ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

A Spectral Based Clustering Algorithm for Categorical Data with Maximum Modularity

Lazhar labiod and Younès bennani

LIPN-UMR 7030, Université Paris 13,
99, av. J-B Clément, 93430 Villetaneuse, France
email: {firstname.secondname}@lipn.univ-paris13.fr

Abstract. In this paper we propose a spectral based clustering algorithm to maximize an extended Modularity measure for categorical data; first, we establish the connection with the Relational Analysis criterion. Second, the maximization of the extended modularity is shown as a trace maximization problem. A spectral based algorithm is then presented to search for the partitions maximizing the extended Modularity criterion. Experimental results indicate that the new algorithm is efficient and effective at finding a good clustering across a variety of real-world data sets

1 Introduction

Clustering is a method of unsupervised learning allowing the assignment of a set of observations into groups. Data clustering is a data analysis technique and has been considered as a primary data mining method for knowledge discovery. Clustering is defined as the process of partitioning a finite set of points in a multi-dimensional space into classes (called clusters)[14]. The problem of clustering becomes more challenging when the data is categorical, that is, when there is no inherent distance measures between data values. Many algorithms have been developed for clustering categorical data, e.g, 1998; Huang [13], 1998; Ganti et al. [1], 1999). The modularity measure has been used recently for graph clustering [15] [6]. In this paper, we show that the Modularity clustering criterion can be formally extended for categorical data clustering. We also establish the connections between the extended Modularity criterion and the Relational Analysis (RA) approach [2][3] which is based on Condorcet's criterion. We then develop an efficient spectral based procedure to find the optimal partition maximizing the extended Modularity criterion. The rest of the paper is organized as follows: Section 2 introduces some notations and definitions, in Section 3 provides the proposed extended modularity measure and its connection with the RA criterion. Some discussions on the spectral connection and optimization procedure are given in Section 4. Section 5 shows our experimental results and finally, Section 6 presents the conclusions and some future works.

2 Definitions and Notations

Let D be a dataset with a set I of N objects $(O_1, O_2, ..., O_N)$ described by the set V of M categorical attributes (or variables) $V^1, V^2., V^m, .., V^M$ each

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

one having $p_1, .., p_m, .., p_M$ categories respectively and let $P = \sum_{m=1}^{M} p_m$ denote the full number of categories of all variables. Each categorical variable can be decomposed into a collection of indicator variables. For each variable V^m, let the p_m values naturally correspond to the numbers from 1 to p_m and let $V_1^m, V_2^m, ..., V_{p_m}^m$ be the binary variables such that for each j, $1 \leq j \leq p_m$, $V_k^m = 1$ if and only if the V^m takes the j-th value. Then the dataset can be expressed as a collection of M $N \times p_m$ matrices K^m, $(m = 1, .., M)$ of the general term k_{ij}^m such as: $k_{ij}^m = 1$ if the object i takes the attribute j of V^m and 0 otherwise. Which gives the N by P binary disjunctive matrix $K = (K^1|K^2|...|K^m|...|K^M)$. For each variable V^m, the similarity matrix S^m can be expressed as $S^m = K^m(K^m)^t$, the global similarity matrix (Condocet's matrix) $S = KK^t$ where $(K^m)^t$ and K^t are the transposed K^m and the transposed K matrix, respectively.

3 Extended Modularity Measure.

Modularity is a recently quality measure for graph clustering, it has immediately received a considerable attention in several disciplines [15] [6]. Given the graph $G = (V, E)$, let A be a binary, symmetric matrix with (i, j) as entry; and $a_{ij} = 1$ if there is an edge between the nodes i and j. If there is no edge between nodes i and j, a_{ij} is equal to zero. We note that in our problem, A is an input having all information on the given graph G and is often called an adjacency matrix. Finding a partition of the set of nodes V into homogeneous subsets leads to the resolution of the following integer linear program: $\max_X Q(A, X)$ where

$$Q(A, X) = \frac{1}{2|E|} \sum_{i,i'=1}^{n} (a_{ii'} - \frac{a_{i.}a_{i'.}}{2|E|})x_{ii'} = \frac{1}{2|E|} Tr[(A - \delta)X] \quad (1)$$

is the modularity measure, $2|E| = \sum_{i,i'} a_{ii'} = a_{..}$ is the total number of edges and $a_{i.} = \sum_{i'} a_{ii'}$ the degree of i and $\forall i, i'$ $\delta_{ii'} = \frac{a_{i.}a_{i'.}}{2|E|}$. X is the solution we looking for which must satisfy the following properties of an equivalence relation defined on $I \times I$; $x_{ii} = 1$, $\forall i$ (reflexivity), $x_{ii'} - x_{i'i} = 0, \forall (i, i')$ (symmetry), $x_{ii'} + x_{i'i''} - x_{ii''} \leq 1, \forall (i, i', i'')$ (transitivity) and $x_{ii'} \in \{0, 1\}, \forall (i, i')$ (binarity)

We shows now how to adapt the Modularity measure for categorical data clustering. The basic idea consist in a direct combination of graphs from all variables into a single dataset (graph) before applying the learning algorithm. Let us consider the Condorcet's matrix S where each entry is denoted as $s_{ii'} = \sum_{m=1}^{M} s_{ii'}^m$, which can be viewed as a weight matrix associated to the graph $G = (I, E)$, where each edge $e_{ii'}$ have the weight $s_{ii'}$. Similarly to the classical Modularity measure, we define the extension $Q_1(S, X)$ as follows:

$$Q_1(S, X) = \frac{1}{2|E|} \sum_{i,i'=1}^{n} (s_{ii'} - \frac{s_{i.}s_{i'.}}{2|E|})x_{ii'} = \frac{1}{2|E|} Tr[(S - \delta)X] \quad (2)$$

where $2|E| = \sum_{i,i'} s_{ii'} = s_{..}$ is the total weight of edges and $s_{i.} = \sum_{i'} s_{ii'}$ - the degree of i. We can establish a relationship between the extended modularity

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

measure and the RA criterion, indeed the function $Q_1(S,X)$ can be expressed as a modified RA criterion in the following way:

$$Q_1(S,X) = \frac{1}{2|E|}(\mathcal{R}_{RA}(S,X) + \psi_1(S,X)) \quad (3)$$

where

$$\psi_1(S,X) = \sum_{i=1}^{n}\sum_{i'=1}^{n}(m_{ii'} - \frac{s_{i.}s_{i'.}}{2|E|})x_{ii'} \quad (4)$$

is the weighted term that depends on the profile of each pair of objects (i,i'), $\mathcal{R}_{RA}(S,X)$ is the relational analysis criterion(see Marcotorchino [3][2] for further details),

$$\mathcal{R}_{RA}(S,X) = \sum_{i}\sum_{i'}(s_{ii'} - m_{ii'})x_{ii'} \quad (5)$$

with $M = [m_{ii'} = \frac{s_{ii}+s_{i'i'}}{4} = \frac{M}{2}]_{i,i'=1,\ldots,N}$, the matrix of threshold values,s_{ii} and $s_{i'i'}$ are the self similarities of objects i and i'.

4 Maximizing the Normalized Extended Modularity with Spectral Algorithm

The original Modularity criterion is not balanced by the cluster size, meaning that a cluster might become small when affected by outliers. Thus we define the new measure which we call normalized extended modularity whose objective function is given as follows:

$$\tilde{\mathcal{Q}}_1(S,X) = Tr[(S-\delta)V^{-1}X] \quad (6)$$

where $V = diag(Xe)$ is a N by N diagonal matrix such that $v_{ii} = x_{i.}$ the number of objects in the same cluster with the object i. and $e = \mathbf{1}$ is the vector of appropriate dimension which all its values are 1.

4.1 Spectral connection

On one hand, it's well known that the largest eigenvalue of $\tilde{S} = D^{-\frac{1}{2}}SD^{-\frac{1}{2}}$(where $D = diag(Se)$) and its eigenvector are $\lambda_0 = 1$, $U_0 = \frac{D^{-\frac{1}{2}}e}{S_{..}}$ [16] [17], where $S_{..} = \sqrt{e^tSe}$.
We apply the spectral decomposition of the scaled matrix $\tilde{S}$ instead on S directly, leading to $S = D^{\frac{1}{2}}\sum_{k=0}U_k\lambda_kU_k^tD^{\frac{1}{2}}$. Subtract the trivial eigenvector corresponding the largest eigenvalue $\lambda_0 = 1$ give $S = \frac{Dee^tD}{e^tDe} + D^{\frac{1}{2}}\sum_{k=1}U_k\lambda_kU_k^tD^{\frac{1}{2}}$ we would consider instead the $(K-1)$th principals eigenvectors of

$$S - \frac{Dee^tD}{e^tDe} = D^{\frac{1}{2}}\sum_{k=1}U_k\lambda_kU_k^tD^{\frac{1}{2}} \quad (7)$$

this matrix multiplied by the constant $\frac{1}{e^tDe}$ (which has no effect on the position of the maximum of the modularity criterion)) is exactly the matrix $(S-\delta)$ used

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

in the modularity measure (see equation (1)).
On the other hand the problem of maximizing the normalized extended modularity can be modelled as a trace maximization problem. Consider the division of the data set I into $\mathcal{K}$ non overlapping clusters, where $\mathcal{K}$ may now be greater or equal to 2. Let us define an $N \times \mathcal{K}$ index matrix Z with one column for each cluster; $Z = (Z_1|Z_2|...|Z_{\mathcal{K}})$. Each column is an index vector now of $(0, 1)$ elements such that $Z_{ik} = (1$ if object i belongs to cluster k, 0 otherwise). The equivalence relation X and the weighted equivalence relation $V^{-1}X$ can now be factorized as follows: $X = ZZ^t$ and $V^{-1}X = \tilde{Z}\tilde{Z}^t$ where $\tilde{Z} = Z(Z^tZ)^{-1/2}$. It's easy to show that $\tilde{Z}$ satisfies the orthogonality constraint, then the maximization of the normalized extended modularity is equivalent to the following trace optimization problem

$$\max_{\tilde{Z}^t\tilde{Z}=I_K} Tr[\tilde{Z}^t(S-\delta)\tilde{Z}] \tag{8}$$

The matrix $S - \delta$ used in the modularity is expressed in term $(K - 1)$th largest eigenvectors of the scaled matrix $\tilde{S}$. After solving the spectral decomposition of $\tilde{S}$ we have the resultant $(K - 1)$ eigenvectors with largest eigenvalues. That is, we can have the $N \times (K - 1)$ matrix $U = [U_1, ..., U_{K-1}]$, where U_k is the $k - th$ eigenvector of the selected $K-1$ eigenvectors we then normalize this matrix into $N \times (K - 1)$ matrix $\tilde{U}$ in which $\tilde{U}_k = \frac{D^{\frac{1}{2}}U_k}{||D^{\frac{1}{2}}U_k||}$. This eigenmatrix $\tilde{U}$ can be an input of K-means, below the pseudo code of the proposed algorithm.

Algorithm2- SpectMod Algorithm :Given a set of data object that we want cluster into $\mathcal{K}$ clusters

1. Form the affinity matrix S
2. Define D to be the diagonal matrix $D = diag(Se)$
3. Find U the $\mathcal{K} - 1$ largest eigenvectors of $\tilde{S} = D^{-1/2}SD^{-1/2}$
4. Form the matrix $\tilde{U}$ from U by $\tilde{U}_k = \frac{D^{\frac{1}{2}}U_k}{||D^{\frac{1}{2}}U_k||}$, $\forall k = 1, ..., \mathcal{K} - 1$
5. Considering each row of $\tilde{U}$ as a point in $\mathbb{R}^{\mathcal{K}}$, cluster them into $\mathcal{K}$ clusters using k-means
6. Finally assign object i to cluster C_k if and only if the corresponding row $\tilde{U}_i$ of the matrix $\tilde{U}$ was assigned to cluster C_k.

5 Experimental Results

A performance study has been conducted to evaluate our method. In this section, we describe those experiments and the results. We ran our algorithm on real-life data set obtained from the UCI Machine Learning Repository to test its clustering performance against other algorithms. Validating clustering results is a non-trivial task. In the presence of true labels, as in the case of the data set

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

we used, the clustering purity is used to measure the quality of clustering. The description of the used data sets is given in Table 1:

Table 1: description of the data set

Data set	# of Objects	# of Attributes	Classes
Soybean small	47	21	4
Mushroom	8124	22	2
Congressional votes	435	16	2
Zoo	101	16	7
Hayes-roth	132	4	3
Balance Scale	625	4	3
Car evaluation	1728	6	4
Soybean large	307	35	19
SPECTF	267	22	2
Post-Operative	90	8	3

5.1 Results analysis

We studied the clusterings found by different algorithms, we first compare the proposed SpectMod algorithm, RA algorithm based on the extended modularity measure and the RA algorithm [3]. Second, we compare our SpectMod algorithm with standard k-modes algorithm introduced in [13], K-representative algorithm proposed in [4], weighted k-modes algorithm [10].

Table 2: Purity measure (%)for $\mathcal{R}_{RA}(S,X)$ et $Q_1(S,X)$ and $SpectMod$

BD	Taille	$\mathcal{R}_{RA}(S,X)$	$Q_1(S,X)$	$SpectMod$
Soybean small	47x21	78	100	100
Zoo	101x16	83	88	90
Soybean large	307x35	70	72	76
SPECTF	267x22	61	72	76
Post-Operative	90x8	71	73	73
Balance Scale	625x4	63	63	65

Table 3: Purity measure (%) for K-modes, K-representatives, weighted k-modes and SpectMod

Data set	K-Modes	K-Representatives	WK-Modes	SpectMod
Soybean small	66	89	89	100
Mushroom	59	61	61	61
Congressional votes	62	87	88	88
Zoo	88	89	90	90
Hayes-roth	41	42	42	54
Balance Scale	50	52	52	65
Car evaluation	70	70	71	70

The Tables 2 and 3 summaries the result of clustering purity, the proposed method SpectMod brought better or similar clustering purity than the other algorithms, which means that the proposed approach improves the clustering purity.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

6 Conclusions

In this paper, we have studied the spectral interpretation of the the extended modularity maximization for categorical data clustering. An efficient spectral procedure for optimization is presented, the experimental results obtained using different data sets showed that our method worked favorably for categorical data. Our method can be easily extended to more general spectral framework for combining multiples heterogenous data sets for clustering. Thus, an interesting future work is to apply the approach on a variety of heterogenous data sets; numerical data, categorical data and graph data.

References

[1] V. Ganti, J. Gehrke & Ramakrishnan, R. (1999). CACTUS - clustering categorical data using summaries. Proceedings of the Fifth ACM SIGKDD Conference (pp. 73- 83).

[2] J. F. Marcotorchino, (2006). *Relational analysis theory as a general approach to data analysis and data fusion.* In Cognitive Systems with interactive sensors, 2006.

[3] J. F. Marcotorchino, P. Michaud, (1978). *Optimisation en analyse ordinale des données* In Masson, 1978.

[4] Ohn Mar San, Van-Nam Huynh, Yoshiteru Nakamori, "An Alternative Extension of The K-Means algorithm For Clustering Categorical Data", J. Appl. Math. Comput. Sci, Vol. 14, No. 2, 2004, 241-247.

[5] R. S. Wallace, (1989). Finding natural clusters through entropy minimization (Technical Report CMU-CS-89- 183). Carnegie Mellon University.

[6] S. White and P. Smyth, (2005). A spectral clustering approach to finding communities in graphs. In SDM, pages 76-84.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, August 2000.

[8] A. Y. Ng, M. Jordan and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Proc. of NIPS-14, 2001.

[9] P. Chan, M. Schlag and J. Zien, "Spectral k-way ratio cut partitioning," IEEE Trans. CAD-Integrated Circuits and Systems, vol. 13, pp. 1088-1096, 1994

[10] S. Aranganayagi and K. Thangavel, "Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure", International Journal of Engineering and Mathematical Sciences. vol5-2-19, 2009.

[11] U. Von Luxburg, (2006): A Tutorial on Spectral Clustering. Technical Report at MPI Tuebingen, 2006.

[12] H. Bock, (1989). Probabilistic aspects in cluster analysis. In O. Opitz (Ed.), Conceptual and numerical analysis of data, 12-44. Berlin: Springer-verlag .

[13] Z. Huang, (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2, 283-304.

[14] T. Li, S. Ma and M. Ogihara, (2004). Entropy-based criterion in categorical clustering. Proceedings of The 2004 IEEE International Conference on Machine Learning (ICML 2004). 536-543.

[15] Newman, M. and Girvan, M.(2004). Finding and evaluating community structure in networks. Physical Review E, 69, 026113.

[16] Ding, Chris H.Q. He, Xiaofeng, Zha, Hongyuan and Simon, Horst D, (2001). Self-aggregation in scaled principal component space. Technical Report LBNL–49048. Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US)

[17] Francis R. Bach and Michael I. Jordan, (2005). Learning spectral clustering, with application to speech separation. Journal of Machine Learning Research.