A Spectral View on the Relational Analysis Approach for Clustering Categorical Data

Lazhar Labiod∗

University Department, University Name, Address
City, State ZIP/Zone,Country†
author_id@domain_name‡
http://<webaddress>

Younès Bennani
Group, Laboratory, Address
City, State ZIP/Zone, Country
author_id@domain_name

This paper introduces a spectral view on the Relational Analysis (RA) approach for categorical data clustering. It first, establishes the theoretical connection between the RA approach and classical spectral clustering technique. In particular, the RA problem is shown as a trace maximization problem, Thus usually this problem is converted by spectral relaxation into an optimization problem with a constraint which can be solved by Lagrange multipliers, and the solution is given by an eigenvalue problem. Second, we propose a new spectral clustering procedure which combine the RA data representation and the spectral Algorithm. Experiments are conducted to show the effectiveness of the proposed procedure.

Keywords: Relational Analysis; Spectral clustering; Inertia criteria; Categorical data

1. Introduction

Clustering has received a significant amount of attention as an important problem with many applications, and a number of different algorithms and methods have emerged over the years. Clustering is the process of organizing objects into classes/groups such that the objects within the same cluster have a high degree of similarity, while objects belonging to other clusters have a high degree of dissimilarity. The clustering algorithms usually employ some heuristic processes to find local optimal results. Most clustering techniques use a pairwise similarity for measuring the distance of two data points. The widely employed examples are the

∗Typeset names in 8 pt roman, uppercase. Use the footnote to indicate the present or permanent address of the author.
†State without abbreviations, the affiliation and mailing address, including country. Typeset in 8 pt italic.
‡Typeset author e-mail address in single line
distance metric, e.g., the Euclidean distance and the Hamming distance. By using the inherent metric property within numerical data, some clustering algorithms perform very well to find sensible geometric partitions. Recently, researches have been elaborated upon the clustering approaches for categorical data, where categorical data are those whose attribute values are nominal and unordered, e.g., color and human hobby. It is noted that difficulties arise in clustering categorical data due to the absence of inherently ordered property of categorical data. Most clustering techniques based on the metric distance measure are thus not applicable to acquire clusters from categorical data domain.

On the other hand, Spectral methods have been used effectively for solving a number of graph partitioning objectives, including ratio cut and normalized cut. Such an approach has been useful in many areas, such as circuit layout and image segmentation. Researchers often show that their clustering method finds a minimum cut in a weighted undirected graph where the nodes are data objects and the weights associated with the edges reflect the similarity between data objects. Given a data set, minimizing normalized cut (or ratio cut) is a trace optimization problem which is NP-hard. Thus usually this problem is converted by spectral relaxation into an optimization problem with a constraint which can be solved by Lagrange multipliers, and the solution is given by an eigenvalue problem. This is called spectral graph clustering which is difficult to assign a cluster label to each node definitely. Thus usually k-means is finally applied to cluster assignments using the resultant eigenvectors. A key issue in spectral clustering, and one that differentiates many of the methods, is the construction of the similarity graph. In recent years spectral clustering has becomes one of the most popular modern clustering algorithms. Its simple to implement, can be solved efficiently by standard linear algebra software and often outperforms traditional algorithms such as the k-means algorithm.

The Relational Analysis approach has been used for categorical data clustering. In this paper, we show that the RA approach can formally modelled as a trace maximization problem. We also establish the connections between the spectral methods and the Relational Analysis (RA) approach which is based on Condorcet’s criterion. We then develop an efficient spectral procedure inspired from Ng Jordan and Weiss heuristic to find the optimal partition to maximizing the RA criteria. The experimental results demonstrate the efficacy and effectiveness of our approach.

The first contribution of the paper is the presentation of the weighted Condorcet criterion maximization as a trace maximization problem, and showing the RA criterion as a Min Max inertia criterion. The second contribution is the spectral relaxation of the RA approach. In particular, we propose a spectral algorithm for categorical data which combine the relational data representation and the principle of the spectral algorithm based on the eigenvalues computation.

The following is an outline of the paper. Section 2 introduces the classical Re-
ional Analysis approach (RA). Section 3 provides the normalized Condorcet’s criteria and his equivalence with the inertial criteria. Some discussions on the spectral connection of the RA approach and the proposed 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. Relational Analysis approach

Relational Analysis was developed in 1977 by F. Marcotorchino and P. Michaud, inspired by the work of Marquis de Condorcet, which was interested in the 18th century with the result of collective vote starting from individual votes. This methodology is based on the relational representation (pairwise comparison) of data objects and the optimization under constraints of the Condorcet criterion. Among many fields of the Relational Analysis methodology application, one side and two side unsupervised clustering, are two fields in which the Relational Analysis have made considerable theoretical and practical advances. Generally, the objective function corresponds to the criterion which measure the adequacy of the solution to the data. The choice of this criterion is a fundamental point since it induces the nature of the resemblances intensity which we want to emphasis. Among a vast range of criteria, the relational approach makes possible to choose the best one answering to the problem arising from the involved data. Some criteria operate on binary data, others are appropriate to frequencies data; but the most of them are based on majority rule which determine the level of the threshold value. Beyond this threshold it is considered that two objects are assigned to the same cluster. The relational analysis includes a set of techniques for analyzing data to solve problems under the following general formula:

*Search for a particular relation S that fits "best" a single (or set of ) given relations \( (R^1, R^2, \ldots, R^M) \).*

Let us recall that one of the major advantages of the relational approach resides in the fact that the number of clusters should not be fixed a priori. This parameter characteristic of the solution is directly resulting from the processing (in an unsupervised way).

2.1. Definitions and notations

Let \( D \) be a dataset with a set \( I \) of \( N \) objects \((O_1, O_2, \ldots, O_N)\) described by the set \( V \) of \( M \) categorical attributes (or variables) \( V^1, V^2, \ldots, V^M \), each one having \( p_1, \ldots, p_m, \ldots, p_M \) categories respectively and let \( P = \sum_{m=1}^{M} p_m \) to 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 to naturally correspond to the numbers from 1 to \( p_m \) and let \( V^m_1, V^m_2, \ldots, V^m_{p_m} \) be the binary variables such that for each \( j, 1 \leq j \leq p_m, V^m_k = 1 \) if and only if the \( V^m \) takes the j-th value. Then the data set can be expressed as a collection of M
matrices $K^m (N \times p_m)$ (for $m = 1, \ldots, M$) of general term $k^m_{ij}$ such as:

$$k^m_{ij} = \begin{cases} 1 & \text{if the object } i \text{ takes the categorie } j \text{ of } V^m \\ 0 & \text{otherwise} \end{cases}$$

(1)

which gives the $N$ by $P$ binary disjunctive matrix $K = (K^1|K^2|\ldots|K^m|\ldots|K^M)$.

2.1.1. Weighted disjunctive matrix

The weighted disjunctive matrix $\tilde{K}$ is obtained by devising each entry $k_{ij}$ of $K$ by the square root of the product of the marginal row sum $k_i$, and column sum $k_j$. In other words, each entry $\tilde{k}_{ij} = \frac{k_{ij}}{\sqrt{k_i k_j}}$, with matrix notation we write:

$$\tilde{K} = R^{-\frac{1}{2}} K C^{-\frac{1}{2}}$$

(2)

where

$$R = \text{diag}(K e) \quad \text{and} \quad C = \text{diag}(K^t e)$$

(3)

where $e = 1$ is the vector of appropriate dimension which all its values are 1 and $\text{diag}(\cdot)$ denote the diagonal matrix.

2.2. Similarity and dissimilarity matrices

If the data is made up of $N$ objects ($O_1, O_2, \ldots, O_N$) on which $M$ attributes (or variables) ($V^1, V^2, \ldots, V^M$) have been measured then the "pairwise comparison principle" consists in transforming the data, which is usually, represented by a $N \times M$ rectangular matrix into two squared $N \times N$ matrices $\tilde{S}$ and $\overline{S}$. The matrix $\tilde{S}$, which is called the global relational weighted Condorcet matrix, of general term $\tilde{s}_{ii'}$ representing the global similarity measure between the two objects $O_i$ and $O_{i'}$ over all the $M$ attributes and matrix $\overline{S}$ of general term $\overline{s}_{ii}$ which represent the global dissimilarity measure of these two objects.

$$\tilde{S} = \tilde{K} \tilde{K}^t$$

(4)

$$\overline{S} = \tilde{s}_{\text{max}} - \tilde{S}$$

(5)

To get matrix $\overline{S}$, a dissimilarity measure $\tilde{s}_{ii'}$ of objects $O_i$ and $O_{i'}$ with regards to the set of attributes $V$ is then computed as the complement to the maximum possible similarity measure between these two objects. As the similarity between two different objects is less or equal to their self-similarities: $\tilde{s}_{ii'} \leq \min(\tilde{s}_{ii}, \tilde{s}_{ii'})$ then we have

$$\tilde{s}_{\text{max}} = \frac{1}{2}(\tilde{s}_{ii} + \tilde{s}_{ii'})$$

(6)
and
\[ \tilde{s}_{ii'} = \tilde{s}_{ii'}^{\text{max}} - \tilde{s}_{ii'} = \frac{1}{2} (\tilde{s}_{ii} + \tilde{s}_{ii'}) - \tilde{s}_{ii'} \]  

(7)

This leads to a dissimilarity measure matrix \( \tilde{S} \) (see appendix for more details on similarity and dissimilarity computation).

The figure below shows the transition from a linear data coding to a relational coding.

Fig. 1. Data coding

2.3. Weighted Condorcet criterion maximization

To cluster a data set of \( N \) objects described by \( M \) variables, the relational analysis approach maximizes the weighted Condorcet criterion:

\[
\max_{X \in E(X)} R_{RA}(\tilde{S}, X) 
\]

(8)

where

\[
R_{RA}(\tilde{S}, X) = Tr(\tilde{S}X) + Tr(\tilde{S}X) 
\]

(9)

\[
= Tr[(\tilde{S} - \tilde{S})X] + e'e\tilde{S}e 
\]

(10)

\[
= Tr[(\tilde{S} - \tilde{S})X] + \beta 
\]

(11)

Where \( \beta \) is a constant term (where here and henceforth we suppress the constant \( \beta \), which has no effect on the position of the maximum of the RA criterion). \( X \) is the searched solution which models a partition in a relational space (an equivalence relation), and must check the following properties:
\[
\begin{align*}
    x_{ii} &= 1, \quad \forall i \quad \text{reflexivity} \\
    x_{ii'} - x_{i'i} &= 0, \forall (i, i') \quad \text{symmetry} \\
    x_{ii'} + x_{i'i''} - x_{i'i''} &\leq 1, \forall (i, i', i'') \quad \text{transitivity} \\
    x_{ii'} &\in \{0, 1\}, \forall (i, i') \quad \text{binarity}
\end{align*}
\]

(12)

\(E(X)\) is the set of square matrices that satisfies the properties given in (12). \(X\) is the complementary matrix of \(X\) where each entry \(\overline{x}_{ii'} = 1 - x_{ii'}\).

The figure 2 show the RA process, given a similarity matrix, the basic idea consist in permuting the data rows and columns in order to show a block diagonal structure of the data matrix with a high similarity density inside the blocks and low similarity density outside.

2.4. **Relational analysis heuristic**

The heuristic process starts from an initial cluster (a singleton cluster) and build in an incremental way, a partition of the set \(I\) by increasing the value of Condorcet criterion \(\mathcal{R}_{RA}(S, X)\) at each assignment. We give below the description of the Relational Analysis algorithm which was used by the Relational Analysis methodology (see Marcotorchino and Michaud \(^{15}\) for further details). The presented algorithm aims at maximizing the criterion given in (11) based on the contribution computation.
Let us consider $C = \{C_1, ..., C_K\}$ a partition of the set $I$ into $K$ clusters, the Condorcet criterion breaks up into terms of contributions where the contribution $\text{cont}(i, k)$ of an object $i$ in a cluster $C_k$ of the partition is written:

$$\text{cont}(i, k) = \sum_{i' \in C_k} [\hat{s}_{ii'} - (\hat{s}_{ii'} + \hat{s}_{ii'})]$$ (13)

we have

$$\mathcal{R}_RA(\tilde{S}, X) = \sum_{i=1}^{N} \sum_{k=1}^{K} \text{cont}(i, k)$$ (14)

That we can express in terms of the object profile $\hat{K}_i$ representing the $i^{th}$ row of the complete disjunctive table $\hat{K}$ and $P_k$ the prototype of cluster $C_k$ in the following way based on the transformations:

$$\hat{s}_{ii'} = <\hat{K}_i, \hat{K}_{i'}> \quad \text{and} \quad P_k = \sum_{i' \in C_k} \hat{K}_{i'}$$ (15)

Then, we have

$$\text{cont}(\hat{K}_i, P_k) = <\hat{K}_i, P_k> - \frac{|C_k| <\hat{K}_i, \hat{K}_i> + \sum_{i' \in C_k} <\hat{K}_{i'}, \hat{K}_{i'}>}{2}$$ (16)

This new formula of the contribution avoids the computation of the square matrices $\hat{S}$ and $\overline{S}$ (Condorcet matrix and its complementary) which reduces considerably the computational cost related to the contributions computation.

The RA clustering problem is related to (0-1) linear programming. The number of variables is $N \times N$, the number of constrains is in the order of $O(N^3)$. Theoretically, the problem can be solved exactly by a linear programming technique, but unfortunately only for the small problems where the size $N \leq 100$; hence only the heuristic approach can be deal with large data set.
Algorithm 1: RA heuristic

Inputs:
\( K_{\text{max}} \) = maximal number of clusters, \( N_{\text{iter}} \) = number of iterations, \( N \) = number of examples (objects), \( \alpha \) = similarity threshold
- take the first object as the first element of the first cluster.
- \( l = 1 \) where \( l \) is the current number of clusters

for \( t = 1 \) to \( N_{\text{iter}} \) do
  for \( i = 1 \) to \( N \) do
    for \( j = 1 \) to \( l \) do
      Compute the contribution of object \( i : \text{cont}(i, j) \)
    end for
    \( l^* = \arg\max_j \text{cont}(i, j) \),
    where \( l^* \) is the cluster id which has the highest contribution with the object \( i \)
    \( \text{cont}(i, l^*) \leftarrow \) the computed contribution
    if \( \text{cont}(i, l^*) < 0 \) and \( l < K_{\text{max}} \) then
      create a new cluster where the object \( i \) is the first element
      \( l \leftarrow l + 1 \)
    else
      assign object \( i \) to cluster \( C_{l^*} \).
    endif
  end for
end for

Output:
at most \( K_{\text{max}} \) clusters

We have to fix a number of iterations and the similarity threshold in order to have an approximate solution in a reasonable processing time. Besides, it is also required a maximum number of clusters, but since we don’t need to fix this parameter, we put by default \( K_{\text{max}} = N \). Basically, this algorithm has \( O(N_{\text{iter}} \times K_{\text{max}} \times N) \) computation cost. In general term, we can assume that \( N_{\text{iter}} \ll N \), but not \( K_{\text{max}} \ll N \). Thus, in the worst case, the algorithm has \( O(K_{\text{max}} \times N) \) computation cost.

3. Normalized Condorcet criterion

The Condorcet measure is defined by the (weighted) number of objects only, meaning that the weighted Condorcet criterion considers the number of categories shared by each pair of objects; most the shared categories by two objects are rare in the dataset more their similarity is high. More importantly, the original Condorcet 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 Condorcet criteria whose cost (or the objective function) is given as follows:
\[ \tilde{R}_{RA}(\tilde{S}, X) = Tr[(\tilde{S} - \overline{\tilde{S}})V^{-1}X] \]  

(17)

where \( V \) is a \( N \) by \( N \) diagonal matrix such that \( v_{ii} = x_i \), the number of objects in the same cluster with the object \( i \).

\[ V = diag(Xe) \]  

(18)

and \( e = 1 \) is the vector of appropriate dimension which all its values are 1. We illustrate below, on a small data set, the structure of the relational matrices \( X, V \) and \( V^{-1}X \)

\[ X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad V = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad V^{-1}X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \]  

(19)

3.1. Normalized Condorcet as a Max-Min inertial criteria

According to the Huygens principle, we know that the total inertia of a partition \( X \) denoted \( I_T(X) \) is the sum of its within cluster inertia \( I_W(X) \) and its between cluster inertia \( I_B(X) \); consider below the Huygens formula (see the appendix for more details on the inertia formulas):

\[ I_T(X) = I_B(X) + I_W(X) \]  

(20)

In terms of operator matrix trace, denoted \( (Tr) \), those inertial formulas are expressed:

\[ I_T = \frac{1}{N} Tr(\tilde{S}1_{N \times N}) = \frac{P}{M} - 1, \text{ with } 1_{N \times N} = ee^t \]  

(21)

\[ I_B(X) = Tr(\tilde{S}V^{-1}X) - 1 \]  

(22)

and

\[ I_W(\tilde{X}) = Tr(\tilde{S}V^{-1}X) \]  

(23)

We give below the explicit expressions of the different inertias 16:

\[ I_W(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \tilde{s}_{ii'} \frac{x_{ii'}}{x_i} \]  

(24)
and

\[ I_B(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \tilde{s}_{ii'} \frac{x_{ii'}}{x_{ii'}} - 1 \]  \tag{25}

We can observe that the normalized Condorcet criteria express the inertia difference \( I_B(X) - I_W(X) \):

\[ \tilde{R}_{RA}(\tilde{S}, X) = I_B(X) - I_W(X) + 1 \]  \tag{26}

Following the Huygens formula, we have the next equivalences:

\[
\begin{align*}
\max_X \tilde{R}_{RA}(\tilde{S}, X) &\iff \max_X I_B(X) \\
&\iff \min_X I_W(X) \\
&\iff \max_X [I_B(X) - I_W(X)]
\end{align*}
\]  \tag{27-29}

This is a well known result to maximize \( \max_X I_B(X) \) or to minimize \( \min_X I_W(X) \). Also a trivial solution exists if we do not place constraints upon the number of cluster; the solution consists of a partition in \( N \) clusters, where all objects are isolated. In the case of categorical data and for a trivial partition \( C \), solution of the inertia maximization problem, without constraints upon the number of clusters, the between cluster inertia keeps the value \( P^* - 1 \). To avoid the trivial partition presented above, it necessary to add a constraint fixing the desired number of clusters.

\[
\max_{\{X \in \mathcal{E}(X). Tr(V^{-1}X) = K\}} \tilde{R}_{RA}(\tilde{S}, X)
\]  \tag{30}

As \( X \) satisfies an equivalence relation, it has the form of a \( K \) blocks diagonal matrix where the sum of diagonal elements of each block of \( V^{-1}X \) is equal to 1. Indeed we have

\[
\text{Tr}(V^{-1}X) = \sum_{i=1}^{N} v_{ii}^{-1} x_{ii} = \sum_{i=1}^{N} \frac{x_{ii}}{x_{ii}} = \sum_{k=1}^{K} (\sum_{i \in C_k} x_{ii}) = \sum_{k=1}^{K} 1 = K
\]  \tag{31}

4. Spectral connection

In this section we give a spectral interpretation of the normalized Condorcet criterion by exploiting some relational algebraic properties of the equivalence relation \( X \). The problem of finding the set of clusters which maximizes this cost is NP-hard. We then apply the spectral clustering approach to maximizes the normalized Condorcet criterion.
4.1. **Decomposition of the equivalence relation** $X$

Before detailing the spectral interpretation of the RA approach, recall here some properties of the relation $X$. Consider the division of the data set $I$ into $L$ non overlapping clusters, where $K$ may now be greater or equal to 2. Let us define an $N \times K$ index matrix $Z$ with one column for each cluster; $Z = (Z_1|Z_2|...|Z_K)$. 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 matrix satisfies the normalization condition $Tr(Z^t Z) = N$, and we have also $Z^t Z = N = diag(n_1,...,n_l,...,n_K)$ where $n_k$ is the cardinality of the class $k$. The equivalence relation $X$ and the weighted equivalence relation $V^{-1}X$ can now be factorized as follows:

$$X = ZZ^t \quad (32)$$

and

$$V^{-1}X = Z(Z^t Z)^{-1}Z^t \quad (33)$$

The relation $X$ is an equivalence relation, it can be decomposed into a product of three matrices as follows:

$$X = \tilde{Z}N\tilde{Z}^t \quad (34)$$

That we can also write:

$$\forall i, i'; \quad x_{ii'} = \sum_{k=1}^{K} n_k \tilde{z}_{ik} \tilde{z}_{i'k} \quad (35)$$

with $||X||_F^2 = ||N||_F^2$, $N = diag(n_1,...,n_l,...,n_K)$ and $\{\tilde{Z}_k, k = 1,...K\}$ are the eigenvectors associated with eigenvalues $\{n_1,...,n_l,...,n_K\}$ of $X$ where $\tilde{Z}_k = \frac{1}{\sqrt{n_k}}Z_k$, $\tilde{z}_{ik} = \frac{1}{\sqrt{n_k}}z_{ik}$, $Z_k$ is the $k^{th}$ vector of the partition matrix $Z$.

We can check that

$$\sum_{k=1}^{K} n_k \tilde{z}_{ik} \tilde{z}_{i'k} = \sum_{k=1}^{K} n_k (\frac{1}{\sqrt{n_k}}z_{ik})(\frac{1}{\sqrt{n_k}}z_{i'k}) = x_{ii'}; \quad \forall i, i' \quad (36)$$

Using matrix notations, we write:

$$ZN^{-\frac{1}{2}}NN^{-\frac{1}{2}}Z^t = \tilde{Z}N\tilde{Z}^t = ZI_KZ^t = X \quad (37)$$

where $I_K$ is the identity matrix of order $K$. The matrix $\tilde{Z}$ satisfies the orthogonality property ($\tilde{Z}^t \tilde{Z} = I_K$), using the decomposition of $X$ we can rewrite the RA program as follows:
\[ \max_{Z^t Z = I_{K}} \text{trace}[NZ^t (\hat{S} - \bar{S}) \hat{Z}] \tag{38} \]

The above program is the spectral equivalent of the \( \max_{X} R_{AR}(\hat{S}, X) \) problem, this spectral formulation requires a priori knowledge of the matrix \( N \), ie, the number of classes and size of each class of the partition. We now give the spectral equivalent of the normalized Condorcet maximization problem \( \max_{X} R_{AR}(\hat{S}, X) \):

\[ \max_{Z^t Z = I_{K}} \text{Tr}[\hat{Z}^t (\hat{S} - \bar{S}) \hat{Z}] \tag{39} \]

Since the equivalences presented in (27), we consider later in this paper the problem of maximizing the between clusters inertia \( I_{B} (X) \), i.e.

\[ \max_{Z^t Z = I_{K}} \text{Tr}[\hat{Z}^t \bar{S} \hat{Z}] \tag{40} \]

5. The proposed procedure: The SpectCat Algorithm

The algorithm described in the previous section has two obvious shortcomings.

First, the sensitivity to the order of data objects presentation and secondly the ignorance of the clusters cardinalities in the weighted Condorcet criterion, which affects the outcome of the algorithm. Second, the optimization algorithm is run in a discrete manner, resulting in high computational costs. These two deficiencies are corrected by the spectral relaxation of the relational method. Following the spectral proposed procedure based upon the algorithm proposed by Ng, Jordan and Weiss 20. Assume that we seek up to a maximum of \( K \) clusters and that we have a similarity matrix \( \hat{S} \in \mathbb{R}_{+}^{N \times N} \). Our proposed algorithm below begin by computing the top \( K - 1 \) eigenvectors ignoring the trivial all ones vector. This algorithm is similar in spirit to the one developed by Ng Jordan and Weiss. The algorithm embed the input data into Euclidean space by eigen decomposing a suitable similarity matrix and then cluster the embedding using a geometric clustering algorithm.

Algorithm2: SpectCat Algorithm

Given a set of data object that we want cluster into \( K \) clusters

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

The SpectCat algorithm contains two majors components: Computing the eigen-
vectors and executing k-means to partition the data objects. We run the k-means $t$ times, where in each case the dimensionality $d = K - 1$. Standard k-means with Euclidean distance metric has time complexity $O(Ndkt)$, where $N$ is the number of data points, $d$ is the dimensionality of each point ($d = K - 1$), and $t$ is the number of iteration required for k-means to converge. In addition, for SpectCat algorithm there is the additional complexity for computing the matrix eigenvectors $U$; for computing the largest eigenvectors using the power method or Lanczos method $^7$, the running time is $O(N^2M)$ per iteration where $M$ is the number of data matrix vector and $N$ is the number of objects. Similar to other spectral graph clustering method, the time complexity of SpectCat can be significantly reduced if the affinity matrix $\tilde{S}$ is sparse.

6. 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.

6.1. Performance Measures

Validating clustering results is a non-trivial task. In the presence of true labels, as in the case of the data set we used, the clustering purity for measuring the clustering results. It was computed as follows:

Cluster purity: One of the ways of measuring the quality of a clustering solution is the cluster purity. Let there be $L$ clusters of the dataset $I$ and size of cluster $\mathcal{C}_l$ be $|\mathcal{C}_l|$. The purity of this cluster is given by $\text{purity}(\mathcal{C}_l) = \frac{1}{|\mathcal{C}_l|} \max_k(|\mathcal{C}_l\text{\_cluster} = k|)$ where $|\mathcal{C}_l\text{\_cluster} = k|$ denote the number of items for the cluster $k$ assigned to cluster $l$. The overall purity of a clustering solution could be expressed as a weighted sum of individual cluster purities

$$\text{purity} = \sum_{l=1}^{L} \frac{|\mathcal{C}_l|}{|I|}\text{purity}(\mathcal{C}_l)$$  \hspace{1cm} (41)

In general, if the values of purity are larger, the clustering solution is better.

6.2. The datasets for validation

We conduct comparisons on several public domain data downloaded from UCI machine learning repository $^4$. Here we select seven categorical data, including the soybean small database, mushrooms database, the congressional voting records database), the zoo database, the Hayes-Roth database, the Balance Scale database, the Car Evaluation and the Audiology database. The description of the used data sets is given in Table 2:
Table 1. Description of the data set

<table>
<thead>
<tr>
<th>Data set</th>
<th># of Objects</th>
<th># of Attributes</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean small</td>
<td>47</td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>Mushroom</td>
<td>8124</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Congressional votes</td>
<td>435</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Zoo</td>
<td>101</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>Hayes-roth</td>
<td>132</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Balance Scale</td>
<td>625</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Car evaluation</td>
<td>1728</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Audiology</td>
<td>200</td>
<td>69</td>
<td>24</td>
</tr>
</tbody>
</table>

6.3. Results analysis

We studied the clusterings found by four algorithms, our SpectCat algorithm, standard k-modes algorithm introduced in \(^1\), K-representative algorithm proposed in \(^2\), weighted k-modes algorithm \(^22\). Until now, there is no well-recognized standard methodology for categorical data clustering experiments. However, we observed that most clustering algorithms require the number of clusters as an input parameter, so in our experiments, we cluster each data set (soybean small database, the zoo database, the Hayes-Roth database and the Balance Scale database) into different number of clusters, varying from the real number of classes for each data to 10. For each fixed number of clusters, the clustering purity of different algorithms were compared.

We have compared the performance of the proposed algorithm with others categorical data algorithms. From the Table 2 and the Figure 3, it is clear that the performance of the proposed method SpectCat is the best or similar to the other approaches for all data sets. This means that the introduced proposed approach improves the purity clustering.

For Scale data set, the proposed method is efficient for all the most values of K, the proposed method produces the less good result than K-representative algorithm in cases when K= 7. For the remaining cases the proposed method produces the clusters with high purity as shown in figure 4 (left).

For Zoo data set, the proposed method gives comparable or less good results compared to the others algorithms for k=8 and k=10. Results are tabulated in figure 4 right. The proposed method is efficient in all the cases for the Hayes-Roth data set. All the three algorithms generate clusters with the less purity for all K =1,..., 9. The results are shown in figure 5(left).

For Soybean Small data set, the proposed algorithm excels in all the cases when compared to others methods, the proposed method generates clusters with equal purity with wk-modes for K=5,6,7, 8 and K=10, and generates better clusters when K = 9 and K = 7. Results are presented figure 5 (right). The proposed and
Table 2. Purity measure (%) for K-modes, K-representatives, weighted k-modes and SpectCat

<table>
<thead>
<tr>
<th>Data set</th>
<th>K-Modes</th>
<th>K-Representatives</th>
<th>WK-Modes</th>
<th>SpectCat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean small</td>
<td>66</td>
<td>89</td>
<td>89</td>
<td>100</td>
</tr>
<tr>
<td>Mushroom</td>
<td>59</td>
<td>61</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>Congressional votes</td>
<td>62</td>
<td>87</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>Zoo</td>
<td>88</td>
<td>89</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Hayes-roth</td>
<td>41</td>
<td>42</td>
<td>42</td>
<td>54</td>
</tr>
<tr>
<td>Balance Scale</td>
<td>50</td>
<td>52</td>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>Car evaluation</td>
<td>70</td>
<td>70</td>
<td>71</td>
<td>70</td>
</tr>
<tr>
<td>Audiology</td>
<td>62</td>
<td>61</td>
<td>62</td>
<td>61</td>
</tr>
</tbody>
</table>

Fig. 3. Purity measure

K-representative generates clusters with the same purity, except when K = 10.

6.4. SpectCat with different number of eigenvectors and different number of clusters

6.4.1. k-1 eigenvectors versus the good number of clusters for each data set

In general, for the four considered data sets, as the number of eigenvectors (k) used in SpectCat increases, the clustering purity increases. Figure 5 show the results of execution of SpectCat using different number of eigenvectors. From these plots, it
can be seen that the clustering purity increases as more eigenvectors are used for the all data set and decrease once the number of eigenvectors becomes greater than the real number of clusters for each data set.
6.4.2. \textit{k-1 eigenvectors versus k clusters}

In this situation, as the number of eigenvectors increases at the same time with the number of the desired number of clusters. The Figure 7 show the results of execution of SpectCat using different number of eigenvectors $k$ and $k - 1$ number of clusters for the four considered data sets. From these plots, it can be seen that the clustering purity increases as more eigenvectors are used for the all data sets. As for the Soybean data set, clustering purity remains at 100% for $k$ greater than or equal to 2.

7. Conclusions

In this paper, we have studied the spectral interpretation of the RA approach for categorical data clustering. An efficient, combined procedure for optimization is presented, which combine the spectral algorithm and the RA data presentation. 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.
8. Appendix

The similarity between two data objects is given by

\[ \tilde{s}_{ii'} = \sum_{j=1}^{P} \frac{K_{ij}K_{i'j}}{\sqrt{K_iK_j}} = \frac{1}{M} \sum_{j=1}^{P} \frac{K_{ij}K_{i'j}}{\sqrt{K_j}} \]  (42)

Where \( K_i = M \) \( \forall i \), (in a disjunctive matrix, the marginal sum of each row is equal to \( M \), the number of attributes). The dissimilarity or (distance) between two data objects is then given

\[ \tilde{s}_{ii'} = \frac{1}{2} ||\tilde{K}_i - \tilde{K}_{i'}||^2 \]  (43)

\[ = \frac{1}{2} \left( \sum_{j=1}^{P} \frac{K_{ij}}{K_j} + \sum_{j=1}^{P} \frac{K_{i'j}}{K_j} - 2 \sum_{j=1}^{P} \frac{K_{ij}K_{i'j}}{K_j} \right) \]  (44)

\[ = \frac{1}{2} (\tilde{s}_{ii} + \tilde{s}_{i'i'}) - \tilde{s}_{ii'} \]  (45)

\[ = \tilde{s}_{ii'}^{\text{max}} - \tilde{s}_{ii'} \]  (46)

Another possible writing of the total inertia is to compute the sum of squared
distances between each data objects pairwise. the relationship between the two writing is given by the following equation:

$$\sum_{i=1}^{N} ||\tilde{K}_i - g||^2 = \frac{1}{2N} \sum_{i=1}^{N} \sum_{i'=1}^{N} ||\tilde{K}_i - \tilde{K}_{i'}||^2$$

(47)

where \(g\) is the data mean vector; \(g = \frac{1}{N} \sum_{i=1}^{N} \tilde{K}_i\)

\[I_T(X) = \frac{1}{2N} \sum_i \sum_{i'} ||\tilde{K}_i - \tilde{K}_{i'}||^2\]

(48)

\[= \frac{1}{2NM} \sum_{i=1}^{N} \sum_{i'=1}^{N} \sum_{j=1}^{P} \left( \frac{k_{ij}}{\sqrt{K_j}} - \frac{K_{ij}}{\sqrt{K_j}} \right)^2\]

(49)

\[= \frac{1}{2NM} \sum_{i=1}^{N} \sum_{i'=1}^{N} \sum_{j=1}^{P} \left( \frac{K_{ij} + K_{ij'}}{K_j} - \frac{2K_{ij}K_{ij'}}{K_j} \right)\]

(50)

\[= \frac{1}{2NM} \left( NP + NP - 2NM \right)\]

(51)

\[= \frac{P}{M} - 1\]

(52)

\[I_W(X) = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in C_k} \sum_{i' \in C_k} ||\tilde{K}_i - \tilde{K}_{i'}||^2\]

(53)

\[= \sum_i \sum_{i'} \frac{1}{2} ||\tilde{K}_i - \tilde{K}_{i'}||^2 \sum_{k=1}^{K} \frac{x_{ik}}{x_i}\]

(54)

\[= \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \sum_{k=1}^{K} \frac{x_{ik}}{x_i}\]

(55)

where \(\sum_{k=1}^{K} \frac{x_{ik}}{x_i} = \frac{x_{ik'}}{x_i},\ x_i^k = n_k\ and\ x_i^{k'} = 1\ if\ i, i' \in C_k\ and\ 0\ otherwise.\n
The explicit formula of the between cluster inertia can be obtained using the Huygens formula

\[I_B(X) = I_T(X) - I_W(X)\]

References

[http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine", CA: University of Cal-
ifornia, School of Information and Computer Science.


