Machine Learning for Data Mining

Cluster Validity

Andres Mendez-Vazquez

August 3, 2020

Outline

Outline

What is a Good Clustering?

Internal criterion

A good clustering will produce high quality clusters in which:

- The **intra-class** (that is, intra-cluster) similarity is high.
- The **inter-class** similarity is low.
- The measured quality of a clustering depends on both the document representation and the similarity measure used.

What is a Good Clustering?

Internal criterion

A good clustering will produce high quality clusters in which:

- The **intra-class** (that is, intra-cluster) similarity is high.
- The **inter-class** similarity is low.
- The measured quality of a clustering depends on both the document representation and the similarity measure used.

What is a Good Clustering?

Internal criterion

A good clustering will produce high quality clusters in which:

- The **intra-class** (that is, intra-cluster) similarity is high.
- The **inter-class** similarity is low.
- The measured quality of a clustering depends on both the document representation and the similarity measure used.

What is a Good Clustering?

Internal criterion

A good clustering will produce high quality clusters in which:

- The **intra-class** (that is, intra-cluster) similarity is high.
- The **inter-class** similarity is low.
- The measured quality of a clustering depends on both the document representation and the similarity measure used.

Problem

Many of the Clustering Algorithms

They impose a clustering structure on the data, even though the data may not posses any.

This is why

Cluster analysis is not a panacea.

Therefore

It is necessary to have an indication that the vectors of X form clusters before we apply a clustering algorithm.

Problem

Many of the Clustering Algorithms

They impose a clustering structure on the data, even though the data may not posses any.

This is why

Cluster analysis is not a panacea.

Therefore

It is necessary to have an indication that the vectors of X form clusters before we apply a clustering algorithm.

Problem

Many of the Clustering Algorithms

They impose a clustering structure on the data, even though the data may not posses any.

This is why

Cluster analysis is not a panacea.

Therefore

It is necessary to have an indication that the vectors of X form clusters before we apply a clustering algorithm.

Then

The problem of verifying whether X possesses a clustering structure

Without identifying it explicitly, this is known as clustering tendency.

What can happen if X posses a cluster structure?

A different kind of problem is encountered now

- All the previous algorithms require knowledge of the values of specific parameters.
- Some of them impose restrictions on the shape of the clusters.

What can happen if X posses a cluster structure?

A different kind of problem is encountered now

- All the previous algorithms require knowledge of the values of specific parameters.
- Some of them impose restrictions on the shape of the clusters.

It is more

Poor estimation of these parameters and inappropriate restrictions on the shape of the clusters may lead to incorrect conclusions about the clustering structure of X.

- This task is known as cluster validity.

What can happen if X posses a cluster structure?

A different kind of problem is encountered now

- All the previous algorithms require knowledge of the values of specific parameters.
- Some of them impose restrictions on the shape of the clusters.

It is more

Poor estimation of these parameters and inappropriate restrictions on the shape of the clusters may lead to incorrect conclusions about the clustering structure of X.

Thus, it is necessary to discuss

- Methods suitable for quantitative evaluation of the results of a clustering algorithm.
- This task is known as **cluster validity**.

What can happen if X posses a cluster structure?

A different kind of problem is encountered now

- All the previous algorithms require knowledge of the values of specific parameters.
- Some of them impose restrictions on the shape of the clusters.

It is more

Poor estimation of these parameters and inappropriate restrictions on the shape of the clusters may lead to incorrect conclusions about the clustering structure of X.

Thus, it is necessary to discuss

- Methods suitable for quantitative evaluation of the results of a clustering algorithm.
- This task is known as **cluster validity**.

What can happen if X posses a cluster structure?

A different kind of problem is encountered now

- All the previous algorithms require knowledge of the values of specific parameters.
- Some of them impose restrictions on the shape of the clusters.

It is more

Poor estimation of these parameters and inappropriate restrictions on the shape of the clusters may lead to incorrect conclusions about the clustering structure of X.

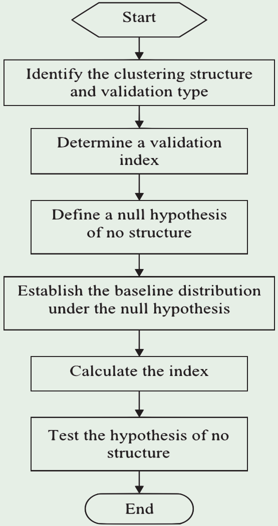
Thus, it is necessary to discuss

- Methods suitable for quantitative evaluation of the results of a clustering algorithm.
- This task is known as **cluster validity.**

Outline

The Cluster Flowchart

Flowchart of the validity paradigm for clustering structures



Outline

Hypothesis Testing Revisited

Hypothesis

$$
\begin{aligned}
H_1 :& \theta \neq \theta_0 \\
H_0 :& \theta = \theta_0
\end{aligned}
$$

In addition

Also let $\overline{D}_{\rho}$ be the critical interval corresponding to significance level ρ of a test statistics q.

Now

Given Θ_1, the set of all values that θ may take under hypothesis H_1.

Hypothesis Testing Revisited

Hypothesis

$$
\begin{aligned}
H_1 :& \theta \neq \theta_0 \\
H_0 :& \theta = \theta_0
\end{aligned}
$$

In addition

Also let $\overline{D}_\rho$ be the critical interval corresponding to significance level ρ of a test statistics q.

Now

Given Θ_1, the set of all values that θ may take under hypothesis H_1.

Hypothesis Testing Revisited

Hypothesis

$$
\begin{aligned}
H_1 :& \theta \neq \theta_0 \\
H_0 :& \theta = \theta_0
\end{aligned}
$$

In addition

Also let $\overline{D}_\rho$ be the critical interval corresponding to significance level ρ of a test statistics q.

Now

Given Θ_1, the set of all values that θ may take under hypothesis H_1.

Power Function

Definition (Power Function)

$$W(\theta) = P\left(q \in \overline{D}_{\rho} | \theta \in \Theta_1\right) \tag{1}$$

Power Function

Definition (Power Function)

$$W\left(\theta\right) = P\left(q \in \overline{D}_{\rho} | \theta \in \Theta_1\right) \quad (1)$$

Meaning

- $W\left(\theta\right)$ is the probability that q lies in the critical region when the value of the parameter vector θ.

- The test whose power under the alternative hypothesis is greater is always preferred.

Power Function

Definition (Power Function)

$$W(\theta) = P\left(q \in \overline{D}_{\rho} | \theta \in \Theta_1\right) \quad (1)$$

Meaning

- $W(\theta)$ is the probability that q lies in the critical region when the value of the parameter vector θ.

Thus

- The power function can be used for the comparison of two different statistical tests.
- The test whose power under the alternative hypotheses is greater is always preferred.

Power Function

Definition (Power Function)

$$W(\theta) = P\left(q \in \overline{D}_\rho | \theta \in \Theta_1\right) \quad (1)$$

Meaning

- $W(\theta)$ is the probability that q lies in the critical region when the value of the parameter vector θ.

Thus

- The power function can be used for the comparison of two different statistical tests.
- The test whose power under the alternative hypotheses is greater is always preferred.

There are two types of errors associated with a statistical test

Suppose that H_0 is true

- If $q(\boldsymbol{x}) \in \overline{D}_\rho$, H_0 will be rejected even if it is true.
- It is called a type error I.
- The probability of such error is ρ.
- The probability of accepting H_0 when it is true is $1-\rho$.

There are two types of errors associated with a statistical test

Suppose that H_0 is true

- If $q(\boldsymbol{x}) \in \overline{D}_{\rho}$, H_0 will be rejected even if it is true.
- It is called a type error I.
- The probability of such error is ρ.
- The probability of accepting H_0 when it is true is $1 - \rho$.

Suppose that H_0 is false

- If $q(\boldsymbol{x}) \notin \overline{D}_{\rho}$, H_0 will be accepted even if it is false.
- It is called a type error I.
- The probability of such error is $1 - W(\theta)$.
- This depends on the specific value of θ.

There are two types of errors associated with a statistical test

Suppose that H_0 is true

- If $q(\boldsymbol{x}) \in \overline{D}_\rho$, H_0 will be rejected even if it is true.
- It is called a type error I.
- The probability of such error is ρ.
- The probability of accepting H_0 when it is true is $1 - \rho$.

Suppose that H_0 is false

- If $q(\boldsymbol{x}) \notin \overline{D}_\rho$, H_0 will be accepted even if it is false.
- It is called a type error I.
- The probability of such error is $1 - W(\theta)$.
- This depends on the specific value of θ.

There are two types of errors associated with a statistical test

Suppose that H_0 is true

- If $q(\boldsymbol{x}) \in \overline{D}_\rho$, H_0 will be rejected even if it is true.
- It is called a type error I.
- The probability of such error is ρ.
- The probability of accepting H_0 when it is true is $1 - \rho$.

Suppose that H_0 is false

- If $q(\boldsymbol{x}) \notin \overline{D}_\rho$, H_0 will be accepted even if it is false.
- It is called a type error I.
- The probability of such error is $1 - W(\theta)$.
- This depends on the specific value of θ.

There are two types of errors associated with a statistical test

Suppose that H_0 is true

- If $q(\boldsymbol{x}) \in \overline{D}_\rho$, H_0 will be rejected even if it is true.
- It is called a type error I.
- The probability of such error is ρ.
- The probability of accepting H_0 when it is true is $1 - \rho$.

Suppose that H_0 is false

- If $q(\boldsymbol{x}) \notin \overline{D}_\rho$, H_0 will be accepted even if it is false.
- It is called a type error I.
- The probability of such error is $1 - W(\theta)$.
- This depends on the specific value of θ.

There are two types of errors associated with a statistical test

Suppose that H_0 is true

- If $q(\boldsymbol{x}) \in \overline{D}_\rho$, H_0 will be rejected even if it is true.
- It is called a type error I.
- The probability of such error is ρ.
- The probability of accepting H_0 when it is true is $1 - \rho$.

Suppose that H_0 is false

- If $q(\boldsymbol{x}) \notin \overline{D}_\rho$, H_0 will be accepted even if it is false.
- It is called a type error I.
- The probability of such error is $1 - W(\theta)$.
- This depends on the specific value of θ.

There are two types of errors associated with a statistical test

Suppose that H_0 is true

- If $q(\boldsymbol{x}) \in \overline{D}_\rho$, H_0 will be rejected even if it is true.
- It is called a type error I.
- The probability of such error is ρ.
- The probability of accepting H_0 when it is true is $1 - \rho$.

Suppose that H_0 is false

- If $q(\boldsymbol{x}) \notin \overline{D}_\rho$, H_0 will be accepted even if it is false.
- It is called a type error I.
- The probability of such error is $1 - W(\theta)$.
- This depends on the specific value of θ.

There are two types of errors associated with a statistical test

Suppose that H_0 is true

- If $q(\boldsymbol{x}) \in \overline{D}_\rho$, H_0 will be rejected even if it is true.
- It is called a type error I.
- The probability of such error is ρ.
- The probability of accepting H_0 when it is true is $1 - \rho$.

Suppose that H_0 is false

- If $q(\boldsymbol{x}) \notin \overline{D}_\rho$, H_0 will be accepted even if it is false.
- It is called a type error I.
- The probability of such error is $1 - W(\theta)$.
- This depends on the specific value of θ.

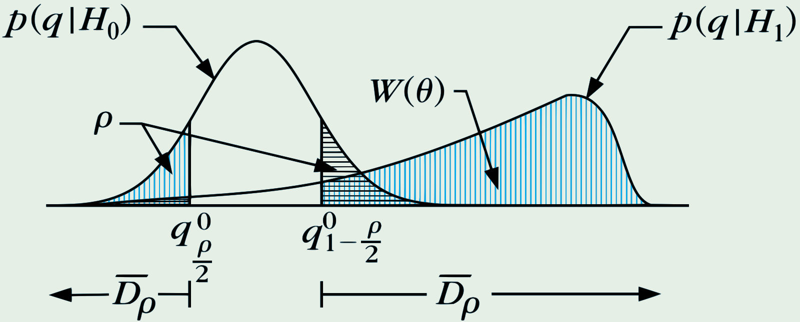
Thus

Something Notable

- The probability density function (pdf) of the statistic q, under H_0 , for most of the statistics used in practice has a single maximum.
- In addition, the region $\overline{D}_{\rho}$, is either a half-line or the union of two half-lines.

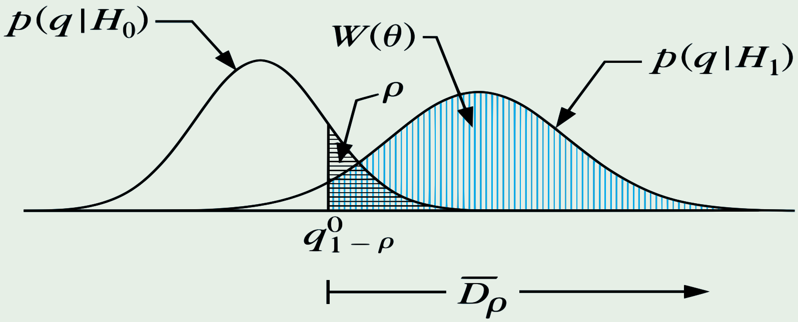
Example

$\overline{D}_\rho$ is the union of two half-lines (A two-tailed statistical test)



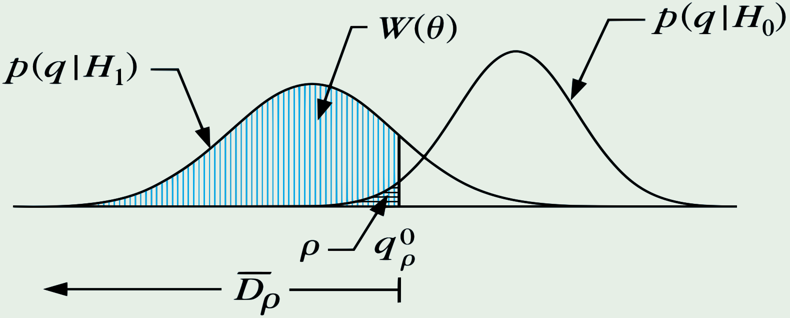
Example

A right-tailed test



Example

A left-tailed test



Problem

In many practical cases

The exact form of the pdf of a statistic q, under a given hypothesis, is not available and it is difficult to obtain.

Problem

In many practical cases

The exact form of the pdf of a statistic q, under a given hypothesis, is not available and it is difficult to obtain.

However, we can do the following

- Monte Carlo techniques.
- Bootstrapping techniques.

Problem

In many practical cases

The exact form of the pdf of a statistic q, under a given hypothesis, is not available and it is difficult to obtain.

However, we can do the following

- Monte Carlo techniques.
- Bootstrapping techniques.

Outline

Monte Carlo techniques

What do they do?

They rely on simulating the process at hand using a sufficient number of computer-generated data.

- Given enough data, we can try to learn the pdf for q.
- Then, using that pdf we simulate samples of q.

Monte Carlo techniques

What do they do?

They rely on simulating the process at hand using a sufficient number of computer-generated data.

- Given enough data, we can try to learn the pdf for q.
- Then, using that pdf we simulate samples of q.

Thus

For each of the say r, data sets, X_i, we compute the value of q, denoted by q_i.

Monte Carlo techniques

What do they do?

They rely on simulating the process at hand using a sufficient number of computer-generated data.

- Given enough data, we can try to learn the pdf for q.
- Then, using that pdf we simulate samples of q.

Thus

For each of the say r, data sets, X_i, we compute the value of q, denoted by q_i.

Monte Carlo techniques

What do they do?

They rely on simulating the process at hand using a sufficient number of computer-generated data.

- Given enough data, we can try to learn the pdf for q.
- Then, using that pdf we simulate samples of q.

Thus

For each of the say r, data sets, X_i, we compute the value of q, denoted by q_i.

Monte Carlo techniques

What do they do?

They rely on simulating the process at hand using a sufficient number of computer-generated data.

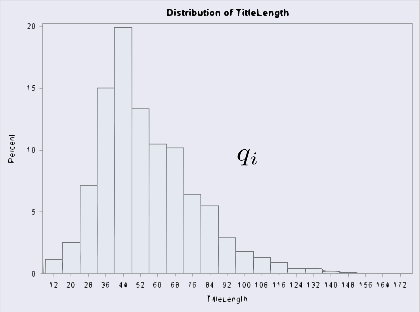
- Given enough data, we can try to learn the pdf for q.
- Then, using that pdf we simulate samples of q.

Thus

For each of the say r, data sets, X_i, we compute the value of q, denoted by q_i.

Then

We construct the corresponding histogram of these values



Using this approximation

Assume

- q corresponds to a right-tailed statistical test.
- A histogram is constructed using r values of q corresponding to the r data sets.

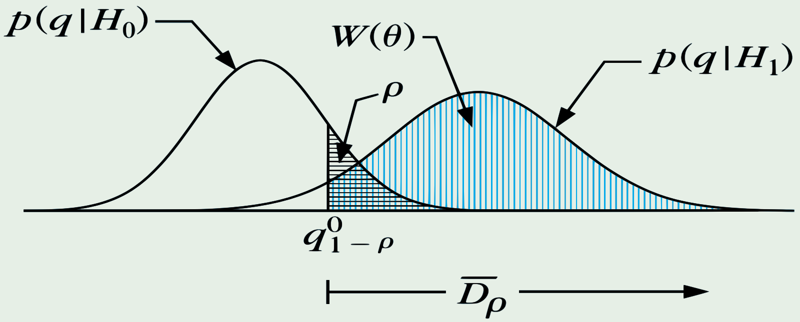
Using this approximation

Assume

- q corresponds to a right-tailed statistical test.
- A histogram is constructed using r values of q corresponding to the r data sets.

Using this approximation

A right-tailed test



Thus

Then, acceptance or rejection may be based on the rules

- Reject H_0, if q is greater than $(1-\rho)\,r$ of the q_i values.
- Accept H_0, if q is smaller than $(1-\rho)\,r$ of the q_i values.

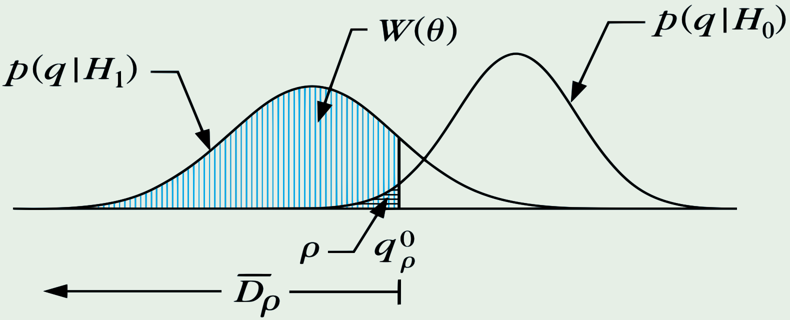
Thus

Then, acceptance or rejection may be based on the rules

- Reject H_0, if q is greater than $(1-\rho)\,r$ of the q_i values.
- Accept H_0, if q is smaller than $(1-\rho)\,r$ of the q_i values.

Next

For a left-tailed test



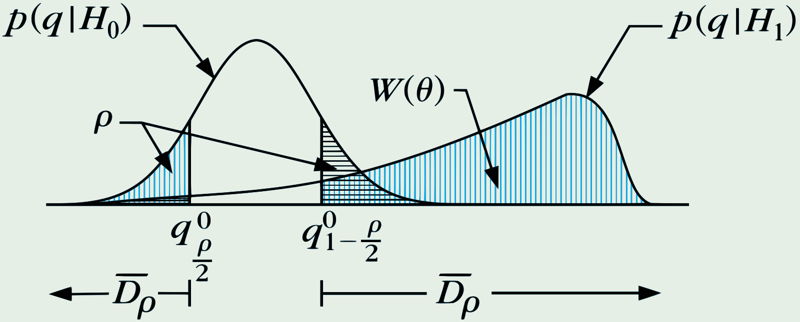
Next

For a left-tailed test, rejection or acceptance of the null hypothesis is done on the basis

1. Reject H_0, if q is smaller than ρr of the q_i values.
2. Accept H_0, if q is greater than ρr of the q_i values

Now, for two-tailed statistical test

A two-tailed statistical test



Thus

For a two-tailed test we have

Accept H_0, if q is greater than $\left(\frac{\rho}{2}\right) r$ of the q_i values and less than $\left(1 - \frac{\rho}{2}\right) r$ of the q_i values.

Outline

Bootstrapping Techniques

Why?

- They constitute an alternative way to cope with a limited amount of data.

Then

- The idea here is to parameterize the unknown pdf in terms of an unknown parameter.

How

- To cope with the limited amount of data and in order to improve the accuracy of the estimate of the unknown pdf parameter.

Bootstrapping Techniques

Why?

- They constitute an alternative way to cope with a limited amount of data.

Then

- The idea here is to parameterize the unknown pdf in terms of an unknown parameter.

How

- To cope with the limited amount of data and in order to improve the accuracy of the estimate of the unknown pdf parameter.

Bootstrapping Techniques

Why?

- They constitute an alternative way to cope with a limited amount of data.

Then

- The idea here is to parameterize the unknown pdf in terms of an unknown parameter.

How

- To cope with the limited amount of data and in order to improve the accuracy of the estimate of the unknown pdf parameter.

The Process

For this, we create

- Several "fake"data sets $X_1, ..., X_r$ are created by sampling X with replacement.

Thus

- By using this sample, we estimate the desired pdf for q.

Then

- Typically, good estimates are obtained if r is between 100 and 200.

The Process

For this, we create

- Several “fake”data sets $X_1, ..., X_r$ are created by sampling X with replacement.

Thus

- By using this sample, we estimate the desired pdf for q.

Then

- Typically, good estimates are obtained if r is between 100 and 200.

The Process

For this, we create

- Several “fake”data sets $X_1, ..., X_r$ are created by sampling X with replacement.

Thus

- By using this sample, we estimate the desired pdf for q.

Then

- Typically, good estimates are obtained if r is between 100 and 200.

Outline

Which Hypothesis?

Random position hypothesis

H_0: All the locations of N data points in some specific region of a d-dimensional space are equally likely.

Random graph hypothesis

H_0: All $N \times N$ rank order proximity matrices are equally likely.

Random label hypothesis

H_0: All permutations of the labels on N data objects are equally likely.

Which Hypothesis?

Random position hypothesis

H_0: All the locations of N data points in some specific region of a d-dimensional space are equally likely.

Random graph hypothesis

H_0: All $N \times N$ rank order proximity matrices are equally likely.

Random label hypothesis

H_0: All permutations of the labels on N data objects are equally likely.

Which Hypothesis?

Random position hypothesis

H_0: All the locations of N data points in some specific region of a d-dimensional space are equally likely.

Random graph hypothesis

H_0: All $N \times N$ rank order proximity matrices are equally likely.

Random label hypothesis

H_0: All permutations of the labels on N data objects are equally likely.

Thus

We must define an appropriate statistic

- Whose values are indicative of the structure of a data set, and compare the value that results from our data set X against the value obtained from the reference (random) population.

Random Population

- In order to obtain the baseline distribution under the null hypothesis, statistical sampling techniques like Monte Carlo analysis and bootstrapping are used (Jain and Dubes, 1988).

Thus

We must define an appropriate statistic

- Whose values are indicative of the structure of a data set, and compare the value that results from our data set X against the value obtained from the reference (random) population.

Random Population

- In order to obtain the baseline distribution under the null hypothesis, statistical sampling techniques like Monte Carlo analysis and bootstrapping are used (Jain and Dubes, 1988).

For example

Random position hypothesis - appropriate for ratio data

- All the arrangements of N vectors in a specific region of the d-dimensional space are equally likely to occur.

How?

- One way to produce such an arrangement is to insert each point randomly in this region of the d-dimensional space, according to a uniform distribution.

The random position hypothesis

- It can be used with either external or internal criterion.

For example

Random position hypothesis - appropriate for ratio data

- All the arrangements of N vectors in a specific region of the d-dimensional space are equally likely to occur.

How?

- One way to produce such an arrangement is to insert each point randomly in this region of the d-dimensional space, according to a uniform distribution.

The random position hypothesis

- It can be used with either external or internal criterion.

For example

Random position hypothesis - appropriate for ratio data

- All the arrangements of N vectors in a specific region of the d-dimensional space are equally likely to occur.

How?

- One way to produce such an arrangement is to insert each point randomly in this region of the d-dimensional space, according to a uniform distribution.

The random position hypothesis

- It can be used with either external or internal criterion.

Outline

Internal criteria

What do we do?

- In this case, the statistic q is defined so as to measure the degree to which a clustering structure, produced by a clustering algorithm, matches the proximity matrix of the corresponding data set.

Internal criteria

What do we do?

- In this case, the statistic q is defined so as to measure the degree to which a clustering structure, produced by a clustering algorithm, matches the proximity matrix of the corresponding data set.

We apply our clustering algorithm to the following data set

1. Let X_i be a set of N vectors generated according to the random position hypothesis.
2. P_i be the corresponding proximity matrix.
3. C_i the corresponding clustering.

Internal criteria

What do we do?

- In this case, the statistic q is defined so as to measure the degree to which a clustering structure, produced by a clustering algorithm, matches the proximity matrix of the corresponding data set.

We apply our clustering algorithm to the following data set

1. Let X_i be a set of N vectors generated according to the random position hypothesis.
2. P_i be the corresponding proximity matrix.
3. C_i the corresponding clustering.

Then, we apply our algorithm over the real data X to obtain C

- Now, we compute the statistics q for each clustering structure.

Internal criteria

What do we do?

- In this case, the statistic q is defined so as to measure the degree to which a clustering structure, produced by a clustering algorithm, matches the proximity matrix of the corresponding data set.

We apply our clustering algorithm to the following data set

1. Let X_i be a set of N vectors generated according to the random position hypothesis.
2. P_i be the corresponding proximity matrix.
3. C_i the corresponding clustering.

Then, we apply our algorithm over the real data X to obtain C

- Now, we compute the statistics q for each clustering structure.

Internal criteria

What do we do?

- In this case, the statistic q is defined so as to measure the degree to which a clustering structure, produced by a clustering algorithm, matches the proximity matrix of the corresponding data set.

We apply our clustering algorithm to the following data set

1. Let X_i be a set of N vectors generated according to the random position hypothesis.
2. P_i be the corresponding proximity matrix.
3. C_i the corresponding clustering.

Then, we apply our algorithm over the real data X to obtain C

- Now, we compute the statistics q for each clustering structure.

Then, we use our hypothesis H_0

The random hypothesis H_0

- It is rejected if the value q, resulting from X lies in the critical interval $\overline{D}_\rho$ of the statistic pdf of the reference random population.

Meaning

- if q is unusually small or large.

Then, we use our hypothesis H_0

The random hypothesis H_0

- It is rejected if the value q, resulting from X lies in the critical interval $\overline{D}_\rho$ of the statistic pdf of the reference random population.

Meaning

- if q is unusually small or large.

Also a External Criteria can be used

Definition

- The statistic q is defined to measure the degree of correspondence between a **prespecified structure** $\mathcal{P}$ imposed on X.
- And the clustering that results after the application of a specific clustering algorithm.

Also a External Criteria can be used

Definition

- The statistic q is defined to measure the degree of correspondence between a **prespecified structure** $\mathcal{P}$ imposed on X.
- And the clustering that results after the application of a specific clustering algorithm.

Then

- Then, the value of q corresponding to the clustering C resulting from the data set X is tested against the q_i's.
- These q_i's correspond to the clusterings resulting from the reference population generated under the random position hypothesis.

Also a External Criteria can be used

Definition

- The statistic q is defined to measure the degree of correspondence between a **prespecified structure** $\mathcal{P}$ imposed on X.
- And the clustering that results after the application of a specific clustering algorithm.

Then

- Then, the value of q corresponding to the clustering C resulting from the data set X is tested against the q_i's.
- These q_i's correspond to the clusterings resulting from the reference population generated under the random position hypothesis.

Again

- The random hypothesis is rejected if q is unusually large or small.

Also a External Criteria can be used

Definition

- The statistic q is defined to measure the degree of correspondence between a **prespecified structure** $\mathcal{P}$ imposed on X.
- And the clustering that results after the application of a specific clustering algorithm.

Then

- Then, the value of q corresponding to the clustering C resulting from the data set X is tested against the q_i's.
- These q_i's correspond to the clusterings resulting from the reference population generated under the random position hypothesis.

Again

- The random hypothesis is rejected if q is unusually large or small.

Also a External Criteria can be used

Definition

- The statistic q is defined to measure the degree of correspondence between a **prespecified structure** $\mathcal{P}$ imposed on X.
- And the clustering that results after the application of a specific clustering algorithm.

Then

- Then, the value of q corresponding to the clustering C resulting from the data set X is tested against the q_i's.
- These q_i's correspond to the clusterings resulting from the reference population generated under the random position hypothesis.

Again

- The random hypothesis is rejected if q is unusually large or small.

Nevertheless

There are more examples

- In chapter 16 in the book of Theodoridis.

Outline

1. Introduction
   - What is a Good Clustering?
2. Cluster Validity
   - The Process
   - Hypothesis Testing
     - Monte Carlo techniques
     - Bootstrapping Techniques
   - Which Hypothesis?
   - Hypothesis Testing in Cluster Validity
   - External Criteria
   - Relative Criteria
     - Hard Clustering

External Criteria Usage

First

- For the comparison of a clustering structure C, produced by a clustering algorithm, with a partition $\mathcal{P}$ of X drawn independently from C.

Second

- For measuring the degree of agreement between a predetermined partition $\mathcal{P}$ and the proximity matrix of X, P.

External Criteria Usage

First

- For the comparison of a clustering structure C, produced by a clustering algorithm, with a partition $\mathcal{P}$ of X drawn independently from C.

Second

- For measuring the degree of agreement between a predetermined partition $\mathcal{P}$ and the proximity matrix of X, P.

Comparison of $\mathcal{P}$ with a Clustering $\mathcal{C}$

First

- In this case, $\mathcal{C}$ may be either a specific hierarchy of clusterings or a specific clustering.

However

- The problem with the hierarchical clustering is the cutting in the correct level of the dendogram.

Then

- We will concentrate on clustering that does not imply hierarchical clustering.

Comparison of $\mathcal{P}$ with a Clustering $\mathcal{C}$

First

- In this case, $\mathcal{C}$ may be either a specific hierarchy of clusterings or a specific clustering.

However

- The problem with the hierarchical clustering is the cutting in the correct level of the dendogram.

Then

- We will concentrate on clustering that does not imply hierarchical clustering.

Comparison of $\mathcal{P}$ with a Clustering $\mathcal{C}$

First

- In this case, $\mathcal{C}$ may be either a specific hierarchy of clusterings or a specific clustering.

However

- The problem with the hierarchical clustering is the cutting in the correct level of the dendogram.

Then

- We will concentrate on clustering that does not imply hierarchical clustering.

Thus

Setup

- Consider a clustering $\mathcal{C}$ given by a specific clustering algorithm.
- This clustering is done in a independently drawn partition $\mathcal{P}$.

Thus

Setup

- Consider a clustering $\mathcal{C}$ given by a specific clustering algorithm.
- This clustering is done in a independently drawn partition $\mathcal{P}$.

Thus, we obtain the following sets

- $\mathcal{C} = \{C_1, C_2, ..., C_m\}$
- $\mathcal{P} = \{P_1, P_2, ..., P_s\}$

Note that the number of clusters in $\mathcal{C}$ need not be the same as the number of groups in $\mathcal{P}$.

Thus

Setup

- Consider a clustering $\mathcal{C}$ given by a specific clustering algorithm.
- This clustering is done in a independently drawn partition $\mathcal{P}$.

Thus, we obtain the following sets

- $\mathcal{C} = \{C_1, C_2, ..., C_m\}$
- $\mathcal{P} = \{P_1, P_2, ..., P_s\}$

Note that the number of clusters in $\mathcal{C}$ need not be the same as the number of groups in $\mathcal{P}$.

Thus

Setup

- Consider a clustering $\mathcal{C}$ given by a specific clustering algorithm.
- This clustering is done in a independently drawn partition $\mathcal{P}$.

Thus, we obtain the following sets

- $\mathcal{C} = \{C_1, C_2, ..., C_m\}$
- $\mathcal{P} = \{P_1, P_2, ..., P_s\}$

Note that the number of clusters in $\mathcal{C}$ need not be the same as the number of groups in $\mathcal{P}$.

Thus

Setup

- Consider a clustering $\mathcal{C}$ given by a specific clustering algorithm.
- This clustering is done in a independently drawn partition $\mathcal{P}$.

Thus, we obtain the following sets

- $\mathcal{C} = \{C_1, C_2, ..., C_m\}$
- $\mathcal{P} = \{P_1, P_2, ..., P_s\}$

Note that the number of clusters in $\mathcal{C}$ need not be the same as the number of groups in $\mathcal{P}$.

Thus

Setup

- Consider a clustering $\mathcal{C}$ given by a specific clustering algorithm.
- This clustering is done in a independently drawn partition $\mathcal{P}$.

Thus, we obtain the following sets

- $\mathcal{C} = \{C_1, C_2, ..., C_m\}$
- $\mathcal{P} = \{P_1, P_2, ..., P_s\}$

Note that the number of clusters in $\mathcal{C}$ need not be the same as the number of groups in $\mathcal{P}$.

Now

Consider the following

- Let n_{ij} denote the number of vectors that belong to C_i and P_j simultaneously.
- Also $n_i^C = \sum_{j=1}^{s} n_{ij}$.
  - It is the number of vectors that belong to C_i.
- Similarly $n_j^P = \sum_{i=1}^{m} n_{ij}$
  - The number of vectors that belong to P_j.

Now

Consider the following

- Let n_{ij} denote the number of vectors that belong to C_i and P_j simultaneously.
- Also $n_i^C = \sum_{j=1}^{s} n_{ij}$.
  - It is the number of vectors that belong to C_i.
- Similarly $n_j^P = \sum_{i=1}^{m} n_{ij}$
  - The number of vectors that belong to P_j.

Now

Consider the following

- Let n_{ij} denote the number of vectors that belong to C_i and P_j simultaneously.
- Also $n_i^C = \sum_{j=1}^{s} n_{ij}$.
  - It is the number of vectors that belong to C_i.
- Similarly $n_j^P = \sum_{i=1}^{m} n_{ij}$
  - The number of vectors that belong to P_j.

Consider the a pair of vectors $(\boldsymbol{x}_v, \boldsymbol{x}_u)$

Thus, we have the following cases

- Case 1: If both vectors belong to the same cluster in $\mathcal{C}$ and to the same group in $\mathcal{P}$.
- Case 2: if both vectors belong to the same cluster in $\mathcal{C}$ and to different groups in $\mathcal{P}$.
- Case 3: if both vectors belong to the different clusters in $\mathcal{C}$ and to the same group in $\mathcal{P}$.
- Case 4: if both vectors belong to different clusters in $\mathcal{C}$ and to different groups in $\mathcal{P}$.

Consider the a pair of vectors $(\boldsymbol{x}_v, \boldsymbol{x}_u)$

Thus, we have the following cases

- Case 1: If both vectors belong to the same cluster in $\mathcal{C}$ and to the same group in $\mathcal{P}$.
- Case 2: if both vectors belong to the same cluster in $\mathcal{C}$ and to different groups in $\mathcal{P}$.
- Case 3: if both vectors belong to the different clusters in $\mathcal{C}$ and to the same group in $\mathcal{P}$.
- Case 4: if both vectors belong to different clusters in $\mathcal{C}$ and to different groups in $\mathcal{P}$.

Consider the a pair of vectors $(\boldsymbol{x}_v, \boldsymbol{x}_u)$

Thus, we have the following cases

- Case 1: If both vectors belong to the same cluster in $\mathcal{C}$ and to the same group in $\mathcal{P}$.
- Case 2: if both vectors belong to the same cluster in $\mathcal{C}$ and to different groups in $\mathcal{P}$.
- Case 3: if both vectors belong to the different clusters in $\mathcal{C}$ and to the same group in $\mathcal{P}$.
- Case 4: if both vectors belong to different clusters in $\mathcal{C}$ and to different groups in $\mathcal{P}$.

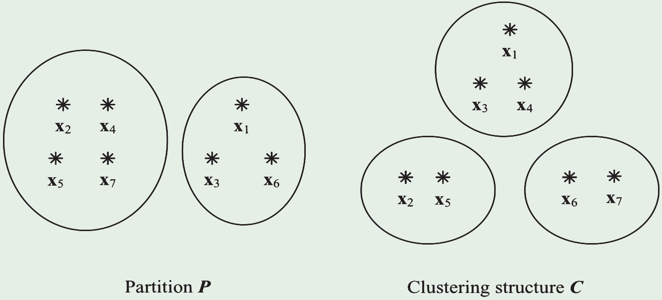
Consider the a pair of vectors $(\boldsymbol{x}_v, \boldsymbol{x}_u)$

Thus, we have the following cases

- Case 1: If both vectors belong to the same cluster in $\mathcal{C}$ and to the same group in $\mathcal{P}$.
- Case 2: if both vectors belong to the same cluster in $\mathcal{C}$ and to different groups in $\mathcal{P}$.
- Case 3: if both vectors belong to the different clusters in $\mathcal{C}$ and to the same group in $\mathcal{P}$.
- Case 4: if both vectors belong to different clusters in $\mathcal{C}$ and to different groups in $\mathcal{P}$.

Example

The numbers of pairs of points for the four cases are denoted as a, b, c, and d



Case	Pairs of data points	Total
1	$\mathbf{x}_1$ and $\mathbf{x}_3$; $\mathbf{x}_2$ and $\mathbf{x}_5$	2
2	$\mathbf{x}_1$ and $\mathbf{x}_4$; $\mathbf{x}_3$ and $\mathbf{x}_4$; $\mathbf{x}_6$ and $\mathbf{x}_7$	3
3	$\mathbf{x}_1$ and $\mathbf{x}_6$; $\mathbf{x}_2$ and $\mathbf{x}_4$; $\mathbf{x}_2$ and $\mathbf{x}_7$; $\mathbf{x}_3$ and $\mathbf{x}_6$; $\mathbf{x}_4$ and $\mathbf{x}_5$; $\mathbf{x}_4$ and $\mathbf{x}_7$; $\mathbf{x}_5$ and $\mathbf{x}_7$	7
4	$\mathbf{x}_1$ and $\mathbf{x}_2$; $\mathbf{x}_1$ and $\mathbf{x}_5$; $\mathbf{x}_1$ and $\mathbf{x}_7$; $\mathbf{x}_2$ and $\mathbf{x}_3$; $\mathbf{x}_2$ and $\mathbf{x}_6$; $\mathbf{x}_3$ and $\mathbf{x}_5$; $\mathbf{x}_3$ and $\mathbf{x}_7$; $\mathbf{x}_4$ and $\mathbf{x}_6$; $\mathbf{x}_5$ and $\mathbf{x}_6$	9

Thus

The total number of pairs of points is $\frac{N(N-1)}{2}$ denoted as M

$$a+b+c+d=M \tag{2}$$

Now

We can give some commonly used external indices for measuring the match between $\mathcal{C}$ and $\mathcal{P}$.

Thus

The total number of pairs of points is $\frac{N(N-1)}{2}$ denoted as M

$$a + b + c + d = M \tag{2}$$

Now

We can give some commonly used external indices for measuring the match between $\mathcal{C}$ and $\mathcal{P}$.

Commonly used external indices

Rand index (Rand, 1971)

$$R = \frac{a + d}{M} \tag{3}$$

Jaccard coefficient

$$J = \frac{a}{a + b + c} \tag{4}$$

Fowlkes and Mallows index (Fowlkes and Mallows, 1983)

$$FM = \sqrt{\frac{a}{a + b} \times \frac{a}{a + c}} \tag{5}$$

Commonly used external indices

Rand index (Rand, 1971)

$$R = \frac{a + d}{M} \tag{3}$$

Jaccard coefficient

$$J = \frac{a}{a + b + c} \tag{4}$$

Fowlkes and Mallows index (Fowlkes and Mallows, 1983)

$$FM = \sqrt{\frac{a}{a + b} \times \frac{a}{a + c}} \tag{5}$$

Commonly used external indices

Rand index (Rand, 1971)

$$R = \frac{a+d}{M} \tag{3}$$

Jaccard coefficient

$$J = \frac{a}{a+b+c} \tag{4}$$

Fowlkes and Mallows index (Fowlkes and Mallows, 1983)

$$FM = \sqrt{\frac{a}{a+b} \times \frac{a}{a+c}} \tag{5}$$

Explanation

Given $a + d$

The Rand statistic measures the fraction of the total number of pairs that are either case 1 or 4.

Something Notable

The Jaccard coefficient follows the same philosophy as the Rand except that it excludes case 4.

Properties

The values of these two statistics are between 0 and 1.

Explanation

Given $a + d$

The Rand statistic measures the fraction of the total number of pairs that are either case 1 or 4.

Something Notable

The Jaccard coefficient follows the same philosophy as the Rand except that it excludes case 4.

Properties

The values of these two statistics are between 0 and 1.

Explanation

Given $a + d$

The Rand statistic measures the fraction of the total number of pairs that are either case 1 or 4.

Something Notable

The Jaccard coefficient follows the same philosophy as the Rand except that it excludes case 4.

Properties

The values of these two statistics are between 0 and 1.

Commonly used external indices

Hubert's Γ statistics

$$\Gamma = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 \left(M - m_1\right)\left(M - m_2\right)}} \tag{6}$$

Property

Unusually large absolute values of suggest that $\mathcal{C}$ and $\mathcal{P}$ agree with each other.

Commonly used external indices

Hubert's Γ statistics

$$\Gamma = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m_2)}} \tag{6}$$

Property

Unusually large absolute values of suggest that $\mathcal{C}$ and $\mathcal{P}$ agree with each other.

How we use all this?

Assuming a Random Hypothesis

It is possible using a Monte Carlo Method of sampling

Example: Gibbs Sampling

1. Initialize x to some value
2. Sample each variable in the feature vector x and resample $x_i \sim P\left(x_i|x_{(i \neq j)}\right)$

How we use all this?

Assuming a Random Hypothesis

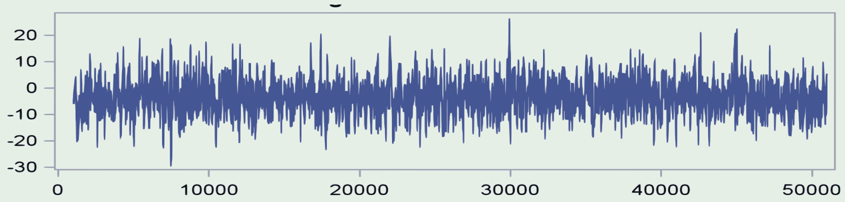
It is possible using a Monte Carlo Method of sampling

Example: Gibbs Sampling

1. Initialize $\boldsymbol{x}$ to some value
2. Sample each variable in the feature vector $\boldsymbol{x}$ and resample $x_i \sim P\left(x_i | \boldsymbol{x}_{(i \neq j)}\right)$

Example

A trace of sampling for a single variable



Algorithm and Example

Please

Go and read the section 16.3 in the Theodoridis' Book page 871 for more.

Outline

Outline

Hard Clustering

The Dunn and Dunn-like indices

Given a dissimilarity function:

$$d\left(C_i, C_j\right) = \min_{x \in C_i, y \in C_j} d\left(\boldsymbol{x}, \boldsymbol{y}\right) \tag{7}$$

The it is possible to define the diameter of a cluster

$$diam\left(C\right) = \max_{x, y \in C} d\left(\boldsymbol{x}, \boldsymbol{y}\right) \tag{8}$$

Then the Dunn Index

$$D_m = \min_{i=1,\dots,m} \left\{ \min_{j=i+1,\dots,m} \left(\frac{d\left(C_i, C_j\right)}{\max_{k=1,\dots,m} diam\left(C_k\right)} \right) \right\} \tag{9}$$

Hard Clustering

The Dunn and Dunn-like indices

Given a dissimilarity function:

$$d\left(C_i, C_j\right) = \min_{x \in C_i, y \in C_j} d\left(\boldsymbol{x}, \boldsymbol{y}\right) \tag{7}$$

The it is possible to define the diameter of a cluster

$$diam\left(C\right) = \max_{\boldsymbol{x}, \boldsymbol{y} \in C} d\left(\boldsymbol{x}, \boldsymbol{y}\right) \tag{8}$$

Then the Dunn Index

$$D_m = \min_{i=1,\ldots,m} \left\{ \min_{j=i+1,\ldots,m} \left(\frac{d\left(C_i, C_j\right)}{\max_{k=1,\ldots,m} diam\left(C_k\right)} \right) \right\} \tag{9}$$

Hard Clustering

The Dunn and Dunn-like indices

Given a dissimilarity function:

$$d\left(C_i, C_j\right) = \min_{x \in C_i, y \in C_j} d\left(\boldsymbol{x}, \boldsymbol{y}\right) \tag{7}$$

The it is possible to define the diameter of a cluster

$$diam\left(C\right) = \max_{\boldsymbol{x}, \boldsymbol{y} \in C} d\left(\boldsymbol{x}, \boldsymbol{y}\right) \tag{8}$$

Then the Dunn Index

$$D_m = \min_{i=1,..,m} \left\{ \min_{j=i+1,...,m} \left(\frac{d\left(C_i, C_j\right)}{\max_{k=1,...,m} diam\left(C_k\right)} \right) \right\} \tag{9}$$

Thus

It is possible to prove that

If X contains compact and well-separated clusters, Dunn's index will be large

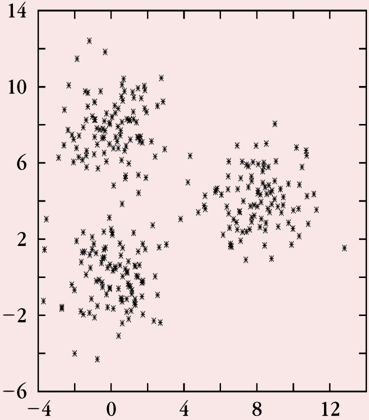
Example

Thus

It is possible to prove that

If X contains compact and well-separated clusters, Dunn's index will be large

Example



Although there are more

Please

Look at chapter 16 in the Theodoridis' book for more examples