1 Salamander Gene Expression Data

Salamander gene expression data are used in the Limb Regeneration project. The ultimate goal of this project is to gain cellular and molecular knowledge from the limb regeneration process in certain creatures (e.g. salamander), and explore the same capacity in mammals, especially human beings [109]. This project started with an intense study of salamander genomics. Various experiments using different microarray technologies at different time points during limb regeneration process have generated a large quantity of time series data. Exploratory analyses of these data are currently our major tasks.

This case study focuses on comparison of two fore limb gene expression data from different chips with close time points. One dataset is called Preliminary (P) data containing time points at 1 day, 3 days, and 6 days. And another is called Agilent (A) data containing time points at 1 day, 3 day, and 5 days. Both datasets contain 500 records with matched internal identification numbers, called NC. Multiple NC numbers may refer to the same gene or EST. This case indicates that the gene or EST has multiple spots in the Preliminary chip. The two datasets are normalized with the same scale, so their expression levels are directly comparable.

The first step of this case study is to run clustering algorithms for each of the two datasets separately. The algorithms we choose are K-means, EM (Expectation Maximization), XM (Extended K-means), and DB (Density-based clustering built on K-means). All of these programs are available in Weka project [www.cs.waikato.ac.nz]. The number of clusters is decided by using an automatic function presented in the Weka EM program. This function uses a 10-folder cross-validation based search to identify the
first local optima as the best number. The search starts from 1 and incrementally continues until the average 10-folder logarithmic likelihood value decreases. With this function, the best number for the Preliminary dataset is 11 and for the Agilent dataset, it is 6. To make the comparison easy, we choose 11 as the common number for both datasets to run all clustering algorithms.

Thinking of the two dataset being like datasets with different noise added, the comparison of clustering results from individual run of different clustering algorithms on both datasets is analogous to perturbation-based cluster analysis. So, computing record stability (RS) for each record based on the total 8 clustering results is meaningful to cluster analysis. Fig. 6-1 contains projection of these 8 clustering results and the RS dimension in a single CComViz. The dimension display order is sorted by sequential similarity. This figure demonstrates CComViz’s capability to display density distributions and hierarchies vs. different hot dimensions in an interactive manner. From Fig. 6-1, we see that the clustering results of the Preliminary dataset are largely different from that of the Agilent dataset. The reason is because the Preliminary dataset contains an excessive amount of noise. The big variance of expression levels in the Preliminary
Fig. 6-1  CComViz projection of DB, KM, XM and EM clustering results of the Agilent (A) and the Preliminary (P) data, RS dimension as well.

Fig. 6-2  Parallel Coordinates for the Agilent (A) and the Preliminary (P) data
dataset, shown in Fig. 6-2, illustrates this fact. In Fig. 6-2, the scale of each axis is the same, so the expression levels can be easily compared.

Based on previous interpretation, stable genes are within dense regions that are intersections of partitioning groups mostly compliant with a label correspondent relationship, whereas unstable records are within sparse regions that are intersections of partitioning groups rarely compliant with a label correspondent relationship. Because of the binding of the label correspondent relationship, stable genes are expected to show similar patterns and similar expression levels across the two datasets. This expectation is actually confirmed in most of the cases, for example, the clusters shown in Fig. 6-3 (b). However, it is not surprising that some stable genes do not show this expectation because of bias or variance of experimental conditions. These genes are intriguing and deserve further investigation, for instance, NC105, NC15, NC245, and NC88 (NC15 and NC245 refer to the same gene) in Fig. 6-3 (e).

As discussed previously, the characteristics of unstable records are often questionable and uncertain due to their geometrical locations. Fig. 6-4 contains a set of unstable records with a threshold \( rs \leq 0.25 \). One questionable record is NC255 (\( rs = 0.21 \)). As shown in Fig. 6-4 (b), the expression level of NC255 drops at the 3d in the Agilent data, but goes up at the 3d in the Preliminary data. Another example is NC470 (\( rs = 0.25 \)), its expression level patterns are quite different in the two datasets.

Finally, Fig. 6-5 illustrates the metaphor of representing stable records as being smooth and unstable records as being fluctuating in CComViz.
Fig. 6-3  Patterns and expression levels of stable records

(a) All stable records with rs > 0.6
(b) Patterns and expression levels are close
(c) Expression levels change
(d) Small patterns change
(e) Expression levels change
Fig. 6-4 Patterns and expression levels of unstable records

(a) All unstable records with rs ≤ 0.25

(b) Different patterns

Fig. 6-5 CComViz projections of stable records (rs > 0.6) (upper) and unstable records (rs ≤ 0.25) (lower)
6.2 MGH Breast Cancer Data

The goal of MGH (Massachusetts General Hospital) breast cancer data analysis is to recognize high risk factors for breast cancer, evaluate existing risk models, and develop more accurate and intuitive risk models using statistical and cluster analysis techniques. MGH breast cancer data are collected from patient surveys, clinical practices, and affiliate programs. These data contain a wide range of information including patients’ physical and feminine data, lifestyle, ethnicity, cancer-related personal and family history, breast cancer risk measurement, cancer status, etc. The cancer status is the class label that indicates whether the patient has been diagnosed with breast cancer. Among breast cancer risk measurements are a few of numerical scores based on breast cancer risk models, such as Gail model [54], Claus model [26, 27], Myriad model [45, 46], and BRCAPRO model [100].

The biggest challenge for MGH breast cancer data analysis lies under the difficulties to recognize data patterns which are associated with breast cancer prevalence. This challenge makes it difficult to develop accurate prediction models and rigorous evaluation methods. Although Gail, Claus, Myriad and BRCAPRO prediction models have been widely used in clinical practice for many years, their prediction accuracy is still too marginal to satisfy prognosis requirement. More concerns arise as they and many other prediction models do not keep consistent. This challenge calls for the application of not only traditional clustering, classification and regression techniques, but also novel methods able to find useful information from data. This study attempts to investigate the impact of record stability on breast cancer prevalence, and seek data regions within
clusters that are significantly different from the clusters. Motivations for making this attempt come from the features of stable records. As stable records are within dense regions of clusters and reflect their main characteristics, if clustering algorithms are able to capture data class structure (breast cancer prevalence in this case), stable records are expected to present clearer data patterns than clusters in favor of the class structure. The clearer patterns may indicate whether or not the population of the stable records is at a higher risk for breast cancer with improved accuracy.

The clustering algorithms used in this study include EM (Expectation Maximization), KM (Kmeans), XM (Extended Kmeans), DB( Density based clustering), and SOM (Self-organized Map). The first four algorithms rely on the implementation in Weka project, while SOM relies on that in MIT whitehead GeneCluster. In this case study, variables have been carefully selected for clustering. These variables should either be widely recognized as risk measures or have potential to become risk factors depending on domain expert’s opinions. We choose three risk measures based on Gail, Myriad and BRCAPRO models plus three recommended risk factors: rrAH, rrLCIS, and Density (mammogram density). The dataset we queried from database and used for clustering contains 17668 records (patients) with no missing values.

We still use the automatic function in the Weka EM program to estimate the number of clusters. This number is 9. We applied this number to the five clustering algorithms and computed the RS dimension based on the five clustering results.

The investigation of the impact of record stability on breast cancer prevalence should be with regards to a clustering solution. Ideally, this solution is a cluster ensemble
based on the five clustering results. In this study, we focus on investigation method rather than result, and therefore simply use one of the five clustering results, XM result, to perform the analysis. Table 6-1 decomposes the XM result to subsets constrained by record stability threshold. In this table, cluster C XM5 seems to have certain degree of indication that record stability impacts on prevalence by its trend. Next we perform statistical analysis to measure the degree of statistical significance.

<table>
<thead>
<tr>
<th>RST</th>
<th>C XM1</th>
<th>C XM2</th>
<th>C XM3</th>
<th>C XM4</th>
<th>C XM5</th>
<th>C XM6</th>
<th>C XM7</th>
<th>C XM8</th>
<th>C XM9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥0 (all)</td>
<td>S 207</td>
<td>303</td>
<td>2515</td>
<td>2511</td>
<td>2736</td>
<td>3720</td>
<td>2641</td>
<td>2799</td>
<td>236</td>
<td>17668</td>
</tr>
<tr>
<td></td>
<td>% 1.93</td>
<td>1.65</td>
<td>1.67</td>
<td>1.27</td>
<td>1.28</td>
<td>1.24</td>
<td>1.55</td>
<td>1.24</td>
<td>1.39</td>
<td>1.36</td>
</tr>
<tr>
<td>≥0.4</td>
<td>S 207</td>
<td>276</td>
<td>2512</td>
<td>2108</td>
<td>2487</td>
<td>3656</td>
<td>2328</td>
<td>2740</td>
<td>81</td>
<td>16395</td>
</tr>
<tr>
<td>P</td>
<td>4</td>
<td>4</td>
<td>42</td>
<td>28</td>
<td>33</td>
<td>44</td>
<td>36</td>
<td>27</td>
<td>3</td>
<td>221</td>
</tr>
<tr>
<td>%</td>
<td>1.93</td>
<td>1.45</td>
<td>1.67</td>
<td>1.33</td>
<td>1.33</td>
<td>1.2</td>
<td>1.55</td>
<td>0.99</td>
<td>3.7</td>
<td>1.35</td>
</tr>
<tr>
<td>≥0.6</td>
<td>S 207</td>
<td>227</td>
<td>1476</td>
<td>1753</td>
<td>1745</td>
<td>3111</td>
<td>1764</td>
<td>2595</td>
<td>80</td>
<td>13958</td>
</tr>
<tr>
<td>P</td>
<td>4</td>
<td>3</td>
<td>42</td>
<td>23</td>
<td>25</td>
<td>40</td>
<td>25</td>
<td>23</td>
<td>2</td>
<td>187</td>
</tr>
<tr>
<td>%</td>
<td>1.93</td>
<td>1.32</td>
<td>1.7</td>
<td>1.31</td>
<td>1.43</td>
<td>1.29</td>
<td>1.42</td>
<td>0.89</td>
<td>2.5</td>
<td>1.34</td>
</tr>
<tr>
<td>≥0.7</td>
<td>S 73</td>
<td>69</td>
<td>2201</td>
<td>953</td>
<td>1018</td>
<td>2604</td>
<td>1115</td>
<td>2216</td>
<td>0</td>
<td>10249</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>1</td>
<td>33</td>
<td>11</td>
<td>21</td>
<td>32</td>
<td>15</td>
<td>21</td>
<td>0</td>
<td>136</td>
</tr>
<tr>
<td>%</td>
<td>2.74</td>
<td>1.45</td>
<td>1.5</td>
<td>1.15</td>
<td>2.06</td>
<td>1.23</td>
<td>1.35</td>
<td>0.95</td>
<td>1.33</td>
<td></td>
</tr>
<tr>
<td>≥0.8</td>
<td>S 73</td>
<td>67</td>
<td>2147</td>
<td>953</td>
<td>890</td>
<td>2426</td>
<td>763</td>
<td>735</td>
<td>0</td>
<td>8075</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>1</td>
<td>31</td>
<td>11</td>
<td>18</td>
<td>32</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>113</td>
</tr>
<tr>
<td>%</td>
<td>2.74</td>
<td>1.49</td>
<td>1.44</td>
<td>1.15</td>
<td>2.02</td>
<td>1.32</td>
<td>1.31</td>
<td>1.06</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>S 0</td>
<td>0</td>
<td>2147</td>
<td>0</td>
<td>0</td>
<td>2426</td>
<td>763</td>
<td>756</td>
<td>0</td>
<td>6092</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
<td>0</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>81</td>
</tr>
<tr>
<td>%</td>
<td>0</td>
<td>0</td>
<td>1.44</td>
<td>0</td>
<td>0</td>
<td>1.32</td>
<td>1.31</td>
<td>1.06</td>
<td>1.33</td>
<td></td>
</tr>
</tbody>
</table>

RST: subset constrained by threshold of record stability
S: size of subset
P: count of positive records (cancer patients)
%: prevalence = P/S;

In clinical practice, a number of statistical tests are used to describe reliability of a diagnostic or prognostic result [36]. In this study, we apply a few of them. The first statistical test used is chi-square that measures the degree of significant difference of
categorical distributions between different independent sample groups or one sample group and expected reference [90]. It is the characteristic of the categorical distributions that we are interested in this analysis. Chi-square is an appropriate test to measure whether the stable subset (890 records with 18 positive and ≥0.8 record stability), highlighted in Table 6-1 with bold border, is different from the whole CXM5 cluster. In the test, the prevalence of CXM5 cluster, 1.28%, is considered as expected prevalence.

Below is the test output from GraphPad online calculator

[www.graphpad.com/quickcalcs/chisquared1.cfm]

<table>
<thead>
<tr>
<th>Row #</th>
<th>Category</th>
<th>Observed</th>
<th>Expected #</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>18</td>
<td>11.392</td>
<td>1.280%</td>
</tr>
<tr>
<td>2</td>
<td>Negative</td>
<td>872</td>
<td>878.608</td>
<td>98.720%</td>
</tr>
</tbody>
</table>

Chi squared equals 3.883 with 1 degrees of freedom. The two-tailed P value equals 0.0488. By conventional criteria, this difference is considered to be statistically significant.

Table 6-3 lists additional statistical test results, along with interpretations and calculation formulas in conjunction with a contingency table (Table 6-2). Note that in Table 6-2, Positive test corresponds to Stable status. This is because we intend to test the impact of stable records on prevalence. If the test is about the impact of unstable records, Positive test should correspond to Unstable status.

<table>
<thead>
<tr>
<th></th>
<th>Positive Diagnose</th>
<th>Negative Diagnose</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Test (Stable)</td>
<td>a: 18</td>
<td>b: 872</td>
<td>a+b: 890</td>
</tr>
<tr>
<td>Negative Test (Unstable)</td>
<td>c: 17</td>
<td>d: 1829</td>
<td>c+d: 1846</td>
</tr>
<tr>
<td>SUM</td>
<td>a+c: 35</td>
<td>b+d: 2701</td>
<td></td>
</tr>
</tbody>
</table>

109
Table 6-3 Statistic calculation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Formula</th>
<th>Value (%)</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>sensitivity</td>
<td>( a / (a+c) )</td>
<td>51.4</td>
<td>51.4% patients with breast cancer in ( C_{XM5} ) are in ( \geq 0.8 ) stable category</td>
</tr>
<tr>
<td>specificity</td>
<td>( d / (b+d) )</td>
<td>67.7</td>
<td>67.7% patients without breast cancer in ( C_{XM5} ) are in &lt;0.8 stable (unstable) category</td>
</tr>
<tr>
<td>positive predictive value</td>
<td>( a / (a+b) )</td>
<td>2.02</td>
<td>Given ( \geq 0.8 ) record stability, 2.02% patients in ( C_{XM5} ) have breast cancer</td>
</tr>
<tr>
<td>negative predictive value</td>
<td>( d / (c+d) )</td>
<td>99.1</td>
<td>Given &lt;0.8 record stability, 99.1% patients in ( C_{XM5} ) do not have breast cancer</td>
</tr>
<tr>
<td>positive likelihood ratio</td>
<td>sensitivity / (1-specificity)</td>
<td>106</td>
<td>Very minimal increase in the likelihood of breast cancer in subset of ( C_{XM5} ) with ( \geq 0.8 ) record stability</td>
</tr>
<tr>
<td>negative likelihood ratio</td>
<td>(1-sensitivity) / specificity</td>
<td>94.6</td>
<td>Very minimal decrease in the likelihood of breast cancer in subset of ( C_{XM5} ) with &lt; 0.8 record stability</td>
</tr>
</tbody>
</table>

Fig. 6-6 and Fig. 6-7 use MDS plot and parallel coordinates to visualize the discovery. Table 6-4 gives variable ranges for entire \( C_{XM5} \) and the stable subset.
**Fig. 6-6** MDS projection of $C_{XM5}$ cluster

**Fig. 6-7** Parallel coordinates projection of $C_{XM5}$ cluster
We select one stable subset from a cluster to perform statistical test. This subset appears to be more different from its cluster than others. Although visualizations in Fig. 6-6 and Fig. 6-7, as well as chi-square test suggest a certain degree of significance in the impact of record stability on breast cancer prevalence, other statistics in Table 6-4 indicate that the impact is insignificant. There are two possible reasons to account for the insignificance. The first is that no significant data patterns exist, so clustering algorithms are not able to generate high quality of clusters. And the second reason is related to the assumption that cluster stability analysis is based on. We assume clustering results, to some extent, are close to data class structure. In fact, this assumption is not easy to satisfy for complex data, such as the MGH data. Firstly, it is less likely to have unsupervised learning results highly correspond to data class structure. And secondly, feature selections for clustering are difficult as they involve intensive and sufficient understanding of domain knowledge.

This case study is not intended to build a decision-making system or derive clinical conclusions due to lack of two important analysis processes: data-driven crossing-validation and clinical validation. Without data-driven crossing-validation, any outcomes may overfit training data. Conclusions depending on these outcomes are often

<table>
<thead>
<tr>
<th>Variable</th>
<th>$C_{X_X5}$ Minimum</th>
<th>$C_{X_X5}$ Maximum</th>
<th>Stable Subset Minimum</th>
<th>Stable Subset Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gail</td>
<td>1</td>
<td>32.46</td>
<td>1</td>
<td>4.35</td>
</tr>
<tr>
<td>Myriad</td>
<td>0.03</td>
<td>0.09</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>BRCAPRO</td>
<td>0.0004</td>
<td>0.23</td>
<td>0.0004</td>
<td>0.01</td>
</tr>
<tr>
<td>rrAH</td>
<td>0.93</td>
<td>1</td>
<td>0.93</td>
<td>1</td>
</tr>
<tr>
<td>rrLCIS</td>
<td>1</td>
<td>10.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Density</td>
<td>93133</td>
<td>124052</td>
<td>93359</td>
<td>107156</td>
</tr>
</tbody>
</table>
unreliable, even useless. Clinical validation is critical to draw conclusions based on findings from cluster analysis. As a rule of thumb, "without clinical expertise, practice risks becoming tyrannized by external evidence, for even excellent external evidence may be inapplicable to or inappropriate for an individual patient." [107]. Above issues are beyond the scope of this research and are left for future work.
7 CONCLUSION

We defined partition comparison and evaluation to address mutual comparison of multiple information sources, whether they are computational or domain specific, categorical or numerical, internal, external or relative. Mutual comparison often leads to more consolidated and robust results.

Label correspondent relationship is useful in comparative cluster analysis. Its polygamous, asymmetric, intransitive properties, and non-correspondence feature characterize the partition stability metrics. The partition stability metrics provide record stability, group stability, and partition similarity measures. Using record stability measure, stability or membership migration rate of individual records across multiple partitions is measured. Record stability is also a measure of record interestingness. Stable records tend to stay compliant with a label correspondent relationship. Because of the binding of label correspondent relationship, stable records within a cluster reflect main characteristic of the cluster. Unstable records, on the other hand, have a strong migration tendency to break the label correspondent relationship. They are often unusual and questionable. Group stability and partition similarity not only serve as characteristics measures, but also provide information for record and dimension display order rearrangement in CComViz.
Partition comparison and evaluation take advantage of visualization, since human visual perception is exceptionally good at image comparison. CComViz is designed to perform visual and interactive comparison and evaluation of a large number of partitions within a single graphic display. CComViz focuses on achieving visual analytics rather than geometrical data mapping. In a data and tool linking environment, CComViz and other visualization tools that feature topology-preserving mapping provide a complementary combination to perform exploratory data analysis.

Because of its highly sophisticated ability to rearrange record and dimension display order, CComViz is extremely useful for displaying data stability, data flow, density distribution and hierarchy across multiple partitions. The methodology used to develop a layout algorithm for record and dimension display order rearrangement is of remarkable originality. Through utilizing knowledge of human perception and cognition, user interaction, and information from partition stability measures, this methodology has achieved visual aesthetics, crossing reduction, as well as visual metaphors (Smooth record projections represent stable records while fluctuating record projections represent unstable records).

In addition to the record display order rearranged by the layout algorithm, the record display order of each partition can also be independently rearranged according to partition internal criteria, such as Silhouette index. Freezing record display order for continuous play is another interactive function useful for observing data density distribution and hierarchy under various conditions.
Record rendering layer control is a critical interactive function to relieve record occlusion. Boolean operation on multiple record selections in CComViz provides an effective way to quickly identify interesting records and gain insight into their characteristics from different aspects in a data and tool linking environment.

Our proposed partition stability metric and CComViz are useful for real-world data analysis. Their application to Salamander data enables not only discovery of interesting genes but also characterization of data patterns across different datasets. They also provide a possible solution, called cluster stability analysis, to MGH data analysis attempting to suggest a new approach to predicting risks of breast cancer.

In summary, through this research, we made four major contributions:

1. Define partition comparison and evaluation, as well as partition stability analysis to address mutual comparison of multiple partitions in order to identify near-optimal structure, build ensembles, or conduct validation.

2. Extend a perturbation-based cluster stability metric system to the partition stability metric system. The feature of label correspondent relationship is explored. And three fundamental properties and one feature of label correspondent relation are recognized.

3. Develop CComViz visualization tool to visualize data stability, data flow, density distribution and hierarchy, and data correlation at the record, at the partition and at the dimension levels within a single graphical interactive display.

4. Present a novel methodology used to develop a layout algorithm for informatively rearranging the order of the records and dimensions in CComViz in order to achieve
visual aesthetics and crossing reduction, as well as visual metaphor for stable and unstable records.
Visualization for partition comparison and evaluation is still a new research area with plenty of room for development of additional new techniques. We are particularly eager to see a continuing research effort in the following areas.

8.1 Label Correspondent Relationship

In this dissertation, the label correspondent relationship is built based on label correlation coefficient. As we described in chapter 3, many other cluster similarity and distance measures can also be used for this purpose. Therefore, a systematic evaluation will be worthwhile to reveal their difference and give guidelines to explore this relationship from different aspects.

8.2 Interpretation and Validation of Stability

CDM measure has a similar purpose to RS measure. We used a synthetic dataset and Iris dataset to illustrate the correlation between standard deviation dimension of CDM rows and the RS dimension. Several real-world datasets, including Salamander gene expression data used in our case study, also show correlations between them. Mathematical proof or robust validation to this correlation will help understand and interpret stability defined in our metric. Additionally, validation to RS stability measure from domain aspect should be conducted based on data with “ground truth”. This work needs extensive collaboration with domain experts.
8.3 CComViz Layout

In CComViz, record and dimension display order rearrangement is a graph layout problem. Different prospects lead to different layout approaches. It will be very helpful if taxonomy for the prospects, followed by various layout algorithms correspondent to relevant prospects, can be developed.

8.4 User Study

CComViz is expected to be a highly interactive visualization with good usability for exploratory data analysis. User study should be carefully planned and conducted to verify this expectation. It also is an effective way to define guidance for the design of various interactive functions employed in CComViz.
LITERATURE CITED


55. Gao, B. Cluster Description and Related Problems, DDM Lab Seminar 2005 Spring


125. Yang, J. 2004. Clustering and Classification, Statistics for Microarray Data, IBC Short Course


