Proceedings of the 9^{th} International Conference on Applied Informatics
Eger, Hungary, January 29–February 1, 2014. Vol. 2. pp. 121–130
doi: 10.14794/ICAI.9.2014.2.121

Utilization of constrained spectral clustering for clustering of graph nodes containing record data*

László Gonda

University of Debrecen, Faculty of Informatics
gonda.laszlo@inf.unideb.hu

Abstract

Clustering is one of the most common and most widely used methods of data mining. Many clustering algorithms can be utilized for various purposes, but most of these methods can only deal with one data type at a time. There are few methods in existence that can deal with data of different origins and/or different types simultaneously, but most of these are iterative [6], which means that they consume large amounts of computation time, and therefore refreshing the clustering can be a serious issue as well.

However, by using the method of constrained spectral clustering, which was originally established to define graph nodes that surely fall into the same cluster or surely not, we can incorporate the influence of the similarities of regular record data associated to the nodes without using complex iterative methods. Furthermore, using this system, we can also introduce different "constraint matrices" that can represent different aspects of the objects, or different subsets of their attributes grouped together based on importance, etc. and incorporate them in the forming of the final clustering without the need for an even lengthier iteration, or serious alterations to the original system. The original constrained spectral clustering method in [7] can also be weighted to determine whether the constraints or the graph edges will be of higher importance in the resulting clustering, thus we can say that the method is customizable in many aspects.

Keywords: data mining, clustering, constrained spectral clustering, graph data, record data, co-clustering

MSC: 68U99

*This research is supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.
supported by the European Union, co-financed by the European Social Fund.

1. Introduction

Clustering is an important data mining method that can be utilized in various use cases, for example in smart city or smart campus applications, but due to the nature of these notions, "heterogeneous data sources for related data need to be integrated" [1], as various kinds of data are collected about the users of such systems, e.g. relationships that can be stored as graphs, in which the users are represented by nodes, record type data about their personal data, the paths that they have moved along on the area of the campus, etc.

However, although the basic method of clustering, i.e. classifying entities into a predefined number of clusters based on their properties is one of the most widely used data mining methods, literature about utilizing multiple, heterogeneous data sources for clustering is rather sparse.

In this paper, the utilization of a method originally designed to introduce constraints into the clustering of nodes in a graph as a method for clustering heterogeneous data is proposed, and the possibilities and limitations of this are presented.

Thus, the aim of the paper is to explore whether, and to what extent this existing constrained spectral clustering method [7] can be utilized for the clustering of data from heterogeneous sources. In this case, this means graph data and record data, and thus to yield a method that can be utilized in such settings.

2. Related work

In the literature of clustering data from different data sources, only a few examples of handling graphs and record data, i.e. interconnected nodes containing record data together can be found, but however, even these have some drawbacks.

In one solution [2], the relationships between the types of data are handled, but the clustering is done basically relying on the record data. Thus, this solution does not use the graph of the individual entities represented as nodes, it only draws up the connections between different data types, so this is not suitable for the purposes heterogeneous clustering algorithms need to be utilized for in smart campus settings.

In a solution that handles record data and their relationships in web pages [6], an iterative approach is used to merge separate clusterings calculated from the different data types, which means that this algorithm might be suitable for the clustering of interconnected smart campus users with individual record data, but due to its iterative nature, the computational costs and running time of it might rise too high.

However, looking at methods for clustering based on a single data source or data type, it can be seen that the method of spectral clustering can be utilized easily for either graph or record type data, which raises the question whether it can be used for the clustering of data from these two different data sources at the same time.

Spectral clustering is “simple to implement, can be solved efficiently by standard linear algebra software, and very often outperforms traditional clustering algorithms such as the k-means” [5].

```
Input: similarity matrix $S \in R^{n \times n}$, the number of clusters to be created $k$
    • Create the similarity graph and let $W$ be the weighted affinity matrix
    • Calculate the normalized Laplace-matrix $L_{sym}$
    • Calculate the first $k$ eigenvectors $u_1, \ldots, u_k$ of $L_{sym}$
    • Place these vectors $u_1, \ldots, u_k$ as columns in the matrix $U \in R^{n \times k}$
    • Create the matrix $T \in R^{n \times k}$ by normalizing the rows of $U$ to 1
    • For $i = 1, \ldots, n$, let $y_i \in R_k$ be the vector corresponding to the $i$-th row of $T$
    • Cluster the vectors $(y_i)_{i=1,\ldots,n}$ into the clusters $C_1, \ldots, C_k$ using the $k$-means algorithm
Output: the clusters $A_1, \ldots, A_k$, for which $A_i = \{j : y_j \in C_i\}$ holds
```

Figure 1: The algorithm of spectral clustering [3]

The algorithm of spectral clustering shows that the original method is designed for the clustering of nodes in a graph, but by a simple modification of replacing the affinity matrix of a graph with the similarity matrix of record data, it can be utilized for the clustering of record data as well.

Also, several extensions of the original algorithm exist, including constrained spectral clustering [7], in which a separate constraint matrix is introduced, originally to assist the clustering performed on the edges of a graph, by defining constraints that indicate that two given nodes either must be linked, i.e. should be in the same cluster, or cannot be linked, i.e. cannot be in the same cluster.

```
Input: affinity matrix $A \in R^{n \times n}$, constraint matrix $Q \in R^{n \times n}$, parameter $\beta$ (within the range $(\lambda_{min}\left(\overline{Q}\right) * vol$, $\lambda_{max}\left(\overline{Q}\right) * vol)$, the number of clusters to be created $k$
    • Calculate the volume of the affinity matrix: $vol \leftarrow \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij}$
    • Calculate the diagonal of the affinity matrix: $D \leftarrow diag\left(\sum_{j=1}^{N} A_{ij}\right)$
    • Calculate the normalized Laplacian of the affinity matrix:$\overline{L} \leftarrow I - \sqrt{D}\ A\ \sqrt{D}$
    • Normalize the constraint matrix: $\overline{Q} \leftarrow \sqrt{D}\ Q\ \sqrt{D}$
    • Solve the generalized eigenvalue system $\overline{L}v = \left(\overline{Q} - \frac{}{vol} I\right) v$
    • Normalize the eigenvectors: $v \leftarrow \frac{v}{\|v\|}\ \sqrt{vol}$
    • $V^* \leftarrow argmin_{V \in N \times (K-1)}\ trace\left(V^T \overline{L} V\right)$
    • $u^* \leftarrow kmeans\left(\sqrt{D}\ V^*,\ \ K\right)$
```

Figure 2: The algorithm of constrained spectral clustering [7]

This extension shows that in constrained spectral clustering, the clustering of graph data can be augmented with the use of additional data, such as a constraint matrix, that holds additional information about the relationships between given nodes besides the edges between them that are described in the affinity matrix of the graph.

3. Method

The above extension of the original spectral clustering algorithm may be utilized, taking into consideration the fact that constraints are not necessarily values clearly indicating that two given graph nodes must or cannot be linked, but they can also be real numbers between 0 and 1 that can be interpreted as probabilities of the given two nodes belonging to the same cluster. This may lead to a method of clustering based on graph and record type data at the same time, if the constraint matrix is built from similarities calculated between the individual records belonging to each graph node.

Thus, the algorithm shown in Figure 2 was used with the modification of having n records corresponding to the n nodes in the graph as an input instead of the constraint matrix Q, and calculating the constraint matrix from these records using a similarity function. During the test of this setting, the Gaussian similarity function was used. Thus the cluster labels were obtained following the steps depicted in Figure 3.

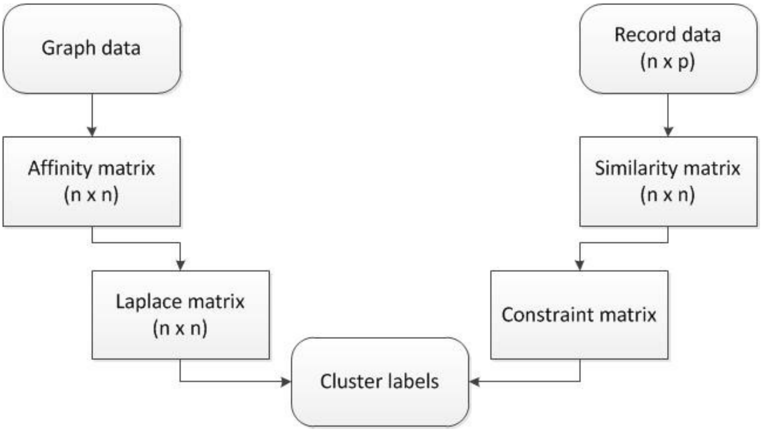

Figure 3: Using constrained spectral clustering to cluster graph and record data simultaneously

This alternative usage of the constrained spectral clustering algorithm was tested both on a synthetic dataset and samples of a real dataset.

4. Experimental results

Firstly, the modified algorithm was tested on a synthetic dataset consisting of 50 nodes in a graph, and a record containing two real numbers corresponding to each node. The data were generated in such a way that both the graph and the record data contain three predefined clusters. Nodes number 1 to 15, 16 to 30, and 41 to 50 respectively have a significantly higher number of connections among each other than to nodes in the other clusters. As for the record data, the numbers generated are similar to each other and relatively dissimilar to the ones in the other nodes in nodes number 1 to 20, 21 to 40, and 41 to 50 respectively. Thus the clusters that can be formed based on the graph and the record data have some overlaps, but they do not coincide completely, as can be seen of Figure 4.

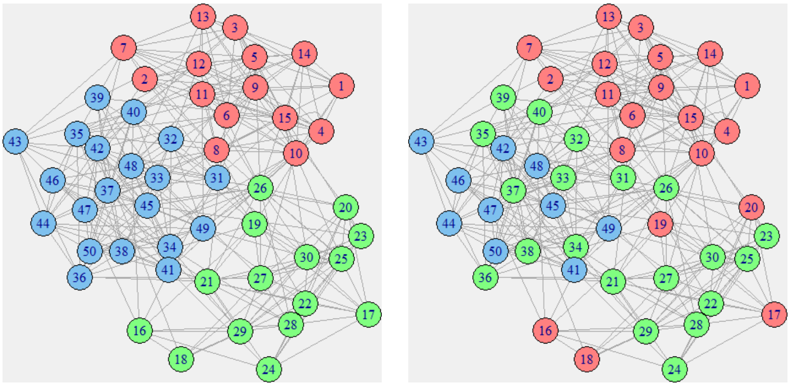

Figure 4: Graphic representation of the synthetic dataset with colors indicating cluster labels based on graph edges (left) and record values (right)

When not only one dimension of the data was used, but the similarities of the record data corresponding to the nodes of the graph were indeed added as constraints to the modified constrained spectral clustering algorithm, different clustering results could be obtained.

At this point, it has to be highlighted that a property of the original algorithm can be exploited here as well - namely that the parameter β influences the weight of the constraints when calculating the cluster labels, i.e. the higher the value of β is, the higher the effect of the constraints [7].

Based on the two end values of the range $(\lambda_{min}\left(\overline{Q}\right) * vol, \lambda_{max}\left(\overline{Q}\right) * vol)$ defined for the value of β, which were in this case $\beta = 10681.73$ for $\lambda_1\left(\overline{Q}\right) * vol$ and $\beta = -343623.1$ for $\lambda_{50}\left(\overline{Q}\right) * vol$ respectively, the results obtained are visualized in Figure 5.

These results show that when using higher β values, the original graph clustering can be influenced successfully using the constraints derived from similarities between the records corresponding to the individual nodes, but if the value of β is

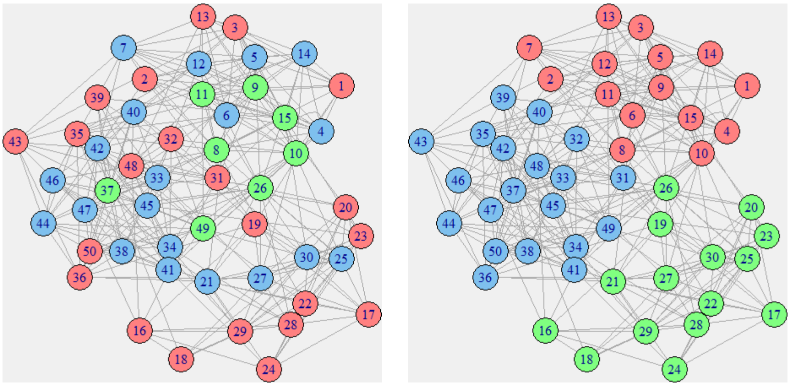

Figure 5: Graphic representation of the synthetic dataset with colors indicating cluster labels based on calculations using the highest (left) and the lowest (right) possible β value

decreased sufficiently, the effect of the constraints can even be removed entirely.

However, it has to be noted that, as can be seen in Figure 4, even when using the highest possible β value for the constraints, the record data will not dominate the resulting clustering completely, as the graph data do when β is set to a lower value. This is due to the fact that the algorithm is not designed to be able to completely leave out the influence of the affinity matrix of the graph, as when the data are normalized in the initial steps of the method, the normalized constraint matrix is computed using the formula $\overline{Q} \leftarrow D^{-1/2}QD^{-1/2}$, which relies on the diagonal matrix computed from the affinity matrix of the graph.

This leads to the conclusion that although the constraint and affinity matrices are computed from two different data sources, the constraint matrix cannot be handled completely separately from the graph. Thus, if the nodes of the graph are difficult to distinguish, e.g. the graph is very sparse, the ability of the constraint matrix to enhance the clustering is limited.

To show this, samples of the Amazon product co-purchasing network metadata dataset [8] was used, as this dataset, when considering the graph representation of the similarities of the given products, is rather sparse, only approximately 6% of all possible connections is present between the nodes representing the individual products.

Two samples were selected from the dataset, both consisting of 100 nodes, but one of the samples was assembled from randomly selected nodes, while the other was created using a crawler-like method, by adding an initial node, than the nodes connected to the initial node, the nodes connected to these nodes, and so on. Then constraint matrices were calculated for both samples based on the number of reviews and the average rating of users for the products represented by the graph nodes.

The randomly assembled sample contains only one edge, so the graph forming part of it can be considered extremely sparse, while the other sample contains 184 edges.

As the random example contains only one edge, clustering could not be carried out successfully based on the graph, which was evident. Also however, adding the constraint matrix did not have any effect whatsoever on this unsuccessful clustering either. This happened even though performing the clustering based on solely the similarities of the records yielded usable results, as can be seen on Figure 6.

This was due to the fact that, as described above, the normalization of the constraint matrix relies heavily on the affinity matrix, and as a result of this, the sparser the affinity matrix is, the more the normalized constraint matrix resembles an identity matrix.

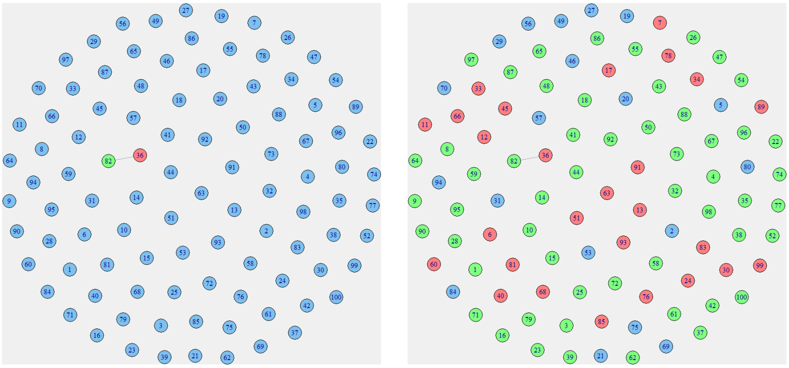

Figure 6: Graphic representation of the random sample with colors indicating cluster labels based on the graph edges without or with using constraints (left) and based on solely the record data (right)

In contrast, on the connected sample, the clustering was performable on the graph representing the similarities between the products, both using and omitting the constraint matrix based on the number of ratings and average ratings for the individual products. For the connected sample, the separate results yielded when performing clustering based only on the graph data, and the record data respectively, are shown in Figure 7.

Based on both data sources, the constrained spectral clustering could be performed as well, and as this sample dataset was appropriate for the usage of the constraint matrix, another notion was tested on it as well. One of the important questions raised when the research began concerning the alternative utilization of the original constrained spectral clustering algorithm was whether only one constraint matrix can be used, or multiple constraint matrices can be added to the method as well.

Without modifying the eigenvalue system defined in the original algorithm, this can be done by adding different constraint matrices together, and then using the thus created cumulative Q matrix as a constraint matrix. This was tested for a case in which two separate records are present for each graph node in the connected

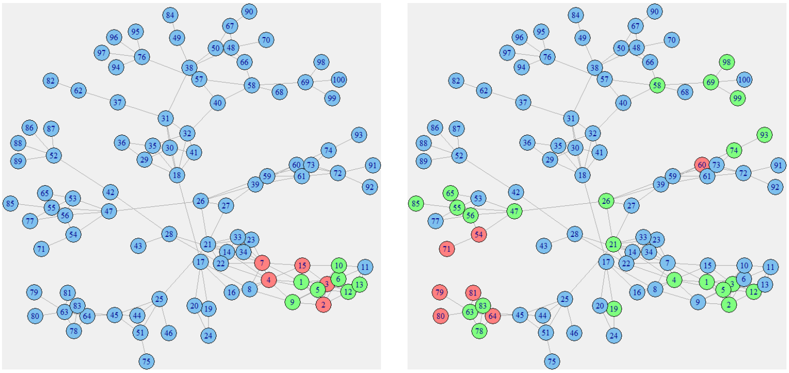

Figure 7: Graphic representation of the connected sample with colors indicating cluster labels based on the graph edges (left) and based on the record data (right)

sample, one containing the number of ratings and the average rating of the given product, and the other containing the salesrank of the given product. In this case, the two similarities of the two records for each node were computed in two separate matrices Q_1 and Q_2, and the sum of the two matrices formed the final constraint matrix Q. The results of the constraint spectral clustering performed on the connected sample using one and two constraint matrices respectively are presented in Figure 8.

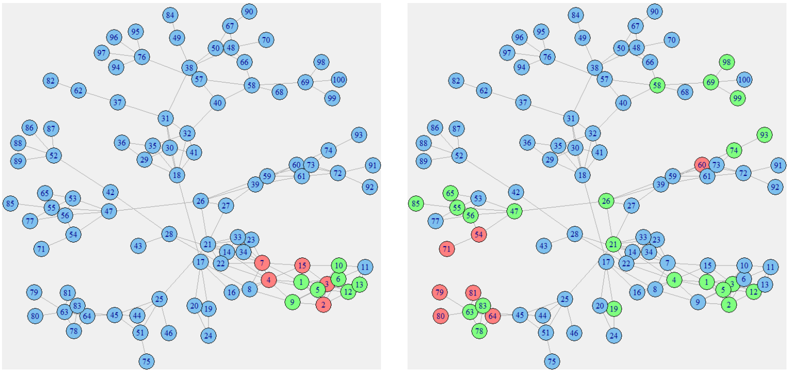

Figure 8: Graphic representation of the connected sample with colors indicating cluster labels based on the graph edges using a constraint matrix (left) and based on two constraint matrices (right)

It can be seen that using these different constraint matrices had different influences on the clustering based on the graph and its affinity matrix. However, it has to be noted, that although in the latter case, two different constraint matrices were used, the joint influence of these can only be as high as that of a single constraint matrix, as in the end, one constraint matrix is formed from them to fit the original algorithm.

Furthermore, this highlights again that using this algorithm, clustering based solely on the record values the similarities of which can be utilized as constraints cannot be carried out, as the algorithm relies heavily on the graph structure.

5. Conclusion

This research takes into consideration the possibilities of performing clustering on data that consist of heterogeneous data types. The most important finding of the present work is that constrained spectral clustering [7] can be utilized for such purposes.

It was shown that in the case of nodes of a graph with nodes containing record type descriptive data, the similarities computed between the record data of the individual nodes can be utilized as constraints, using which the clustering of the graph nodes can be influenced based on the similarities of the record data in the individual nodes.

In some practical applications, this can be a powerful tool, but unfortunately, it has several drawbacks that limit its utilization. The most important issue is that using the constraints is dependent on the graph structure, thus if the graph structure is unsuitable for clustering, the method cannot yield valid results regardless of whether constraints are available or not.

Furthermore, the effects of the constraints on the clustering are limited. Yet again due to the design of the algorithm, that relies mainly on the graph structure, a clustering that omits the graph structure entirely cannot be performed.

These limitations lead to the conclusion that the special utilization of constrained spectral clustering described in the present article needs to be improved further if possible, and further analysis has to be done regarding its usability and the conditions under which it yields appropriate results.

In addition, further methods that can be used for the clustering of heterogeneous data, e.g. graphs with nodes containing record type data should be sought out, or devised from existing methods, to the results of which the results yielded by this algorithm can be compared.

References

[1] A. Boran, I. Bedini, C. J. Matheus, P. F. Patel-Schneider, J. Keeney: A Smart Campus Prototype for Demonstrating the Semantic Integration of Heterogeneous Data. RR 2011, LNCS 6902, 238–243 (2011).

[2] A. Banerjee, S. Basu, S. Merugu: Multi-way Clustering on Relation Graphs. Proceedings of the 2007 SIAM International Conference on Data Mining, 145–156 (2007).

[3] A. Y. Ng, M. I. Jordan, Y. Weiss: On spectral clustering: analysis and an algorithm. Advances in Neural Information Processing Systems, 14, 849–856 (2002).

[4] S. E. Schaeffer: Graph clustering. Survey, Computer Science Review. I, 27−64 (2007).

[5] U. von Luxburg: A tutorial on spectral clustering. Stat Comput, 17, 295−416 (2007).

[6] J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, W-Y. Ma: ReCoM: Reinforcement Clustering of Multi-Type Interrelated Data Objects. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 274−281 (2003).

[7] X. Wang: Spectral Clustering in Complex Settings. Dissertation, University of California, Davis (2013).

[8] Amazon product co-purchasing network metadata – Source: J. Leskovec, L. Adamic, B. Adamic: The Dynamics of Viral Marketing. ACM Transactions on the Web (ACM TWEB), 1(1), 2007.