Two-Step Heterogeneous Finite Mixture Model Clustering for Mining Healthcare Databases

Ahmed Najjar and Christian Gagné
Laboratoire de vision et systèmes numériques
Dép. 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
Dép. de médecine sociale et préventive
Université Laval, Québec (Québec), Canada
Email: daniel.reinharz@fmed.ulaval.ca

***Abstract*—Dealing with real-life databases often implies handling sets of heterogeneous variables. We are proposing in this paper a methodology for exploring and analyzing such databases, with an application in the specific domain of healthcare data analytics. We are thus proposing a two-step heterogeneous finite mixture model, with a first step involving a joint mixture of Gaussian and multinomial distribution to handle numerical (i.e., real and integer numbers) and categorical variables (i.e., discrete values), and a second step featuring a mixture of hidden Markov models to handle sequences of categorical values (e.g., series of events). This approach is evaluated on a real-world application, the clustering of administrative healthcare databases from Québec, with results illustrating the good performances of the proposed method.**

I. Introduction

Healthcare systems are characterized by an increasing number of medical disciplines and specialized departments. Such modern organizations include information systems keeping information on patients and services provided in various administrative databases. Extracting useful knowledge from these large heterogeneous healthcare databases is usually hard to achieve with traditional methods (e.g., SQL queries). This limitation comes both from the complex nature and the size of these databases. For instance, on average between 80 and 86 million medical services were provided each year to the population of Québec [8]. Moreover, the various databases are made of observational information of various purposes (e.g., insurance, patient records), which are not designed nor adapted to the application of conventional analytics methods.

In the current work, we are considering the use of clustering methods producing probabilistic models describing the data. The model-based approach assumes data generated by a finite mixture with some probability distributions. More specifically, in order to deal with administrative databases, we are proposing an algorithm for clustering with a heterogeneous finite mixture model of complex entities characterized by numerical, categorical, and multivalued categorical variables.

The paper is organized as follows. An overview of relevant clustering approaches is first presented in Sec. II. We then present our methodology and the corresponding algorithm in Sec. III. Follows in Sec. IV the evaluation of the method for the clustering of administrative healthcare databases from Québec, before concluding the paper in Sec. V.

II. Related Work

Model-based clustering (or mixture model) is one of the two main families of approaches used for clustering – the other being distance-based methods. Models are usually based on the use of mixture probability densities. The choice of the probability densities depends on the types of variables at hand. Gaussian mixture models are widely used to handle quantitative variables whereas a mixture of multinomial distributions are more suited for categorical variables, assuming independence between these variables (see Clogg [1]). Furthermore, mixture models can also be applied to process discrete sequences. For example, Tino et al. [13] have used a constrained mixture of discrete hidden Markov models for the clustering of Web logs data. For general reviews of finite mixture models, see McLachlan and Peel [7].

In practice, data sets can contain many types of variables. Jorgensen and Hunt [5], [6] proposed a mixture model to handle data having both continuous and categorical variables. In this model, they assumed that variables are pairwise independent. The component of each variable depends on its type, while the distribution in the cluster for each individual is the product of the distributions of each variable. Moreover, this approach is not applicable for datasets composed of multivariate and sequential data. In contrast, Smyth [12] has demonstrated that it is possible to make a finite mixture model that handles both multivariate and sequential variables, assuming independence between these different types of variables. This model extends Jorgensen and Hunt's model by adding the finite mixtures of sequential models. In the application of his model, Smyth used finite mixtures of two-dimension Gaussian components coupled to first order Markov chains.

Over the last decade, model-based clustering has also been applied to healthcare. Garg et al. [3] developed a mixed distribution survival tree to cluster patients

according to the length of their hospital stay. Their model-based clustering is a finite mixture of Gaussian components to model the length of a hospital stay in each node making a decision tree over the categorical data. In opposition to our proposal, their model was not built in a purely unsupervised way nor aimed at modeling arbitrary mix of heterogeneous variables. Rebuge and Ferreira [10] applied a finite mixture of first order Markov chains to cluster the logs event sequences of clinical and administrative care flows for the Hospital of São Sebastião. These models have been applied for one type of variables. To our knowledge, the approach proposed in this paper is the first application to healthcare of probabilistic clustering able to handle three important types of variables.

III. Proposed Two-Step Heterogeneous Mixture Model

Relational databases are organized into many tables whose variables are interconnected by links. By querying the relational databases, we can build a set of n heterogeneous objects $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ characterized by the various linked variables $V_1, \ldots, V_m$. Each object $\mathbf{x}_i$ is 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 (domain is over numbers), the next $(q-r)$ elements are categorical values (domain is finite and unordered), and the remaining ones are multivalued categorical values (concatenation of some categorical values). For instance, one value of the diagnosis variable of one hospital stay is a series of diagnostic (e.g., {O48001,Z370,O62101}). Our method aims at clustering objects composed of a mix of these three types of values based on mixture model clustering. By proceeding this way, we aim at working on the original representation of objects.

The proposed two-step finite mixture model algorithm has been designed to work on such a representation. Our work extends the proposal of Smyth [12] in two ways. First, we introduced the use multinomial distributions and hidden Markov models (HMM) to handle categorical and multivalued categorical values, respectively. Second, the algorithm is organized as iterations over two steps. During the first step, we determine the model for numerical and categorical variables of the data. Depending on the clustering results obtained, the second step determines a HMM for each multivalued categorical values in each cluster. In the end, the individual observation membership is calculated as the product of memberships generated in those two steps. At each step, we determine a model fitting data of specific variable types.

A. Step 1: EM for Numerical and Categorical Variables

Considering the formalization given above, the purpose of the first step is to cluster in K groups the objects $\mathbf{x}_i \in \mathcal{X}$, considering only numerical and categorical variables, modeling them as a finite mixture model. The proportion for the groups (priors) are given by $\{\omega_1, \ldots, \omega_K\}$, each group following a probability distribution $f(\mathbf{x}_i|\phi_k)$, with a distinct parametrization ϕ_k for each group $k = 1, \ldots, K$. We thus assume that each observation $\mathbf{x}_i$ is generated by a finite mixture model with probability given by $f(\mathbf{x}_i) = \sum_{k=1}^{K} \omega_k f(\mathbf{x}_i|\phi_k)$. We assume that the variables are independent such that the distribution probability $f(\cdot|\phi_k)$ of group k is the product of distribution probability $f(\cdot|\varphi_{k,l})$ for all numerical and categorical variables and is given by $f(\mathbf{x}_i|\phi_k) = \prod_{l=1}^{q} f(x_{i,l}|\varphi_{k,l})$. For numerical variables, $f(x_{i,l}|\varphi_{k,l})$ is modeled as a Gaussian distribution, $f(x_{i,l}|\varphi_{k,l}) \sim \mathcal{N}(\mu_{k,l}, \sigma^2_{k,l})$, where $\mu_{k,l}$ and $\sigma_{k,l}$ are respectively the mean and standard deviation of the l-th variable in cluster k. For categorical variables, we are using a multinomial density, $f(x_{i,l}|\varphi_{k,l}) \sim \mathrm{Mutl}(1, \boldsymbol{\lambda}_{k,l})$, where $\lambda_{k,l,e}$ is the probability that the l-th variable takes modality e when the individual i belongs to cluster k. In such a setting, $x_{i,l,e}$ takes 1 if individual i has modality e for the l-th variable, taking 0 otherwise. L_l is the number of modalities for variable V_l and $\sum_{e=1}^{L_l} \lambda_{k,l,e} = 1$. This modeling is the first novelty of our proposal.

Determining the parameters of a finite mixture model is usually done by using the Expectation Maximization (EM) algorithm [2]. The principle of the algorithm is to iteratively evaluate the membership expectations $z_{i,k}^{(t)}$ of each instance for each group according to the model parameters $\phi^{(t-1)}$ (E-step), followed by re-evaluating the model parameters $\phi^{(t)}$ by maximizing the log-likelihood expectation (M-step). The algorithm starts from some initial estimate of parameter $\phi^{(0)}$ and then proceeds by iteratively evaluating membership expectations $z_{i,k}^{(t)}$ and updating $\phi^{(t)}$ until convergence is achieved. At the E-step, each individual membership probability belongs to cluster k with respect to the current model parameter $\phi^{(t-1)}$, that is:

$$z_{i,k}^{(t)} = \frac{\omega_k^{(t-1)} f(\mathbf{x}_i|\phi_k^{(t-1)})}{\sum_{j=1}^{K} \omega_j^{(t-1)} f(\mathbf{x}_i|\phi_j^{(t-1)})}. \quad (1)$$

At the M-step, parameters that maximize the log-likelihood $\mathcal{L}(\phi)$ are calculated, usually analytically (i.e., by evaluating $\partial\mathcal{L}(\phi_k)/\partial\phi_{k,l} = 0$ for all $\phi_{k,l}$ making ϕ_k). For the mixed Gaussian-multinomial model proposed, the expression of parameters that maximize expectation of the complete log-likelihood is:

$$\mu_{k,l}^{(t)} = \frac{\sum_{i=1}^{n} z_{i,k}^{(t)} x_{i,l}}{\sum_{i=1}^{n} z_{i,k}^{(t)}}, \quad (2)$$

$$\left(\sigma_{k,l}^{(t)}\right)^2 = \frac{\sum_{i=1}^{n} z_{i,k}^{(t)} (x_{i,l} - \mu_{k,l}^{(t)})^2}{\sum_{i=1}^{n} z_{i,k}^{(t)}}, \quad (3)$$

$$\lambda_{k,l,e}^{(t)} = \frac{\sum_{i=1}^{n} z_{i,k}^{(t)} x_{i,l,e}}{\sum_{i=1}^{n} z_{i,k}^{(t)}}. \quad (4)$$

The first step of the method requires choosing initial parameters for the Gaussian and multinomial distributions. For that purpose, we apply several times the k-prototypes method for mixed numeric and categorical values proposed by Huang [4], choosing the partition which minimizes the total error. Then, we compute means and standard deviation for each variable in each cluster and we assign those values to the initial $\mu_{k,l}^{(0)}$ and $\sigma_{k,l}^{(0)}$ of the Gaussian distributions. Similarly, we compute frequencies of each modality for each categorical attribute and assign it to the corresponding $\lambda_{k,l,e}^{(0)}$.

After parameter initialization, the clustering process over numerical and categorical variables is launched. This clustering process is based on algorithm EM described above. The expression of parameters in the maximization step is given by Eq. 2-4.

As a result of this first step, in which we deal only with numerical and categorical variables, we obtain individual memberships computed by Eq. 1. Furthermore, we can compute crisp partition by assigning each object i to the cluster that maximizes $z_{i,k}$:

$$b_{i,k} = \begin{cases} 1 & \text{if } k = \operatorname{argmax}_j z_{i,j} \\ 0 & \text{otherwise} \end{cases} . \quad (5)$$

The values of the multivalued categorical variable for this partition is used as input for the second step.

B. Step 2: HMM for Multivalued Categorical Variables

The second novelty of our method is the use of HMM [9] to handle multivalued categorical variables values. The purpose of the second step of the algorithm is to fit a HMM for each multivalued variable values in each cluster obtained from the first step, using the crisp partitions of the data instances given by Eq. 5. For simplification purpose, we provide details in the following for the l-th multivalued variable.

Obtaining as input the crisp partition of objects given by the first step, we have a set of individual variable values in each cluster k. Each individual variable value $x_{i,l}$ is given by $x_{i,l} = \{x_{i,l,1}, x_{i,l,2}, \ldots, x_{i,l,\zeta_i}\}$, and is a sequence of observed symbols $x_{i,l,e}$ that can take O_l possibles values in an observation space. For example, diagnosis multivalued variable can take as individual value {O48001,Z370,O62101}. So, its observations space is the set of diagnosis codes. From this, we train K HMMs for each multivalued variable V_l, one for each cluster, by the well-known Baum-Welch algorithm, in order to obtain a model parameter $\varphi_{k,l} = \{\boldsymbol{\pi}_{k,l}, \mathbf{A}_{k,l}, \mathbf{B}_{k,l}\}$ where $\boldsymbol{\pi}_{k,l}$ is the initial probability vector, $\mathbf{A}_{k,l}$ is the transition probability matrix, and $\mathbf{B}_{k,l}$ is the emission probability matrix. Since the convergence of this algorithm depends on its initialization, we run it many times where parameters $\varphi_{k,l}$ are initialized to values coming from a Dirichlet distribution, varying the α parameter (same for all dimensions) of the distribution between 0.1 and 1 by 0.1 steps. For each cluster, we select the best HMM obtained with this method according to the log-likelihood.

The next phase in the second step is to compute variable value $x_{i,l}$ the emission probability $P(x_{i,l}|\varphi_{k,l})$ for each individual given a HMM of parameters $\varphi_{k,l}$. This probability is computed by the so-called forward algorithm. To provide a value that is comparable with the membership values computed in the first step, we normalize this probability by the sum of probabilities of all clusters, as in:

$$\xi_{i,k,l} = \frac{P(x_{i,l}|\varphi_{k,l})}{\sum_{k=1}^{K} P(x_{i,l}|\varphi_{k,l})}. \quad (6)$$

So, we obtain individual memberships for each cluster given a variable of the object value.

C. Two-step EM+HMM Algorithm

One iteration of the two-step algorithm proposed terminates by computing individual membership probabilities as the product of the individual membership probability obtained in the first step using the EM algorithm with the individual membership probabilities obtained from the HMMs of each multivalued variable:

$$h_{i,k} = z_{i,k} \prod_{l=q+1}^{m} \xi_{i,k,l}. \quad (7)$$

For the next iteration, the crisp partition obtained with this membership $h_{i,k}$ is used to compute initial parameters of the Gaussian and multinomial distribution in order to make a new iteration of the proposed two-step algorithm. We set a fixed number of iterations as the stop criterion for the algorithm. The general algorithm of the method is presented as Algo. 1, while the specific version of EM used for the numerical and categorical variables is presented as Algo. 2.

D. Interpreting Multivalued Results

After choosing the best allocation, we need to interpret the results obtained for the multivalued variables. Let V_l be the l-th multivalued variable and $o_{j,l}$ be one categorical value in the space of V_l. We define the support for each state $o_{j,l}$ in each cluster as $\text{support}_k(o_{j,l}) = \frac{\sum_{i=1}^{n} b_{i,k}\, \mathcal{J}_{i,l}(o_{j,l})}{|C_k|}$ where $\mathcal{J}_{i,l}(o_{j,l}) = 1$ if $o_{j,l} \in x_{i,l}$ and 0 otherwise, for $l = q+1, \ldots, m$ and $j = 1, \ldots, O_l$ and $|C_k|$ is the cardinality of the k-th cluster. So the support for one observation in a cluster is the ratio of the number of the multivalued variable values that contains this observation by the number of objects in the cluster. We analyze the distribution of state supports in the clusters for each multivalued variable to interpret their variability.

Algorithm 1 Two-step EM+HMM algorithm for complex objects

input $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$: set of objects to cluster; T: maximum number of iterations

output $b_{i,k}^{(T)}$: final labels of objects

1: Compute $b_{i,k}^{(0)}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ with k-prototypes, by making a clustering partition of objects in $\mathcal{X}$ using only numerical and categorical variables
2: **for** $t = 1, \ldots, T$ **do**
3: Compute labels $b_{i,k}^{(t)}$ and membership $z_{i,k}^{(t)}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ applying EM over numerical and categorical variables as described in Sec. III-A, using Algo. 2
4: Generate HMM models, $\varphi_{k,l}^{(t)}$ for $k = 1, \ldots, K$ and $l = q+1, \ldots, m$, using Baum-Welch as described in Sec. III-B
5: Compute emission probabilities $P(x_{i,l}|\varphi_{k,l}^{(t)})$ using the forward algorithm and membership individual probability $\xi_{i,k,l}^{(t)}$ using Eq. 6, for $k = 1, \ldots, K$ and $l = q+1, \ldots, m$
6: Compute two-step membership individual probabilities $h_{i,k}^{(t)}$ with Eq. 7 for $i = 1, \ldots, n$ and $k = 1, \ldots, K$
7: Compute labels $b_{i,k}^{(t)}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ as:

$$b_{i,k}^{(t)} = \begin{cases} 1 & \text{if } k = \operatorname{argmax}_j h_{i,j}^{(t)} \\ 0 & \text{otherwise} \end{cases}$$

8: **end for**

IV. Case study: Heart Failure of Elderly People in Québec

In this section, an evaluation of the proposed methodology presented in Sec. III was carried out on the clustering of medical records of patients over 65 years old with diagnosed heart failure diseases and who live in the province of Québec (Canada). We have thus 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 Québec 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 Québec. Our aim is to extract information from these data that allow us 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 of the domain.

Algorithm 2 EM for numerical and categorical variables

input $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$: set of objects to cluster described only by numerical and categorical variables; $b_{i,k}^{(0)}$: initial labels of objects; T^{EM}: maximum number of iterations

output $b_{i,k}$: final object labels; $z_{i,k}^{(T^{\mathrm{EM}})}$: final membership probabilities

1: Compute $\mu_{k,l}^{(0)} \leftarrow \frac{\sum_{i=1}^{n} b_{i,k}^{(0)} x_{i,l}}{n_k}$ for $k = 1, \ldots, K$ and $l = 1, \ldots, r$
2: Compute $\sigma_{k,l}^{(0)} \leftarrow \frac{\sum_{i=1}^{n} b_{i,k}^{(0)} (x_{i,l} - \mu_{k,l}^{(0)})^2}{n_k}$ for $k = 1, \ldots, K$ and $l = 1, \ldots, r$
3: Compute $\lambda_{k,l,e}^{(0)} \leftarrow \frac{\sum_{i=1}^{n} b_{i,k}^{(0)} x_{i,j,e}}{n_k}$ for $k = 1, \ldots, K$, $l = r+1, \ldots, q$, and $e = 1, \ldots, L_l$
4: $t \leftarrow 1$
5: **while** $\left(\frac{|\mathcal{L}(\phi^{(t)}) - \mathcal{L}(\phi^{(t-1)})|}{\mathcal{L}(\phi^{(t-1)})} \geq \epsilon \right) \wedge (t \leq T^{\mathrm{EM}})$ **do**
6: E-step: compute $z_{i,k}^{(t)}$ using Eq. 1 for $i = 1, \ldots, n$ and $k = 1, \ldots, K$
7: M-step: compute $\mu_{k,l}^{(t)}$ and $\sigma_{k,l}^{(t)}$ for $k = 1, \ldots, K$ and $l = 1, \ldots, r$ using respectively Eq. 2 and 3, and $\lambda_{k,l,e}^{(t)}$ for $k = 1, \ldots, K$, $l = r+1, \ldots, q$, and $e = 1, \ldots, L_l$ using Eq. 4
8: $t \leftarrow t + 1$
9: **end while**
10: Compute labels $b_{i,k}$ from $z_{i,k}^{(T^{\mathrm{EM}})}$ using Eq. 5 for $i = 1, \ldots, n$ and $k = 1, \ldots, K$

A. Data Preprocessing

We have preprocessed these databases by gathering the various medical services to obtain hospital stays category. Hospital stays are 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 1 January 2000 and 31 December 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 between 1 January 2000 and 31 December 2009. We then associated the patient information joined to the diagnostics and interventions information. Each hospital stay of a patient is considered as a 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.

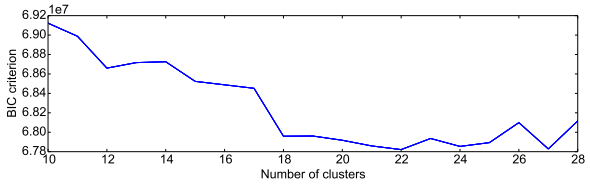
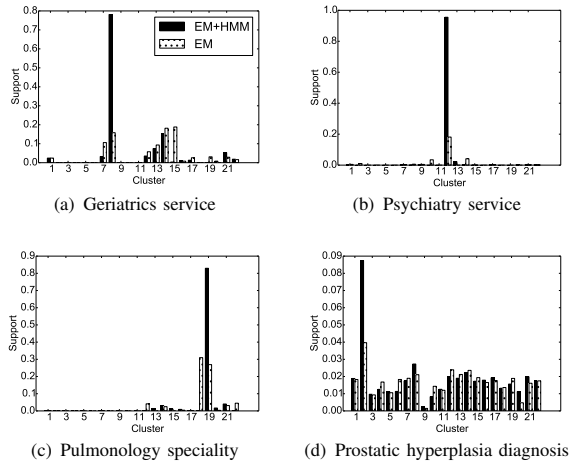


Fig. 1. BIC value according to the number of clusters.



(a) Geriatrics service (b) Psychiatry service

(c) Pulmonology speciality (d) Prostatic hyperplasia diagnosis

Fig. 2. Comparison of results obtained with the Expectation Maximization method applied only on numerical and categorical variables (EM) with the proposed method handling also multivariate categorical values (EM+HMM).

B. Preliminary Experiments

For the experiments presented in this paper, we run our two-step method for $T = 3$ iterations globally (see Algo. 1), $T^{\mathrm{EM}} = 100$ iterations for EM at the first step (see Algo. 2), and we train HMMs having 10 hidden states. Four repetitions are conducted, where the best value according to Schwarz's Bayesian inference criterion (BIC) [11] is kept. The GHMM library (http://ghmm.org) was used as implementation of the HMM algorithms.

In order to determine the number of clusters to use, we executed the clustering of hospital stays entities with a varying number of clusters $k \in \{10, 11, \ldots, 28\}$, evaluating clustering results with the BIC, as illustrated in Fig. 1 Results obtained suggest the use of 22 clusters, where the BIC value is minimum.

According to this choice, we compare the final clustering given by our two-step EM+HMM algorithm with the clustering given by EM algorithm in the first cycle of our algorithm. These results showed that the proposed approach allows the generation of more homogeneous clusters and other specialized clusters, justifying the addition of multivalued variables and the effectiveness of our algorithm at discovering trends in clusters. The analysis of the distribution of medical services, specialists and diagnosis within clusters presented in Fig. 2 confirms that the proposed EM+HMM algorithm improves EM results for several groups, but also allows other more homogeneous groups to be identified, which EM cannot detect as it uses only numerical and categorical variables.

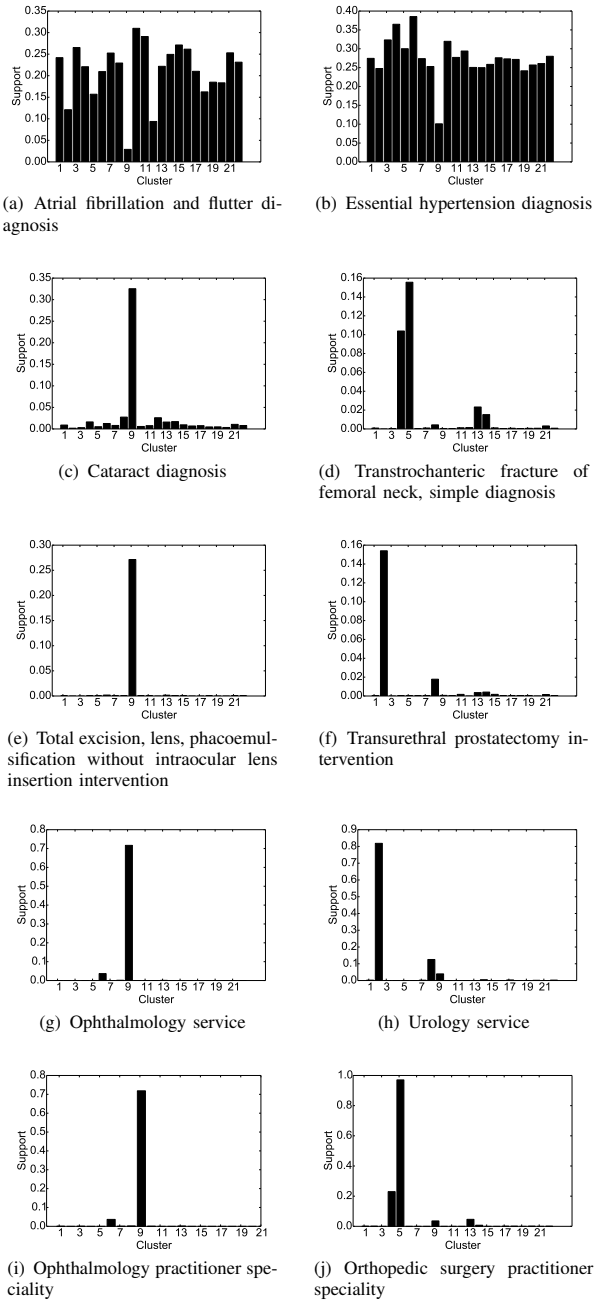
C. Results and Analysis

In the following, results are presented through a description of the families of hospital stay clusters obtained and a description of variability of multivalued categorical variables (diagnoses variables and intervention variables). It is interesting to note that, although patients with heart failure have concurrent illnesses and healthcare services are often provided by specialists and non-specialists, our method has allowed us to discover several large families of hospital stays. This illustrates that huge amount of hospital stays can be clustered into several categories. Clusters included hospital stays related to: i) cataracts disease (cluster 9); ii) heart problems (clusters 3, 6, 10, 11, and 20); iii) disease of the musculoskeletal system (clusters 4 and 5); iv) physical ailments, mental, functional and social related to ageing (cluster 8); v) kidney and prostate disease, urinary tract (cluster 2); vi) respiratory tract (cluster 19); vii) mental disorder (cluster 12); viii) infections (cluster 13).

In addition, it is possible to make a description of cluster centers obtained with our method. For instance, the center of cluster 9 is characterized by 1) ophthalmology as a most frequent service and speciality, 2) cataract as the most frequent diagnosis, and 3) insertion of an intraocular prosthesis + cataracts extraction as the most frequent intervention.

On the other hand, the distributions of some diagnostics in clusters demonstrate the difficulty of extracting a homogeneous and dissociated cluster. This seems to be logical according to the comorbidity of heart insufficiency with other diseases and the conditions used for extracting the relevant instances from the databases (i.e., diagnosis of heart failure diseases), as highlighted in Fig. 3(a)-3(b). Despite this comorbidity, in each detected cluster we have some specific diseases that differentiate it from others, as illustrated by Fig. 3(c)-3(d).

As noticed in Fig. 3(e)-3(f), the analysis of intervention distributions in clusters strengthens clustering results and comes in line with the distribution of the diagnostic obtained. In Fig. 3(g)-3(h), we can see also that each hospital stay is characterized by medical services which signifies that the analysis of the distribution of services within clusters can denote the capacity of our method to cluster hospitalizations with homogeneous services. Similar conclusions were supported by the variability of practitioner speciality within clusters, as illustrated in Fig. 3(i)-3(j). These results demonstrate that the proposed two-level heterogeneous finite mixture model clustering is capable of extracting hospital stay cluster regardless of the complexity of variables and comorbidity on diagnoses.



(a) Atrial fibrillation and flutter diagnosis

(b) Essential hypertension diagnosis

(c) Cataract diagnosis

(d) Transtrochanteric fracture of femoral neck, simple diagnosis

(e) Total excision, lens, phacoemulsification without intraocular lens insertion intervention

(f) Transurethral prostatectomy intervention

(g) Ophthalmology service

(h) Urology service

(i) Ophthalmology practitioner speciality

(j) Orthopedic surgery practitioner speciality

Fig. 3. Support values illustrating the variability according to the different hospital stay clusters.

V. Conclusion

In this paper, we are proposing a two-step heterogeneous finite mixture model clustering algorithm based on mixture and hidden Markov models. Our algorithm handles objects characterized by numerical, categorical, and multivalued categorical variables. We apply it on administrative health care databases from the province of Québec, to cluster hospital stay services. Results with those real-life databases showed that our method can identify large families of health services. In addition, our approach is not limited to exploring health care databases, it can be applied to other types of complex entities. We plan to further extend our work by constructing patient pathways according to service labels obtained by clustering and analyzing those pathways with process mining methods.

Acknowledgments

This work was funded through grants from the CIHR Institute of Genetics (Canada), CIHR Institute of Health Services Research (Canada), NSERC (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] C. C. Clogg. Latent class models. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 311–359. Springer, 1995.

[2] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[3] 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 *Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems, 2009. CBMS 2009.*, pages 1–7. IEEE, 2009.

[4] 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.

[5] L. Hunt and M. Jorgensen. Theory & methods: Mixture model clustering using the multimix program. *Australian & New Zealand Journal of Statistics*, 41(2):154–171, 1999.

[6] M. Jorgensen and L. Hunt. Mixture model clustering of data sets with categorical and continuous variables. In *Proceedings of the Conference ISIS96, Australia*, pages 375–84, 1996.

[7] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

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

[9] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[10] Á. Rebuge and D. R. Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, 37(2):99–116, 2012.

[11] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[12] P. Smyth. Probabilistic model-based clustering of multivariate and sequential data. In *Proceedings of the Seventh International Workshop on AI and Statistics*, pages 299–304. San Francisco, CA: Morgan Kaufman, 1999.

[13] P. Tiňo, A. Kabán, and Y. Sun. A generative probabilistic approach to visualizing sets of symbolic sequences. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–706. ACM, 2004.