Individuals tell a fascinating story: using unsupervised text mining methods to cluster policyholders based on their medical history

Romain Gauchon, Jean-Pascal Hermet

Individuals tell a fascinating story: using unsupervised text mining methods to cluster policyholders based on their medical history

Romain Gauchon[1]

Jean-Pascal Hermet[2]

Background and objective: Classifying people according to their health profile is crucial in order to propose appropriate treatment. However, the medical diagnosis is sometimes not available. This is for example the case in health insurance, making the proposal of custom prevention plans difficult. When this is the case, an unsupervised clustering method is needed. This article aims to compare three different methods by adapting some text mining methods to the field of health insurance. Also, a new clustering stability measure is proposed in order to compare the stability of the tested processes.

Methods : Nonnegative Matrix Factorization, the word2vec method, and marginalized Stacked Denoising Autoencoders are used and compared in order to create a high-quality input for a clustering method. A self-organizing map is then used to obtain the final clustering. A real health insurance database is used in order to test the methods.

Results: the marginalized Stacked Denoising Autoencoder outperforms the other methods both in stability and result quality with our data.

Conclusions: The use of text mining methods offers several possibilities to understand the context of any medical act. On a medical database, the process could reveal unexpected correlation between treatment, and thus, pathology. Moreover, this kind of method could

[1] Université de Lyon, Université de Lyon 1, Laboratoire de Sciences Actuarielles et Financière, Institut de Science Financières et d'Assurances ; ADDACTIS en France
[2] Addactis en France

exploit the refund dates contained in the data, but the tested method using temporality, word2vec, still needs to be improved since the results, even if satisfying, are not as better as the one offered by other methods.

Keywords : unsupervised learning; clustering; health insurance; word embedding; Nonnegative Matrix Factorization; Self Organizing Map

1. Introduction

Healthcare circuits are a keystone of the medical system in most countries. They represent the succession of treatments every patient goes through. Thus, they tell the story of the patient’s medical history : previous and current diseases, the way they have been treated, the effects on the patient’s life… Obviously, the current patient’s health status depends on past cares and on their efficiency. Thus, studying healthcare circuits could improve the understanding of the reasons for a treatment failure, for example.

It could also indicate how a patient should be treated in the future. For instance, the success of a tertiary prevention action could depend on the different steps the patient has gone through. Thus, the healthcare circuit seems to be a useful input for an algorithm seeking to target a custom prevention plan based on specific patient needs.

Targeting such a prevention plan is a major challenge for health insurers. By improving policyholders’ health, they improve their reputation and reduce their health costs efficiently. However, they are not aware of the health status of their policyholders. They are only aware of a crude vision of the healthcare circuit, due to every health expense’s being refunded.

Thus, health insurers offer a prime example of how healthcare circuit data could be used for medical purposes, since they barely possess other data due to regulations (such as the GDPR).

A way to use this data is to create clusters of people. Classifying is a common practice in the medical field since formulating a diagnosis is already a classification task (e.g. **[1], [2]**, **[3]**). It consists in finding homogeneous groups underlying a given population, in order to later compare two populations or target as a specific cluster for appropriate care.

A clustering method based on the healthcare circuit in the insurance industry has already been proposed in **[4].** It is based on a two-steps method : the dimension is first reduced using Non-negative Matrix Factorization (NMF), then individuals are clustered using a self-organizing Kohonen's map. However, this method suffers some limitation: for instance, it cannot create a pregnancy cluster.

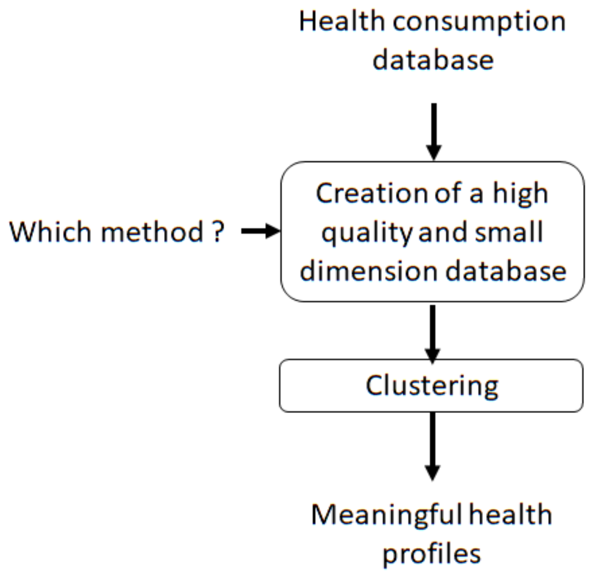

Health consumption database
Which method ?
Creation of a high quality and small dimension database
Clustering
Meaningful health profiles

Figure 1 : The main stages of the process

The goal of this article is to challenge the NMF method with two other unsupervised methods inspired by text mining algorithms, based respectively on Word2Vec word embedding (W2V) and autoencoders (mSDA) methods, as proposed by Baldassini and Serrano **[5].** Both improve cluster quality. Like the NMF method, mSDA offers meaningful medical acts clusters. Moreover, W2V can take into account the timeline of the healthcare circuit to reduce dimensions. Methods are compared using cluster consistency and stability. In order to quantify the stability, a new similarity measure between two clusters is proposed.

2. Materials and methods

2.1 Dimension reduction methods

Dimension reduction makes it possible to identify patterns in a dataset, and thus improves clustering quality by allowing to deal with the dimension curse. These methods have been particularly successful at studying languages: dimension reduction methods are often able to understand general word contexts. It could also improve spectacularly the quality of a clustering method [4]. In this section, we present three dimension reduction methods: NMF, W2V and mSDA. Since these methods are mainly used in the native language processing field, we assume in this section that we work with a text corpus of *m* different texts counting *n* different words. The reader could refer to Simpson and Demner-Fushman **[6]** for a survey of text-mining applications in biomedicine.

Before explaining these methods, we need to introduce a common concept in text mining: frequency matrices. Given a text corpus, a frequency matrix *A* counts every occurrence of each word for each text. These matrices are usually very large non-negative sparse matrices. Dimension reduction methods aims to find a matrix *A'* of dimension *m*k*, with *k<n*, minimizing the loss of information from the matrix *A*.

2.2.1 Non-negative Matrix Factorization

Lee and Seung **[7]** proposed a method for factorizing the matrix *A* into a produce of two matrices *W* and *H* with *m* (resp. *k*) rows and *k* (resp. *n*) columns. Since it is not always possible to find *W* and *H* such as $WH = A$, NMF usually amounts to finding *W* and *H* minimizing *d(WH,A)*, with *d()* a function measuring differences between two matrices.

The matrix *H* is the basis of the reduced dimensional space. It can be seen as a fuzzy word clustering: it represents the frequency of apparition of word pairs in the same text. Since this basis is not orthogonal, it makes it possible for a word to belong to two different clusters, and thus takes into account the different meanings of a word. Matrix *W* is the projection of matrix *A* in the reduced dimensional space. *W* shows if a text is well represented by each word cluster created this way.

NMF has been used both in text mining **[8]**, **[9]** and in medicine **[10], [11].**

Since it has performed better than other algorithms in the field of insurance **[4],** the snmf/l algorithm[3] as proposed by Kim and Park has been used **[12].**

2.2.2 Word2Vec

W2V is a neural network proposed by Mikolov et al. **[15]**. It aims to create a word vector space of k dimensions, called word embedding, capturing each component of each word and

[3] The R package "NMF", developed by Gaujoux and Seoighe has been used to obtain results of this article **[13].**

allowing operations on words[4]. This vector space can be seen as a reduction in k dimensions of the initial space. W2V has been used in many fields, such as medical literature study **[16], [17]** or text clustering **[18].**

We implemented the Continuous Bag of Word (CBOW) architecture for this article. It is a neural network with one hidden layer of k neurons. Contrary to mSDA and NMF algorithms, CBOW does not use frequency matrices and focuses on the context of each word: given a studied word, CBOW takes the neighbors of this word as an input.

In order to counter the imbalance between the rare and frequent words, Mikolov et al. proposed to improve the model by subsampling based on the overall word frequency **[19]**.

The neural network is trained to guess the studied word by only knowing its context. Usually, it performs badly at this task. However, the weights of the hidden layer of neurons form the desired word embedding.

2.2.3 Marginalized stacked denoising autoencoders

Autoencoders are usually a kind of neural network trying to compress then reconstruct an input. To do so, the compression must be as efficient as possible. Thus, autoencoders make it possible to capture relationships between data features and to create a reduced dimensional space.

The method called marginalized stacked denoising autoencoders (mSDA) has been proposed by Chen et al. in order to speed up traditional algorithms **[20]**. Denoising autoencoders work

[4] As an example, Mikolov et al. wrote that in that kind of space: King – Man + Women = Queen.

the same way as classical autoencoders, except that the input is deliberately corrupted. For example, it is possible to set each feature to *0* with probability *p*. Chen et al. pointed out that for this kind of noise, the law of large numbers makes it possible to explicitly determine the result of an infinite amount of corrupted input. Thus, mSDA is a deterministic method and only depends on two parameters: the probability of noise p and the number of stacked autoencoders.

2.3 Clustering with self organizing maps

2.3.1 Self organizing map

Self Organizing Maps[5] (SOM) are a clustering neural network architecture popularized by Kohonen **[21]**. Its clustering quality is similar to k-means algorithms one, with the added benefit of a natural result visualization. However, results are very sensitive to the input dimension, which justifies the need for a previous dimension reduction step **[23]**.

In this neural network, each neuron is arranged following a given topology (for example, a two-dimensional hexagonal grid). It is thus possible to define a neighborhood for each neuron, using a so-called neighborhood function *h*.

Suppose that one wants to cluster *M* individuals into a *m* neurons SOM. Each neuron *i* is initialized with initial random weights w_i^0. The first individual *k* is then presented to each neuron, in order to determine the neuron with the weights closest to the individual features,

[5] The R package ”Kohonen”, developed by Wehrens et al. has been used to obtain the results of this article **[22]**.

called the Best Machine Unit (BMU). Then the BMU weights, but also neighboring neurons' weights are adjusted, in order to create an area which attracts similar individuals. The process is then iterated.

Once the map is trained, we reduce the cluster number by proceeding to a hierarchical clustering **[24]**.

2.3.2 Projection

The NMF method results in two matrices, one of them being the data projected into a reduced dimensional space. However, W2V only provides an embedding, which can be seen as the basis W of the reduced dimensional space. Thus, it is still needed to project the inputs on this space.

In order to do that, we first normalize the basis H. We also normalize A row by row. We then project all individuals by computing $A' = AH$. This is the classical way to project individuals using the Euclidean distance. However, since both the basis and A have been normalized, this is equivalent to projecting individuals using the cosine similarity.

Moreover, mSDA does not directly give an embedding. Instead, it results in a matrix Z with m rows and m+1 columns (the last column being called the bias). To compute the embedding, we have to drop the bias and to take the absolute value of this matrix. This way, we can use a NMF[6] in order to reduce this matrix into a matrix with k rows and m columns, which play the role of H.

[6] PCA and SVD have also been tested, resulting in worst results. Also, it is possible to duplicate the rows of the matrix Z a few times with a very small noise, which makes it possible for NMF to find a better correlation between medical acts and result in a better embedding.

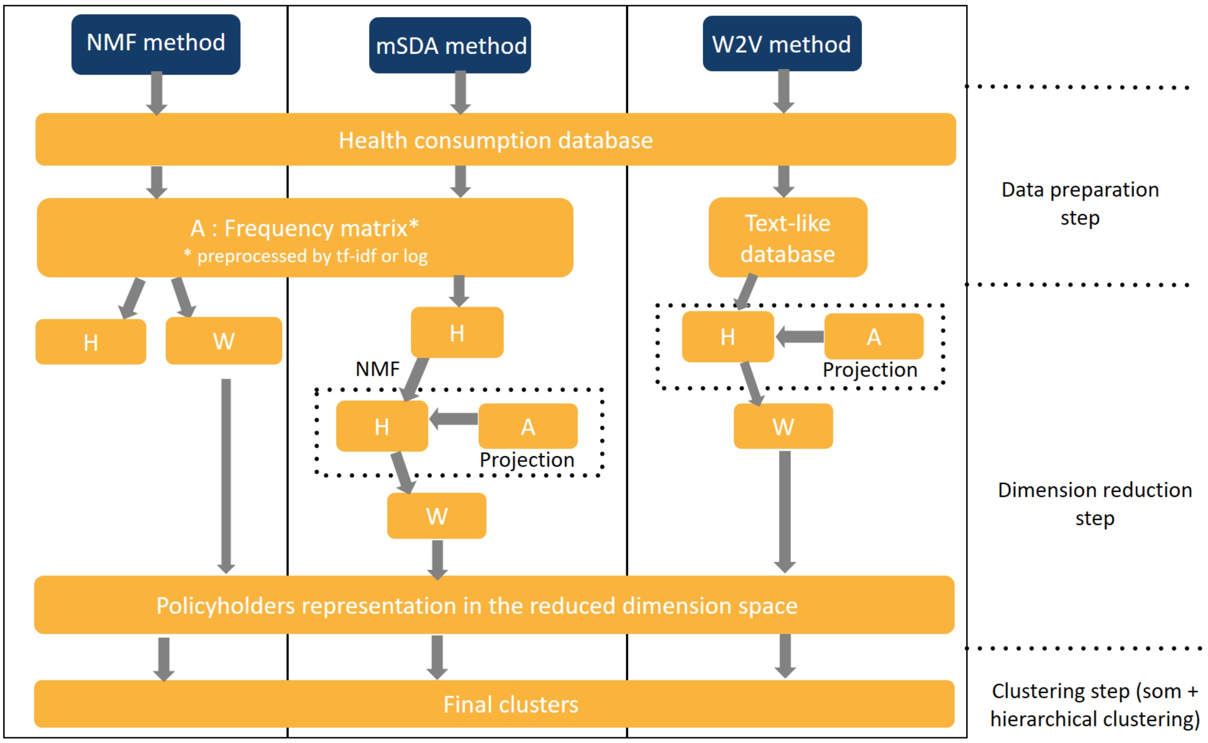

NMF method
mSDA method
W2V method
Health consumption database
A : Frequency matrix*
* preprocessed by tf-idf or log
Text-like database
H
W
H
NMF
H
A
Projection
W
H
A
Projection
W
Policyholders representation in the reduced dimension space
Final clusters
Data preparation step
Dimension reduction step
Clustering step (som + hierarchical clustering)

Figure 2 : simplified display of tested methods.

2.4 From health insurance data to text

A health insurer usually possesses two databases. The first one groups global information about all policyholders, such as age or sex. Due to European data regulation, this database often does not contain a lot of features. For the purpose of focusing only on healthcare circuit, we use this database only to analyze results, and not for clustering.

This database is joined to a second one: the health consumption database. This database contains all the information needed by the health insurer to refund policyholders. It contains the date, the cost and the nature of all policyholder health expenditures (called medical act).

It is then possible to transform these data into a database similar to a text mining dataset. Each policyholder becomes a text, where words are the reimbursed medical acts, and where text order is given by the chronological order of each medical act. For example, a policyholder using drugs on the 01/02 and hospitalization on the 05/05 becomes the text “drugs hospitalization”.

As in text mining, some medical acts are much more common than others (for example, pharmaceutical spending), and a given medical act could carry numerous meanings (e.g. ultrasound can be used both for maternity and cardiology).

This dataset does not take into account whether two successive medical acts are separated by one day or by one month. To address this point, it is possible to add a “word” capturing the temporal difference. Hence, the previous example can be represented by the text “drugs 3_6_months hospitalization”.

As in text mining, it is still possible to construct a frequency matrix, and then to use NMF or mSDA methods. Since some words are so common that they carry almost no information, text mining frequency matrices are usually transformed using tf-idf method[7]. Since we meet the same problems with drugs, and, to a lesser extent, generalist and specialist practitioners, we also implement this pretreatment, which improves results effectively. An example of application of tf-idf for text classification in biomedicine can be found in Srivastava et al **[25]**. Also, as shown by Huang, cosine similarity is more adapted to text comparison than the Euclidean distance **[26],** thus, henceforth all presented metrics are based on the cosine similarity.

Database : the database used in these studies comes from a private French health insurer. It is composed of 28,540 women aged between 16 and 62 years. Most of them are working women. There are more than 1,300,000 medical acts observed over one year, distributed

[7] For the NMF algorithm, taking the logarithm of the frequency matrix can lead to better results [Gau 2019]. Thus, we tested both logarithm and tf-idf preprocessing. We kept the logarithm for the NMF method and tf-idf for W2V and mSDA methods.

into a hundred different medical acts. Women without at least one medical consumption have been removed.

3 Results

3.1 Dimension reduction

The first step of the proposed method consists in creating a reduced dimensional space by obtaining a health consumption embedding. In order to make comparison easier, the chosen dimension of the final space is 20, regardless of the dimension reduction method. By normalizing the embedding by columns, it is possible to see each embedding vector as a Medical Act Clusters (MAC) (see Figure 2).

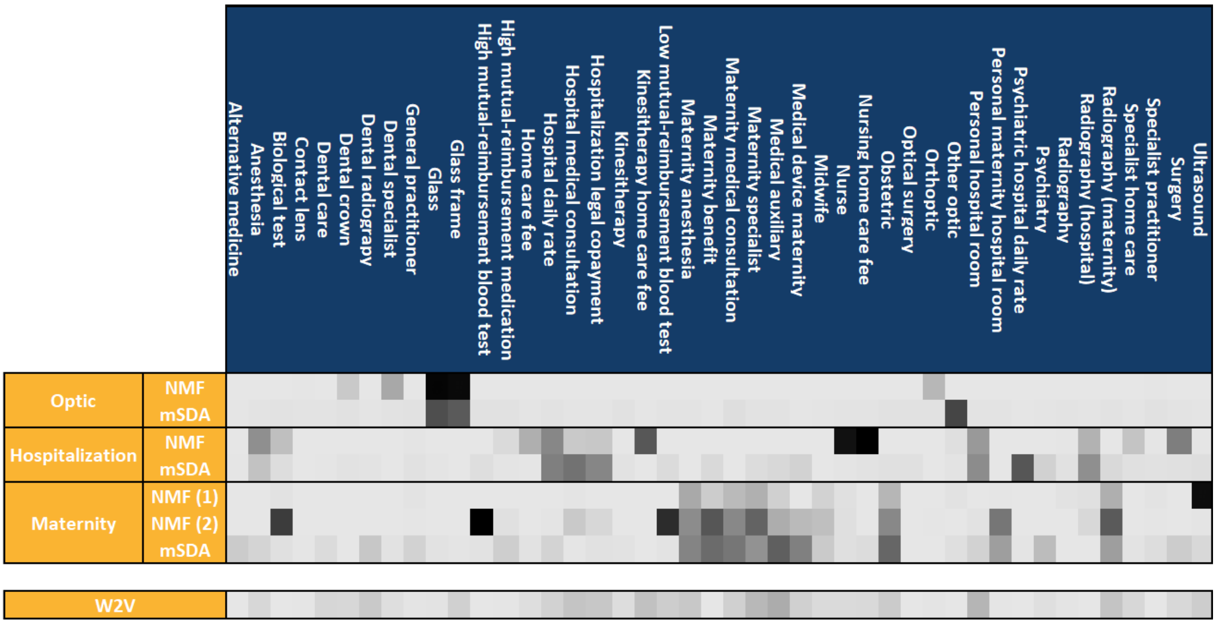


Figure 3 : Dimension reduction leads to Medical Act Clusters. The last line is a vector of the embedding offered by W2V, illustrating that this embedding is not meaningful. NMF(1) and NMF(2) are two medical act classes produced by the NMF method and probably linked to maternity.

First, the embedding obtained with the W2V dimension reduction is not meaningful. It is due to the faculty of W2V to capture nonlinear relationships between medical acts.

On the contrary, MACs obtained with NMF and mSDA are easily understandable. However, MACs obtained with NMF seem to be less accurate than the ones obtained with mSDA, especially for uncommon medical acts, like "Other optic" or "optical surgery" (consumed less than 200 times).

Both algorithms do not focus on the same effects. For example, NMF clusters hospitalization and physical therapy together, whereas mSDA combines psychiatric and classical hospitalization. As for maternity MACs, NMF is not able to dissociate maternity medical acts from biological analysis or ultrasound. Instead, mSDA is able to dissociate maternity from both ultrasounds and biological analysis. However, it is interesting to notice that mSDA puts psychiatry into maternity MAC, this likely being due to postpartum depression.

Finally, notice that for both algorithms, a same medical act can belong to multiple MACs.

3.2 Clustering

The dimension reduction makes it possible to obtain a small number of high quality features before clustering. The SOM method has been chosen to take advantage of its the natural visualization (an example is given in Appendix A). To make the comparison between each algorithm easier, 20 clusters have been done for each classification. For the sake of concision, we call the whole process by the name of "dimension reduction method".

To understand the meaning of a given cluster, it is possible to compute the rate of people consuming these acts for every medical act (Example given in Figure 4).

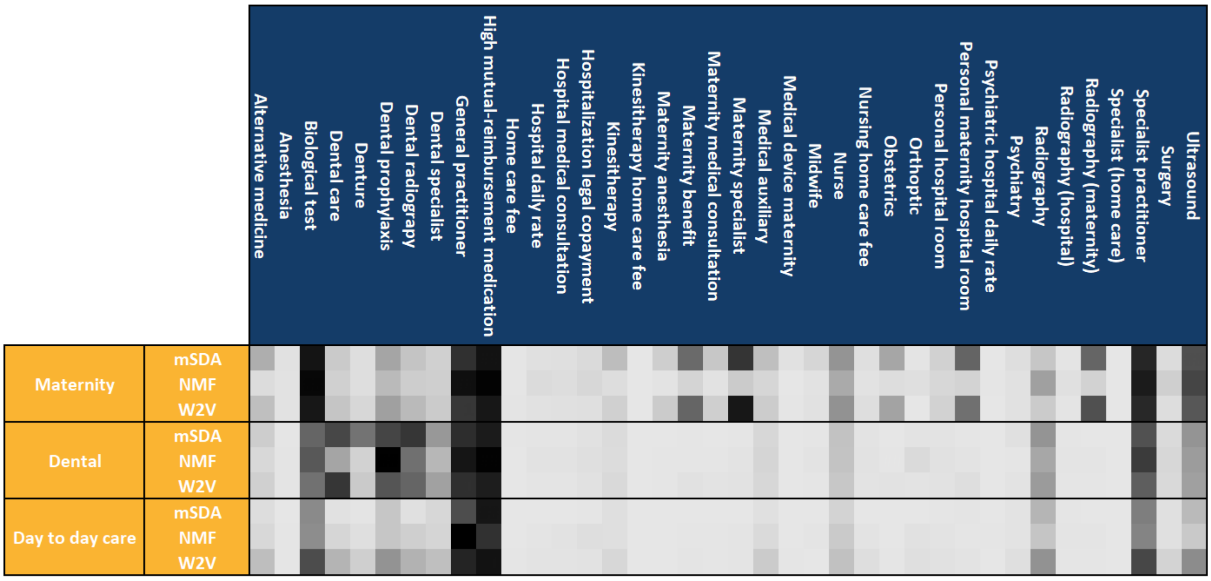


Figure 4 : Comparison of final clusters obtained. A dark square means that most of the policyholders in the associated cluster are consuming the associated medical act.

It is of interest to note that W2V gives understandable final results, even if the embedding is not meaningful. However, the clusters created this way usually seem to be less pure. NMF systematically fails to create a maternity cluster, whereas mSDA rarely creates a dentures cluster. All three methods create several day-to-day care clusters. They also create a huge low-consuming policyholder cluster of around 8 000 people, containing healthy women.

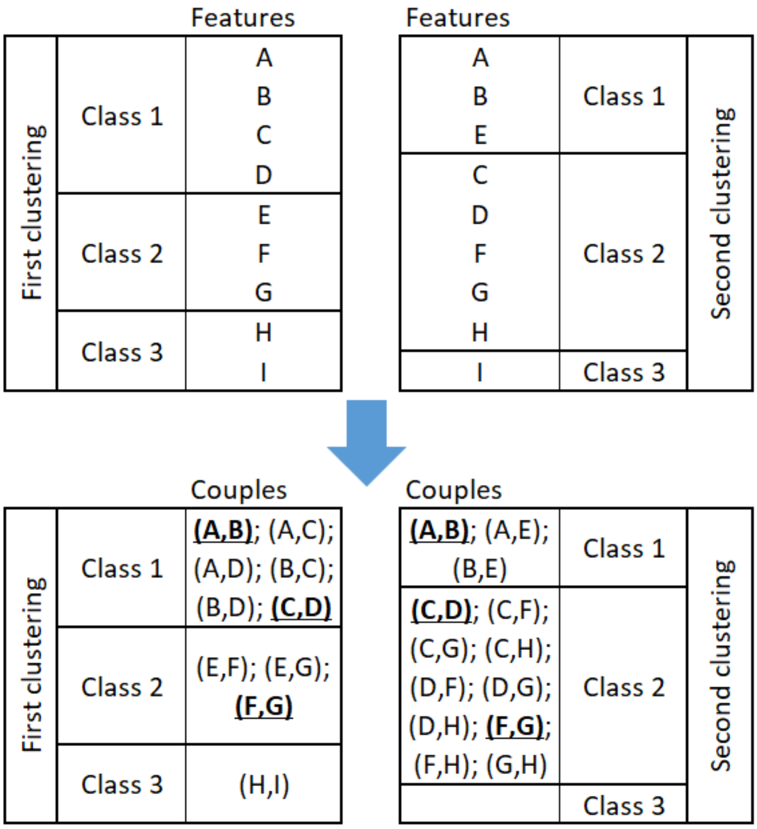
There are two main qualities for a clustering method: clustering quality and stability. Clustering quality are compared using the cluster inertia. Classical similarity measure, such as Gower's distance, cannot be used to compare two partitions. To compare stability, we introduce a new similarity, inspired by the Jaccard coefficient as suggested by Hennig **[27]**. To the best of the authors knowledge, this similarity has not been seen in this form before. It can be used for a totally unsupervised context, with a nondeterministic algorithm such as a Kohonen's map.

Let N be a set of features. Let C_1 and C_2 two partitions in k subsets (clusters) of N. Intuitively, for each couple of features in a subset in C_1, the clustering method is stable if they also belong to the same subset in C_2.

Formally : Let n_i^1 (resp n_i^2) be the cardinal of subset number i in the partition C_1 (resp C_2). The number of possible couple in a subset is given by $\binom{n_i^1}{2}$. Let $n_{i,j}^{1,2}$ be the number of features in the subset number i of C_1 and in the subset j of C_2.

We compute $sim(C_1, C_2) = \dfrac{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{i,j}^{1,2}}{2}}{\sum_{i=1}^{k}\binom{n_i^1}{2}} + \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{i,j}^{2,1}}{2}}{\sum_{i=1}^{k}\binom{n_i^2}{2}}}{2}$. It is easy to verify that sim is a partition similarity.

This way, we can test if two clusters are similar, and thus test if the method is stable. (see Figure 5 for an example).



First clustering		Features
	Class 1	A B C D
	Class 2	E F G
	Class 3	H I

Features		Second clustering
A B E	Class 1	
C D F G H	Class 2	
I	Class 3	

First clustering		Couples
	Class 1	**(A,B)**; (A,C); (A,D); (B,C); (B,D); **(C,D)**
	Class 2	(E,F); (E,G); **(F,G)**
	Class 3	(H,I)

Couples		Second clustering
(A,B); (A,E); (B,E)	Class 1	
(C,D); (C,F); (C,G); (C,H); (D,F); (D,G); (D,H); **(F,G)**; (F,H); (G,H)	Class 2	
	Class 3	

Figure 5 : Two clusterings in 3 classes are made on a set of 9 features (A,B,…,I). It appears that A and B; C and D; F and G are on the same class in the first AND in the second clustering. Thus, $\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{i,j}^{1,2}}{2} = \sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{i,j}^{2,1}}{2}$ = 3. Moreover, $\sum_{i=1}^{k}\binom{n_i^1}{2} = 6+3+1 = 10$ and $\sum_{i=1}^{k}\binom{n_i^2}{2} = 3+10 = 13$. Finally, the similarity beetween the two clusters is $\frac{\frac{3}{10}+\frac{3}{13}}{2}$.

Since the NMF and mSDA dimension reduction are almost deterministic, we have constructed 10 different SOMs. We have also constructed 5 different embeddings using W2V, and each of them were used to construct 5 different SOMs. Two stability measures have been computed for W2V : one comparing all 25 SOMs, and one comparing two SOMs only if they are obtained from the same embedding, in order to test the variability caused by the SOM algorithm alone.

	NMF	mSDA	W2V	W2V (for a given embedding)
Stability mean	47%	**56%**	46%	**58%**
R² mean	35%	**43%**	45%	
R² maximum	36%	45%	**48%**	

Table 1 : clustering quality and stability

Both mSDA and W2V offer better cluster quality than NMF (in terms of inertia). Moreover, mSDA is the most stable process. It is interesting to notice that a big part of W2V's variability comes from the dimension reduction, which has also been observed by Baldassini and Serrano **[5]**.

3.3 Numerical analysis

Analyzing each cluster is necessary for a better understanding of their underlying population. Here, we take the example of clusters mainly described by a high consumption of biological tests. Purity score shows the number of people in the clusters consuming biological tests. mSDA produces a single biological test cluster whereas W2V and NMF produce two biological test clusters.

The mSDA cluster is quite centered on biological test consumption. These profiles could be linked to either chemotherapy control, chronic fatigue or urinary infection. Maternity profiles are contained in another cluster. This is not the case for NMF, which fails to create a maternity cluster. Thus, maternity profiles can be found in the NMF (B) cluster. NMF (A) and W2V (A) are similar to the mSDA cluster, while penalizing optic and denture consumption less. Finally, W2V (B) is probably a cluster of people subject to disease monitoring.

	People having at least one biological test	W2V (A)	W2V (B)	mSDA	NMF (A)	NMF (B)	All basis average
Biological tests	**105**	**154**	**113**	**163**	**112**	**157**	58
Medical device	51	27	65	22	47	31	42
Others	21	19	24	10	25	16	16
Medical auxiliary	12	5	8	4	7	4	8
Surgery	43	15	26	9	47	36	33
Dental	60	57	76	23	71	31	59
General practitioner	103	107	146	73	**211**	117	78
Hospitalisation	54	14	23	10	71	36	40
Kinesitherapy	73	11	22	6	89	6	58
Maternity	43	15	2	6	19	**126**	29
Optic	216	294	286	110	263	32	203
Drugs	174	147	**413**	119	242	173	132
Dentures	129	211	228	11	165	45	126
Psychiatry	28	2	8	2	23	4	23
Radiography	84	76	81	67	102	113	60
Specialist practioner	100	119	110	78	122	135	78
Sum	1296	1272	1631	712	1617	1063	1042
Headcount	15686	971	858	1505	814	1362	28540
Average age	41.5	38.8	47.7	38.5	40.2	39.2	40.8
Purity	100%	90%	86%	100%	80%	97%	

Table 2 : “biological tests” clusters, numerical details. Both NMF and W2V produce two “biological test” clusters.

Conclusion :

Three clustering processes based on the use of the healthcare circuit have been presented and compared. All three processes are in twofold: first the dimension is reduced, and then a clustering is performed using a self-organizing map. The dimension reduction step makes it possible for the self-learning process to understand the correlation between medical acts, and creates high quality features for the clustering step. The method based on the W2V principle offer the best results. However, results offered by the one based on mSDA are closed, but this method is much more meaningful. All methods offer a useful clustering and a more precise understanding of the makeup of a population’s health.

Moreover, a new stability measure has been proposed, in order to compare all three methods. It results that the mSDA method is the most stable methods for our data.

The method has been tested on a health insurance database. Moreover, no other medical information has been added in the clustering process. However, it would be of interest to use them in a more general medical context, such as for studying and clustering low back pain diseases. Adding features to the clustering algorithm could improve results easily and effectively.

Conflict of interest statement : Jean-Pascal Hermet and Romain Gauchon work for ADDACTIS France, an actuarial consulting firm.

Acknowledgements : The authors would like to thank ADDACTIS France for providing the database. This paper was realized within the framework of the Chair Prevent'Horizon, supported by the risk foundation Louis Bachelier and in partnership with Claude Bernard Lyon 1 University, ADDACTIS France, AG2R La Mondiale, G2S, Covea, Groupama Gan Vie, Groupe Pasteur Mutualité, Harmonie Mutuelle, Humanis Prévoyance and La Mutuelle Générale.

References :

[1] : S. Ibrahim, P. Chowriappa, S. Dua, U. Acharya, K. Noronha, S. Bhandary, H. Mugasa, Classification of diabetes maculopathy images using data-adaptive neuro-fuzzy inference classifier, Medical & biological engineering & computing 53 (2015) 1345-1360.

[2] : A.T. Azar, S.A. El-Said, A.E. Hassanien, Fuzzy and hard clustering analysis for thyroid disease, Computer methods and programs in biomedicine 11 (2013) 1-16.

[3] : L.F. Silva, A.A.S. Santos, R.S. Bravo, A.C. Silva, D.C. Muchaluat-Saade, A. Conci, Hybrid analysis for indicating patients with breast cancer using temperature time series, Computer methods and programs in biomedicine 130 (2016) 142-153.

[4] : R. Gauchon, S. Loisel, J.L. Rullière, Health-policyholder clustering using health consumption, HAL (2019).

[5] : L. Baldassini, J.A.R. Serrano, client2vec: towards systematic baselines for banking applications, arXiv preprint arXiv:1802.04198 (2019).

[6] : M.S. Simpson, D. Demner-Fushman, Biomedical text mining: a survey of recent progress, Mining text data (2012) 465-517.

[7] : D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999), 788.

[8] : C. Ding, T. Li, W. Peng, On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing, Computational Statistics & Data Analysis 52 (2008) 3913-3927.

[9] : D. Kuang, J. Choo, H. Park, Nonnegative matrix factorization for interactive topic modeling and document clustering, Partitional Clustering Algorithms (2015) 215-243.

[10] : J.P. Brunet, P. Tamayo, T.R. Golub, J.P. Mesirov, Metagenes and molecular pattern discovery using matrix factorization, Proceedings of the national academy of sciences 101 (2004) 4164–4169

[11] : Z. Chen, A. Cichocki, T.M. Rutkowski, Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer disease, IEEE International Conference on Acoustics Speech and Signal Processing Proceedings 5 (2006).

[12] : H. Kim, H. Park, Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, Bioinformatics 23 (2007) 1495–1502.

[13] : R. Gaujoux, C. Seoighe, A flexible r package for nonnegative matrix factorization, Bioinformatics 11 (2010) 367.

[14] : M.H. Van Benthem, M.R. Keenan, Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems, Journal of Chemometrics 10 (2004) 441–450.

[15] : T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[16] : J.A. Minarro-Giménez, O. Marin-Alonso, M. Samwald, Exploring the application of deep learning techniques on medical text corpora, Studies in health technology and informatics 205 (2014) 584-588.

[17] : L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, P. Bruza, Medical semantic similarity with a neural language model, ACM international conference on information and knowledge management 23 (2014) 1819-1822.

[18] : J. Lilleberg, Y. Zhu, Y. Zhang, Support vector machines and word2vec for text classification with semantic features, IEEE International Conference on Cognitive Informatics & Cognitive Computing 14 (2015) 136-140.

[19] : T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems (2013) 3111-3119.

[20] : M. Chen, Z. Xu, K. Weinberger, F. Sha, Marginalized denoising autoencoders for domain adaptation, arXiv preprint arXiv:1206.4683 (2012).

[21] : T. Kohonen, The self-organizing map, Proceedings of the IEEE 78 (1990) 1464–1480.

[22] : R. Wehrens, L.M. Buydens et al., Self-and super-organizing maps in r: the kohonen package, Journal of Statistical Software 21 (2007) 1–19.

[23] : S.A. Mingoti, J.O. Lima, Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms, European Journal of Operational Research 174 (2006) 1742-1759.

[24] : F. Murtagh, Interpreting the kohonen self-organizing feature map using contiguity-constrained clustering, . Pattern Recognition Letter 1 (1995) 399–408.

[25] : S.K. Srivastava, S.K. Singh, J.S. Suri, Effect of incremental feature enrichment on healthcare text classification system: A machine learning paradigm, Computer methods and programs in biomedicine 172 (2019) 35-51.

[26] : A. Huang, Similarity measures for text document clustering, Proceedings of the sixth new zealand computer science research student conference (2008) 49–56.

[27] : C. Hennig, Cluster-wise assessment of cluster stability, Computational Statistics & Data Analysis 52 (2007) 258-271.

Appendix A : Example of a Kohonen's map obtained following the whole process

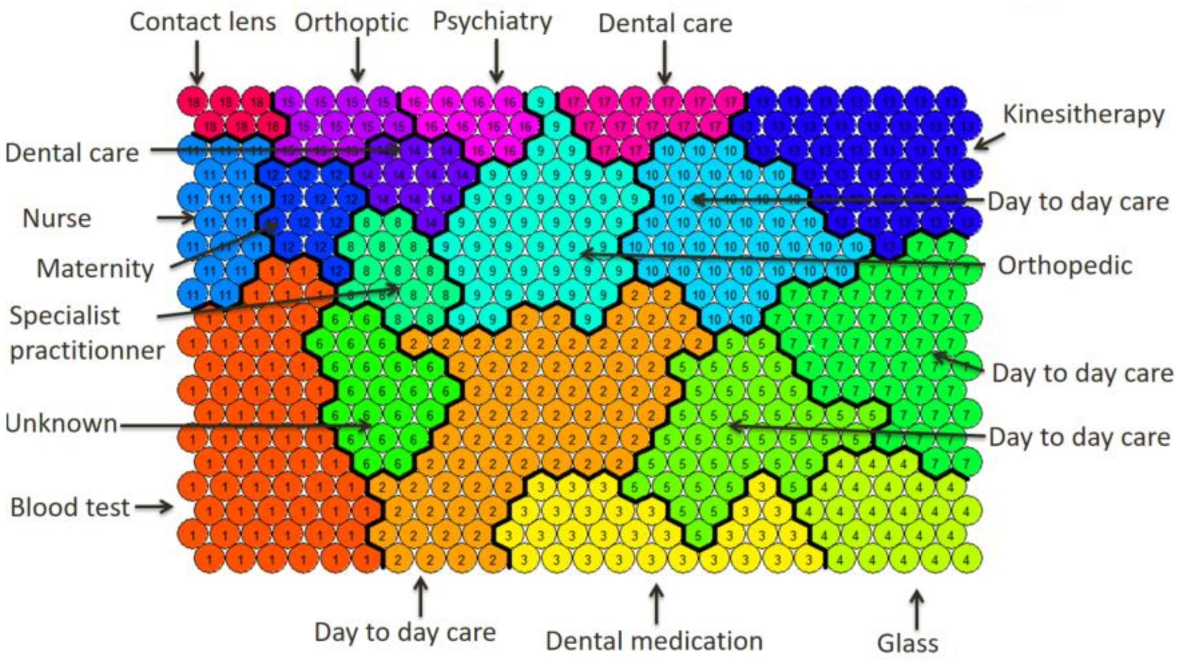


Figure 4 : Example of a Kohonen's map. Data has been preprocessed using the W2V method.