2014 IEEE 26th International Conference on Tools with Artificial Intelligence

# An entropy-based subspace clustering algorithm for categorical data

Joel Luis Carbonera

*Institute of Informatics*
*Universidade Federal do Rio Grande do Sul – UFRGS*
*Porto Alegre, Brazil*
*jlcarbonera@inf.ufrgs.br*

Mara Abel

*Institute of Informatics*
*Universidade Federal do Rio Grande do Sul – UFRGS*
*Porto Alegre, Brazil*
*marabel@inf.ufrgs.br*

***Abstract*—The interest in attribute weighting for soft subspace clustering have been increasing in the last years. However, most of the proposed approaches are designed for dealing only with numeric data. In this paper, our focus is on soft subspace clustering for categorical data. In soft subspace clustering, the attribute weighting approach plays a crucial role. Due to this, we propose an entropy-based approach for measuring the relevance of each categorical attribute in each cluster. Besides that, we propose the EBK-modes (entropy-based k-modes); an extension of the basic k-modes that uses our approach for attribute weighting. We performed experiments on five real-world datasets, comparing the performance of our algorithms with four state-of-the-art algorithms, using three well-known evaluation metrics: accuracy, f-measure and adjusted Rand index. According to the experiments, the EBK-modes outperforms the algorithms that were considered in the evaluation, regarding the considered metrics.**

***Keywords*-clustering; subspace clustering; categorical data; attribute weighting; data mining; entropy;**

## I. Introduction

Clustering is a widely used technique in which *objects are partitioned into groups, in such a way that objects in the same group (or cluster) are more similar among themselves than to those in other clusters* [1]. Most of the clustering algorithms in the literature were developed for handling data sets where objects are defined over *numerical attributes*. In such cases, the similarity (or dissimilarity) of objects can be determined using well-studied measures that are derived from the geometric properties of the data [2]. However, there are many data sets where the objects are defined over *categorical attributes*, which are neither numerical nor inherently comparable in any way. Thus, *categorical data clustering* refers to the clustering of objects that are defined over *categorical attributes* (or discrete-valued, symbolic attributes) [2], [3].

One of the challenges regarding categorical data clustering arises from the fact that categorical data sets are often *high-dimensional* [4], i.e., records in such data sets are described according to a large number of attributes. In high-dimensional data, the dissimilarity between a given object $x$ and its nearest object will be close to the dissimilarity between $x$ and its farthest object. Due to this loss of the dissimilarity discrimination in high dimensions, discovering meaningful separable clusters becomes a very challenging task.

For handling the high-dimensionality, some works take advantage of the fact that clusters usually occur in a *subspace* defined by a subset of the initially selected attributes [5], [6], [7], [8]. In this work, we are interested in the so-called *soft subspace clustering approaches* [9], [4]. In these approaches, different *weights* are assigned to each attribute in each cluster, for measuring their respective contributions to the formation of each clusters. In this way, soft subspace clustering can be considered as an extension of the conventional attribute weighting clustering [10] that employs a common *global* (and usually fixed) weight vector for the *whole* data set in the clustering procedure. However, it is different, since *different weight vectors are assigned to different clusters*. In this approaches, the strategy for *attribute weighting* plays a crucial role.

Most of the recent results in soft subspace clustering for categorical data [4], [11] propose modifications of the k-modes algorithm [12], [4], [13]. In general, in these approaches the contribution of each attribute is measured considering only the frequency of the mode category or the average distance of data objects from the mode of a cluster. In this paper, adopting a different approach, we explore a strategy for measuring the contribution of each attribute considering the notion of *entropy*, which measures the uncertainty of a given random variable. As a consequence, we propose the EBK-modes (entropy-based k-modes)[1]; an extension of the basic k-modes algorithm that uses the notion of entropy for measuring the relevance of each attribute in each cluster. Basically, our algorithm assumes that a given value $v$ of the attribute $a$ of the mode of a cluster $c_i$ determines a subset $c'_i$ of $c_i$, i.e., the set of all objects in $c_i$ that have the categorical value $v$ for the attribute $a$. Then, we can determine the uncertainty of another attribute $b$, regarding $v$, calculating the entropy of $c'_i$, considering the attribute $b$. In this way, we assume that the relevance of a given attribute $b$ is inversely proportional to the average of the entropies induced by the values of each attribute of the

[1]The source codes of our algorithms are available in http://www.inf.ufrgs.br/bdi/wp-content/uploads/EBK-Modes.zip

1082-3409/14 $31.00 © 2014 IEEE
DOI 10.1109/ICTAI.2014.48

mode of $c_i$ in $b$.

In this paper, we also compare the performance of EBK-modes against the performance of four other algorithms available in the literature, using five real data sets. According to the experiments, EBK-modes outperforms the considered algorithms.

In Section II we discuss some related works. Before presenting our approach, in Section III we introduce the notation that will be used throughout the paper. Section IV presents our entropy-based attribute weighting approach for categorical attributes. Section V presents the EBK-modes. Experimental results are presented in Section VI. Finally, section VII presents our concluding remarks.

## II. Related works

In *subspace clustering*, objects are grouped into clusters according to subsets of *dimensions* (or *attributes*) of a data set [9]. These approaches involve two mains tasks, identification of the subsets of dimensions where clusters can be found and discovery of the clusters from different subsets of dimensions. According to the ways with which the subsets of dimensions are identified, we can divide subspace clustering methods into two categories: *hard subspace clustering* and *soft subspace clustering*. The approaches of hard subspace clustering determine the *exact subsets of attributes* where clusters are discovered. On the other hand, approaches of soft subspace clustering determine the subsets of dimensions according to the *contributions of the attributes* in discovering the corresponding clusters. The contribution of a dimension is measured by a *weight* that is assigned to the dimension in the clustering process. The algorithm proposed in this paper can be viewed as a soft subspace clustering approach.

In [12], for example, it is proposed an approach in which each weight is computed according to the average distance of data objects from the mode of a cluster. That is, it is assigned a larger weight to an attribute that has a smaller sum of the within cluster distances and a smaller weight to an attribute that has a larger sum of the within cluster distances. An analysis carried out by [4] have shown that this approach is sensitive to the setting of the parameter $\beta$. In [4], it is assumed that the weight of a given attribute for a given cluster is a function of the frequency of the categorical value of the mode of the cluster for that attribute. This approach requires the setting of three parameters ($\beta$,$T_v$ and $T_s$) for determining the attribute weights. In [11], the authors use the notion of *complement entropy* for weighting the attributes. 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 [9] the authors noticed that the decrease of the entropy in a cluster implies the increase of certainty of a subset of dimensions with larger weights in determination of the cluster. According to this, their approach simultaneously minimize the within cluster dispersion and maximize the negative weight entropy to stimulate more dimensions to contribute to the identification of a cluster.

In our approach, as in [9], we also use the notion of entropy for measuring the relevance of each attribute. However, here we assume that the relevance of a given attribute, for a given cluster, is inversely proportional to the average of the entropy that is induced by each attribute value of the mode of the cluster.

## III. Notations

In this section, we introduce the following notation that 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 $1 \leq q \leq m$.

## IV. An entropy-based approach for categorical attribute weighting

In *information theory*, *entropy* is a measure of the *uncertainty* in a random variable [14]. The larger the entropy of a given random variable, the larger is the uncertainty associated to it. When taken from a finite sample, the entropy $H$ of a given random variable $X$ can be written as

$$H(X) = -\sum_{i} P(x_i) \log P(x_i) \quad (1)$$

where $P(x_i)$ is the probability mass function.

In this work, we use the notion of entropy for measuring the relevance of a given attribute for a given partition $c_i \in C$. In order to illustrate the main notions underlying our approach, let us consider the categorical data set represented in the Table I, where:

- $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$.
- $A = \{a_1, a_2, a_3, a_4, a_5\}$.
- $dom(a_1) = \{a, b, c, d, e, f, g, h, i, j\}$, $dom(a_2) = \{k, l, m\}$, $dom(a_3) = \{n, o, p\}$, $dom(a_4) = \{q, r\}$ and $dom(a_5) = \{s, t, u\}$.

Let us suppose, in our example, that the universe $U$ represented in Tale I is partitioned in three disjoint partitions $c_1 = \{x_1, x_2, x_3, x_4\}$, $c_2 = \{x_5, x_6, x_7\}$, $c_3 = \{x_8, x_9, x_{10}\}$, with their respective centers, as follows:

| Data objects | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| $x_1$ | $a$ | $k$ | $n$ | $q$ | $s$ |
| $x_2$ | $b$ | $k$ | $n$ | $r$ | $s$ |
| $x_3$ | $c$ | $k$ | $n$ | $q$ | $s$ |
| $x_4$ | $d$ | $k$ | $o$ | $r$ | $t$ |
| $x_5$ | $e$ | $l$ | $o$ | $q$ | $t$ |
| $x_6$ | $f$ | $l$ | $o$ | $r$ | $t$ |
| $x_7$ | $g$ | $l$ | $o$ | $q$ | $t$ |
| $x_8$ | $h$ | $m$ | $o$ | $r$ | $u$ |
| $x_9$ | $i$ | $m$ | $p$ | $q$ | $u$ |
| $x_{10}$ | $j$ | $m$ | $p$ | $r$ | $u$ |

Table I
REPRESENTATIONS OF AN EXAMPLE OF CATEGORICAL DATA SET.

$c_1$: $(d, k, n, r, s)$.
$c_2$: $(g, l, o, q, t)$.
$c_3$: $(j, m, p, r, u)$.

Firstly, let us consider the function $\phi_i : V \rightarrow 2^{c_i}$ that maps a given categorical value $a_h^{(l)} \in dom(a_h)$ to a set $s \subseteq c_i$, which contains every object in the partition $c_i \in C$ that has the value $a_h^{(l)}$ for the attribute $a_h$. Notice that $2^{c_i}$ represents the *powerset* of $c_i$, i.e., the set of all subsets of $c_i$, including the empty set and $c_i$ itself. Thus:

$$\begin{aligned}\phi_i(a_h^{(l)}) = &\{x_q | x_q \in c_i \\ &\textit{and } x_{qh} = a_h^{(l)}\}\end{aligned} \tag{2}$$

Considering our example, we have that $\phi_1(s) = \{x_1, x_2, x_3\}$.

Also, let us consider the function $\alpha_i : V \times A \rightarrow 2^V$ that maps a given categorical value $a_h^{(l)}$ and a given attribute $a_j \in A$ to a set $V' \subseteq V$, which represents the set of categorical values of attribute $a_j$ that co-occur with the categorical value $a_h^{(l)}$, in the partition $c_i \in C$. That is:

$$\alpha_i(a_h^{(l)}, a_j) = |\{a_j^{(p)} | \forall x_q \in \phi_i(a_h^{(l)}), x_{qj} = a_j^{(p)}\}| \tag{3}$$

Thus, in our example, we have that $\alpha_1(s, a_1) = \{a, b, c\}$.

Moreover, let us consider $\psi_i : V \times V \rightarrow \mathbb{N}$ as a function that 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:

$$\begin{aligned}\psi_i(a_h^{(l)}, a_j^{(p)}) = &|\{x_q | x_q \in c_i \\ &\textit{and } x_{qh} = a_h^{(l)} \\ &\textit{and } x_{qj} = a_j^{(p)}\}|\end{aligned} \tag{4}$$

In our example, we have that $\psi_1(s, q) = |\{x_1, x_3\}| = 2$.

Furthermore, let us consider the function $\mathcal{E}_i : V \times A \rightarrow \mathbb{R}$ that maps a given categorical value $a_h^{(l)} \in dom(a_h)$ and a categorical attribute $a_j \in A$ to the entropy of the set $\phi_i(a_h^{(l)})$, regarding the attribute $a_j$. That is:

$$\mathcal{E}_i(a_h^{(l)}, a_j) = -\sum_{a_j^p \in \alpha(a_h^{(l)}, a_j)} \frac{\psi_i(a_h^{(l)}, a_j^{(p)})}{|\phi_i(a_h^{(l)})|} \log \frac{\psi_i(a_h^{(l)}, a_j^{(p)})}{|\phi_i(a_h^{(l)})|} \tag{5}$$

According to our example, we have that

$$\mathcal{E}_1(s, a_1) = -\left(\frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3}\right) = 1.10 \tag{6}$$

On the other hand, for example, we also have that

$$\mathcal{E}_1(s, a_3) = -\left(\frac{3}{3}\log\frac{3}{3}\right) = 0 \tag{7}$$

At this point, let us consider $z_i$ as the mode of the partition $c_i \in C$, such that $z_i = \{z_{i1}, ..., z_{im}\}$, where $z_{iq} \in dom(a_q)$ for $1 \leq i \leq k$ and $1 \leq q \leq m$.

Notice that we can use the function $\mathcal{E}$ for measuring the entropy associated to a given categorical attribute $a_h \in A$, in a given partition $c_i \in C$, regarding the value $z_{ij}$ of the attribute $a_j \in A$ of the mode $z_i$. Intuitively, this would measure the uncertainty associated to the attribute $a_h$, given the categorical value $z_{ij}$ of the mode.

Using the function $\mathcal{E}$, we can define a function $e_i : A \rightarrow \mathbb{R}$ that maps a given categorical attribute $a_h \in A$ to the average of the value of $\mathcal{E}_i(z_{ij}, a_h)$, for all $a_j \in A$, considering a partition $c_i \in C$. That is:

$$e_i(a_h) = \frac{\sum\limits_{z_{ij} \in z_i} \mathcal{E}_i(z_{ij}, a_h)}{|A|} \tag{8}$$

Intuitively, $e_i(a_h)$ measures the average of the uncertainty associated to $a_h$, considering all the categorical values of the mode $z_i$. In our example, we have: $\mathcal{E}_1(d, a_5) = 0$, $\mathcal{E}_1(k, a_5) = 0.56$, $\mathcal{E}_1(n, a_5) = 0$, $\mathcal{E}_1(r, a_5) = 0.69$ and $\mathcal{E}_1(s, a_5) = 0$. As a consequence, we have:

$$e_1(a_5) = \frac{0 + 0.56 + 0 + 0.69 + 0}{5} = 0.25 \tag{9}$$

We also have: $e_1(a_1) = 0.86$, $e_1(a_2) = 0$, $e_1(a_3) = 0.25$ and $e_1(a_4) = 0.39$.

At this point, we are able to introduce the *entropy-based relevance index* (*ERI*); the main notion underlying our approach.

**Definition 1.** *Entropy-based relevance index*: The *ERI* measures the relevance of a given categorical attribute $a_h \in A$, for a partition $c_i \in C$. The *ERI* can be measured through the function $ERI_i : A \rightarrow \mathbb{R}$, such that:

$$ERI_i(a_h) = \frac{\exp(-e_i(a_h))}{\sum\limits_{a_j \in A} \exp(-e_i(a_j))} \tag{10}$$

According to our example, we have that: $ERI_1(a_1) = 0.12$, $ERI_1(a_2) = 0.27$, $ERI_1(a_3) = 0.21$, $ERI_1(a_4) = 0.18$ and $ERI_1(a_5) = 0.21$. Notice that $ERI_i(a_h)$ is inversely proportional to $e_i(a_h)$. The smaller the $e_i(a_h)$, the larger the $ERI_i(a_h)$, the more important the corresponding categorical attribute $a_h \in A$.

## V. EBK-MODES: AN ENTROPY-BASED K-MODES

The EBK-modes extends the basic K-modes algorithm [15] by considering our entropy-based attribute weighting approach for measuring the relevance of each attribute in each cluster. Thus, the EBK-modes can be viewed as a soft subspace clustering algorithm. Our algorithm uses the *k-means* paradigm to search a partition of $U$ into $k$ clusters that minimize the objective function $P(W, Z, \Lambda)$ with unknown variables $W$, $Z$ and $\Lambda$ as follows:

$$\min_{W,Z,\Lambda} P(W, Z, \Lambda) \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} d(x_i, z_l) \tag{11}$$

subject to

$$\begin{cases} w_{li} \in \{0, 1\} & 1 \leq l \leq k, 1 \leq i \leq n \\ \sum_{l=1}^{k} w_{li} = 1, & 1 \leq i \leq n \\ 0 \leq \sum_{i=1}^{n} w_{li} \leq n, & 1 \leq l \leq k \\ \lambda_{lj} \in [0, 1], & 1 \leq l \leq k, 1 \leq j \leq m \\ \sum_{j=1}^{m} \lambda_{lj} = 1, & 1 \leq l \leq k \end{cases} \tag{12}$$

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:

$$d(x_i, z_l) = \sum_{j=1}^{m} \theta_{a_j}(x_i, z_l) \tag{13}$$

where

$$\theta_{a_j}(x_i, z_l) = \begin{cases} 1, & x_{ij} \neq z_{lj} \\ 1 - \lambda_{lj}, & x_{ij} = z_{lj} \end{cases} \tag{14}$$

where

$$\lambda_{lj} = ERI_l(a_j) \tag{15}$$

The minimization of the objective function 11 with the constraints in 12 forms a class of constrained nonlinear optimization problems whose solutions are unknown. The usual method towards optimization of 11 is to use partial optimization for $Z$, $W$ and $\Lambda$. In this method, following [11], we first fix $Z$ and $\Lambda$ and find necessary conditions on $W$ to minimize $P(W, Z, \Lambda)$. Then, we fix $W$ and $\Lambda$ and minimize $P(W, Z, \Lambda)$ with respect to $Z$. Finally, we then fix $W$ and $Z$ and minimize $P(W, Z, \Lambda)$ with respect to $\Lambda$. The process is repeated until no more improvement in the objective function value can be made. The Algorithm 1 presents the EBK-modes algorithm, which formalizes this process, using the *entropy-based relevance index* for measuring the relevance of each attribute in each cluster.

**Algorithm 1:** EBK-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 $newmodes$;
  Set all initial weights $\lambda_{lj}$ to $\frac{1}{|A|}$, where $1 \leq l \leq k$, $1 \leq j \leq m$;
  **while** $oldmodes \neq newmodes$ **do**
    **foreach** $i = 1$ *to* $|U|$ **do**
      **foreach** $l = 1$ *to* $k$ **do**
        Calculate the dissimilarity 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;
    **foreach** $l = 1$ *to* $k$ **do**
      Find the mode $z_l$ of each cluster and assign to $newmodes$;
      Calculate the weight of each attribute $a_h \in A$ of the $l-th$ cluster, using $ERI_l(a_h)$;

## VI. EXPERIMENTS

For evaluating our approach, we have compared the EBK-modes algorithm, proposed in Section V, with state-of-the-art algorithms. For this comparison, we considered three well-known evaluation measures: *accuracy* [15], [13], *f-measure* [16] and *adjusted Rand index* [4]. The experiments were conducted considering five real-world data sets: congressional voting records, mushroom, breast cancer, soybean[2] and genetic promoters. All the data sets were obtained from the UCI Machine Learning Repository[3]. Regarding the data sets, missing value in each attribute was considered as a special category in our experiments. In Table II, we present the details of the data sets that were used.

| Dataset | Tuples | Attributes | Classes |
|---|---|---|---|
| Vote | 435 | 17 | 2 |
| Mushroom | 8124 | 23 | 2 |
| Breast cancer | 286 | 10 | 2 |
| Soybean | 683 | 36 | 19 |
| Genetic promoters | 106 | 58 | 2 |

Table II
DETAILS OF THE DATA SETS THAT WERE USED IN THE EVALUATION PROCESS.

In our experiments, the EBK-modes algorithm were compared with four algorithms available in the literature: standard k-modes (KM) [15], NWKM [4], MWKM [4] and WK-modes (WKM) [11]. For the NWKM algorithm, following the recommendations of the authors, the parameter $\beta$ 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 $T_s = 1$.

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

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

For each data set in Table II, 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. Besides that, the parameter $k$ was set to the corresponding number of clusters of each data set. For the considered data sets, the expected number of clusters $k$ is previously known.

The resulting measures of accuracy, f-measure and adjusted Rand index that were obtained during the experiments are summarized, respectively, in Tables III, IV and V. Notice that the tables present both, the best performance (at the top of each cell) and the average performance (at the bottom of each cell), for each algorithm in each data set. Moreover, the best results for each data set are marked in bold typeface.

| Algorithm | Vote | Mushroom | Breast cancer | Soybean | Promoters | Average |
|---|---|---|---|---|---|---|
| KM | 0.86<br>0.86 | 0.89<br>0.71 | 0.73<br>0.70 | 0.70<br>0.63 | 0.80<br>0.59 | 0.80<br>0.70 |
| NWKM | **0.88**<br>0.86 | 0.89<br>0.72 | **0.75**<br>0.70 | 0.73<br>0.63 | 0.77<br>0.61 | 0.80<br>0.70 |
| MWKM | 0.87<br>0.86 | 0.89<br>0.72 | 0.71<br>0.70 | 0.72<br>0.63 | 0.81<br>0.61 | 0.80<br>0.70 |
| WKM | **0.88**<br>**0.87** | 0.89<br>0.73 | 0.74<br>0.70 | 0.74<br>0.65 | 0.78<br>**0.62** | 0.81<br>0.71 |
| EBKM | **0.88**<br>**0.87** | 0.89<br>**0.76** | 0.74<br>0.70 | **0.75**<br>**0.66** | **0.83**<br>**0.62** | **0.82**<br>**0.72** |

Table III
COMPARISON OF THE ACCURACY PRODUCED BY EACH ALGORITHM.

| Algorithm | Vote | Mushroom | Breast cancer | Soybean | Promoters | Average |
|---|---|---|---|---|---|---|
| KM | 0.77<br>0.76 | 0.81<br>0.64 | 0.68<br>0.54 | 0.52<br>0.42 | 0.69<br>0.53 | 0.69<br>0.58 |
| NWKM | **0.80**<br>**0.78** | 0.81<br>0.64 | **0.70**<br>**0.56** | 0.55<br>0.42 | 0.66<br>0.54 | 0.70<br>0.59 |
| MWKM | 0.78<br>0.76 | 0.81<br>0.64 | 0.67<br>0.54 | 0.53<br>0.42 | 0.70<br>0.54 | 0.70<br>0.58 |
| WKM | 0.79<br>**0.78** | 0.81<br>0.66 | 0.69<br>0.55 | 0.55<br>0.45 | 0.68<br>**0.55** | 0.70<br>0.60 |
| EBKM | 0.79<br>**0.78** | 0.81<br>**0.68** | 0.69<br>**0.56** | **0.57**<br>**0.47** | **0.73**<br>**0.55** | **0.72**<br>**0.61** |

Table IV
COMPARISON OF THE F-MEASURE PRODUCED BY EACH ALGORITHM.

The Tables III, IV and V show that EBK-modes achieves high-quality overall results, considering the selected data sets and measures of performance. It has performances that are better than the performance of state-of-the-art algorithms, such as WKM. As shown in the column that represents the average performance, considering all the data sets, the EBK-modes algorithm outperforms all the considered algorithm, in the three considered metrics of performance.

## VII. CONCLUSION

In this paper, we propose a subspace clustering algorithm for categorical data called EBK-modes (entropy-based k-modes). It modifies the basic k-modes by considering the *entropy-based relevance index* (ERI) as a measure of the relevance of each attribute in each cluster. The ERI of a given attribute is inversely proportional to the average of the entropy induced to this attribute for each attribute value of the mode of a cluster. We conducted experiments on five real-world datasets, comparing the performance of our algorithms with four state-of-the-art algorithms, using three well-known evaluation metrics: accuracy, f-measure and adjusted Rand index. The results have shown that the EBK-modes outperforms the state-of-the-art algorithms. In the next steps, we plan to investigate how the *entropy-based relevance index* can be calculate for mixed data sets, i.e., data sets with both categorical and numerical attributes.

| Algorithm | Vote | Mushroom | Breast cancer | Soybean | Promoters | Average |
|---|---|---|---|---|---|---|
| KM | 0.52<br>0.51 | 0.61<br>0.26 | 0.19<br>0.01 | 0.48<br>0.37 | 0.36<br>0.06 | 0.43<br>0.24 |
| NWKM | 0.56<br>**0.54** | **0.62**<br>0.26 | **0.21**<br>0.02 | 0.51<br>0.37 | 0.30<br>0.07 | 0.44<br>0.25 |
| MWKM | 0.56<br>0.52 | **0.62**<br>0.28 | 0.14<br>0.01 | 0.49<br>0.37 | 0.39<br>0.07 | 0.44<br>0.25 |
| WKM | **0.57**<br>0.54 | **0.62**<br>**0.28** | 0.18<br>0.02 | 0.51<br>0.41 | 0.32<br>0.08 | 0.44<br>0.27 |
| EBKM | **0.57**<br>**0.54** | **0.62**<br>**0.33** | 0.18<br>**0.03** | **0.53**<br>**0.42** | **0.44**<br>**0.09** | **0.47**<br>**0.28** |

Table V
COMPARISON OF THE ADJUSTED RAND INDEX (ARI) PRODUCED BY EACH ALGORITHM.


## ACKNOWLEDGMENT

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


## REFERENCES


[1] D. Barbará, Y. Li, and J. Couto, "Coolcat: an entropy-based algorithm for categorical clustering," in *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 2002, pp. 582–589.

[2] P. Andritsos and P. Tsaparas, "Categorical data clustering," in *Encyclopedia of Machine Learning*. Springer, 2010, pp. 154–159.

[3] J. L. Carbonera and M. Abel, "Categorical data clustering:a correlation-based approach for unsupervised attribute weighting," in *Proceedings of ICTAI 2014*, 2014.

[4] L. Bai, J. Liang, C. Dang, and F. Cao, "A novel attribute weighting algorithm for clustering high-dimensional categorical data," *Pattern Recognition*, vol. 44, no. 12, pp. 2843–2861, 2011.

[5] G. Gan and J. Wu, "Subspace clustering for high dimensional categorical data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 2, pp. 87–94, 2004.

[6] M. J. Zaki, M. Peters, I. Assent, and T. Seidl, "Clicks: An effective algorithm for mining subspace clusters in categorical datasets," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 51–70, 2007.

[7] E. Cesario, G. Manco, and R. Ortale, "Top-down parameter-free clustering of high-dimensional categorical data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 12, pp. 1607–1624, 2007.

[8] H.-P. Kriegel, P. Kröger, and A. Zimek, "Subspace clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 351–364, 2012.

[9] L. Jing, M. K. Ng, and J. Z. Huang, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 8, pp. 1026–1041, 2007.

[10] A. Keller and F. Klawonn, "Fuzzy clustering with weighting of data variables," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 8, no. 06, pp. 735–746, 2000.

[11] F. Cao, J. Liang, D. Li, and X. Zhao, "A weighting k-modes algorithm for subspace clustering of categorical data," *Neurocomputing*, vol. 108, pp. 23–30, 2013.

[12] E. Y. Chan, W. K. Ching, M. K. Ng, and J. Z. Huang, "An optimization algorithm for clustering using weighted dissimilarity measures," *Pattern recognition*, vol. 37, no. 5, pp. 943–952, 2004.

[13] Z. He, X. Xu, and S. Deng, "Attribute value weighting in k-modes clustering," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15 365–15 369, 2011.

[14] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[15] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[16] B. Larsen and C. Aone, "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*. ACM, 1999, pp. 16–22.