A Novel Mixed Values k-Prototypes Algorithm with Application to Health Care Databases Mining

Ahmed Najjar and Christian Gagné
Laboratoire de vision et systèmes numériques (LVSN)
Département de génie électrique et de génie informatique
Université Laval, Québec (Québec), Canada
Emails: ahmed.najjar.1@ulaval.ca, christian.gagne@gel.ulaval.ca

Daniel Reinharz
Laboratoire de simulation du dépistage (LSD)
Département de médecine sociale et préventive
Université Laval, Québec (Québec), Canada
Email: daniel.reinharz@fmed.ulaval.ca

***Abstract*—The current availability of large datasets composed of heterogeneous objects stresses the importance of large-scale clustering of mixed complex items. Several algorithms have been developed for mixed datasets composed of numerical and categorical variables, a well-known algorithm being the k-prototypes. This algorithm is efficient for clustering large datasets given its linear complexity. However, many fields are handling more complex data, for example variable-size sets of categorical values mixed with numerical and categorical values, which cannot be processed as is by the k-prototypes algorithm. We are proposing a variation of the k-prototypes clustering algorithm that can handle these complex entities, by using a bag-of-words representation for the multivalued categorical variables. We evaluate our approach on a real-world application to the clustering of administrative health care databases in Quebec, with results illustrating the good performances of our method.**

I. Introduction

Many organizations produce, collect, and store an increasing amount of data. The need to make sense of this information is obviously essential. The nature of the data mining field makes this possible for two main reasons. First, the main focus of data mining is on methods to process observational data, that is data collected for purposes other than data analysis (e.g., keeping a history of medical services). This is important as, in most cases, we cannot limit ourselves to a specific format of data and it is unrealistic to organize a new data collection according to the methods at hand. The second reason is that data mining provides techniques and methods that can extract information from large data sets, with methods having a good efficiency (computation and storage) and scaling well according to the database size.

As mentioned by Yoo et al. [23], classification and association are the most frequently used techniques for health care data, while clustering is more commonly used in genomic. However, application of clustering to health care is becoming increasingly interesting given the availability of electronic medical records, which is a rich source of information on health care practises. For example, in Quebec for 2005-2006, there were nearly 714,000 hospitalizations for acute care and more than 465,000 one day surgeries. On average, between 80 to 86 million medical services were provided each year to the population of Quebec [19]. This is clearly an example of big data [18].

We are proposing here a scalable algorithm for clustering complex entities characterized by numerical, categorical, and multivalued categorical variables. We want to develop a method that would allow public health professionals to gain insight into real-life practises through the large amount of data available in administrative health care databases. Indeed, in health care, clinical studies assume ideal conditions that are not always the real-life conditions health professionals are finding in their practises. Administrative health care databases can unveil the constraints of reality, as they are capturing elements from a great variety of real medical care situations. These relational databases have many tables and many links connecting them. Our aim is to use those tables and links to construct complex entities describing the various medical services provided. The complex entities are characterized by numerical, categorical, and multivalued variables, and we would like to categorize them into homogeneous clusters of medical services.

Over the last decade, some researchers have proposed techniques and methodologies to handle such health care databases. For instance, Garg et al. [8] developed a mixed distribution survival tree to cluster patients according to the length of their hospital stay. For this purpose, hierarchical clustering was accomplished by recursively splitting nodes using one of three co-variates (age, sex, or diagnosis) as long as the Akaike Information Criterion (AIC) improves. Elghazel [7] presented a graph b-colouring method to cluster hospital stays which are described by categorical, numerical, and multivalued categorical variables. The author follows an approach similar to ours, by considering the complexity of health care entities and the importance of describing these entities and including all of the important variables concerned. However, his method is not scalable as it uses pairwise dissimilarity rather than defining some description of the clusters of the complex entities used as instances (e.g., using centers as clusters definition). The quadratic complexity of this approach makes it difficult to apply to large databases.

The paper is organized as follows. We first present the problem definition in Sec. II, followed by an overview of relevant clustering approaches in Sec. III. We then present our methodology and the proposed algorithm in Sec. IV. A case study on the clustering of an administrative health care database validating our methodology and algorithm follows in Sec. V. Finally, we conclude the paper in Sec. VI with some extension proposals of this work.

II. Problem Definition

Let us assume that objects from a database are represented by variables $V_1, \ldots, V_m$. When the domain value of a variable

is over numbers (integer or real), this variable is stated as *numeric*, while for finite and unordered domain values, the variable is stated as *categorical*. When we concatenate some categorical variables, we obtain a *multivalued categorical* variable. An example of this is the set of diagnostic codes corresponding to one hospital stay, which can have values such as {O48001,Z370,O62101}. Possible values for these multivalued variables correspond to the power set of the single categorical values composing them.

We can formalize a variable as:

$$\begin{aligned} V_l : \mathcal{X} &\longrightarrow D_l \\ x_i &\longmapsto V_l(x_i), \end{aligned}$$

where D_l is the domain of a variable, $\mathcal{X}$ is a set of objects, x_i is the object i, and $V_l(x_i) = x_{i,l}$ is the value of this object for variable V_l. So when, D_l is in $\mathbb{R}$ or $\mathbb{N}$, this variable is numeric. If D_l is a finite and unordered set, the variable is categorical. If $D_l = \mathcal{P}(S)$, where S is a finite and unordered set, and $\mathcal{P}(S)$ is the power set of S, then V_l is a multivalued categorical variable.

Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a set of n objects. We represent an object x_i as a vector $(x_{i,1}, \ldots, x_{i,r}, \ldots, x_{i,q}, \ldots, x_{i,m})$, where the first r elements are numeric values, the next $(q-r)$ elements are categorical values, and the remaining ones are multivalued categorical values.

The method we are proposing aims at clustering objects composed of a mix of these three types of values, in opposition to map the original values into some representation fitting the clustering method (e.g., fixed-size real-valued vectors). By using the objects as is, we expect to avoid removing any useful information from the data that may happen during conversions. We thus want to fit the clustering method to the data, rather than fitting the data to the method.

III. Related Work

Clustering algorithms can be categorized into two groups: hierarchical methods and partitioning algorithms. Traditional hierarchical algorithms are not suitable for large datasets given the huge computation required (i.e., quadratic processing complexity), such that partitioning algorithms are now mostly used for that purpose. The classical k-means algorithm of MacQueen [16] is one of the most commonly used clustering algorithms for processing large volumes of numerical data, thanks to its linear complexity in terms of the dataset size. Moreover, the method has been extended to many aspects. For instance, several proposals have been made to extend it to support other variables types. Huang [10] has proposed an extension of k-means for categorical data. The mode value of categorical variables is used as the centers of each cluster while a matching dissimilarity is employed as a cost function. An extension of this work has been proposed as the k-prototypes algorithm, which combines the dissimilarity measure used in the k-means and the k-modes algorithm [11]. In their paper, Chan et al. [6] have proposed an improvement to the k-prototype algorithm where the dissimilarity measure includes a weighting of the variables, where the weighting is also optimized by the method. A further improvement of this algorithm has been proposed by Bai et al. [3] to address a weakness in the computation of categorical data weights. Liang et al. [14] modified the latest k-prototypes algorithm version by defining the weights in a given distance for categorical and numerical data. All these algorithms are not designed to handle complex data that include variable-length sequences of discrete values (e.g., set of categories). However, such data are common in real-life databases in general, and in administrative health case databases in particular. We are thus proposing to extend the last evolution of k-prototypes [14] in order to support multivalued categorical values mixed with numerical and categorical data.

Algorithm 1 Apriori algorithm

input ST: set of all transactions; θ: threshold
output F: set of frequent itemsets in ST

```
F_1 ← {Frequent 1-itemsets}
k ← 2
while F_{k-1} ≠ ∅ do
    S_k ← {p ∪ {q} | p ∈ F_{k-1} ∧ q ∈ ∪F_{k-1} ∧ q ∉ p}
    supp(c) ← 0, ∀c ∈ S_k
    for all st ∈ ST do
        D_t ← {c | c ∈ S_k ∧ c ⊆ st}
        for all c ∈ D_t do
            supp(c) ← supp(c) + 1
        end for
    end for
    F_k ← {c | c ∈ S_k ∧ supp(c) ≥ θ}
    k ← k + 1
end while
```

IV. Proposed Algorithm

As defined in Sec. II, the input for the algorithm is a set of objects $\mathcal{X}$. The results of the k-prototypes algorithm is a set of prototypes, $c_1, \ldots, c_K$, describing the K clusters. The proposed modification to the k-prototypes algorithm is described in the following.

A. Apriori Algorithm

The Apriori algorithm, proposed by Agrawal et al. [1], aims at mining frequent itemsets from a set of transactions ST. In order to achieve this, it conducts multiple passes over the transaction data. In the first pass, the support of individual items is counted and we determine which itemsets have a support value above a given threshold. Support is the number of transactions in which the item holds. In each subsequent pass, the algorithm starts with the set of itemsets obtained at the previous iteration and agglomerates more items to the set. The new larger itemsets, called candidate itemsets S_k, generated that way have their support evaluated, and the itemsets among these with a support over the threshold are kept, and so on. At the end of the pass, we determine which of the candidate itemsets are actually larger than the threshold, and they become the set for the next pass. This process continues until no new large itemsets are kept at a given iteration. Algorithm 1 presents the pseudo-code of the Apriori method. This algorithm is used in some recent work for feature selection and reducing the large dimension space [12], [22]. So, the idea for using this algorithm in our context is to define a feature projection space formed by the most frequent

itemsets for each multivalued categorical variable. Let X_l be the l-th multivalued categorical variables, $q+1 \leq l \leq m$ and $x_{i,l}$ be the value of object i for this variable. Consider that X_l is ST and calculate the set of most frequent itemsets with the Apriori method, and define it as the projection space for this variable.

This corresponds to a Bag-of-Words (BoW) representation which was first proposed for text document analysis, using the Apriori method to determine the corpus of words (corresponding here to the frequent itemsets) of the representation. This model is used in recent works with machine learning methods in the healthcare field to classify data [4], [20]. Let $\mathbf{T}$ be a set of words, then a representation of the document consists in a vector of word weights $(w_{i,1}, \ldots, w_{i,L})$, where L is the space projection length (i.e., number of different words of interest). Weights $w_{i,j}$ are computed through the *term frequency - inverse document frequency* (tfidf) formula [21], that is $w_{i,j} = \mathrm{tf}(t_j, d_i) \times \mathrm{idf}(t_j)$, where $\mathrm{tf}(t_j, d_i)$ is number of occurrences t_j in document d_i. We calculate $\mathrm{idf}(t_j)$ as:

$$\mathrm{idf}(t_j) = \log\left(\frac{n}{\mathrm{df}(t_j)}\right),$$

with $\mathrm{df}(t_j)$ the number of vectors where t_j is a word in $\mathbf{T}$.

In our case, for each multivalued categorical variable, we extract the most frequent itemsets with the Apriori method to determine the BoW representation. A document is an object value of this variable. The value of each center for this kind of variable is a vector of itemset weights. We use $\mathrm{proj}(x_{i,l})$ as notation which refers to the projection of $x_{i,l}$ (a value of the l-th multivalued categorical variable for object i) on the itemset space for this variable. We are therefore clustering $x_{i,l}$ with its BoW representation based on a set of words determined by the Apriori method.

B. Dissimilarity between Objects

With only numerical and categorical variables, we measure the dissimilarity between object x_i and prototype c_j as [14]:

$$\begin{aligned} d(x_i, c_k) = &\frac{r}{q}\frac{\sum_{l=1}^{r}(x_{i,l} - c_{k,l})^2}{\sum_{j=1}^{K}\sum_{l=1}^{r}(x_{i,l} - c_{j,l})^2} \\ &+ \frac{q-r}{q}\frac{\sum_{l=r+1}^{q} 1 - I(x_{i,l}, c_{k,l})}{\sum_{j=1}^{K}\sum_{l=r+1}^{q} 1 - I(x_{i,l}, c_{j,l})}, \end{aligned} \tag{1}$$

with $I(x, y) = 1$ if $x = y$ and $I(x, y) = 0$ otherwise.

For mixed variables with multivalued categorical variables we propose the following dissimilarity measure:

$$\begin{aligned} d(x_i, c_k) = &\frac{r}{m}\frac{\sum_{l=1}^{r}(x_{i,l} - c_{k,l})^2}{\sum_{j=1}^{K}\sum_{l=1}^{r}(x_{i,l} - c_{j,l})^2} \\ &+ \frac{q-r}{m}\frac{\sum_{l=r+1}^{q} 1 - I(x_{i,l}, c_{k,l})}{\sum_{j=1}^{K}\sum_{l=r+1}^{q} 1 - I(x_{i,l}, c_{j,l})} \\ &+ \frac{m-q}{m}\frac{\sum_{l=q+1}^{m} 1 - \cos(\mathrm{proj}(x_{i,l}), c_{k,l})}{\sum_{j=1}^{K}\sum_{l=q+1}^{m} 1 - \cos(\mathrm{proj}(x_{i,l}), c_{j,l})}, \end{aligned} \tag{2}$$

where $\mathrm{proj}(x_{i,l}) = (w_{i,l,1}, \ldots, w_{i,l,L})$ is the projection of $x_{i,l}$ on the projection space of multivalued variable X_l as described in Sec. IV-A. For this variable, the center for cluster k is $c_{k,l} = (w_{k,l,1}, \cdots, w_{k,l,L})$, where the value $w_{k,l,v}$ is described as the mean of all object weights $w_{i,l,v}$ in the cluster k, $1 \leq v \leq L$, and L is the projection space size (i.e., number of words used in the BoW representation of variable l).

C. Center Computation

Let $c_1, \ldots, c_K$ be K cluster centers. Each center c_k is presented as $(c_{k,1}, \ldots, c_{k,r}, \ldots, c_{k,q}, \ldots, c_{k,m})$, where the first r elements are numerical values, the next $(q-r)$ elements are categorical values, and the remaining elements are BoW representations of multivalued categorical values, using as words the itemsets selected with the Apriori algorithm.

Computation of the center values of each variable depends on its type. When this variable is numerical, $c_{k,l}$ is simply the mean value of the variable for the cluster objects. For categorical variables, $c_{k,l}$ is the mode of all objects in the cluster. And for multivalued categorical variables we have $c_{k,l} = (w_{k,l,1}, \ldots, w_{k,l,L})$, where each $w_{k,l,v}$ is the average weights of the data in cluster k for variable l in dimension v, with $v = 1, \ldots, L$ and L being the projection space size.

Thus, allocation of an object x_i to a cluster C is done with the following equations:

$$c_l = \frac{\left(\sum_{x_j \in C} x_{j,l}\right) + x_{i,l}}{|C|+1}, \qquad l = 1, \ldots, r, \tag{3}$$

$$H_l(x_{i,l}) = H_l(x_{i,l}) + 1, \qquad l = r+1, \ldots, q, \tag{4}$$

$$c_l = \max_h H_l(h), \qquad l = r+1, \ldots, q, \tag{5}$$

$$w_{l,v} = \frac{\left(\sum_{x_j \in C} w_{j,l,v}\right) + w_{i,l,v}}{|C|+1}, \quad \begin{matrix} l = q+1, \ldots, m \\ v = 1, \ldots, L \end{matrix}, \tag{6}$$

$$C = C + \{x_i\}, \tag{7}$$

where $H_l(h)$ is the count of h values for categorical variable l in cluster C, and $|C|$ is the number of objects in the cluster.

Moreover, when the update consists in reallocating the object x_i cluster, the previous cluster C of the object and its center value also need to be updated, using the following equations:

$$c_l = \frac{\left(\sum_{x_j \in C} x_{j,l}\right) - x_{i,l}}{|C|-1}, \qquad l = 1, \ldots, r, \tag{8}$$

$$H_l(x_{i,l}) = H_l(x_{i,l}) - 1, \qquad l = r+1, \ldots, q, \tag{9}$$

$$c_l = \max_h H_l(h), \qquad l = r+1, \ldots, q, \tag{10}$$

$$w_{l,v} = \frac{\left(\sum_{x_j \in C} w_{j,l,v}\right) - w_{i,l,v}}{|C|-1}, \quad \begin{matrix} l = q+1, \ldots, m \\ v = 1, \ldots, L \end{matrix}, \tag{11}$$

$$C = C \backslash \{x_i\}. \tag{12}$$

D. Proposed k-prototypes

The k-prototypes algorithm, proposed by Liang et al. [14], aims at clustering mixed datasets with numerical and categorical variables. It consists in the four following steps.

1) Randomly choose K distinct objects from the dataset as initial cluster centers.

2) Allocate each object to the nearest center according to the dissimilarity measure given in Eq. 1. Update the cluster centers after each allocation.
3) Process all objects by evaluating their nearest center. For each object, if its nearest center belongs to a cluster different from the one to which the object is allocated, reallocate it immediately to the nearest center cluster. Update the centers of the object's previous and current cluster accordingly.
4) Repeat step 3 until no object is reallocated or another stopping criterion is reached.

However, this algorithm cannot handle variables of the multivalued categorical type. Our proposal aims at extending it using the dissimilarity function given in Eq. 2. The result is presented as Algorithm 2.

We use our algorithm to cluster the complex objects described in Sec. II. We run this algorithm for different values of K and choose the one that maximizes the Calinski-Harabasz index (CH index) [5]. This index performs well for this purpose according to Arbelaitz et al. [2]. The CH index is computed as follows:

$$CH(C) = \frac{n-K}{K-1} \frac{\sum_{k=1}^{K} |C_k|\, d(c_k, \overline{x})}{\sum_{k=1}^{K} \sum_{x_i \in C_k} d(x_i, c_k)}, \tag{13}$$

where c_k is the center of the k-th cluster, $\overline{x}$ is the center of the dataset, and $|C_k|$ is the count of objects in cluster C_k.

E. Interpreting Multivalued Results

After choosing the best allocation, we need to interpret the results obtained for the multivalued variables. Let X_l be the l-th multivalued variable and let $S_1 = p_1, \ldots, p_{N_l}$ be the set of first candidate items in the Apriori algorithm. We define the support for each item s in each cluster as:

$$\mathrm{supp}_k(p_j) = \frac{\sum_{x_{i,l} \in C_k} n_{i,l}(p_j)}{|C_k|}, \tag{14}$$

where $n_{i,l}(p_j) = 1$ if $p_j \in x_{i,l}$ and 0 otherwise, $q+1 \leq l \leq m$ and $1 \leq j \leq N_l$. So the support for one item in a cluster is the ratio of a number of the multivalued variable values which contain the item by the number for all objects in this cluster. We analyze the distribution of items' supports in clusters for each multivalued variable to interpret their variability.

V. Case Study: Heart Failure in the Elderly in Quebec

In this section, we are evaluating the proposed k-prototypes method on the clustering of medical records of elderly patients with diagnosed heart failure diseases. This disease, widespread in Western societies [17], is a significant user of several health care resources [15], [17]. Despite the great advancements made in terms of diagnosis and treatment, recent studies emphasize that the care given for heart failure, and particularly to the elderly patients, does not always follow the practise guidelines [9]. This is not surprising given that the management of heart failure is complex and requires many actors from different disciplines, with a large diversity of patients [17]. Indeed, health professionals are generally making decisions by taking into account real-life conditions, which often differ from those of studies made in controlled environments [13].

Algorithm 2 Proposed k-prototypes algorithm

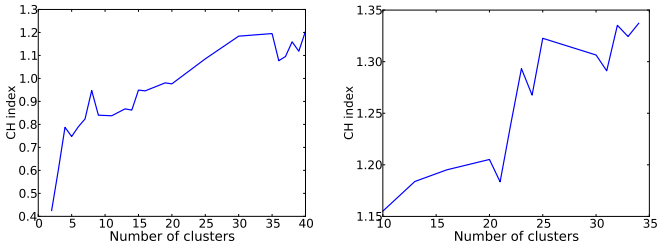
input $\mathcal{X} = \{x_1, \ldots, x_n\}$: set of objects to cluster; t^{max}: maximum number of iterations.
output b_i, $i = 1, \ldots, n$: labels of objects x_i; c_k: cluster centers

```
Compute feature projection space F_l for each multivalued variables X_l, q+1 ≤ l ≤ m using Algorithm 1.
Select K distinct objects randomly from the dataset 𝒳 and use them as the initial cluster centers c_k, k = 1, ..., K.
for all x_i ∈ 𝒳, in random order do
    b_i ← argmin_{j=1..K} (d(x_i, c_j)) (using Eq. 2).
    Update center c_{b_i} and cluster C_{b_i} by adding x_i (using Eq. 3-7).
end for
changed ← true; t ← 1
while (changed = true) ∧ (t ≤ t^max) do
    changed ← false
    for all x_i ∈ 𝒳, in random order do
        y_i ← argmin_{j=1..K} (d(x_i, c_j)) (using Eq. 2).
        if y_i ≠ b_i then
            changed ← true
            Update center c_{b_i} and cluster C_{b_i} by removing x_i (using Eq. 8-12).
            Update center c_{y_i} and cluster C_{y_i} by adding x_i (using Eq. 3-7).
            b_i ← y_i
        end if
    end for
    t ← t + 1
end while
```

We are studying this disease for patients over 65 years old who live in the province of Quebec (Canada). For this purpose, we have been granted access to administrative health care databases of the RAMQ (Régie de l'assurance-maladie du Québec), which acts as the health insurer for Quebec residents covered by the universal public health insurance program (virtually 100 % of the people living in the province), and from the MSSS (Ministère de la Santé et des Services sociaux du Québec), which contains a table of hospital stays and other related tables. These databases record all medical acts from health care professionals that are covered by the RAMQ and all hospital stays in Quebec. Our intent is to exploit these data to reconstruct how patient medical services are given to elderly people suffering from this disease, to cluster this service into homogeneous groups, and, as a first step, to interpret the results with specialists in the field.

A. Data Preprocessing

We have preprocessed these databases by gathering the various medical services according to two categories:



(a) Hospital stays (b) Consultations

Fig. 1. CH index value according to the number of clusters: (a) hospital stays database; (b) consultations database.

- **Consultation**: defined as a service provided by physicians in ambulatory care;
- **Hospital stays**: defined as a service given in the context of a hospitalization of at least one night.

We use the databases described above to integrate information of patient services. We have two types of databases, the first one contains information of all hospital stays and the second one contains physician compensations for medical services provided and drugs purchased in non-hospital settings. For our experiments, we selected individuals in these databases with at least one diagnosis of heart failure (i.e., ICD-10 diagnosis codes 428.0, 428.1, or 428.9) made between January 1, 2000 and December 31, 2005. We also rejected individuals that were not 65 years or older at the earliest consultation date or earliest departure date from hospital stays. By applying these criteria, we have extracted 684,906 hospital stays which took place before December 31, 2009. We then associated the patient information joined to the diagnostics and interventions information. Each hospital stay of a patient is considered as a first category of complex objects described by a set of numerical and categorical variables corresponding to the patient and care information, and multivalued categorical variables corresponding to the diagnostic and intervention values. Also, we extracted 2,864,934 consultations, which are medical services recommended by a referral doctor who oriented a patient to another doctor, before December 31, 2009. These services represent a complex object with patient information joined to medical act information and drug information. The drug information variable for each consultation is multivalued and determined by concatenating a drug consumed by patient between the current consultation date and next consultation date. This process provided us with two sets for two categories of complex objects: hospital stays and consultations.

B. Results and Analysis

The proposed k-prototypes method presented in Sec. IV is applied to the two health care services databases. The hospital stays database contained 684,906 stays and the consultation database contained 2,864,934 consultations in the form described above.

1) Number of Clusters: We begin by clustering all hospital stay entities with various numbers of clusters, to identify the number of clusters that would provide the best results according to the CH index, see Fig. 1(a) for the results obtained. The same procedure is applied for consultation entities and the CH index is given in Fig. 1(b). With 35 clusters we obtain excellent performances for a relatively medium number of clusters for hospital stays. For consultation entities, we choose 25 clusters given the plot presented in Fig. 1(b).

In the following sections, we present the results of hospital stay clustering, consultations clustering, and the analysis for these results. Given the stochastic nature of the algorithm (i.e., random initialization and processing order), we have conducted eight runs for each database. Results reported correspond to the best of 8 runs conducted according to the CH index (Eq. 13).

2) Clustering Results: Table I provides a description of some cluster center values obtained with our algorithm on the consultations.

Following the clustering, we also computed support values for diagnosis and for intervention items using Eq. 14. These support values are helpful for identifying the frequent diagnoses and interventions composing a cluster.

By analyzing consultation clusters, we can expect some conclusions. First, we can observe large families of consultations for elderly people with heart failure. We can also observe that there are six related consultations:

- Consultations for cataracts (cluster 22);
- Consultations concerning disease associated with the skin (cluster 13);
- Consultations concerning respiratory problems (cluster 17);
- Consultations in imagery (radiology and ultrasonography) (clusters 1, 4, 7, 8, 12, 15, 16, 19, 20, 25);
- Consultations concerning heart problems such as chronic ischemic heart disease (clusters 2, 11), heart failure (10, 23), and cardiovascular disease (2);
- Consultations concerning ear, nose, and throat conditions (cluster 24).

Fig. 2 shows that diagnoses allow the identification of large families of consultations for elderly people with heart failure. Diagnoses differentiate one cluster from another. For example, chronic ischemic heart disease can differentiate clusters 2, 11, and 23 from the others. We can also see that heart failure characterizes clusters 9, 10, and 23 compared to others, cardiovascular disease characterizes cluster 2. Cluster 17 includes a diagnosis of chest pain.

We can also detect a variability in physician specialities. For their part, patients in groups 2, 10, 11, and 23 consulted a specialist in cardiology. Also, there are groups that differ from all others by their specific speciality such as cluster 22 (ophthalmology), as shown in Fig. 3. Furthermore, Fig. 3(d) shows drug variability according to cluster. We observe in the consultations that for patients in clusters 1, 3, 4, 15, 16, 19, 21, and 23, no medication is prescribed. This is explained by the fact that these consultations either belong to the family of radiology and ultrasonography (clusters 1, 4, 15, 16, 19) or the time period between two consecutive consultations is short (clusters 3, 21, 23).

In agreement with consultation results and by examining the first and second most frequent diagnoses for hospital stays, it becomes clear that many of the diagnoses are for heart failure diseases, which is expected given that this was the preselection criterion for the individuals database. However, by analyzing diagnosis support in clusters, we note the presence of large families of typical hospital stays for elderly populations:

- Cataracts (clusters 20, 29);
- Hernia (clusters 17, 30);
- Senile dementia (clusters 6, 7, 23);
- Illnesses of the musculoskeletal system (clusters 12, 19, 25);
- Stays related to heart problems such as: aortic valve disease, cardiac complications, other complications due to surgery, atherosclerosis of arteries periph diagnostics (clusters 13, 26); or congestive heart failure or left heart failure (4, 13, 24, 26, 28, 32); or mitral valve disease (4, 13, 26, 32);
- Infections (cluster 27);
- Nervous system (hemiplegia, dysphagia) (cluster 18);
- Kidney problems: hypertensive nephropathy (clusters 3, 8).

Table II provides a description of some cluster center obtained with our algorithm on the hospital stays databases. Fig. 4 shows the variability of the most frequent diagnoses according to each cluster. This variability is still limited given that all patients have a common disease, heart failure. However, some diagnoses allow the identification of the presence of large families of hospital stays for elderly people with heart failure and can characterize one cluster from another. For example, cataracts can differentiate clusters 20 and 29 from the others. We can also see that inguinal hernia without obstruction or gangrene characterizes clusters 17 and 30 (see Fig. 4(b)-4(d)).

Fig. 5 presents the variability of interventions according to clusters. From this figure, it is clear that there is a stronger variability of interventions among clusters. The distribution of diagnoses in clusters seems to correspond with the distribution of the services and interventions. For example, we note that the intervention distinguishing clusters 20 and 29 from the other clusters is the total excision, lens phacoemulsication without intraocular lens insertion. For clusters 13 and 26, we note the presence of the extra-corporeal circulation which is linked to heart complications, etc. We were also able to detect that patients in clusters 4, 24, 28, and 32 were treated primarily in the cardiology service while patients in clusters 2, 7, and 30 were hospitalized in general surgery services. Fig. 6 shows some repartition based on the patient services.

3) Discussion: These results show the potential of this approach which takes all variables into consideration and tries to describe a multivalued categorical variable and search for a dominant intervention and diagnosis in clusters. Moreover, analyzing care pathways is a recent trend to improve the quality of care services. However, services in health care data are large and complex so it is difficult to process them. Indeed, these data sets contain heterogeneous and complex variables which must be considered together. In medical care, a set of variables describes the services. Taking into account each variable separately to make a judgement, leads to a loss of the overall view that determines decisions in the medical field. So, summarizing services in homogeneous clusters allows us to build patients' pathways in future steps. As we presented here, our algorithm succeeded in detecting large families of health care services, characterizing and analyzing them. This should facilitate the categorization of services and highlight pathways of care in real life between these categories.

Like most approaches derived from k-means, this proposed algorithm is sensitive to outliers. In this paper, we purposely decided not to filter out the irregular data of the database beforehand. Indeed, there is a particular interested in our context to handle these cases, in order to allow discovering and processing rare and irregular patterns that may occur within the health care system. But from a clustering perspective, making such a preprocessing can certainly help to improve the results.

VI. Conclusion

This paper presents a new k-prototypes algorithm to handle large scale complex entities. These objects are described by numerical, categorical, and multivalued categorical variables. They are present in a variety of contexts, one of them being administrative health care databases. These databases can reveal real life medical practises. We apply our approach and algorithm on administrative health care databases from the province of Quebec. The interpretation of the results show that we can identify large families of health care services. The modelling and computing approaches proposed in this paper may be applied to other types of complex entities. An extension of this work that we will explore next is to construct patient pathways according to service labels and applying data mining techniques to analyze these pathways.

Acknowledgements

This work was funded through grants from the CIHR Institute of Genetics (Canada), CIHR Institute of Health Services Research (Canada), and APOGEE-Net/CanGèneTest. We acknowledge access to supercomputing facilities of Calcul Québec / Compute Canada. We also thank Annette Schwerdtfeger for proofreading this manuscript.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2):207–216, 1993.

[2] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

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

[4] R. Bouslimi, A. Messaoudi, and J. Akaichi. Using a bag of words for automatic medical image annotation with a latent semantic. *International Journal of Artificial Intelligence & Applications*, 4(3), 2013.

[5] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

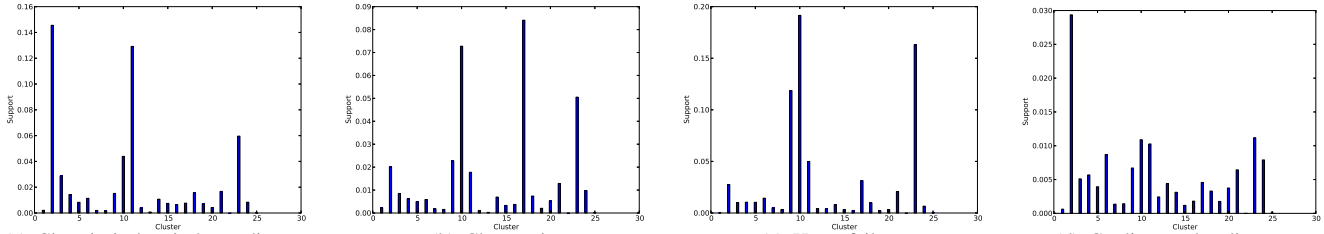
[6] E. Chan, W. Ching, M. Ng, and J. Huang. An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognition*, 37(5):943–952, 2004.

[7] H. Elghazel. *Classification et prévision des données hétérogènes: application aux trajectoires et séjours hospitaliers*. PhD thesis, Unversité Lyon 1, France, 2007.

[8] L. Garg, S. McClean, B. Meenan, E. El-Darzi, and P. Millard. Clustering patient length of stay using mixtures of gaussian models and phase type distributions. In *IEEE Intl Symp. on Computer-Based Medical Systems (CBMS 2009)*, pages 1–7, 2009.

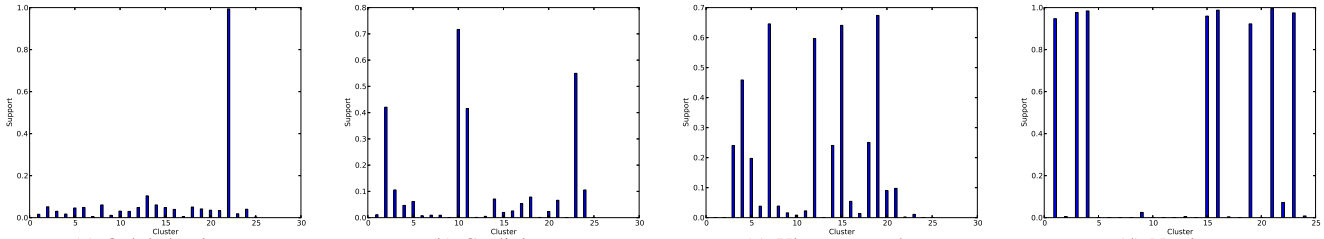
[9] J. G. Howlett, R. S. McKelvie, J. Costigan, A. Ducharme, E. Estrella-Holder, J. A. Ezekowitz, N. Giannetti, H. Haddad, G. A. Heckman, A. M. Herd, et al. The 2010 Canadian cardiovascular society guidelines for the diagnosis and management of heart failure update: heart failure in ethnic minority populations, heart failure and pregnancy, disease management, and quality improvement/assurance programs. *Canadian Journal of Cardiology*, 26(4):185–202, 2010.

[10] Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997.



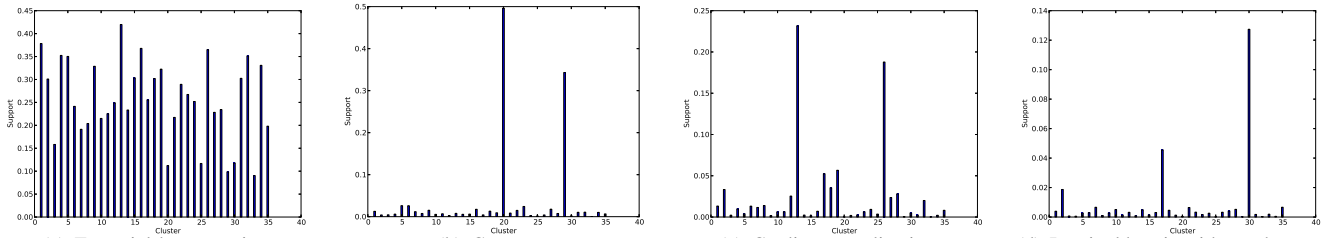
(a) Chronic ischemic heart disease (b) Chest pain (c) Heart failure (d) Cardiovascular disease

Fig. 2. Support values of some diagnoses illustrating their variability according to the different consultation clusters.



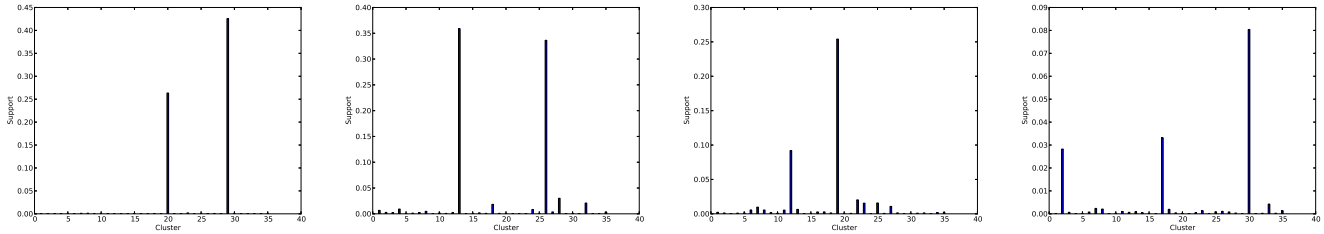
(a) Ophthalmology (b) Cardiology (c) Ultrasonography (d) No drug

Fig. 3. Support values of some practise specialities plus consultations without medication prescribed according to the different consultation clusters



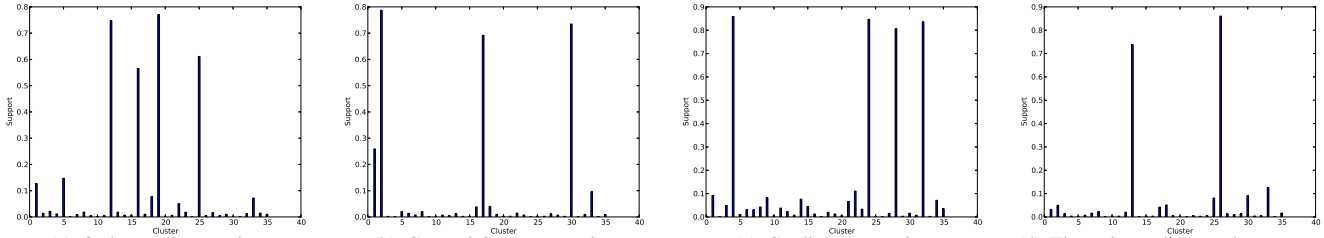
(a) Essential hypertension (b) Cataracts (c) Cardiac complications (d) Inguinal hernia without obstruction or gangrene

Fig. 4. Support values of some diagnoses illustrating their variability according to the different hospital stay clusters.



(a) Total excision, lens, phacoemulsification without intraocular lens insertion (b) Extra-corporeal circulation (c) Open fracture reduction (d) Muscles of thorax and abdomen repair

Fig. 5. Support values of some interventions according to the hospital stay clusters.



(a) Orthopaedics service (b) General Surgery service (c) Cardiology service (d) Thoracic cardiovascular surgery service

Fig. 6. Support values of some medical services according to the hospital stay clusters.

Clust.	#	Patients	Care	Practitioners
C1	408850	**Age**: 75-79 (39.14 %) **Sex**: M (66.5 %)	**Code of act**: Radiology diagnostic (chest, lungs) (25.26 %) **Diagnosis code**: Missing diagnosis or unspecified (64.92 %) **Establishment type**: Diagnostic radiology laboratories (medicine laboratory generally managed by a radiologist doctor) (81.61 %) **Most frequent Drug's AHF class**: No drug (94.73 %)	**Speciality of practitioner**: Radiology diagnostic (86.82 %) **Speciality of referent practitioner**: No speciality (77.37 %)
C10	29006	**Age**: 85 and over (44.31 %) **Sex**: F (82.98 %)	**Code of act**: Emergency Room (consultation) (45.51 %) **Diagnosis code**: Heart failure (19.16 %) **Establishment type**: Emergency (75.01 %) **Most frequent Drug's AHF class**: Diuretics (38.94 %)	**Speciality of practitioner**: Cardiology (71.69 %) **Speciality of referent practitioner**: No speciality (80.81 %)
C13	31004	**Age**: 75-79 (47.41 %) **Sex**: F (82.18 %)	**Code of act**: Dermatologist (private cabinet) (54.68 %) **Diagnosis code**: Contact dermatitis (17.13 %) **Establishment type**: Private cabinet with the municipality number (92.37 %) **Most frequent Drug's AHF class**: Non-steroidal anti-inflammatory (37.27 %)	**Speciality of practitioner**: Dermatology (69.56 %) **Speciality of referent practitioner**: No speciality (79.52 %)
C22	24507	**Age**: 75-79 (34.64 %) **Sex**: F (67.44 %)	**Code of act**: Ophthalmology (visit at the request of an optometrist including report writing) (50.36 %) **Diagnosis code**: Cataracts (52.56 %) **Establishment type**: Private cabinet with the number of municipality (75.56 %) **Most frequent Drug's AHF class**: Non-steroidal anti-inflammatory (39.42 %)	**Speciality of practitioner**: Ophthalmology (99.40 %) **Speciality of referent practitioner**: Optometrist in Quebec (78.99 %)
C24	52756	**Age**: 75-79 (41.11 %) **Sex**: F (88.24 %)	**Code of act**: Private cabinet (consultation) (75.99 %) **Diagnosis code**: Deafness (8.13 %) **Establishment type**: Private cabinet with the municipality number (95.17 %) **Most frequent Drug's AHF class**: Non-steroidal anti-inflammatory (36.86 %)	**Speciality of practitioner**: Otorhinolaryngology (26.92 %) **Speciality of referent practitioner**: No speciality (82.05 %)

TABLE I. DESCRIPTION OF SOME CLUSTER CENTERS OBTAINED ON THE CONSULTATIONS DATABASE WITH THE PROPOSED k-PROTOTYPES METHOD.

Clust.	#	Patients	Care	Practitioners
C13	5571	**Age**: 80-84 (50.39 %) **Sex**: F (83.23 %)	**Type of care**: Physical and psychiatric acute care **Origin**: Home (63.06 %) **Destination**: General and specialized hospital care or hospital center of psychiatric care (68.28 %) **Mean hospitalization length**: 22 days **Most frequent diagnosis**: Coronary atherosclerosis **2nd most frequent diagnosis**: Essential hypertension **Most frequent intervention**: Extra-corporeal circulation **2nd most frequent intervention**: Transfusion, blood cells agglomerated	**Service**: Cardiovascular surgery, thoracic (73.76 %) **Specialist**: Cardiovascular surgery, thoracic (64.01 %)
C19	9317	**Age**: 85 and over (66.48 %) **Sex**: F (87.96 %)	**Type of care**: Physical and psychiatric acute care **Origin**: Home (91.97 %) **Destination**: Funeral home or other hospitalization center for organ harvesting (34.29 %) **Mean hospitalization length**: 22 days **Most frequent diagnosis**: Essential hypertension **2nd most frequent diagnosis**: Coronary atherosclerosis **Most frequent intervention**: Open fracture reduction + femoral external device **2nd most frequent intervention**: Transfusion, blood cells agglomerated	**Service**: Orthopaedics (77.01 %) **Specialist**: Orthopaedic surgery (81.69 %)
C29	29952	**Age**: 80-84 (37.69 %) **Sex**: F (93.00 %)	**Type of care**: One-day surgery care **Origin**: Home (99.08 %) **Destination**: Home (98.30 %) **Mean hospitalization length**: 1 day **Most frequent diagnosis**: Cataracts, unspecified **2nd most frequent diagnosis**: Cataracts **Most frequent intervention**: Excision total, crystalline phacoemulsification without intraocular lens insertion **2nd most frequent intervention**: Insertion of an intraocular prosthesis + cataracts extraction	**Service**: Ophthalmology (91.54 %) **Specialist**: Ophthalmology (91.74 %)
C30	8383	**Age**: 80-84 (44.55 %) **Sex**: M (85.51 %)	**Type of care**: One-day surgery care **Origin**: Home (94.54 %) **Destination**: Home (83.48 %) **Mean hospitalization length**: 1 day **Most frequent diagnosis**: Coronary atherosclerosis **2nd most frequent diagnosis**: Inguinal hernia without obstruction or gangrene **Most frequent intervention**: Repair, muscles of the chest and abdomen, open approach **2nd most frequent intervention**: No intervention	**Service**: General surgery (73.49 %) **Specialist**: General surgery (79.76 %)

TABLE II. DESCRIPTION OF SOME CLUSTER CENTERS OBTAINED ON THE HOSPITAL STAYS DATABASE WITH THE PROPOSED k-PROTOTYPES METHOD.

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

[12] O. Inan, M. S. Uzer, and N. Yılmaz. A new hybrid feature selection method based on association rules and PCA for detection for breast cancer. *International Journal of Innovative Computing, Information and Control*, 9(0):2, 2013.

[13] D. Kent and G. Kitsios. Against pragmatism: on efficacy, effectiveness and the real world. *Trials*, 10(1):48, 2009.

[14] J. Liang, X. Zhao, D. Li, F. Cao, and C. Dang. Determining the number of clusters using information entropy for mixed data. *Pattern Recognition*, 45(6):2251–2265, June 2012.

[15] D. Lloyd-Jones, R. J. Adams, T. M. Brown, M. Carnethon, S. Dai, G. De Simone, T. Ferguson, E. Ford, K. Furie, C. Gillespie, et al. Heart disease and stroke statistics – 2010 update. *Circulation*, 121(12):948–954, 2010. On Behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee.

[16] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. California, USA, 1967.

[17] J. P. Man and B. I. Jugdutt. Systolic heart failure in the elderly: optimizing medical management. *Heart failure reviews*, 17(4-5):563–571, 2012.

[18] J. Manyika, M. Chui, J. Bughin, B. Brown, R. Dobbs, C. Roxburgh, and A. Byers. Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, 2011.

[19] Ministère de la Santé et des Services sociaux du Québec. Statistiques. http://wpp01.msss.gouv.qc.ca/appl/g74web/statistiques.asp, 2013.

[20] P. Ordónez, T. Armstrong, T. Oates, and J. Fackler. Using modified multivariate bag-of-words models to classify physiological data. In *Intl Conf. on Data Mining Workshops (ICDMW)*, pages 534–539. IEEE, 2011.

[21] G. Salton and M. McGill. The SMART and SIRE experimental retrieval systems. *McGraw-Hill, New York*, pages 118–155, 1983.

[22] N. Thangsupachai, P. Kitwatthanathawon, S. Wanapu, and N. Kerdprasop. Clustering large datasets with apriori-based algorithm and concurrent processing. In *Proc. of Intl MultiConference of Engineers and Computer Scientists*, volume 1, 2011.

[23] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua. Data mining in healthcare and biomedicine: A survey of the literature. *Journal of medical systems*, pages 1–18, 2012.