Semi-supervised Clustering using Heterogeneous Dissimilarities

Manuel Martín-Merino
University Pontificia of Salamanca
mmartinmac@upsa.es
Spain

Contents

- Motivation
- Approaches proposed in the literature
- Our approach. Advances
- Challenges still remaining
- Open questions and challenges

Motivation (I)

- Clustering algorithms depend strongly on the distance considered to evaluate the sample proximities
- Choosing a dissimilarity that reflects accurately "which is similar" for the human experts is a critical step
- In certain applications such as Bioinformatics several dissimilarities and data sources are available that provide complementary information about the problem

Motivation (II)

Goal: Learning a linear combination of dissimilarities that reflects better "which is similar" for human experts

- A family of dissimilarities and data sources is available
- Experts provide only incomplete knowledge in the form of which pairs of objects are similar
- The classes and even the number of groups are not known a priory
- Overfitting is likely to occur

Motivation (III)

We are addressing a clustering problem:

- The object proximities should be modeled by a set of heterogeneous dissimilarities and data sources
- Sparse supervision in the form of a set of equivalence constraints will help to determine the relevance of each measure
- Interpretability about the contribution each dissimilarity or data source is quite interesting for biologists

Approaches in the literature (I)

- Classification techniques that learn a linear combination of dissimilarities or kernels have been widely studied
- Several algorithms have been proposed to learn a Mahalanobis distance from a set of equivalence constraints
  - (Xing et. al) is costly computationally
  - (Kwok et. al; Bar-Hillel et. al) are more efficient for high dimensional problems

Approaches in the literature (II)

Drawbacks

- They rely on a single dissimilarity and data source
- Equivalence constraints are incorporated as in the supervised case, without taking into account the topology of the data
- Most of them are prone to overfitting

Our Approach (I)

Goal: Learning a convex combination of dissimilarities or kernels from a small set of equivalence constraints,

$$k_{ij} = \sum_{l=1}^{M} \beta_l k_{ij}^l$$

- The idealized dissimilarity is defined in a semi-supervised way:

$$k_{ij}^* = \begin{cases} \text{máx}_l\{k_{ij}^l\} & \text{If } (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S} \\ \text{mín}_l\{k_{ij}^l\} & \text{If } (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D} \end{cases}$$

Our Approach (II)

A quadratic optimization algorithm is obtained similar to the one solved by the SVM.

$$\min_{\boldsymbol{\beta},\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{C_S}{N_S}\sum_{(x_i,x_j)\in S}\xi_{ij} + \frac{C_D}{N_D}\sum_{(x_i,x_j)\in D}\xi_{ij}$$

$$\text{s. t.} \qquad \boldsymbol{\beta}^T\boldsymbol{K}_{ij} \geq K^*_{ij} - \xi_{ij} \quad \forall\,(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{S}$$

$$\boldsymbol{\beta}^T\boldsymbol{K}_{ij} \leq K^*_{ij} + \xi_{ij} \quad \forall\,(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{D}$$

$$\beta_l \geq 0 \quad \xi_{ij} \geq 0 \quad \forall\, l = 1,\ldots,M$$

Our Approach. Advances (III)

- The regularization term favors dense combinations of several dissimilarities, which is more meaningful for biologists
- The idealized kernel defined helps to reduce the overfitting
- As $\beta_l \geq 0$ the experimental results are more easily interpretable
- The optimization problem can be solved in the dual and the computational burden doesn't depend on the space dimensionality.

Our Approach. Advantages (IV)

Experimental Results

- The combination of heterogeneous dissimilarities improves significantly a standard metric learning algorithm for all the datasets considered.
- The method proposed performs well for high dimensional data
- The combination of dissimilarities improves kernel k-means based on the best dissimilarity
- The best dissimilarity depends on the dataset considered

Challenges Still Remaining

- Development of new algorithms that allow us to combine dissimilarities locally
- Improve the computational efficiency of the learning algorithms
- Estimation of the regularization parameters when a priory knowledge is sparse
- Non-linear combination of dissimilarities should be studied
- Integration of different kind of constraints

Challenges Still Remaining

- Development of new algorithms that allow us to combine dissimilarities locally
- Improve the computational efficiency of the learning algorithms
- Estimation of the regularization parameters when a priory knowledge is sparse
- Non-linear combination of dissimilarities should be studied
- Integration of different kind of constraints

Challenges Still Remaining

- Development of new algorithms that allow us to combine dissimilarities locally
- Improve the computational efficiency of the learning algorithms
- Estimation of the regularization parameters when a priory knowledge is sparse
- Non-linear combination of dissimilarities should be studied
- Integration of different kind of constraints

Challenges Still Remaining

- Development of new algorithms that allow us to combine dissimilarities locally
- Improve the computational efficiency of the learning algorithms
- Estimation of the regularization parameters when a priory knowledge is sparse
- Non-linear combination of dissimilarities should be studied
- Integration of different kind of constraints

Challenges Still Remaining

- Development of new algorithms that allow us to combine dissimilarities locally
- Improve the computational efficiency of the learning algorithms
- Estimation of the regularization parameters when a priory knowledge is sparse
- Non-linear combination of dissimilarities should be studied
- Integration of different kind of constraints

Open Questions and Challenges (I)

- Considering a broad family of meaningful dissimilarities provides more benefits that developing new complex and non-linear clustering algorithms ?
- Modern applications such as genomics and proteomics require new hybrid techniques that exploit the advantages of clustering and classification techniques
- Learning algorithms should be able to work with imprecise knowledge formulated with a certain probability of error

Open Questions and Challenges (I)

- Considering a broad family of meaningful dissimilarities provides more benefits that developing new complex and non-linear clustering algorithms ?
- Modern applications such as genomics and proteomics require new hybrid techniques that exploit the advantages of clustering and classification techniques
- Learning algorithms should be able to work with imprecise knowledge formulated with a certain probability of error

Open Questions and Challenges (I)

- Considering a broad family of meaningful dissimilarities provides more benefits that developing new complex and non-linear clustering algorithms ?
- Modern applications such as genomics and proteomics require new hybrid techniques that exploit the advantages of clustering and classification techniques
- Learning algorithms should be able to work with imprecise knowledge formulated with a certain probability of error

Open Questions and Challenges (I)

- Considering a broad family of meaningful dissimilarities provides more benefits that developing new complex and non-linear clustering algorithms ?
- Modern applications such as genomics and proteomics require new hybrid techniques that exploit the advantages of clustering and classification techniques
- Learning algorithms should be able to work with imprecise knowledge formulated with a certain probability of error

Open Questions and Challenges (II)

- New techniques should be developed that allow us to extract multivariate relationships from clustering algorithms complementing the available multivariate statistical techniques.
  For human experts, knowledge discovery is as relevant as the identification of clusters.

Open Questions and Challenges (II)

- New techniques should be developed that allow us to extract multivariate relationships from clustering algorithms complementing the available multivariate statistical techniques.
  For human experts, knowledge discovery is as relevant as the identification of clusters.

Experimental results (I)

Accuracy for k-means clustering considering several dissimilarities.

Technique	Kernel	Wine	Ionosphere	Breast	Colon
k-means (Euclidean)	linear	0.92	0.72	0.88	0.87
	pol. 3	0.87	0.73	0.88	0.88
k-means (Best diss.)	linear	0.94	0.88	0.90	0.88
	pol. 3	0.94	0.88	0.90	0.88
		χ^2	Maha.	Manha.	Co./euc.
Comb. dissimilarities	linear	0.94	**0.90**	0.92	0.89
	pol. 3	**0.96**	0.89	**0.92**	**0.90**
Metric learning (Xing)	linear	0.87	0.74	0.85	0.87
	pol. 3	0.51	0.74	0.86	0.88

Experimental results (II)

Adjusted RandIndex for k-means clustering considering several dissimilarities.

Technique	Kernel	Wine	Ionosphere	Breast	Colon
k-means (Euclidean)	linear	0.79	0.20	0.59	0.59
	pol. 3	0.67	0.21	0.60	0.59
k-means (Best diss.)	linear	0.82	0.58	0.66	0.59
	pol. 3	0.81	0.58	0.66	0.59
		χ^2	Maha.	Manha.	Co./euc.
Comb. dissimilarities	linear	0.82	**0.63**	0.69	0.60
	pol. 3	**0.85**	0.60	**0.69**	**0.63**
Metric learning (Xing)	linear	0.68	0.23	0.50	0.54
	pol. 3	0.50	0.23	0.52	0.58