Fuzzy Semi-supervised Clustering with Active Constraint Selection

Natalia Novoselova(✉) and Igor Tom

United Institute of Informatics Problems, NAS Belarus, Surganova Street 6,
220012 Minsk, Belarus
{novosel,tom}@newman.bas-net.by

Abstract. The paper presents the approach to semi-supervised fuzzy clustering, based on the extended optimization function and the algorithm of the active constraints selection. The approach is tested on the artificial and real data sets. Clustering results, obtained by the proposed approach, are more accurate relative to the ground truth due to utilization of the additional information about the class labels in the most uncertain regions.

Keywords: Semi-supervised clustering · Pairwise constraints · Active selection

1 Introduction

In the field of machine learning and bioinformatics the semi-supervised methods present the realization of the technology which uses the data with both known and unknown class labels in order to solve some particular task, such as data clustering or classification. As a rule the number of unlabeled data considerably exceeds the number of the data with known labels. It can be explained by high financial and temporal expenses, connected with the manual data classification in such fields of research as natural language processing, text mining, computational biology etc. In order to make use of a tremendous amount of rapidly coming information, which due to the progress in information technologies can be unlimitedly stored in data bases the special methods of semi-supervised learning are currently of great concern. It is recognized that the unlabeled data together with the sufficiently small amount of constrained ones enable the significant improvement in learning accuracy [1].

The semi-supervised clustering is one of the evolving research directions, used for the data exploratory analysis. The main task of the clustering methods is to reveal the groups of similar data points, according to the specified notion of similarity. There are already several approaches to consider the known constraints between the points in order to guide the clustering process [2, 3]. The semi-supervised learning allows improving the efficiency of the clustering using the available expert knowledge in the form of data labels or the relations between the data points.

For the semi-supervised algorithm, processing the great amount of data, as e.g. biological data it is very important to:

(1) automatically determine the number of clusters in the data;
(2) take into consideration the available data constraints;

V.V. Krasnoproshin and S.V. Ablameyko (Eds.): PRIP 2016, CCIS 673, pp. 132–139, 2017.
DOI: 10.1007/978-3-319-54220-1_14

(3) automatically select the constraints in more uncertain, transition regions in order to get the high accuracy of results. It must be emphasized that the defined constraints greatly influence the clustering result. The improper constraint selection can even decrease the clustering performance [4]. Recently the most important topic of research is the development of active selection strategies, which search for the most useful constraints [5, 6]. They can minimize the expenses of getting the labeled information without loss of clustering accuracy.

In the paper we propose the approach to semi-supervised fuzzy clustering, based on the active constraint selection algorithm [7]. The semi-supervised fuzzy clustering algorithm [8] takes into account the pairwise constraints and belongs to the class of optimization clustering methods. The basis of such methods is the construction of some optimization function the minimization of which enables to define the optimal cluster parameter values. The experiments on several datasets have shown the improvement of clustering performance with the inclusion of constraints, especially when they were actively selected.

2 Semi-supervised Fuzzy Clustering

In our research we have adopted the fuzzy clustering algorithm, proposed in [8], which is based on the algorithm of competitive agglomeration. The algorithm can automatically determine the number of clusters in the analyzed data and takes into account the data constraints using the extended clustering optimization function. There are two types of constraints: “must link” constraint and “cannot link” constraint for data points. Let M is the set of “must-link” constraints, i.e. $\left(x_i, x_j\right) \in M$ means the data points x_i and x_j lie in the same cluster. The set Q consists of “cannot-link” constraints, i.e. $\left(\mathrm{x_i}, \mathrm{x_j}\right) \in \mathrm{Q}$ means the data points x_i and x_j lie in the different clusters. The extended optimization function is the following

$$J(V, U) = \sum_{k=1}^{C} \sum_{i=1}^{N} \left(u_{ik}\right)^2 d^2\left(x_i, \mu_k\right) - \alpha\Big(\sum_{\left(x_i, x_j\right) \in M} \sum_{k=1}^{C} \sum_{l=1, l \neq k}^{C} u_{ik} u_{jl} + \sum_{\left(x_i, x_j\right) \in Q} \sum_{k=1}^{C} u_{ik} u_{jk}\Big) - \beta \sum_{k=1}^{C} \sum_{i=1}^{N} \left(u_{ik}\right)^2 \tag{1}$$

where $X = \left\{x_i | i \in \{1, \ldots, N\}\right\}$ is the dataset of size N, $V = \left\{\mu_k | k \in \{1, \ldots, C\}\right\}$ is the centers of C clusters, $U = \left\{u_{ik} | k \in \{1, \ldots, C\}, i \in \{1, \ldots, N\}\right\}$ is the set of membership degrees. The constraint $\sum_{k=1}^{C} u_{ik} = 1, i = \{1, \cdots, N\}$ must be considered.

The cluster centers $(1 \leq k \leq C)$ are calculated in iterative fashion as

$$\mu_k = \frac{\sum_{i=1}^{N} \left(u_{ik}\right)^2 x_i}{\sum_{i=1}^{N} \left(u_{ik}\right)^2} \tag{2}$$

and the cardinalities of the clusters are defined as

$$N_s = \sum_{i=1}^{N} u_{is}. \tag{3}$$

The first component of optimization function (1) presents the FCM optimization term and considers the cluster compactness. The second component consists of two terms: (1) penalty for the violation of the pairwise "must-link" constraints; (2) penalty for the violation of the pairwise "cannot-link" constraints. The weight constant α determines the relative importance of supervision. The third component in (1) is the sum of squares of cardinalities of the individual clusters and corresponds to the regularization term, which controls the number of clusters. The weight function β of the third component provides the balance between the components and is expressed as

$$\beta(t) = \frac{\eta_0 exp\left(-\left|t - t_0\right|/\tau\right)}{\sum_{j=1}^{C} \left(\sum_{i=1}^{N} u_{ij}\right)^2} \times \left[\sum_{j=1}^{C} \sum_{i=1}^{N} u_{ij}^2 d^2\left(x_j, \mu_j\right)\right] \tag{4}$$

Function $\beta(t)$ allows regulating the data memberships u_{ij} to clusters and has the small value at the beginning of the optimization process in order to form the initial clusters. After that the weight rises in order to reduce the number of clusters and again falls to diminish its influence on the cluster formation.

We have reconsidered the derivation of the expressions for the modification of the parameters $u_{rs} = u_{rs}^{FCM} + u_{rs}^{constr} + u_{rs}^{bias}, r = \{1, \cdots, N\}, s = \{1, \cdots, C\}$ in (1). The expressions are calculated using the Lagrange multipliers and are the following:

$$\begin{aligned} u_{rs}^{FCM} &= \frac{1}{d^2\left(x_r, \mu_s\right)} \Bigg/ \sum_{k=1}^{C} \frac{1}{d^2\left(x_r, \mu_k\right)} \\ u_{rs}^{constr} &= \frac{\alpha}{2d^2\left(x_r, \mu_s\right)} \left(\overline{C_{v_r}} - C_{v_{rs}}\right) \\ u_{rs}^{bias} &= \frac{\beta}{d^2\left(x_r, \mu_s\right)} \left(N_s - \overline{N_r}\right) \end{aligned} \tag{5}$$

where $C_{v_{rs}}$ – penalty expression for violation the constraint for the *rth* point in the case of *sth* cluster, $\overline{C_{v_r}}$ – weighted average penalty for all clusters for the *rth* point, $\overline{N_r}$ – weighted average of cluster cardinalities relative to *rth* point.

The term u_{rs}^{FCM} is the same as in FCM; the term $\mathrm{u}_{\mathrm{rs}}^{\mathrm{constr}}$ allows decreasing or increasing the membership according to the pairwise constraints, defined by user; the term u_{rs}^{bias}

allows to reduce the cardinalities of the non-informative clusters and to discard them from consideration when the cluster cardinalities are below the threshold.

The important step of the semi-supervised clustering process is the cluster merging, which is executed at each iteration of the optimization procedure. It allows excluding from consideration not only the small clusters, but also the non-informative clusters of different sizes.

Below is the scheme of the semi-supervised algorithm [8].

Algorithm

```
* Define the maximal cluster number $C$ .
* Initialize the cluster centers randomly.
* Initialize the membership values of data objects to clusters: the equal membership to each cluster.
* Calculate the initial cardinalities of each cluster.
Repeat
* Calculate $\beta$ using expression (4).
* Calculate memberships $u_{ij}$ using (5).
* Calculate the cluster cardinalities $N_j, 1 \le j \le C$ using (3).
* For each cluster if $N_j <$ threshold discard cluster $j$ .
* Update number of clusters $C$ .
      Repeat
    * Merge the nearest clusters using the special procedure.
      Until further merging is required
* Update the cluster centers using expression (2).
    Until the clusters stabilize.
```

3 Active Constraint Selection

In [8] the available constraints, which are selected randomly, significantly increase the performance of data clustering. In our paper we propose to use the active constraint selection algorithm [7], which is able to direct the search for the constraints to the most uncertain (transition) clustering regions. The candidate subset of constraints is constructed on the basis of k-Nearest Neighbor Graph (k-NNG). After that the selection is performed from the list of candidate constraints, sorted according to their ability to separate clusters. As a rule such constraints lie in the most uncertain clustering regions.

The k-NNG graph is constructed using the information about k- nearest neighbors of each data point. The weight $w(x_i, x_j)$ of the graph edge between two points x_i and x_j is defined as the number of their common nearest neighbors

$$w(x_i, x_j) = \left| NN(x_i) \cap NN(x_j) \right|, \tag{6}$$

where $NN(x)$ is the k- nearest neighbors of point x.

The constraint ability to separate clusters is estimated using the following utility measure

$$ASC(x_i, x_j) = \frac{k - w\left(x_i, x_j\right) + \frac{1}{1 + \min\left\{LDS\left(x_i\right),\, LDS\left(x_j\right)\right\}}}{k + 1}, \tag{7}$$

where $LDS(x) = \frac{\sum_{q \in NN(x)} w(x, q)}{k}$ is the local density of point x. The ASC measure of pairwise constraint $\left(x_i, x_j\right)$ depends on the corresponding edge weight $w\left(x_i, x_j\right)$ and the constraint density, which is defined by minimum of local densities of points x_i and x_j. The ASC measure helps to reveal the constraints, which are more informative for clustering, i.e. can improve the clustering performance. The higher ASC value corresponds to more informative constraint.

The candidate constraints are selected using the k-NNG graph as follows

$$C = \{(u, v) | w(u, v) < \theta\}, \tag{8}$$

where u, v – graph vertices, θ – the threshold parameters, defined in the interval $\left[\frac{k}{2} - 2, \frac{k}{2} + 2\right]$.

In order to refine the constraint selection process the authors in [7] propagate the already selected constraints to the whole set of candidate constraints. The propagation procedure helps to exclude from further consideration the constraints, which can be derived from the already selected ones using the strong path and transitive closure concepts.

According to the algorithm [7] the constraints are defined iteratively, starting from the one constraint till the required number. Each constraint can be selected from the candidate subset C using two variants: (1) random choice of constraints; (2) taking from the constraint list, sorted according to ASC measure. In our research we have compared the active constraint selection procedure, based on ASC measure with purely random choice of constraints from the data.

4 Results of Experiments

Several comparative experiments using fuzzy semi-supervised clustering algorithm with active constraints (AS) and with purely random constraints (RS) were conducted on artificial and real data sets. The results were compared with the simple k-means (KM) and competitive agglomeration algorithm (CA).

The artificial data set Data1 consists of 150 objects with two features. The objects are divided into three clusters, generated according to the multivariate normal distribution and are partly overlapped. The real data set Leukemia consists of two classes – 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloid leukemia (AML) [9]. In order to validate our approach we have taken into account the two subtypes of ALL: 38 samples of B-cell ALL and 9 samples of T-cell T-ALL, analyzing the classification into 3 classes (ground truth). All the samples are characterized by the expression of 7129 genes. After data preprocessing with thresholding and filtering the 3571 genes are selected for further analysis.

The clustering quality was estimated with the external validation criterion using the ground truth. The criterion estimates the similarity of two data partitions. The first partition corresponds to the known class labels. The second partition is calculated on the basis of fuzzy clustering results, where the label for each data point corresponds to the cluster to which it has the highest membership. As the labels of data points in two partitions can be permuted it is necessary to find the correspondence between them, solving the following optimization task:

Let α_1 and α_2 are two class label functions, defined by two partitions $\Pi_k^{(1)}, \Pi_k^{(2)}$ of the set X into k groups, i.e. $\alpha_i(x) = j$, if and only if $x \in \pi_j^{(i)}, \; i = 1, 2 \; j = 1, \ldots, k$. For given permutation φ of class labels from set V_k consider the empirical validation criterion:

$$d_k\left(\alpha_1, \alpha_2, \varphi\right) = \frac{1}{|X|} \sum\nolimits_{x \in X} \alpha_1(x) \neq \varphi\left(\alpha_2(x)\right), \tag{9}$$

where δ – indicator function $\alpha_1(x) \neq \varphi\left(\alpha_2(x)\right)$:

$$\delta\left(\alpha_1(x) \neq \varphi\left(\alpha_2(x)\right)\right) = \left\{ \begin{array}{c} 1, \; if \; \alpha_1(x) \neq \varphi\left(\alpha_2(x)\right) \\ 0, \quad otherwise \end{array} \right. . \tag{10}$$

The optimal class label permutation φ^* is defined as

$$\varphi^* = \arg\min\nolimits_{\varphi} d_k\left(\alpha_1, \alpha_2, \varphi\right) \tag{11}$$

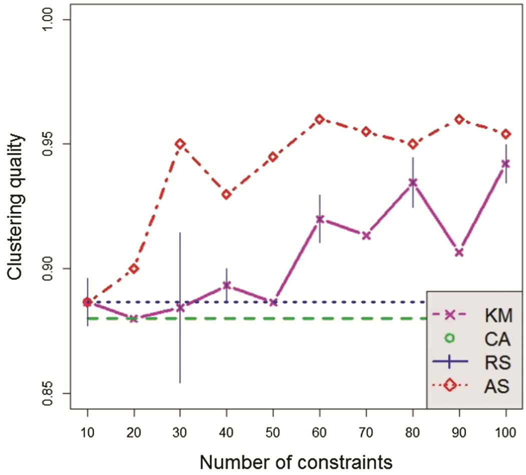

Fig. 1. Clustering results for the artificial dataset using cluster validation criterion

Figures 1 and 2 present the dependence between the validation criterion and the number of pairwise constraints considered for AS and RS algorithms. The number of constraints is in the range from 0 to 100 in increments of 10. For every number of constraints, 100 experiments were performed with different random selections of the constraints in order to estimate the standard errors for the RS approach. KM and CA algorithms don't consider the constraints and are depicted for reference. The algorithms CA, RS and AS were initialized with more than real number of clusters and the search for real clustering structure is performed automatically during algorithm execution. According to Figs. 1 and 2 including the random constraints into clustering allows to improve the clustering quality.

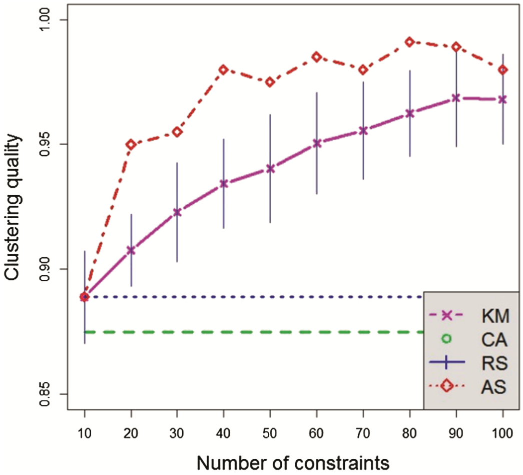

Fig. 2. Clustering results for Leukemia dataset using cluster validation criterion

AS algorithm improves the clustering results even more and requires fewer constraints in order to reach the same clustering quality as the RS algorithm. The cluster centers for the artificial dataset, which are determined by the fuzzy semi-supervised clustering algorithm are shown in Fig. 3.

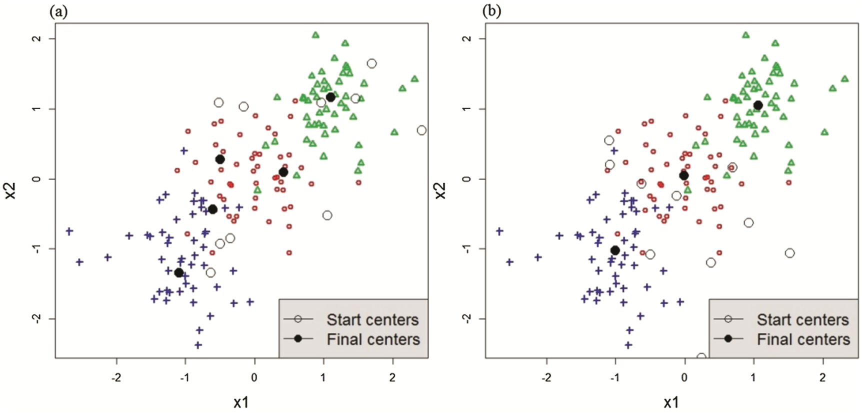

Fig. 3. Initial and final cluster centers for artificial dataset: a) clustering with 10 random constraints; b) clustering with 10 active constraints.

According to Fig. 3 the application of active constraints not only helps to raise the cluster validation measure but also to improve the search for the real number of clusters.

5 Conclusion

The paper presents the approach to semi-supervised fuzzy clustering with active constraints selection. The extended clustering optimization function of the clustering algorithm takes into account the “must link” and “cannot link” constraints on pairwise data positions in the clusters and is based on the scheme, proposed in [8]. We have applied the algorithm of the active constraints selection [7] to generate the constraints for experimental datasets. The clustering results have shown the improved performance with both random and active constraints, included into fuzzy clustering process. The inclusion of active constraints led to better clustering results and to less number of constraints to attain the high level of the clustering quality. Moreover the active constraints help to define the real number of clusters in the competitive agglomeration process. The algorithms’ realization, data modeling and experiments were performed in R Studio environment using the R language [10].

References

1. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning. MIT Press, Cambridge, MA, USA (2006). ISBN 0-262-25589-8
2. Basu, S., Banerjee, A., Mooney, R.J.: Semi-supervised clustering by seeding. In: 19th International Conference on Machine Learning (ICML-2002), pp. 19–26. Morgan Kaufmann Publishers Inc., San Francisco (2002)
3. Basu, S., Davidson, I., Wagstaff, K.: Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall, Boca Raton (2008)
4. Wagstaff, K.L.: Value, cost, and sharing: open issues in constrainted clustering. In: 5th International Workshop on Knowledge Discovery in Inductive Databases, pp. 1–10 (2007)
5. Mallapragada, P.K., Jin, R., Jain, A.K.: Active query selection for semi-supervised clustering. In: 19th International Conference on Pattern Recognition (2008). doi:10.1109/ICPR.2008.4761792
6. Sk, J.A., Prasad, M., Gubbi, A., Rahman, H.: Active Learning of constraints using incremental approach in semi-supervised clustering. Int. J. Comput. Sci. Inf. Technol. **6**(2), 1962–1964 (2015)
7. Vu, V.V., Labroche, N., Bouchon-Meunier, B.: Boosting clustering by active constraint selection. In: 19th European Conference on Artificial Intelligence, ECAI-2010, pp. 297–302 (2010)
8. Grira, N., Crucianu, M., Boujemaa, N.: Semi-supervised fuzzy clustering with pairwise-constrained competitive agglomeration. In: 14th IEEE International Conference on Fuzzy Systems (Fuzz’IEEE 2005), May 2005. doi:10.1109/FUZZY.2005.1452508
9. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of Cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)
10. R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2013). http://www.R-project.org