A Principled and Flexible Framework for Finding Alternative Clusterings

Eike Brechmann

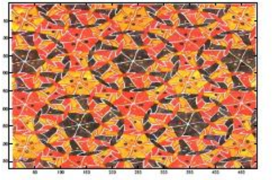
October 29, 2009

Table of Contents

Table of Contents

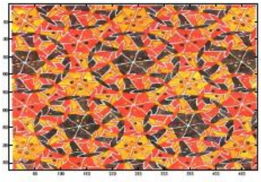
Nothing is as it seems

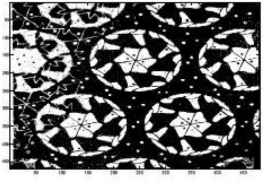
Images by M.C. Escher



Nothing is as it seems

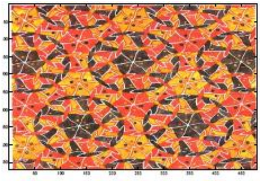
Images by M.C. Escher

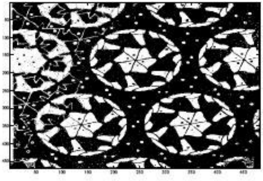


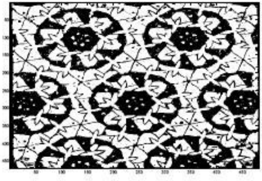


Nothing is as it seems

Images by M.C. Escher

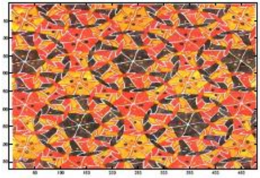


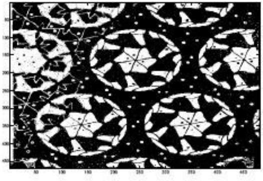


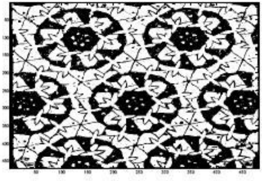


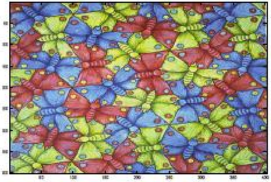
Nothing is as it seems

Images by M.C. Escher



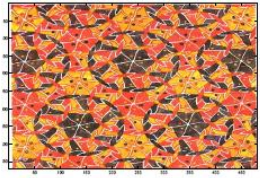


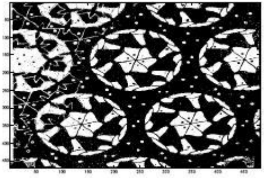


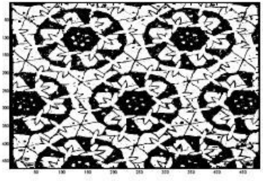


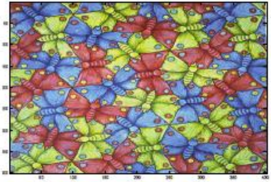
Nothing is as it seems

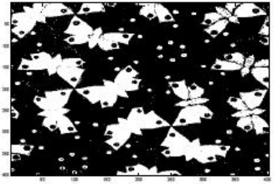
Images by M.C. Escher





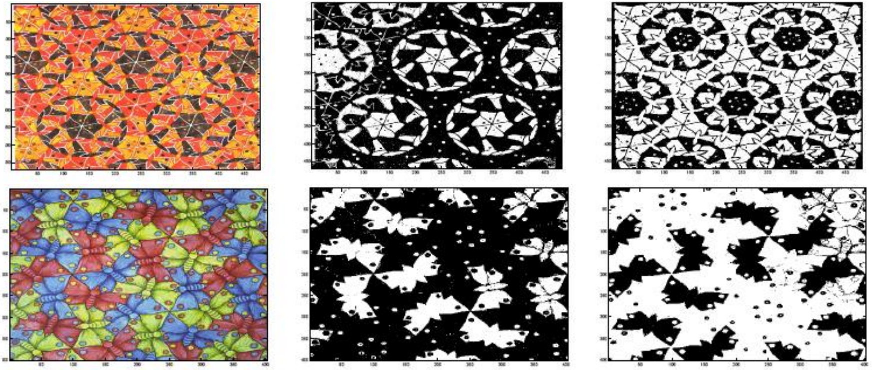






Nothing is as it seems

Images by M.C. Escher



Framework

General framework:
Algorithms typically find a single interpretation of the data.

- Alternative interpretations could exist.

Clustering framework:
Clustering is unsupervised classification and returns a set of clusters.
What if prior knowledge is available?

- Alternative clustering(s) might be desirable.
- Semi-supervised methods.

Framework

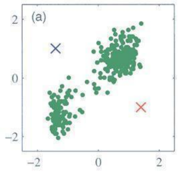
General framework:
Algorithms typically find a single interpretation of the data.

- Alternative interpretations could exist.

Clustering framework:
Clustering is unsupervised classification and returns a set of clusters.
What if prior knowledge is available?

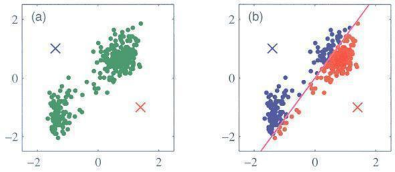
- Alternative clustering(s) might be desirable.
- Semi-supervised methods.

Reminder – k-Means Clustering



(a) Initialise means randomly
(b) Assign points to clusters
(c) Re-estimate Means
(d) Re-assign points to clusters
(e) Re-estimate Means
(f) Re-assign points to clusters
(g) Re-estimate Means
(h) Re-assign points to clusters
(i) Convergence

Reminder – k-Means Clustering



(a) Initialise means randomly

(b) Assign points to clusters

(c) Re-estimate Means

(d) Re-assign points to clusters

(e) Re-estimate Means

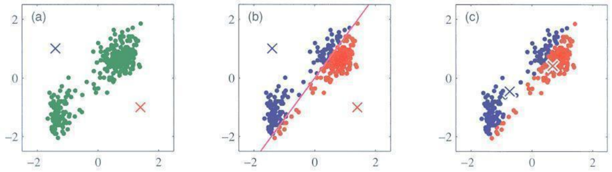
(f) Re-assign points to clusters

(g) Re-estimate Means

(h) Re-assign points to clusters

(i) Convergence

Reminder – k-Means Clustering



(a) Initialise means randomly

(b) Assign points to clusters

(c) Re-estimate Means

(d) Re-assign points to clusters

(e) Re-estimate Means

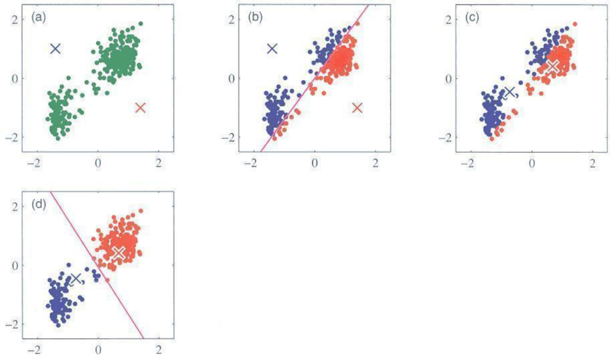
(f) Re-assign points to clusters

(g) Re-estimate Means

(h) Re-assign points to clusters

(i) Convergence

Reminder – k-Means Clustering



(a) Initialise means randomly

(b) Assign points to clusters

(c) Re-estimate Means

(d) Re-assign points to clusters

(e) Re-estimate Means

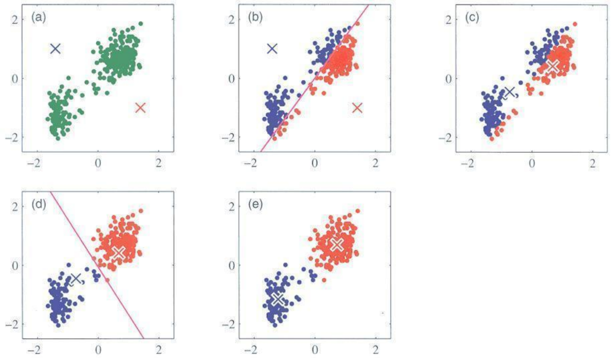
(f) Re-assign points to clusters

(g) Re-estimate Means

(h) Re-assign points to clusters

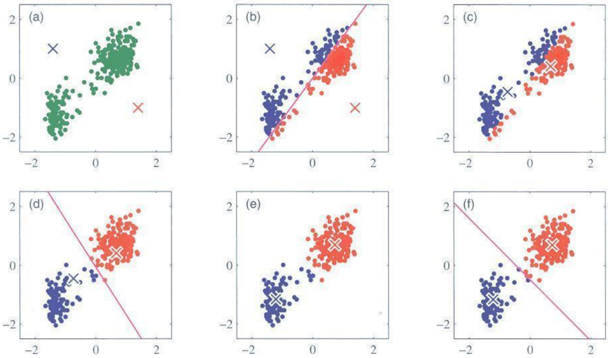
(i) Convergence

Reminder – k-Means Clustering



(a) Initialise means randomly
(b) Assign points to clusters
(c) Re-estimate Means
(d) Re-assign points to clusters
(e) Re-estimate Means
(f) Re-assign points to clusters
(g) Re-estimate Means
(h) Re-assign points to clusters
(i) Convergence

Reminder – k-Means Clustering



(a) Initialise means randomly

(b) Assign points to clusters

(c) Re-estimate Means

(d) Re-assign points to clusters

(e) Re-estimate Means

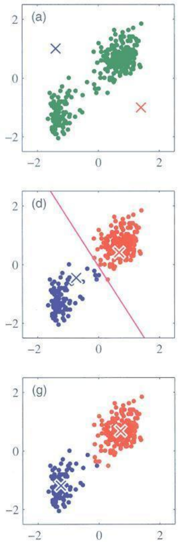
(f) Re-assign points to clusters

(g) Re-estimate Means

(h) Re-assign points to clusters

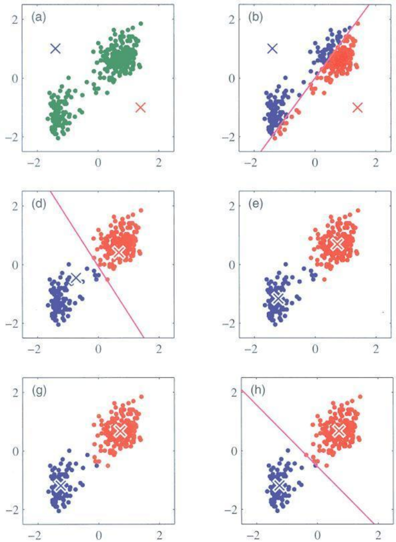
(i) Convergence

Reminder – k-Means Clustering



(a) Initialise means randomly
(b) Assign points to clusters
(c) Re-estimate Means
(d) Re-assign points to clusters
(e) Re-estimate Means
(f) Re-assign points to clusters
(g) Re-estimate Means
(h) Re-assign points to clusters
(i) Convergence

Reminder – k-Means Clustering



(a) Initialise means randomly

(b) Assign points to clusters

(c) Re-estimate Means

(d) Re-assign points to clusters

(e) Re-estimate Means

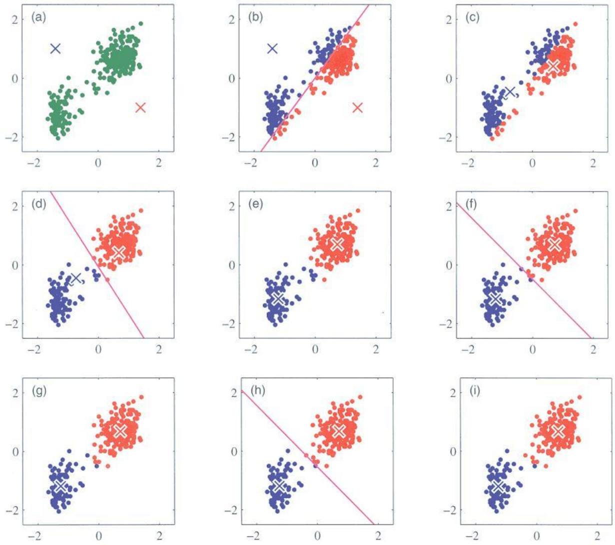
(f) Re-assign points to clusters

(g) Re-estimate Means

(h) Re-assign points to clusters

(i) Convergence

Reminder – k-Means Clustering



(a) Initialise means randomly

(b) Assign points to clusters

(c) Re-estimate Means

(d) Re-assign points to clusters

(e) Re-estimate Means

(f) Re-assign points to clusters

(g) Re-estimate Means

(h) Re-assign points to clusters

(i) Convergence

Example – Automatic Lane Finding from GPS Traces

Where is the lane? [Wagstaff2001]

- Lane-level navigation (e.g. advance notification for taking exits).
- Lane-keeping suggestions (e.g. lane departure warning).

Constraints: width of a lane (maximum separation), points from the same vehicle end on the same lane if there are no lane changes (trace contiguity)

Example – Automatic Lane Finding from GPS Traces

Where is the lane? [Wagstaff2001]

- Lane-level navigation (e.g. advance notification for taking exits).
- Lane-keeping suggestions (e.g. lane departure warning).

Constraints: width of a lane (maximum separation), points from the same vehicle end on the same lane if there are no lane changes (trace contiguity)

Example – Automatic Lane Finding from GPS Traces

Where is the lane? [Wagstaff2001]

- Lane-level navigation (e.g. advance notification for taking exits).
- Lane-keeping suggestions (e.g. lane departure warning).

Constraints: width of a lane (maximum separation), points from the same vehicle end on the same lane if there are no lane changes (trace contiguity)

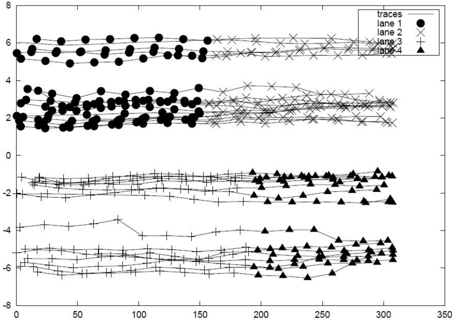


Table of Contents

Problem Description

Singular Alternative Clustering Problem

Given an objective function f, an existing clustering π so that $f(\pi) = x$, does there exist another clustering π' that is different from π and where $f(\pi') \approx f(\pi)$?

Key factors:

- Alternativeness
- Quality

Issues:

- Trade-off between alternativeness and quality of a new clustering.
- Retain certain clusters or chunklets?
- Alternative clustering just for a subspace of the data?

Problem Description

Singular Alternative Clustering Problem

Given an objective function f, an existing clustering π so that $f(\pi) = x$, does there exist another clustering π' that is different from π and where $f(\pi') \approx f(\pi)$?

Key factors:

- Alternativeness
- Quality

Issues:

- Trade-off between alternativeness and quality of a new clustering.
- Retain certain clusters or chunklets?
- Alternative clustering just for a subspace of the data?

Problem Description

Singular Alternative Clustering Problem

Given an objective function f, an existing clustering π so that $f(\pi) = x$, does there exist another clustering π' that is different from π and where $f(\pi') \approx f(\pi)$?

Key factors:

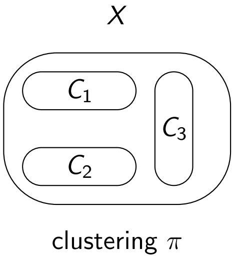
- Alternativeness
- Quality

Issues:

- Trade-off between alternativeness and quality of a new clustering.
- Retain certain clusters or chunklets?
- Alternative clustering just for a subspace of the data?

Algorithm-Independent Approach

Given: data $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and clustering $\pi = \{C_1, \ldots, C_k\}$ (with centroids m_j) found in X

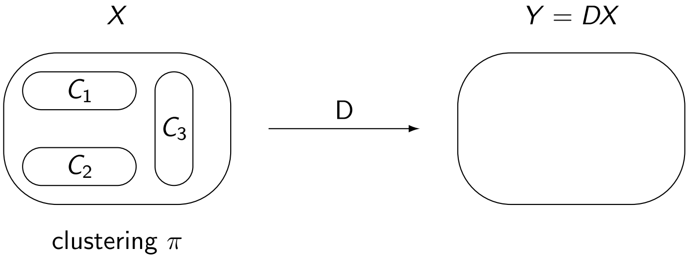


Idea:

- transform X into new space Y with transformation matrix $D \in \mathbb{R}^{d \times d}$
- find a new clustering $\pi' = \{C'_1, \ldots, C'_k\}$ in Y

Algorithm-Independent Approach

Given: data $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and clustering $\pi = \{C_1, \ldots, C_k\}$ (with centroids m_j) found in X

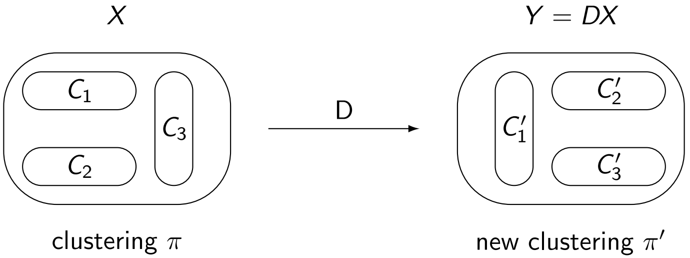


Idea:

- transform X into new space Y with transformation matrix $D \in \mathbb{R}^{d \times d}$
- find a new clustering $\pi' = \{C'_1, \ldots, C'_k\}$ in Y

Algorithm-Independent Approach

Given: data $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and clustering $\pi = \{C_1, \ldots, C_k\}$ (with centroids m_j) found in X



Idea:

- transform X into new space Y with transformation matrix $D \in \mathbb{R}^{d \times d}$
- find a new clustering $\pi' = \{C'_1, \ldots, C'_k\}$ in Y

Solution to the Problem

Key factors:

- *Quality:* retain data properties ⇒ minimise Kullback-Leibler divergence between probability distributions of X and Y: $p_X(x), p_Y(y)$
- *Alternativeness:* properties from π to keep or not keep ⇒ constraints

Constraint Optimisation Problem

$$\min_{B \succeq 0} D_{KL}(p_Y(y) || p_X(x))$$

$$s.t. \ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, x_i \notin C_j}^{k} ||x_i - m_j||_B^2 \leq \beta$$

where $B = D^T D$ and $||x - y||_B = \sqrt{(x-y)^T B (x-y)}$ (Mahalanobis distance).

Solution: $D = \tilde{\Sigma}^{-\frac{1}{2}}$ where $\tilde{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, x_i \notin C_j}^{k} (x_i - m_j)(x_i - m_j)^T$

Solution to the Problem

Key factors:

- *Quality:* retain data properties $\Rightarrow$ minimise Kullback-Leibler divergence between probability distributions of X and Y: $p_X(x), p_Y(y)$
- *Alternativeness:* properties from π to keep or not keep $\Rightarrow$ constraints

Constraint Optimisation Problem

$$\min_{B \succeq 0} D_{KL}(p_Y(y) || p_X(x))$$

$$s.t. \ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, x_i \notin C_j}^{k} ||x_i - m_j||_B^2 \leq \beta$$

where $B = D^T D$ and $||x - y||_B = \sqrt{(x-y)^T B (x-y)}$ (Mahalanobis distance).

Solution: $D = \tilde{\Sigma}^{-\frac{1}{2}}$ where $\tilde{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, x_i \notin C_j}^{k} (x_i - m_j)(x_i - m_j)^T$

Solution to the Problem

Key factors:

- *Quality:* retain data properties $\Rightarrow$ minimise Kullback-Leibler divergence between probability distributions of X and Y: $p_X(x), p_Y(y)$
- *Alternativeness:* properties from π to keep or not keep $\Rightarrow$ constraints

Constraint Optimisation Problem

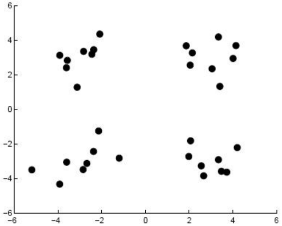
$$\min_{B \succeq 0} D_{KL}(p_Y(y)||p_X(x))$$

$$s.t. \ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, x_i \notin C_j}^{k} ||x_i - m_j||_B^2 \leq \beta$$

where $B = D^T D$ and $||x - y||_B = \sqrt{(x-y)^T B (x-y)}$ (Mahalanobis distance).

Solution: $D = \tilde{\Sigma}^{-\frac{1}{2}}$ where $\tilde{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, x_i \notin C_j}^{k} (x_i - m_j)(x_i - m_j)^T$

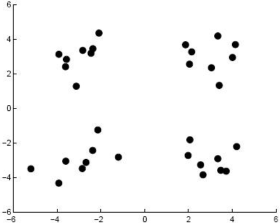
Example

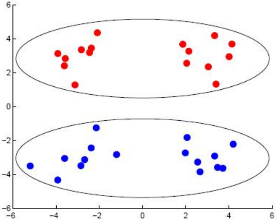


$$\tilde{\Sigma} = \begin{pmatrix} 9.7419 & 0.1801 \\ 0.1801 & 36.6461 \end{pmatrix}$$

$$\Rightarrow D = \begin{pmatrix} 0.3204 & -0.0010 \\ -0.0010 & 0.1652 \end{pmatrix}$$

Example

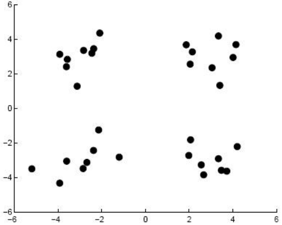


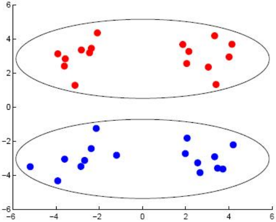


$$\tilde{\Sigma} = \begin{pmatrix} 9.7419 & 0.1801 \\ 0.1801 & 36.6461 \end{pmatrix}$$

$$\Rightarrow D = \begin{pmatrix} 0.3204 & -0.0010 \\ -0.0010 & 0.1652 \end{pmatrix}$$

Example

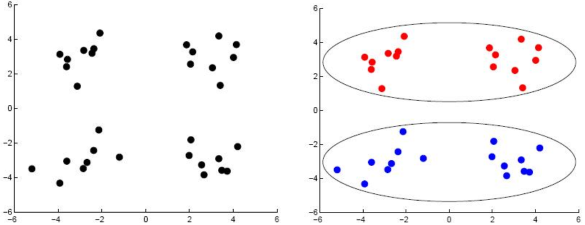




$$\tilde{\Sigma} = \begin{pmatrix} 9.7419 & 0.1801 \\ 0.1801 & 36.6461 \end{pmatrix}$$

$$\Rightarrow D = \begin{pmatrix} 0.3204 & -0.0010 \\ -0.0010 & 0.1652 \end{pmatrix}$$

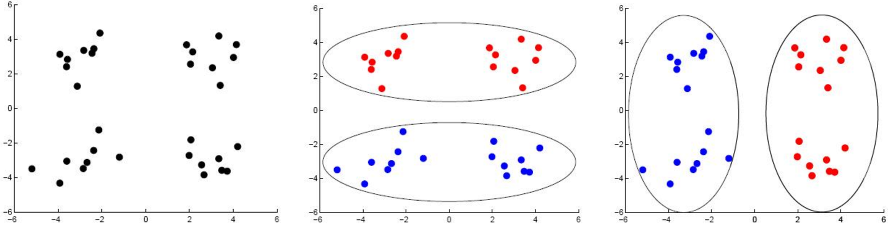
Example



$$\tilde{\Sigma} = \begin{pmatrix} 9.7419 & 0.1801 \\ 0.1801 & 36.6461 \end{pmatrix}$$

$$\Rightarrow D = \begin{pmatrix} 0.3204 & -0.0010 \\ -0.0010 & 0.1652 \end{pmatrix}$$

Example



$$\tilde{\Sigma} = \begin{pmatrix} 9.7419 & 0.1801 \\ 0.1801 & 36.6461 \end{pmatrix}$$

$$\Rightarrow D = \begin{pmatrix} 0.3204 & -0.0010 \\ -0.0010 & 0.1652 \end{pmatrix}$$

Table of Contents

Learning Techniques

unsupervised: e.g. clustering, association analysis

semi-supervised: clustering with constraints

supervised: e.g. decision trees, neural networks, logistic regression, support vector machines

Learning Techniques

unsupervised: e.g. clustering, association analysis

semi-supervised: clustering with constraints

supervised: e.g. decision trees, neural networks, logistic regression, support vector machines

Assets and Drawbacks

Advantages:

- Algorithm-independent and easy to implement (closed-form solution).
- Trade-off between alternativeness and quality can be controlled.
- Easy to specify what properties of a given clusterings to keep or not keep.
- Distance matrix can be used in any distance-based method (cp. ordination methods with distance metrics).
- Approach can be used along with ordination methods in order to analyse classification methods (e.g. reveal additional classes or misclassified points).

Assets and Drawbacks

Advantages:

- Algorithm-independent and easy to implement (closed-form solution).
- Trade-off between alternativeness and quality can be controlled.
- Easy to specify what properties of a given clusterings to keep or not keep.
- Distance matrix can be used in any distance-based method (cp. ordination methods with distance metrics).
- Approach can be used along with ordination methods in order to analyse classification methods (e.g. reveal additional classes or misclassified points).

Assets and Drawbacks

Advantages:

- Algorithm-independent and easy to implement (closed-form solution).
- Trade-off between alternativeness and quality can be controlled.
- Easy to specify what properties of a given clusterings to keep or not keep.
- Distance matrix can be used in any distance-based method (cp. ordination methods with distance metrics).
- Approach can be used along with ordination methods in order to analyse classification methods (e.g. reveal additional classes or misclassified points).

Assets and Drawbacks

Advantages:

- Algorithm-independent and easy to implement (closed-form solution).
- Trade-off between alternativeness and quality can be controlled.
- Easy to specify what properties of a given clusterings to keep or not keep.
- Distance matrix can be used in any distance-based method (cp. ordination methods with distance metrics).
- Approach can be used along with ordination methods in order to analyse classification methods (e.g. reveal additional classes or misclassified points).

Assets and Drawbacks

Advantages:

- Algorithm-independent and easy to implement (closed-form solution).
- Trade-off between alternativeness and quality can be controlled.
- Easy to specify what properties of a given clusterings to keep or not keep.
- Distance matrix can be used in any distance-based method (cp. ordination methods with distance metrics).
- Approach can be used along with ordination methods in order to analyse classification methods (e.g. reveal additional classes or misclassified points).

Assets and Drawbacks

Disadvantages:

- Algorithm-independent approach, i.e. the approach inherits the drawbacks of the algorithm used,
  - e.g. k-means: efficient, but it is sensitive to outliers, it often terminates at a local optimum and an inappropriate choice of k may yield poor results.
- Assumptions: clusters in π' are multivariate Gaussian, same cluster sizes, constant variances, dimensions highly independent,...
- Sometimes a non-linear transformation might be more appropriate ($\rightsquigarrow$ future work).
- Approach is very general; special algorithms such as COP k-means might be more efficient [Wagstaff2001].
- Expert guidance becomes impractical in very large datasets.

Assets and Drawbacks

Disadvantages:

- Algorithm-independent approach, i.e. the approach inherits the drawbacks of the algorithm used,
  - e.g. k-means: efficient, but it is sensitive to outliers, it often terminates at a local optimum and an inappropriate choice of k may yield poor results.
- Assumptions: clusters in π' are multivariate Gaussian, same cluster sizes, constant variances, dimensions highly independent,. . .
- Sometimes a non-linear transformation might be more appropriate ($\rightsquigarrow$ future work).
- Approach is very general; special algorithms such as COP k-means might be more efficient [Wagstaff2001].
- Expert guidance becomes impractical in very large datasets.

Assets and Drawbacks

Disadvantages:

- Algorithm-independent approach, i.e. the approach inherits the drawbacks of the algorithm used,
  - e.g. k-means: efficient, but it is sensitive to outliers, it often terminates at a local optimum and an inappropriate choice of k may yield poor results.
- Assumptions: clusters in π' are multivariate Gaussian, same cluster sizes, constant variances, dimensions highly independent,. . .
- Sometimes a non-linear transformation might be more appropriate ($\rightsquigarrow$ future work).
- Approach is very general; special algorithms such as COP k-means might be more efficient [Wagstaff2001].
- Expert guidance becomes impractical in very large datasets.

Assets and Drawbacks

Disadvantages:

- Algorithm-independent approach, i.e. the approach inherits the drawbacks of the algorithm used,
  - e.g. k-means: efficient, but it is sensitive to outliers, it often terminates at a local optimum and an inappropriate choice of k may yield poor results.
- Assumptions: clusters in π' are multivariate Gaussian, same cluster sizes, constant variances, dimensions highly independent,...
- Sometimes a non-linear transformation might be more appropriate ($\rightsquigarrow$ future work).
- Approach is very general; special algorithms such as COP k-means might be more efficient [Wagstaff2001].
- Expert guidance becomes impractical in very large datasets.

Assets and Drawbacks

Disadvantages:

- Algorithm-independent approach, i.e. the approach inherits the drawbacks of the algorithm used,
  - e.g. k-means: efficient, but it is sensitive to outliers, it often terminates at a local optimum and an inappropriate choice of k may yield poor results.
- Assumptions: clusters in π' are multivariate Gaussian, same cluster sizes, constant variances, dimensions highly independent,...
- Sometimes a non-linear transformation might be more appropriate ($\rightsquigarrow$ future work).
- Approach is very general; special algorithms such as COP k-means might be more efficient [Wagstaff2001].
- Expert guidance becomes impractical in very large datasets.

Table of Contents

Bibliography

Z. Qi, I. Davidson.
A principled and flexible framework for finding alternative clusterings.
KDD 2009, 717-726, 2009.

K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl.
Constrained K-means Clustering with Background Knowledge..
ICML 2001, 577–584, 2001.

Table of Contents

Variations of the Constrained Optimisation Problem

Specifiying the trade-off between alternativeness and quality:
New constraint:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1, x_i \notin C_j}^{k} ||x_i - m_j||_B^{\alpha} \leq \beta \text{ where } \alpha \geq 1$$

$\alpha \uparrow \Rightarrow$ alternativeness $\uparrow$

Specifiying which clusters to keep and not keep:

- Retain cluster C_p: $\sum_{x_i \in C_p} ||x_i - m_p||_B^2 \leq \delta$ with δ small
- Retain clusters $C_Y = \{C_1, \ldots, C_r\}$ $(1 < r < k)$:
  New constraint:

$$\sum_{x_i \in C_Y}\sum_{p=1, x_i \in C_p}^{r} ||x_i - m_p||_B^2 + \sum_{x_i \notin C_Y}\sum_{j=1, x_i \notin C_j}^{k} ||x_i - m_j||_B^2 \leq \beta$$