DATA INTEGRATION SIMULATION USING DATA CONSOLIDATION

Hadaiq R. Sanabila and Ito Wasito

Faculty of Computer Science, Universitas Indonesia, Kampus Baru UI Depok, Jawa Barat, 16424, Indonesia

E-mail: hadaiq@cs.ui.ac.id

Abstract

One of the data integration methods is data consolidation. This method captures data from multiple source systems/data and integrates it into a single persistent data. We examined the performance of data consolidation using k-means and Gaussian mixture clustering. Meanwhile, we use Silhouette index as cluster validation and measure how well of a clustering relative to others. The experiments analyses the data in various data duplication rate and actual number of data cluster. Based on the experimental result, there are two factors affecting the performance of data consolidation. These factors are the rate/percentage of duplicate data and the number of actual cluster contained in a data. The higher percentages of duplicate data and less number of clusters contained in the data would be increasing the performance of clustering algorithm.

Keywords*: data integration, data consolidation, data duplication*

Abstrak

Salah satu metode dari integrasi data adalah konsolidasi data. Metode ini mengambil data dari beberapa sumber data untuk digabungkan menjadi data persisten tunggal. Peneliti memeriksa kinerja konsolidasi data menggunakan beberapa teknik *clustering* yaitu *k-means* dan *gaussian mixture clustering*. Penulis menggunakan Silhouette *index* sebagai metode validasi *cluster* untuk mengukur seberapa baik suatu pengelompokan relatif terhadap data lain. Penelitian ini melakukan analisis data terhadap jumlah rata-rata duplikasi data dan jumlah sebenarnya dari *cluster* data. Berdasarkan hasil percobaan, ada dua faktor yang mempengaruhi kinerja integrasi data dengan menggunakan konsolidasi data. Faktor-faktor tersebut antara lain adalah tingkat atau persentase dari duplikasi data dan jumlah *cluster* sebenarnya yang terkandung dalam data. Persentase duplikasi data yang tinggi dan data yang mengandung jumlah *cluster* yang rendah, akan meningkatkan kinerja dari algoritma *clustering*.

Kata Kunci: *integrasi data, data konsolidasi, duplikasi data*

1. Introduction

Data is a set of attributes that contains information about the facts resulting from the observation or measurement is used to draw conclusions. The data becomes an integral entity of an institution or company as the basis of the decision-making. There are many types of data where inextricably linked each other. The needs for information derived from various data are encouraging the emergence of data integration.

Data integration is the merger of data derived from different types and sources to provide the combined observations from these data. Users expect to perform data integration to obtain more accurate and comprehensive information. Data integration becomes a significant process in various fields along with the gain of frequency and the need of sharing data. In addition, the needs for accurate and rapid data, as decision support information, also affect the increased demand for data integration. However, there are some issues in data integration [1]. The issues in data integration are: data formats, data visualization, and proprietary databases.

Most of data warehouses are designed to support several types of data. When other data type was added, the compromises begin, including possible redesign of the warehouse. It is conducted to accommodate the new types of data. Examples of various formats for integration are internal RDBMS data, geographic data and object data (reports, spreadsheets, notes, and multimedia).

The demand from user for graphs, charts, data animation, data landscapes, cartographic or dashboard representations, or even sophisticated virtual reality interfaces are key marketing

features of specific software products. These visualizations maybe not integrated to the base analytic tool. Actually, there may be no one base analytic tool in the architecture.

Today enterprise are all attempt to attain various data in order to complete their customer, market, competitor, and forecasting picture. This has generated a new industry of data services which ready to integrate data or even full service data analytics.

In this research, researcher want to examine the data integration in small-scale simulation using particular clustering algorithm i.e. k-means clustering and Gaussian mixture clustering. The data that researcher used in data integration is contains various duplication rate and actual number of cluster. This paper is organized as follows. Section 2 gives an overview of data integration meanwhile section 3 present the method which used to examine the data integration. The experimental results in various data described in section 4. Finally, the conclusion discussed in section 5.

Recently, the necessity for data integration is not only in commercial applications but also in science application. Data integration in the commercial sector employed on specific areas, such as: supply chain management [2][3], customer relationship management [4], business intelligence [5], decision support system, and etc. Whereas in science area, data integration is employed in several fields such as: bioinformatics, geology, meteorology, and etc. Concomitant with integration growth in various fields, the needs of integration are not only on the data layer but also can move to application layer integration, which resulted in the need for a robust and accurate algorithm.

There are three main methods used for data integration: consolidation, federation, and propagation. Data Consolidation captures data from multiple source systems and integrates it into a single persistent data store. Data Federation provides a single virtual view of one or more source data files. Data Propagation applications copy data from one location to another.

Tian Mi et.al. were observing hierarchical clustering as a solutions to integrate multiple data sets. They are inspecting complete-linkage Hierarchical clustering. Its distance between two records within one cluster after following their algorithm is smaller than or equal to the threshold. And all the records with a distance smaller than or equal to the threshold will be in one cluster. Meanwhile, [2] show how XML can actually be implemented for a Web-based integration in supply chains.

Furthermore, McGrath et.al. [3] Using PEG (Pharmaceutical Extranet Gateway) to improve pharmaceutical supply chain efficiency within the healthcare sector. PEG is an Internet-based facility, developed to allow the automated passing of common order transactions between all parties and, in the process, to more tightly integrate their disparate systems. Moreover, Berlanga et.al. proposed HeC Integrated Data Model. The HeC Integrated Data Model (IDM) is intended to describe, store, link and annotate patient data. All data is associated to the patient concerned and is integrated structurally by the metadata that defines it, and conceptually by the semantic data beneath. Besides that, Liu et.al present a data integration program based on web services of heterogeneous bioinformatics databases.

In general, information systems are not designed for integration. Thus, whenever integrated access to different source systems is desired, the sources and their data that do not fit together have to be coalesced by additional adaptation and reconciliation functionality. Since the goal is to provide a homogeneous, unified view on data from different sources, the particular integration task may depend on several factor, i.e.: the architectural view of an information system, the content and functionality of the component systems, the kind of information that is managed by component systems (alphanumeric data, multimedia data; structured, semi-structured, unstructured data), requirements concerning autonomy of component systems, intended use of the integrated information system (read-only or write access), performance requirements, and the available resources (time, money, human resources, know-how, etc.).

There are some science applications using data integration. One of the science studies which developing data integration is bioinformatics. The necessity of data integration is widely admitted in bioinformatics. Biological data is spreading across the Internet in a wide various formats. The bioinformatics related research requires an integrated view of all relevant data, including the results of back-end. Each data integration approach has strengths and weaknesses. Furthermore, it quite difficult to evaluate which approach suits a particular needs best without fully understanding the data integration landscape.

In this research, the author performs small-scale of simulation to examine the data consolidation method. Data consolidation is one of data integration method that captures data from multiple source systems and integrates it into a single persistent data. This method help the users to obtain the essence of information contained in various data type and source. By using this

method, the author will measure the data consolidation accuracy and robustness of data integration, whether it will increase or decrease the integrated data contents information.

2. Methodology

Firstly, the author generates duplicate data from various data and save it into data1, data2, data3, and data4. Furthermore, the duplicate data grouped and merged in integrated data. Then, the integrated data analyzed using k-means and Gaussian mixture clustering, then validated using Silhouette index. The experiments illustration is depicted in figure 1.

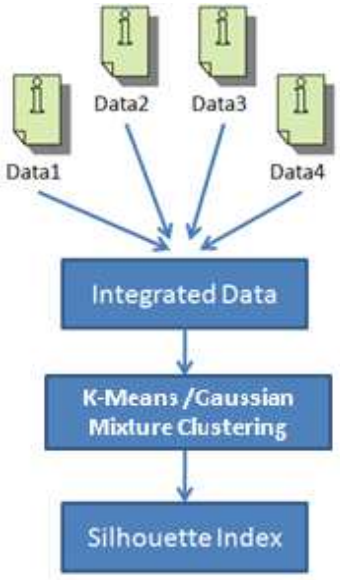

Figure 1. The experiments illustration.

The author generates the duplicate data from two data with particular percentage. The new data is the combination of first data that contain the second data with particular percentage in random order. The result of this procedure is a data which contain duplicate information from the others data. The pseudocode of this procedure depicted in figure 2.

Furthermore, find and group the duplicate data, which the data label is exist in more than one data, and integrate it into a new integrated data by concatenate it as a new column in identical duplicate data label row. The pseudocode of this procedure illustrated in figure 3. The final step is integrating from various data and stores it into single persistent data store. This procedure is conducted by looking for duplicate data in the combination of input data and keeps it single data. The pseudocode of integrating data procedure illustrated in figure 4.

```
function genDuplicateData(File1,
File2, n)

// File1 data size
nFile1  = size (File1)
// number of duplicate data in
particular percentage
ndup = n*nFile1
// gap between duplicate data
gap = nFile1/ndup

for i = 1:gap:nData1
// the index of duplicate data
ind = i + gap *rand(1,1)
// replace the data of file1 with
data of file2
data1.data (ind,:)=
data2.data(ind,:)
// replace the label of data in
file1 with the label of data in
file2
data1.label (ind)= data2.label
(ind)

end
// new data is data in file1 where
containing n percent of data in
file2
newdata = data1

return newdata
```

Figure 2. The pseudocode of generating duplicate data procedure.

Illustrating a higher dimension data set is crucial issues in data mining. Principal Component scatter plot is one of widely used technique to overcome it. The principal components of dataset are represents the data variability and obtained throughout eigenvector-eigenvalue decomposition of its covariance matrix. The scatter plot describes the clustering result into vertical and horizontal axis respectively. The x axis represents the largest eigenvalue meanwhile y axis represent the second largest eigenvalue.

An important issue in cluster analysis is cluster validation. The objective of clustering is to decide intrinsic grouping in a set of unlabeled data. Clustering validation used to measure how well of a clustering relative to others. Silhouette index is one of clustering validation that provides a compact graphical representation of how well each object lies within its cluster. Silhouette index was introduced by Peter J. Rousseeuw in 1986. The incorrect clusters members are visualized as

negative Silhouette index and vice versa. Silhouette index also has more accuracy compared to other cluster validation such as Davies-Bouldin index. The Silhouette index is plotted vertical and horizontal axis respectively. The x axis represents the index value meanwhile y axis represent the cluster number.

3. Result and Analysis

In this experiment, the author used four different data with the similar number of cluster and various duplicate percentages, i.e. 5%, 10%, 20%.

```
function getDupData (File1, File2)

// check the duplicate label in
file1 and file2
dup           = isMember (File1,
File2)
// number of duplicate data
 ndup   = size(dup)

for i = 1:ndup

// concatenate duplicate data as a
new column in identical duplicate
data label row
newdata.data (i,:) =
[Data1.data(dup(i),:)
Data2.data(dup(i), :)]
  // the label of new data
 newdata.label(i) =
Data1.data(dup(i);
end

return newdata
```

Figure 3. The pseudocode of detecting duplicate data procedure.

Furthermore, researcher examined this method in the data that consist of particular cluster i.e. 2.3 and 4 clusters using k-means clustering and Gaussian mixture clustering. The experiments conducted in the same hardware using Intel Centrino Duo 1.67 GHz processor, 2.5 GB memory. Researcher use Windows Vista Basic 32 bit operating system with MATLAB R2009a as the programming and testing environment.

First, researchers want to examine the performance of data consolidation in the data, which consist of 2 clusters. Furthermore, researchers attempt to cluster the data into several cluster, such as 2, 3, 4, 5 clusters. It used to measure the performance and the integrated data accuracy, whether it has high endurance even clustered in improper actual number of data cluster. The experimental result using k-means clustering for the data which consist of 5% duplicate data, the Silhouette index of data classified in 2, 3, 5 clusters shows negative Silhouette index. Nevertheless, there are no negative indexes when the data classified in four clusters. Moreover, for the data that consist of 10% duplicate data, there are no positive Silhouette indexes except for the data that classified into two clusters. Finally, for data, which consist of 20% duplicate data, there are positive Silhouette index in the data, which classified in two and three clusters and negative Silhouette index in data, which classified four and five clusters. The experimental result for k-means clustering depicted in figure 5-7.

```
Function IntegrateData (File1,
File2,…,FileN)

// number of file
n = nargin
k = 1

for i = 1:n-1
   for j = i+1:n
   // find the duplicate data in
each 2 file
   dup{k} =
getDupData(varargin{i},
varargin{j});
    k = k+1
   end
end

for a = 1:k
  //merge the duplicate data
    integrated = [integrated ;
dup{a}]
end
```

Figure 4. The pseudocode of integrating data procedure.

Meanwhile, for the Gaussian mixture clustering, the experimental result for the data which consist of 2 cluster and 5% duplicate data shows negative Silhouette index for the data that classified in 2, 4, 3, 5 cluster.. Moreover, for the data that consist of 10% duplicate data, there are no positive Silhouette indexes. Finally, for data, which consist of 20% duplicate data, there are no positive Silhouette indexes as well. The experimental result for Gaussian mixture clustering depicted in figure 8-10.

Afterwards, researchers want to analyze the performance of data consolidation in the data, which consist of 3 clusters. By using k-means clustering, the data which consist of 5% duplicate data, only the data classified in 2 cluster has a negative Silhouette index while the others have a positive Silhouette index Moreover, for data which consist of 10% duplicate data, there are no positive Silhouette index except for the data which classified in 3 cluster. Finally, for data, which consist of 20% duplicate data, there are no positive Silhouette indexes in the data. The experimental result depicted in figure 11-13.

Besides that, for the Gaussian mixture clustering, the experimental result for the data which consist of 3 cluster and 5% duplicate data shows there are no positive indexes. Furthermore, for the data that consist of 10% duplicate data, there are no positive Silhouette indexes as well. Finally, for data that consist of 20% duplicate data, there are also no positive Silhouette indexes. The experimental result for Gaussian mixture clustering depicted in figure 14-16.

Moreover, the k-means clustering result for the data which consist of 4 clusters and 5% duplication, only the data classified in 2 clusters has a positive Silhouette index while the others have a negative Silhouette index. For the data, which consist of 10% duplicate data, there are positive Silhouette index in the data, which classified in 2 and 5 clusters and negative Silhouette index in data which classified 3 and 4 cluster. Finally, for data that consist of 20% duplicate data, there is no positive Silhouette index in the data except for the data which classified in 2 clusters. The experimental result depicted in figure 17-19.

Meanwhile, for the Gaussian mixture clustering, the experimental result for the data which consist of 4 cluster and 5% duplicate data shows negative Silhouette index for the that classified in 3, 4, 5 cluster. Nevertheless, there are no negative indexes when the data classified in two clusters. Moreover, for the data that consist of 10% duplicate data, there are no positive Silhouette indexes. Finally, for data which consist of 20% duplicate data, there are positive Silhouette index as well. The experimental result for Gaussian mixture clustering depicted in figure 20-22.

Based on the experimental results, several factors affect the performance of data consolidation. These factors are the rate of duplicate data and the number of actual cluster contained in a data. While clustering on the data with a higher rate of duplication, the Silhouette index at the data that contains 10% and 20% duplication are relatively better than the data with the number of duplications in 5%. Increasing the number of data, duplication progressively adds the information accuracy leading to the representation of information from the integrated data. In addition, the actual number of clusters contained in a data also affects how well the data is clustered. The higher number of clusters contained in data, it is more difficult to separate. This is because the higher number of clusters contained in the data contains higher ambiguous data possibility. Its ambiguity gives some impact on the performance of clustering algorithms. Clustering algorithm also determine how well each object lies within its cluster. K-means clustering obtained a better cluster result than Gaussian mixture clustering. This is evidenced by the Silhouette index result of k-means clustering mainly has more positive index than Gaussian mixture clustering.

4. Conclusion

Nowadays, data integration plays a main role in the information integration. Data integration necessities are not only in corporation but also in the science application. One of the data integration methods is data consolidation. Data consolidation is captures data from multiple source systems and integrates it into a single persistent data. There are several factors, which affect the data consolidation performance. These factors are the rate/percentage of duplicate data and the number of actual cluster contained in a data. The higher percentages of duplicate data add the information and give more accurate data integration representation. Whereas for the actual number of clusters contained in a data also affect how well the data is clustered. The higher number of clusters contained in the data contains higher ambiguous data possibility so that decreasing the performance of clustering algorithm.

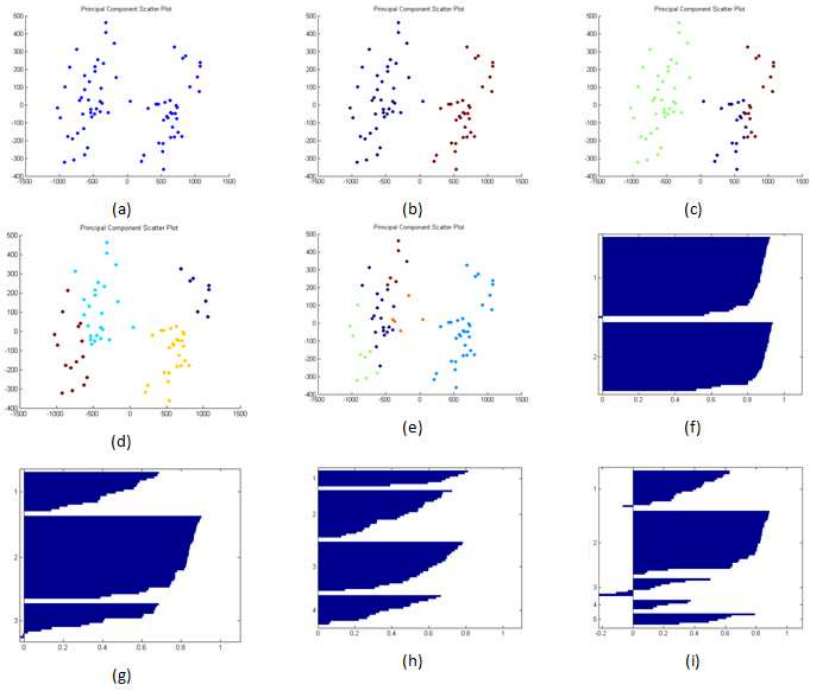

Figure 5. The experiments result of k-means clustering for the data containing 5% of duplicate data and 2 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x-axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x-axis is the index value and the y-axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e.

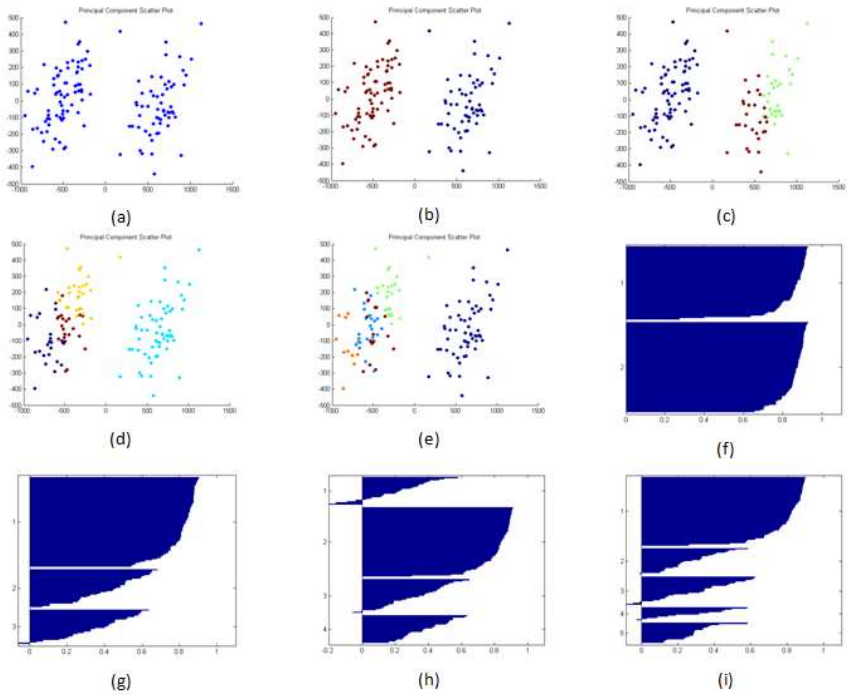

Figure 6. The experiments result of k-means clustering for the data containing 10% of duplicate data and 2 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e.

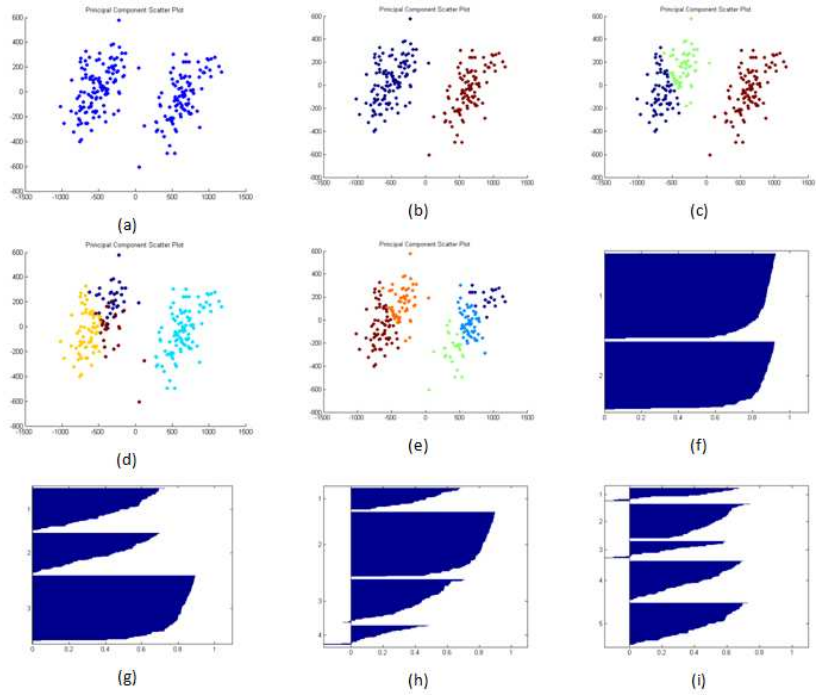

Figure 7. The experiments result of k-means clustering for the data containing 20% of duplicate data and 2 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e.

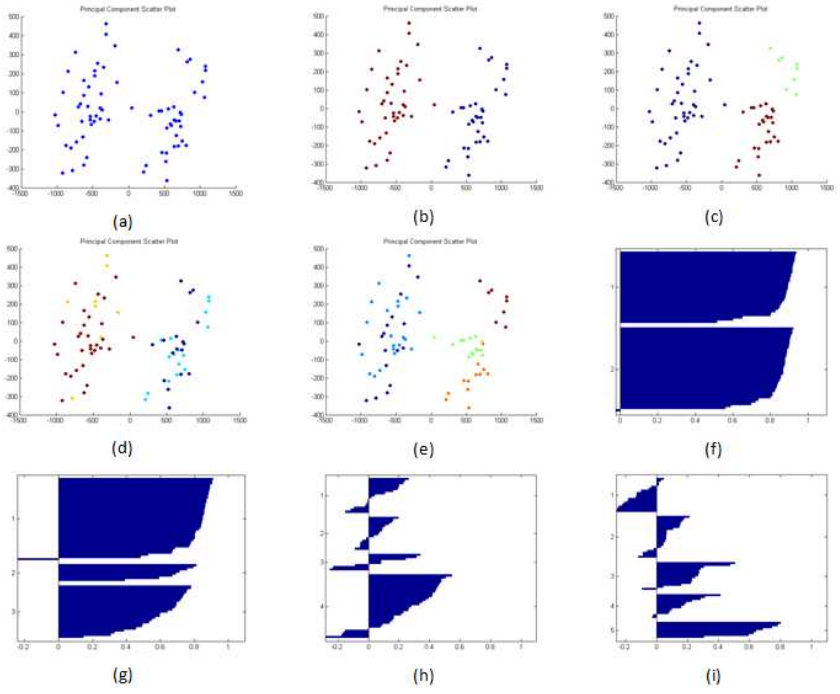

Figure 8. The experiments result of Gaussian mixture clustering for the data containing 5% of duplicate data and 2 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e.

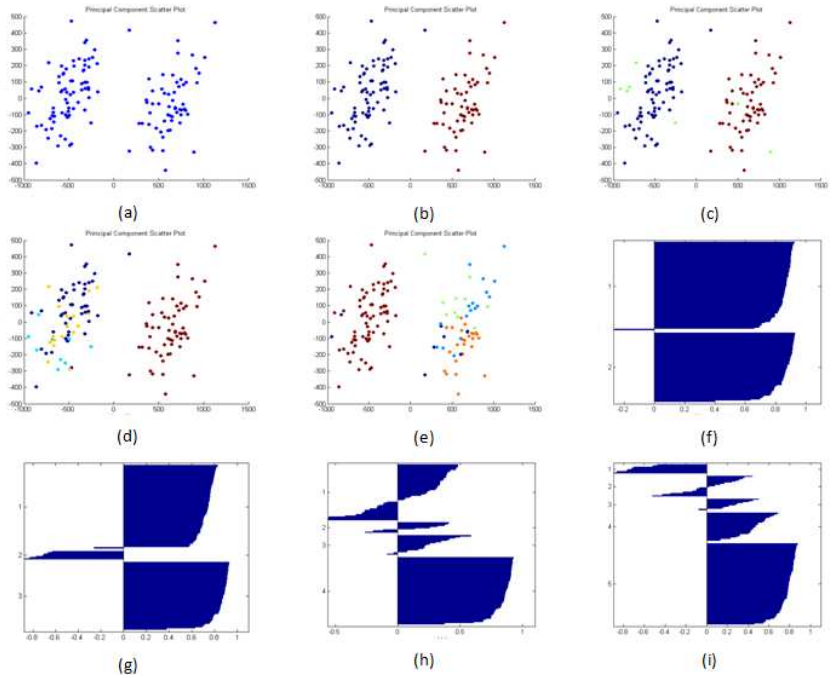

Figure 9. The experiments result of Gaussian mixture clustering for the data containing 10 % of duplicate data and 2 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e.

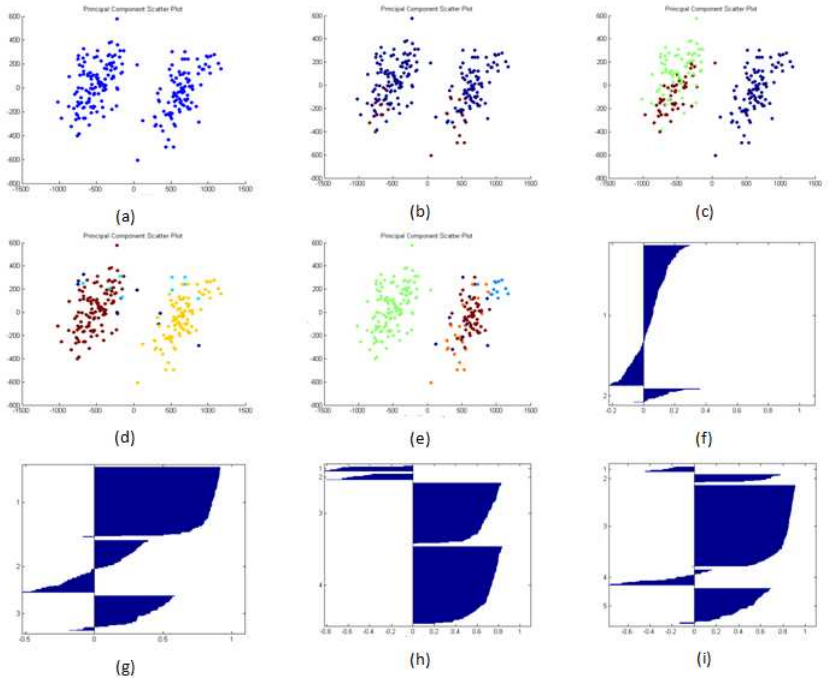

Figure 10. The experiments result of gaussian mixture clustering for the data containing 20 % of duplicate data and 2 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e.

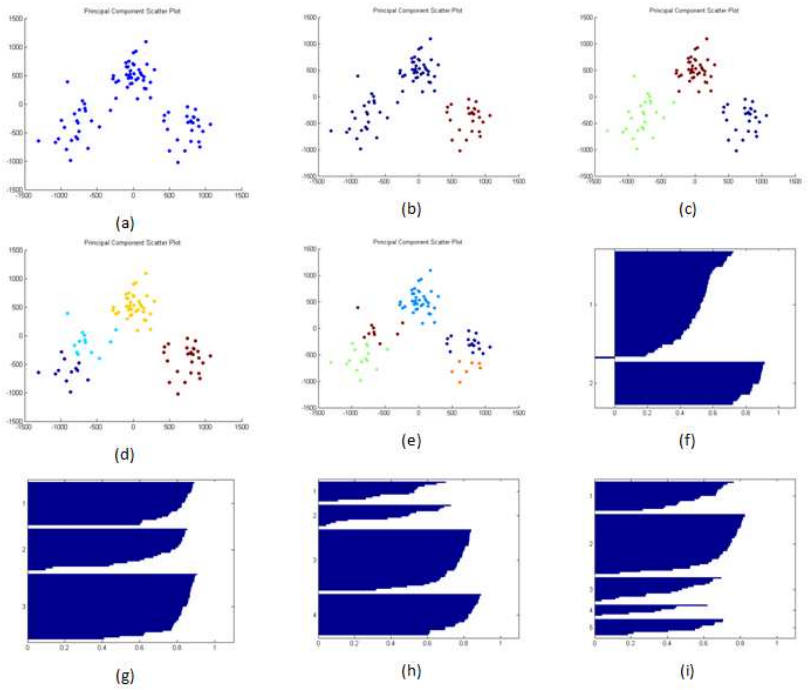

Figure 11. The experiments result of k-means clustering for the data containing 5 % of duplicate data and 3 actual clusters**.** In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e.

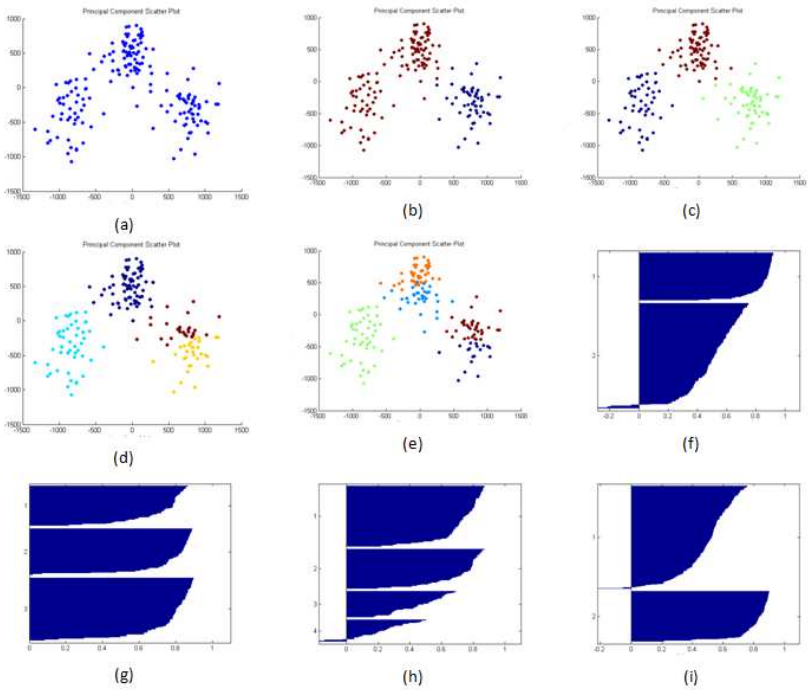

Figure 12. The experiments result of k-means clustering for the data containing 10 % of duplicate data and 3 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e.

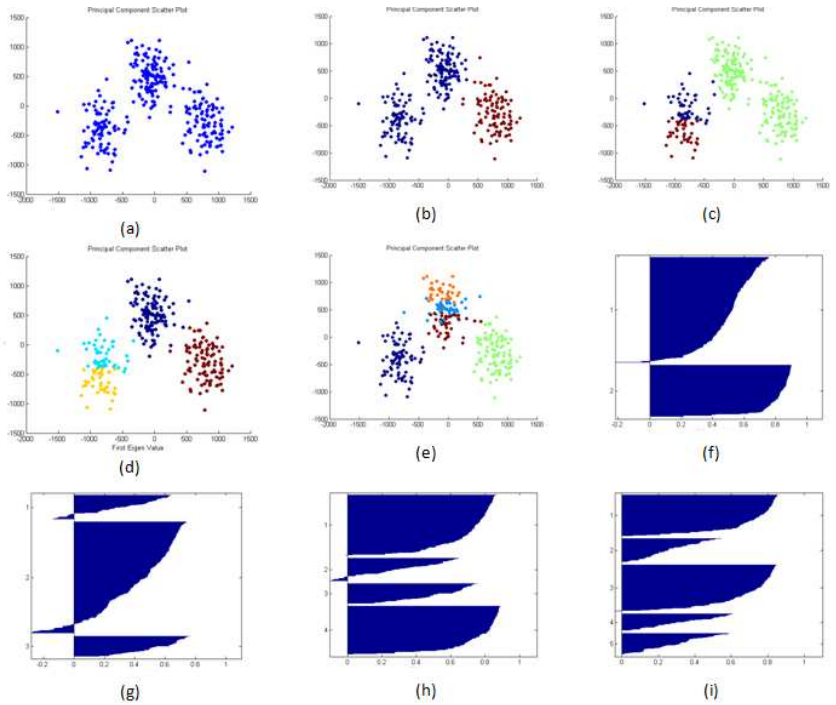

Figure 13. The experiments result of k-means clustering for the data containing 20 % of duplicate data and 3 actual clusters**.** In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

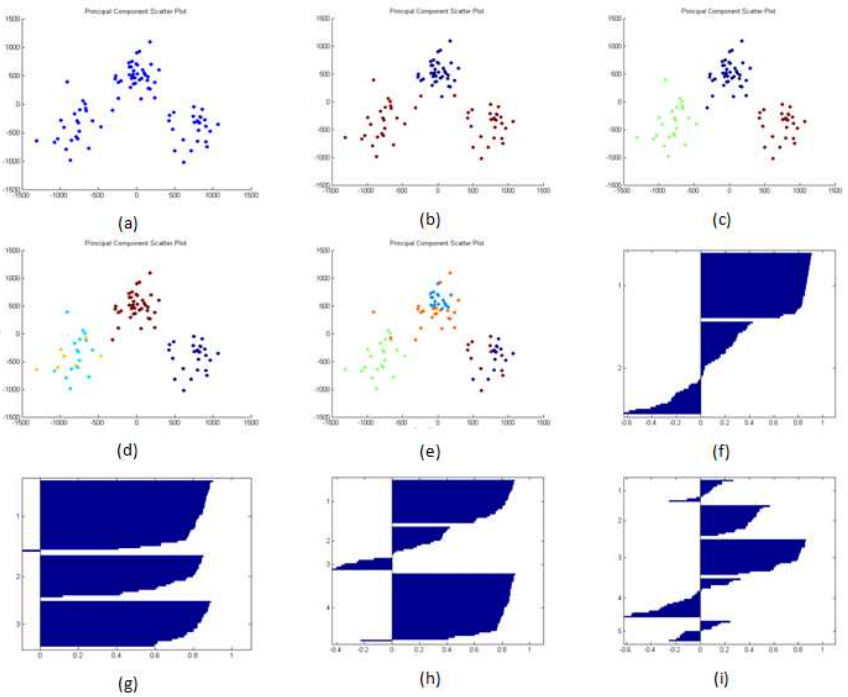

Figure 14. The experiments result of Gaussian mixture clustering for the data containing 5 % of duplicate data and 3 actual clusters**.** In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

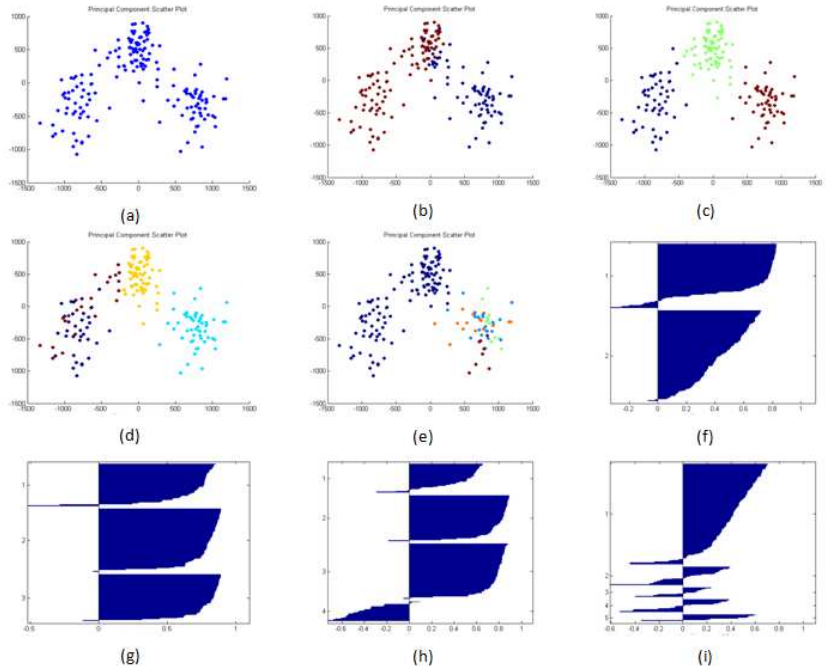

Figure 15. The experiments result of Gaussian mixture clustering for the data containing 10 % of duplicate data and 3 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

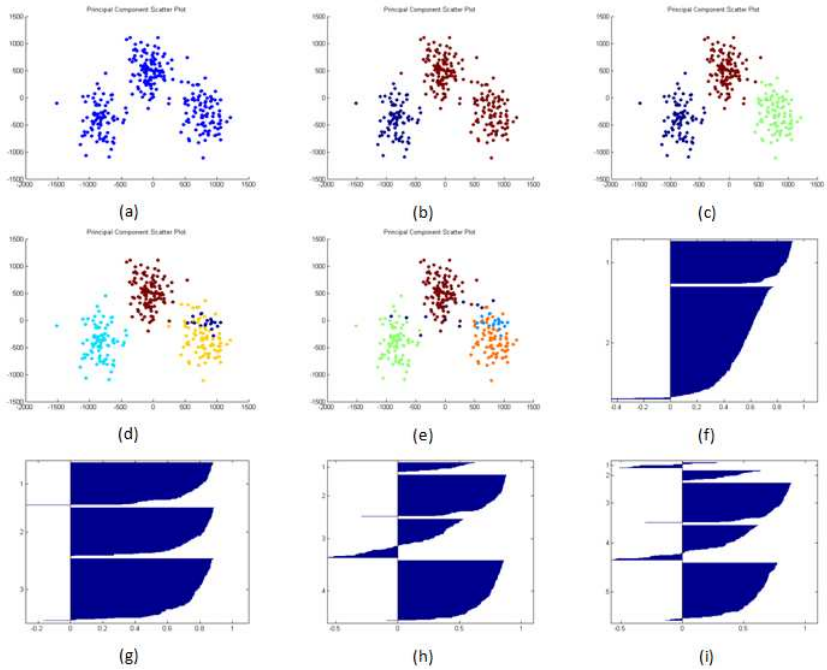

Figure 16. The experiments result of Gaussian mixture clustering for the data containing 20 % of duplicate data and 3 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

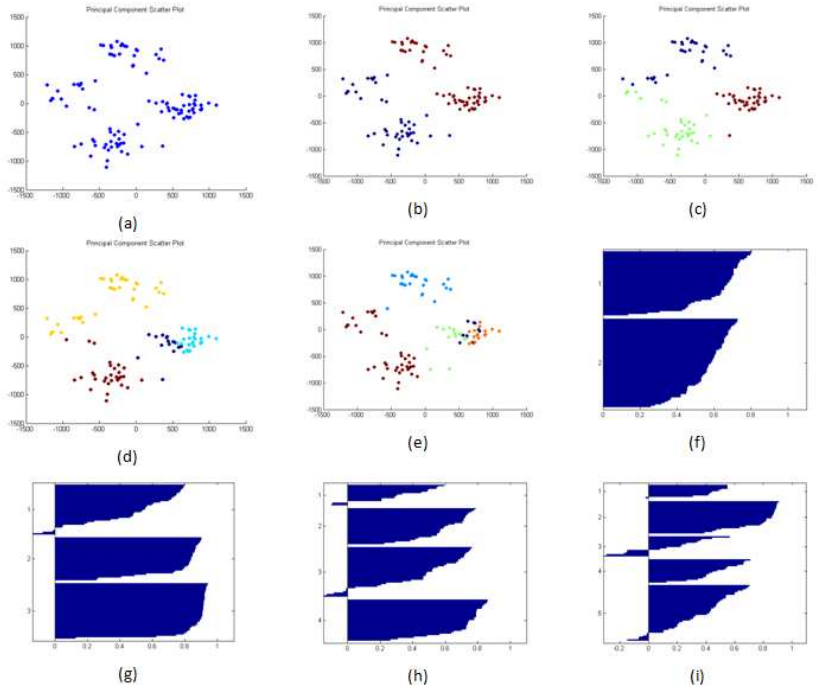

Figure 17. The experiments result of k-means clustering for the data containing 5 % of duplicate data and 4 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

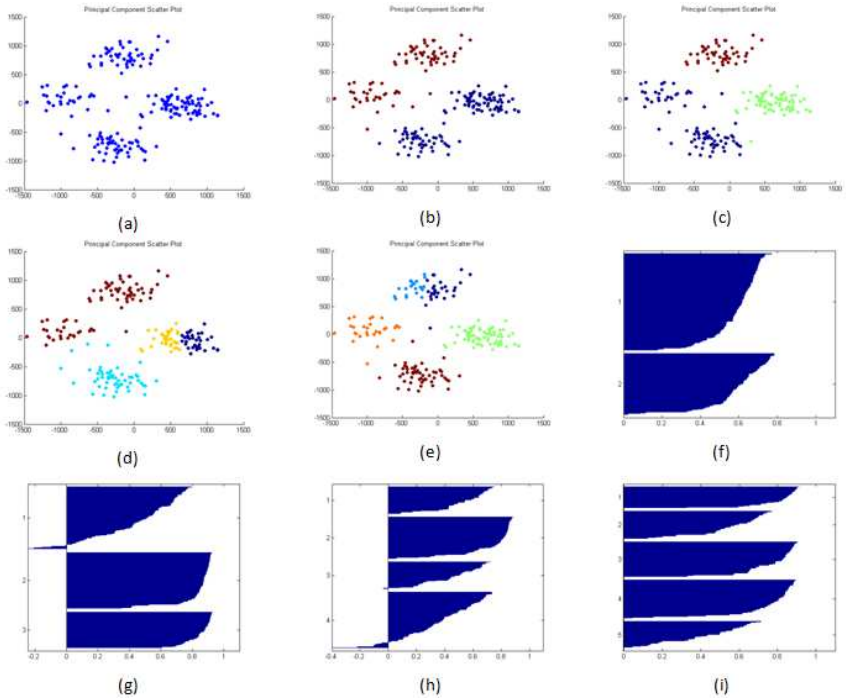

Figure 18. The experiments result of k-means clustering for the data containing 10 % of duplicate data and 4 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

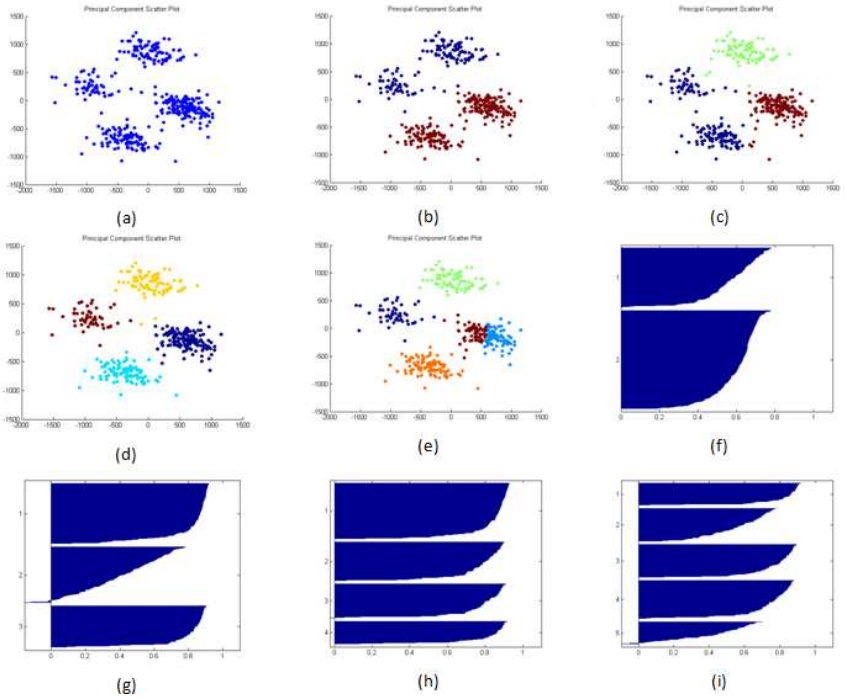

Figure 19. The experiments result of k-means clustering for the data containing 20 % of duplicate data and 4 actual cluster. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

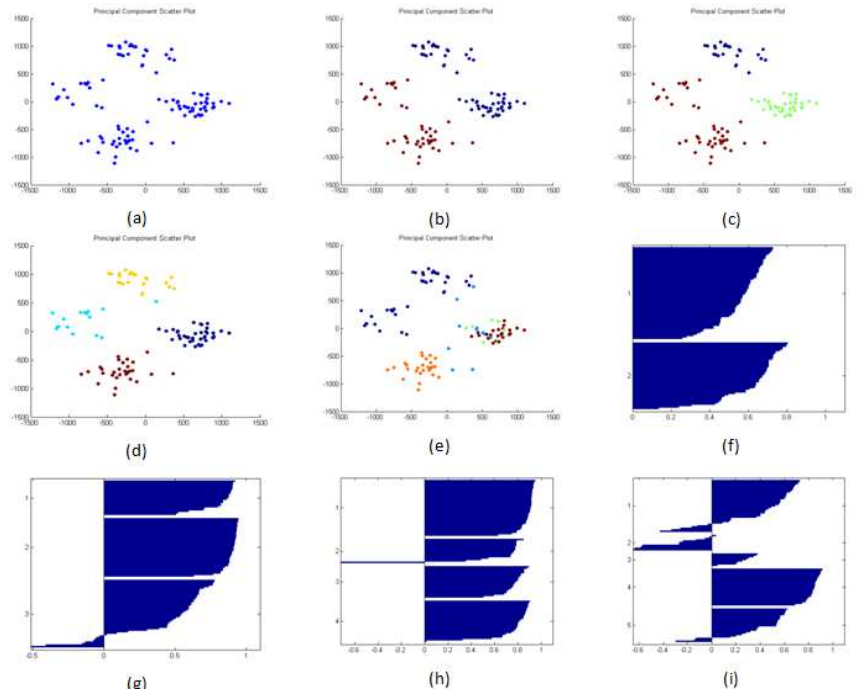

Figure 20. The experiments result of gaussian mixture clustering for the data containing 5 % of duplicate data and 4 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

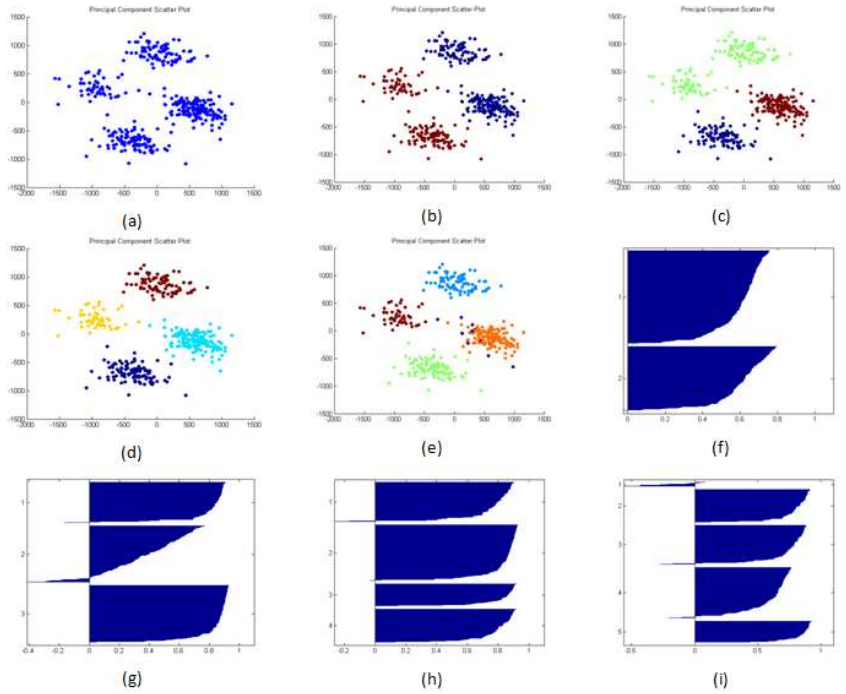

Figure 21. The experiments result of gaussian mixture clustering for the data containing 10 % of duplicate data and 4 actual clusters**.** In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

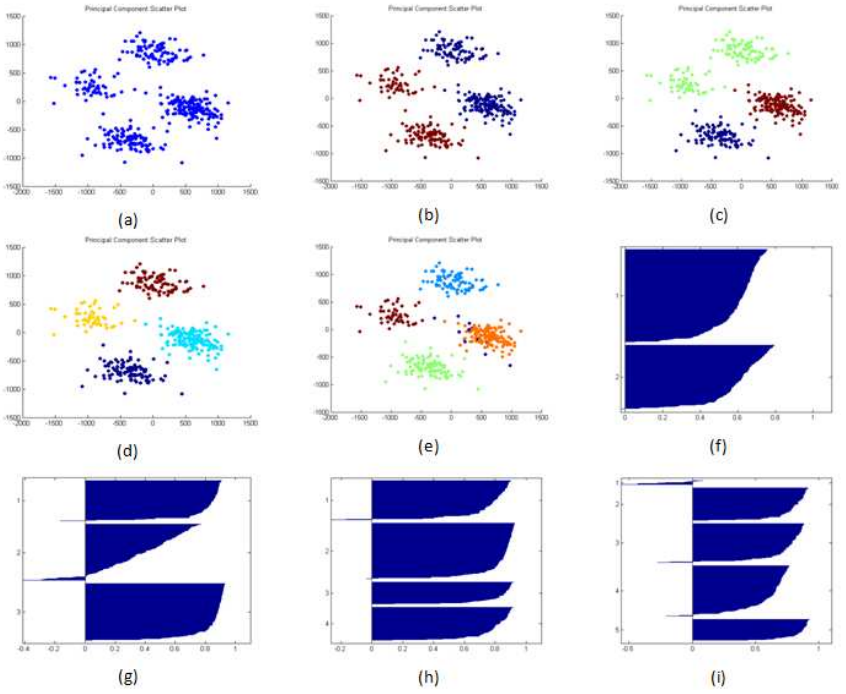

Figure 22. The experiments result of gaussian mixture clustering for the data containing 20 % of duplicate data and 4 actual clusters. In this research we employed Principal Component scatter plot (PCSP) to depict the cluster result and silhouette index to evaluate the cluster quality. The x axis of PCSP describes the first eigenvalue and the y axis describes the second eigenvalue. Meanwhile for silhouette index, the x axis is the index value and the y axis is the cluster number. (a) PCSP of Integrated data, (b) PCSP of data clustered in 2 clusters, (c) PCSP of data clustered in 3 clusters, (d) PCSP of data clustered in 4 clusters, (e) PCSP of data clustered in 4 clusters, (f) Silhouette index of b, (g) Silhouette index of c, (h) Silhouette index of d, and (i) Silhouette index of e

References

[1] D. Henderson, Overcoming Data Integration Problems, InfoManagement Direct, http://www.information-management.com/infodirect/19980701/924-1.html, 1998, retrieved December 2, 2010.

[2] P. Buxmann, L. Díaz, & E. Wüstner, "XML-Based Supply Chain Management--As SIMPLEX as It Is" *In Proceeding 35th Annual Hawaii International Conference on System Sciences (HICSS'02)*, p. 168, 2002.

[3] G. McGrath & E. More, "Data Integration along the Healthcare Supply Chain: The Pharmaceutical Extranet Gateway" *In Proceeding 34th Annual Hawaii International Conference on System Sciences (HICSS-34)*, p. 6004, 2001.

[4] L.He, G. Xin, & G.Yufeng, "Study on the Design of CRM System Based on Business Intelligence" *In First International Workshop on Knowledge Discovery and Data Mining,* pp. 185-189, 2008.

[5] R.T. Herschel & N.E. Jones, "Knowledge Management and Business Intelligence: The Importance of Integration," *Journal of Knowledge Management*, vol. 9, pp. 45-55. 2005.