CBK-Modes: A Correlation-based Algorithm for Categorical Data Clustering

Joel Luis Carbonera and Mara Abel
Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

Keywords: Clustering, Subspace Clustering, Categorical Data, Attribute Weighting, Data Mining.

Abstract: Categorical data sets are often high-dimensional. For handling the high-dimensionality in the clustering process, some works take advantage of the fact that clusters usually occur in a subspace. In soft subspace clustering approaches, different weights are assigned to each attribute in each cluster, for measuring their respective contributions to the formation of each cluster. In this paper, we adopt an approach that uses the correlation among categorical attributes for measuring their relevancies in clustering tasks. We use this approach for developing the CBK-Modes (Correlation-based K-modes); a soft subspace clustering algorithm that extends the basic k-modes by using the correlation-based approach for measuring the relevance of the attributes. We conducted experiments on five real-world datasets, comparing the performance of our algorithm with five state-of-the-art algorithms, using three well-known evaluation metrics: accuracy, f-measure and adjusted Rand index. The results show that the performance of CBK-Modes outperforms the algorithms that were considered in the evaluation, regarding the considered metrics.

1 INTRODUCTION

Clustering is a widely used technique in which *a set of data points is partitioned into a set of groups of objects that are as similar as possible* (Aggarwal, 2014). In this context, according to (Andreopoulos, 2014), *categorical data clustering* refers to the clustering of objects that are defined over *categorical attributes* (or discrete-valued, symbolic attributes).

Traditionally, techniques of clustering are developed for handling objects that are described by *numerical attributes*. In such cases, the similarity (or dissimilarity) of objects and the quality of a cluster can be determined using well-studied measures that are derived from the geometric properties of the data (Andreopoulos, 2014). In the case of categorical data clustering, the categorical attributes are not inherently comparable. Another challenge regarding categorical data clustering arises from the fact that categorical data sets are often *high-dimensional* (Bai et al., 2011). In high-dimensional data, as the number of dimensions in a dataset increases, distance measures become increasingly meaningless, since thet the distance between a given object x and its nearest object will be close to the dissimilarity between x and its farthest object. Due to this problem, which is one of the aspects of the *curse of dimensionality* (Parsons et al., 2004; Zimek, 2014), discovering meaningful separable clusters becomes a very challenging task.

For dealing with the curse of dimensionality, the so-called *subspace clustering approaches* (Gan and Wu, 2004; Zaki et al., 2007; Cesario et al., 2007; Kriegel et al., 2012; Carbonera and Abel, 2014b) take advantage of the fact that clusters usually occur in a *subspace* defined by a subset of the original set of attributes (Zimek, 2014). *Soft* subspace clustering (Jing et al., 2007; Bai et al., 2011) is a special case of subspace clustering, in which different *weights* are assigned to each attribute in each cluster, for measuring their respective contributions for the formation of each cluster (Zimek, 2014). That is, in these techniques, *different weight vectors are assigned to different clusters*. Due to this, the strategy for *attribute weighting* plays a crucial role in soft subspace clustering approaches.

In (Carbonera and Abel, 2014a), the authors explore a strategy for measuring the contribution of each attribute considering its correlations with other attributes. This approach is inspired by cognitive studies that state that humans *spontaneously* learn categories by exploring the *correlations* among the attributes of the perceived objects. However, this approach was not evaluated in practical clustering algorithms yet. In this paper, we address this issue,

Carbonera J. and Abel M..
CBK-Modes: A Correlation-based Algorithm for Categorical Data Clustering.
DOI: 10.5220/0005367106030608
In *Proceedings of the 17th International Conference on Enterprise Information Systems* (ICEIS-2015), pages 603-608
ISBN: 978-989-758-096-3

by developing a novel algorithm called *CBK-modes*[1] (Correlation-based K-modes), which extends the basic k-modes algorithm by adopting the approach proposed by (Carbonera and Abel, 2014a) for attribute weighting. The performance of this algorithm was compared against the performances of five algorithms available in the literature, considering five real data sets. The results show that CBK-Modes has performances that are comparable to the performances of other state-of-the-art algorithms that were considered in the evaluation. The results also show that, in general, CBK-modes has performances that are better than the performances of other algorithms. The experimental analysis also suggest that the correlation-based approach for attribute weighting is a sufficient condition for improving the performance of clustering algorithms.

In Section 2 we discuss some related works. Section 3 presents the formal notation that will be used throughout the paper. Section 4 presents the correlation-based attribute weighting proposed by (Carbonera and Abel, 2014a). Section 5 presents the CBK-modes algorithm. Experimental results are presented in Section 6. Finally, section 7 presents our concluding remarks.

2 RELATED WORKS

In the last few years, several algorithms have been proposed for dealing with categorical data clustering. In this work, our focus of interest is on the so-called *soft subspace clustering* approaches o categorical data clustering, such as (Chan et al., 2004; Bai et al., 2011; Cao et al., 2013; Jing et al., 2007; Carbonera and Abel, 2014b).

According to (Jing et al., 2007), in *subspace clustering*, objects are grouped into clusters considering subsets of the original set of *dimensions* (or *attributes*) of the data set. *Soft subspace clustering* can be viewed as a *specific case* of subspace clustering. Approaches of this type estimate the *contribution of each attribute* for each specific cluster. The contribution of a dimension is measured by a *weight* that is estimated and assigned to the dimension during the clustering process. Thus, the resulting clustering is performed in the *full-dimensional*, though *skewed* data space.

The approach proposed in (Chan et al., 2004), for example, computes each weight according to the average distance of data objects from the mode of a cluster. Thus, a larger weight is assigned to an attribute that has a smaller sum of the within cluster distances and a smaller weight is assigned to an attribute that has a larger sum of the within cluster distances. The approach proposed in (Bai et al., 2011) assumes that the contribution of a given attribute for a given cluster is proportional to the frequency of the categorical value of the mode of the cluster for that attribute. In (Cao et al., 2013), the authors apply the notion of *complement entropy* for developing an approach for attribute weighting. The complement entropy reflects the uncertainty of an object set with respect to an attribute (or attribute set), in a way that the bigger the complement entropy value is, the higher the uncertainty is. In (Jing et al., 2007), the authors propose an approach for attribute weighting based on the notion of *entropy*, which is a measure of the uncertainty of a given random variable. This approach minimizes the within cluster dispersion and maximizes the negative weight entropy to stimulate more dimensions to contribute to the identification of a cluster. In (Carbonera and Abel, 2014b), the authors propose to measure the relevance of categorical attributes in the clustering process through the *entropy-based relevance index* (ERI). The ERI of some attribute a_h (given by $ERI(a_h)$) is inversely proportional to the average of the uncertainty that is projected to the attribute a_h by the modes of all attributes in the dataset.

3 NOTATION

In this section, we will introduce the notation, adopted from (Carbonera and Abel, 2014a), which will be used throughout the paper:

- $U = \{x_1, x_2, ..., x_n\}$ is a non-empty set of n data objects, called a universe.
- $A = \{a_1, a_2, ..., a_m\}$ is a non-empty set of m categorical attributes.
- $dom(a_i) = \{a_i^{(1)}, a_i^{(2)}, ..., a_i^{(l_i)}\}$ describes the domain of values of the attribute $a_i \in A$, where l_i, is the number of categorical values that a_i can assume in U. Notice that $dom(a_i)$ is finite and unordered, e.g., for any $1 \leq p \leq q \leq l_i$, either $a_i^{(p)} = a_i^{(q)}$ or $a_i^{(p)} \neq a_i^{(q)}$.
- V is the union of attribute domains, i.e., $V = \bigcup_{j=1}^{m} dom(a_j)$.
- $C = \{c_1, c_2, ..., c_k\}$ is a set of k disjoint partitions of U, such that $U = \bigcup_{i=1}^{k} c_i$.
- Each $x_i \in U$ is a $m-tuple$, such that $x_i = (x_{i1}, x_{i2}, ..., x_{im})$, where $x_{iq} \in dom(a_q)$ for $1 \leq i \leq n$ and $1 \leq q \leq m$.

[1]The source of the CBK-modes algorithm can be found in http://www.inf.ufrgs.br/∼jlcarbonera/?page_id=87.

4 CORRELATION-BASED APPROACH FOR CATEGORICAL ATTRIBUTE WEIGHTING

In (Carbonera and Abel, 2014a), the authors developed an approach for attribute weighting considering the correlations among the categorical attributes as a measure of their relevance. This proposal was inspired by studies in the Cognitive Sciences (Sloutsky, 2010) that have pointed out that humans *spontaneously* learn categories exploring the *correlations* among the attributes of the perceived objects. The approach proposed by the authors does not require previous supervised labeling of the data set and does not require the setting of any parameter. In the following, we will present this approach.

Since the frequency of the categorical values in the dataset is important for the approach, it is considered the function $freq_i \colon V \rightarrow \mathbb{N}$, which maps a given categorical value $a_h^{(l)}$ to the number of objects in the partition $c_i \in C$ that are characterized by $a_h^{(l)}$ in the corresponding attribute $a_h \in A$. That is

$$freq_i(a_h^{(l)}) = |\{x_q | x_q \in c_i \text{ and } x_{q,h} = a_h^{(l)}\}| \quad (1)$$

where, $\forall a_h^{(l)} \in V; \forall c_i \in C; 0 \leq freq_i(a_h^{(l)}) \leq |c_i|$; let $|c_i|$ be the number of data objects in c_i. Notice that in $freq_i$, the index i means that we are considering all objects in the partition $c_i \in C$. We will adopt the same notation to the other functions.

Also, the function $\psi_i \colon V \times V \rightarrow \mathbb{N}$ maps two given categorical values $a_h^{(l)} \in dom(a_h)$ and $a_j^{(p)} \in dom(a_j)$ to the number of objects, in $c_i \in C$, in which these values co-occur (assigned to the attributes a_h and a_j, respectively). That is:

$$\begin{aligned} \psi_i(a_h^{(l)}, a_j^{(p)}) = |\{x_q | x_q \in c_i \\ \text{and } x_{qh} = a_h^{(l)} \\ \text{and } x_{qj} = a_j^{(p)}\}| \end{aligned} \quad (2)$$

Besides that, the function $\mathcal{M}_i \colon V \times A \rightarrow \mathbb{N}$ maps a given categorical value $a_h^{(l)} \in dom(a_h)$ and a given categorical attribute $a_j \in A$, to the greatest value that $\psi_i(a_h^{(l)}, a_j^{(p)})$ can assume, considering all $a_j^{(p)} \in dom(a_j)$. That is:

$$\mathcal{M}_i(a_h^{(l)}, a_j) = \max_{p \in dom(a_j)} \{\psi_i(a_h^{(l)}, a_j^{(p)})\} \quad (3)$$

Thus, $\mathcal{M}_i(a_h^{(l)}, a_j)$ represents the number of co-occurrences of the value $a_h^{(l)} \in a_h$ and the value $a_j^{\mathcal{M}} \in a_j$ in the partition $c_i \in C$; where $a_j^{\mathcal{M}}$ is the categorical value that has the greatest number of co-occurrences with the value $a_h^{(l)}$.

Finally, it is defined the function $\alpha_i \colon V \times A \rightarrow \mathbb{R}$, in a way that

$$\alpha_i(a_h^{(l)}, a_j) = \frac{\mathcal{M}_i(a_h^{(l)}, a_j)}{freq_i(a_h^{(l)})} \quad (4)$$

where $\forall a_h^{(l)} \in V; \forall a_j \in A; \forall c_i \in C; 0 \leq \alpha_i(a_h^{(l)}, a_j) \leq 1$.

Considering these functions, it is possible to define the *maximum co-occurrence correlation index* (*mcci*) and the *correlational relevance index* (*cri*); the two main notions underlying the approach.

Definition 1. *Maximum co-occurrence correlation index*: The *mcci* is an index that can be measured between two given categorical attributes $a_h \in A$ and $a_j \in A$, considering a given partition $c_i \in C$, through the function $mcci_i \colon A \times A \rightarrow \mathbb{R}$, such that:

$$mcci_i(a_h, a_j) = \frac{\sum_{l=1}^{|dom(a_h)|} \alpha_i(a_h^{(l)}, a_j)}{|dom(a_h)|} \quad (5)$$

where $\forall a_h \in A; \forall a_j \in A; \forall c_i \in C; 0 \leq mcci_i(a_h, a_j) \leq 1$. It is important to notice that $\forall a_h^{(l)} \in V; \forall a_j \in A; \forall c_i \in C; (\mathcal{M}_i(a_h^{(l)}, a_j) = freq_i(a_h^{(l)})) \implies (mcci_i(a_h, a_j) = 1)$; i.e., $mcci_i(a_h, a_j)$ assumes the greatest value possible in this situation. Thus, informally, the *mcci* measured between $a_h \in A$ and $a_j \in A$ is proportional to how much the categorical values $a_j^{(p)} \in a_j$ vary, regarding each categorical value $a_h^{(l)} \in a_h$. Notice also that $mcci_i(a_h, a_j)$ is not necessarily equal to $mcci_i(a_j, a_h)$.

Definition 2. *Correlational relevance index*: The *cri* is an index that can be assigned to a given attribute a_h in a given partition $c_i \in C$ as defined by the function $cri_i \colon A \rightarrow \mathbb{R}$, such that

$$cri_i(a_h) = \frac{\sum_{j=1}^{|A|} mcci_i(a_j, a_h)}{|A|} \quad (6)$$

where $\forall a_j \in A; \forall c_i \in C; 0 \leq cri_i(a_h) \leq 1$.

Thus, the *correlational relevance index* of a given attribute $a_h \in A$, considering a given partition $c_i \in C$ of the data set, is proportional to the average of the *maximum co-occurrence correlation indexes* that are measured between every $a_j \in A$ and a_h. We assume that the *correlational relevance index* of a given attribute can be used as a measure of its relevance, considering a given partition of the data set, for categorical clustering tasks.

In (Carbonera and Abel, 2014a) the authors also proposed an algorithm for computing the *cri* of all attributes a_h, assuming a given partition c_i of the data

set as input. More details and examples regarding this approach can be viewed in (Carbonera and Abel, 2014a).

5 CBK-MODES: A CORRELATION-BASED K-MODES

The CBK-modes extends the basic K-modes algorithm (Huang, 1998) by considering *correlational relevance index* (cri) for measuring the relevance of each attribute in each cluster. Thus, the CBK-modes can be viewed as a soft subspace clustering algorithm. Our algorithm uses the *k-means* paradigm for searching a partition of U into k clusters that minimize the objective function $P(W,Z,V)$, with unknown variables W, Z and V, as follows:

$$\min_{W,Z,V} P(W,Z,V) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li} d(x_i, z_l) \tag{7}$$

subject to

$$\begin{cases} w_{li} \in \{0,1\} & 1 \leq l \leq k, 1 \leq i \leq n \\ \sum_{l=1}^{k} w_{li} = 1, & 1 \leq i \leq n \\ 0 \leq \sum_{i=1}^{n} w_{li} \leq n, & 1 \leq l \leq k \\ v_{lj} \in [0,1], & 1 \leq l \leq k, 1 \leq j \leq m \end{cases} \tag{8}$$

where:

- $W = [w_{li}]$ is a $k \times n$ binary membership matrix, where $w_{li} = 1$ indicates that x_i is allocated to the cluster C_l.
- $Z = [z_{lj}]$ is a $k \times m$ matrix containing k cluster centers.

The dissimilarity function $d(x_i, z_l)$ is defined as follows:

$$d(x_i, z_l) = \sum_{j=1}^{m} \theta_{a_j}(x_i, z_l) \tag{9}$$

where

$$\theta_{a_j}(x_i, z_l) = \begin{cases} 1, & x_{ij} \neq z_{lj} \\ 1 - v_{lj}, & x_{ij} = z_{lj} \end{cases} \tag{10}$$

where

$$v_{lj} = cri_l(a_j) \times cri_U(a_j) \tag{11}$$

Notice that v_{lj} is the result of a *local attribute weight* ($cri_l(a_j)$) multiplied by a *global attribute weight* ($cri_U(a_j)$). In this way, we are considering the contributions of the correlations among attributes in both *local* and *global* levels.

The minimization of the objective function 7 with the constraints in 8 forms a class of constrained nonlinear optimization problems whose solutions are unknown. The usual method towards optimization of 7 is to use partial optimization for Z, W and V. In this method, following (Cao et al., 2013), we first fix Z and V and find necessary conditions on W to minimize $P(W,Z,V)$. Then, we fix W and V and minimize $P(W,Z,V)$ with respect to Z. Finally, we then fix W and Z and minimize $P(W,Z,V)$ with respect to V. The process is repeated until no more improvement in the objective function value can be made. The Algorithm 1 presents the CBK-modes algorithm, which formalizes this process, using the *correlational relevance index* for measuring the relevance of each attribute in each cluster.

Algorithm 1: CK-modes.

Input: A set of categorical data objects U and the number k of clusters.
Output: The data objects in U partitioned in k clusters.
begin
  Initialize the variable *oldmodes* as a $k \times |A|$-ary empty array;
  Randomly choose k distinct objects $x_1, x_2, ..., x_k$ from U and assign $[x_1, x_2, ..., x_k]$ to the $k \times |A|$-ary variable *newmodes*;
  Calculate initial global weights GW_j of $a_j \in A$ using $cri_U(a_j)$;
  Set all initial weights LW_{lj} to GW_j, where $1 \leq l \leq k$, $1 \leq j \leq m$;
  Set all initial weights v_{lj} to $LW_{lj} \times GW_j$, where $1 \leq l \leq k$, $1 \leq j \leq m$;
  while *oldmodes* $\neq$ *newmodes* **do**
    for $i \leftarrow 1$ **to** $|U|$ **do**
      for $l \leftarrow 1$ **to** k **do**
        Calculate the dissimilarity (using the dissimilarity function of equation 9) between the $i-th$ object and the $l-th$ mode and classify the $i-th$ object into the cluster whose mode is closest to it;
    for $l \leftarrow 1$ **to** k **do**
      Find the mode z_l of each cluster and assign to *newmodes*;
      Calculate the weight LW_{lh} of each attribute $a_h \in A$ of the $l-th$ cluster, using $cri_l(a_h)$;
      Set all initial weights vlj to $LW_{lj} \times GW_j$, where $1 \leq l \leq k$, $1 \leq j \leq m$;

6 EXPERIMENTS

The evaluation of our approach was performed by comparing the CBK-modes algorithm with five state-of-the-art algorithms. This comparison was based on three well-known evaluation measures: *accuracy* (or purity) (Huang, 1998; He et al., 2011), *f-measure* (Larsen and Aone, 1999) and *adjusted Rand index* (Bai et al., 2011). Our tests considered six real-world data sets: congressional voting records, mushroom, breast cancer, soybean[2], genetic promoters and splice-junction gene sequences. All the data sets were

[2]This data set combines the large soybean data set and its corresponding test data set

obtained from the UCI Machine Learning Repository[3]. Regarding the data sets, the missing value in each attribute was considered as a special category in our experiments.

Table 1: Comparison of the average accuracy produced by each algorithm in 100 random runs, and the respective standard deviations.

Algorithm	Vote	Mushroom	Breast cancer	Soybean	Promoters
KM	0.86 ±0.02	0.71 ±0.15	0.70 ±0.00	0.63 ±0.03	0.59 ±0.08
NWKM	0.86 ±0.03	0.72 ±0.14	0.70 ±0.01	0.63 ±0.04	0.61 ±0.08
MWKM	0.86 ±0.01	0.72 ±0.14	0.70 ±0.00	0.63 ±0.03	0.61 ±0.08
WKM	0.87 ±0,01	0.73 ±0.13	0.70 ±0.01	0.65 ±0.03	0.62 ±0.08
EBKM	0.87 ±0.00	0.76 ±0.12	0.70 ±0.01	0.66 ±0.03	0.62 ±0.08
CBKM	0.87 ±0.00	0.76 ±0.13	0.71 ±0.01	0.66 ±0.03	0.65 ±0.11
Average	0.87 ±0.01	0.73 ±0.14	0.70 ±0.01	0.64 ±0.03	0.62 ±0.09

We compared the CBK-modes (CBKM) algorithm with five algorithms available in the literature: standard k-modes (KM) (Huang, 1998), NWKM (Bai et al., 2011), MWKM (Bai et al., 2011), WK-modes (WKM) (Cao et al., 2013) and EBK-modes (EBKM) (Carbonera and Abel, 2014b). For the NWKM algorithm, following the recommendations of the authors, the parameter β was set to 2. For the same reason, for the MWKM algorithm, we have used the following parameter settings:$\beta = 2$, $T_v = 1$ and $T_s = 1$.

Table 2: Comparison of the average f-measure produced by each algorithm in 100 random runs, and the respective standard deviations.

Algorithm	Vote	Mushroom	Breast cancer	Soybean	Promoters
KM	0.76 ±0.02	0.64 ±0.13	0.54 ±0.00	0.42 ±0.04	0.53 ±0.05
NWKM	0.78 ±0.03	0.64 ±0.12	0.56 ±0.05	0.42 ±0.05	0.54 ±0.04
MWKM	0.77 ±0.01	0.64 ±0.12	0.54 ±0.02	0.42 ±0.05	0.54 ±0.05
WKM	0.78 ±0,01	0.66 ±0.12	0.55 ±0.04	0.45 ±0.04	0.55 ±0.05
EBKM	0.78 ±0.00	0.67 ±0.11	0.56 ±0.05	0.45 ±0.04	0.55 ±0.05
CBKM	0.79 ±0.01	0.68 ±0.12	0.59 ±0.06	0.47 ±0.04	0.57 ±0.07
Average	0.78 ±0.01	0.66 ±0.12	0.56 ±0.04	0.44 ±0.04	0.55 ±0.05

For each data set, we carried out 100 random runs of each one of the considered algorithms. This was done because all of the algorithms choose their initial cluster centers via random selection methods, and thus the clustering results may vary depending on the initialization. In each run, we computed the performance metrics that were selected: accuracy, f-measure and adjusted Rand index. The Tables 1, 2 and 3 present respectively, the averages (with standard deviation) of accuracy, f-measure and adjusted Rand index. Notice that in these tables, the average performance is presented at the top of each cell and standard deviation is presented at the bottom.

Table 3: Comparison of the average adjusted Rand index (ARI) produced by each algorithm in 100 random runs, and the respective standard deviations.

Algorithm	Vote	Mushroom	Breast cancer	Soybean	Promoters
KM	0.51 ±0.01	0.26 ±0.26	0.01 ±0.02	0.37 ±0.04	0.06 ±0.08
NWKM	0.54 ±0.06	0.26 ±0.25	0.02 ±0.05	0.37 ±0.05	0.07 ±0.08
MWKM	0.52 ±0.01	0.28 ±0.25	0.01 ±0.02	0.37 ±0.05	0.07 ±0.09
WKM	0.54 ±0.02	0.29 ±0.25	0.02 ±0.05	0.41 ±0.05	0.08 ±0.09
EBKM	0.54 ±0.01	0.33 ±0.23	0.03 ±0.05	0.42 ±0.05	0.09 ±0.10
CBKM	0.54 ±0.01	0.33 ±0.25	0.05 ±0.06	0.42 ±0.04	0.13 ±0.13
Average	0.53 ±0.02	0.29 ±0.25	0.02 ±0.04	0.39 ±0.05	0.08 ±0.10

The Tables 1, 2 and 3 show that the CBK-modes algorithm is able to achieve high-quality overall results, considering the selected data sets and measures of performance. Notice that the CBK-modes algorithm have performances that are equivalent to or better than the performances of state-of-the-art algorithms, such as NWKM, MWKM and EBKM. And, since the performance CBK-modes is better than the performance of the basic K-modes, we can conclude that using the *correlational relevance index* as a measure of the relevance of attributes is a sufficient condition for improving the performance of the basic K-modes algorithm.

7 CONCLUSION

In this paper, we propose CBK-modes, an extension of the K-modes algorithm, which uses a correlation-based approach for attribute weighting. Our experiments have shown that the proposed algorithm has a performance comparable to (or even better than) the performance of the state-of-the-art algorithms. The results also suggest that using the *correlational relevance index* as a measure of the relevance of attributes is a sufficient condition for improving the performance of the clustering algorithms. In the next steps, we plan to investigate how the *correlational relevance index* can be used for improving the performance of others algorithms and how this approach can be extended for dealing with mixed data sets (with both categorical and numerical attributes).

[3]http://archive.ics.uci.edu/ml/

ACKNOWLEDGEMENTS

We gratefully thank Brazilian Research Council, CNPq, PRH PB-17 program (supported by Petrobras), for the financial support to this work. In addition, we would like to thank Sandro Fiorini for comments and ideas.

REFERENCES

Aggarwal, C. C. (2014). *Data Clustering: Algorithms and Applications*, chapter An Introduction to Cluster Analysis, pages 1–28. CRC Press.

Andreopoulos, B. (2014). *Data Clustering: Algorithms and Applications*, chapter Clustering Categorical Data, pages 1–28. CRC Press.

Bai, L., Liang, J., Dang, C., and Cao, F. (2011). A novel attribute weighting algorithm for clustering high-dimensional categorical data. *Pattern Recognition*, 44(12):2843–2861.

Cao, F., Liang, J., Li, D., and Zhao, X. (2013). A weighting k-modes algorithm for subspace clustering of categorical data. *Neurocomputing*, 108:23–30.

Carbonera, J. L. and Abel, M. (2014a). Categorical data clustering:a correlation-based approach for unsupervised attribute weighting. In *Proceedings of ICTAI 2014*.

Carbonera, J. L. and Abel, M. (2014b). An entropy-based subspace clustering algorithm for categorical data. In *Proceedings of ICTAI 2014*.

Cesario, E., Manco, G., and Ortale, R. (2007). Top-down parameter-free clustering of high-dimensional categorical data. *Knowledge and Data Engineering, IEEE Transactions on*, 19(12):1607–1624.

Chan, E. Y., Ching, W. K., Ng, M. K., and Huang, J. Z. (2004). An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern recognition*, 37(5):943–952.

Gan, G. and Wu, J. (2004). Subspace clustering for high dimensional categorical data. *ACM SIGKDD Explorations Newsletter*, 6(2):87–94.

He, Z., Xu, X., and Deng, S. (2011). Attribute value weighting in k-modes clustering. *Expert Systems with Applications*, 38(12):15365–15369.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304.

Jing, L., Ng, M. K., and Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *Knowledge and Data Engineering, IEEE Transactions on*, 19(8):1026–1041.

Kriegel, H.-P., Kröger, P., and Zimek, A. (2012). Subspace clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(4):351–364.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–22. ACM.

Parsons, L., Haque, E., and Liu, H. (2004). Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive science*, 34(7):1244–1286.

Zaki, M. J., Peters, M., Assent, I., and Seidl, T. (2007). Clicks: An effective algorithm for mining subspace clusters in categorical datasets. *Data & Knowledge Engineering*, 60(1):51–70.

Zimek, A. (2014). *Data Clustering: Algorithms and Applications*, chapter Clustering High-Dimensional Data, pages 201–229. CRC Press.