Une méthode basée sur des effectifs pour calculer la contribution des variables à un clustering

Oumaima Alaoui Ismaili[*,***], Julien Salotti[**], Vincent Lemaire[***]

[*]AgroParisTech 16, rue Claude Bernard 75005 Paris
[**]INSA Lyon 20, avenue Albert Einstein - 69621 Villeurbanne
[***]Orange Labs 2 avenue Pierre Marzin 22300 Lannion

Résumé. Cet article présente une étude préliminaire effectuée dans un contexte industriel. On dispose d'une typologie de clients que le service marketing souhaite contacter. Cette typologie est une segmentation des clients en groupes de clients dont les profils seront utilisés pour proposer des campagnes marketing différenciées. La constitution des groupes est réalisée à l'aide d'une technique de clustering qui ne permet pas actuellement de connaître l'importance des variables explicatives (qui décrivent les clients). Cet article propose de résoudre ce problème à l'aide d'une méthodologie qui donne dans notre contexte industriel, l'importance des variables explicatives. Cette méthode sera comparée à certaines méthodes de l'état de l'art.

1 Introduction

Lorsqu'on désire contacter un client pour lui proposer un produit on calcule au préalable la probabilité qu'il achète ce produit. Cette probabilité est calculée à l'aide d'un modèle prédictif pour un ensemble de clients. Le service marketing contacte ensuite ceux ayant les plus fortes probabilités d'acheter le produit. En parallèle, et avant le contact commercial, on réalise une typologie des clients auxquels on propose des campagnes différenciées par groupes. Plus formellement, le problème est celui du clustering supervisé, où un clustering est appliqué sur des données étiquetées. Ce problème peut être défini comme étant un processus de regroupement des individus en clusters, tels que les données de chaque cluster soient les plus similaires possibles et appartiennent à la même classe à prédire. Le lecteur pourra trouver une description détaillée de ce problème dans l'article (Lemaire et al., 2012).

Actuellement, la technique de clustering utilisée pour la constitution des groupes ne permet pas d'identifier les variables les plus importantes. Autrement dit, cette technique ne permet pas de connaître les variables qui contribuent le plus lors de la construction des clusters. Par conséquent, le service marketing éprouve des difficultés à adapter sa campagne aux différents profils identifiés. L'objectif de cette étude est donc de proposer une méthode qui permet de mesurer l'importance des variables à la fin de la convergence d'un clustering. Cette dernière doit prendre en compte trois points principaux :

1. Conserver toutes les variables utilisées lors du clustering.
2. Ne pas réapprendre le modèle.

3. Garder l'espace de représentation des données utilisé lors de la phase de prétraitement (discrétisation pour les variables continues et groupage des valeurs pour les variables catégorielles) qui précède la phase de clustering.

Au vu du contexte d'étude, cet article propose de poser le problème de mesure de contribution des variables comme un problème de classification supervisée. C'est à dire apprendre à prédire l'appartenance aux clusters à partir d'une variable explicative donnée, puis d'ordonner les variables selon leur pouvoir prédictif.

La section 2 de cet article décrit la méthode de clustering utilisée qui contraint le problème de calcul d'importance. La section 3 décrit la solution proposée pour trier les variables en fonction de leur importance dans ce contexte. La section 4 présente des résultats préliminaires avant de conclure au cours de la dernière section.

2 L'existant : la méthode de clustering utilisée

L'ensemble des notations qui seront utilisées par la suite, sont les suivantes :

- Une base d'apprentissage, E, comportant N éléments (individus), M variables explicatives et une variable Y à prédire comportant J modalités (les classes à prédire sont C_j).
- Chaque élément D des données est un vecteur de valeurs (continues ou catégorielles) $D = (D_1, D_2, \ldots, D_M)$.
- K est utilisé pour désigner le nombre de classes souhaitées.

2.1 L'algorithme de clustering

L'algorithme de clustering utilisé est décrit dans (Lemaire et al., 2012). Cet article a montré que si on utilise un algorithme de type k-moyennes à l'aide d'une présentation supervisée et de la norme L1, on obtient des clusters où deux individus proches au sens de la distance seront proches au sens de leur probabilité d'appartenance à la classe cible (voir équation 5 dans (Lemaire et al., 2012)). Cet algorithme peut être présenté de la manière suivante :

- Prétraitements des données (voir section 2.2)
- Pour replicate=1 à R [1]
  - initialisation des centres (voir section 2.3)
  - Algorithme usuel des k-moyennes avec comme centre une approximation de la médiane (Kashima et al., 2008) et la norme L1 (Jajuga, 1987)
- choix de la meilleure "replicate" parmi les R solutions obtenues (voir section 2.4)
- présentation des résultats (voir section 2.5)

Cet algorithme est en partie supervisé puisque les prétraitements et le choix du meilleur "replicate" sont basés sur des critères supervisés qui sont décrits ci-dessous.

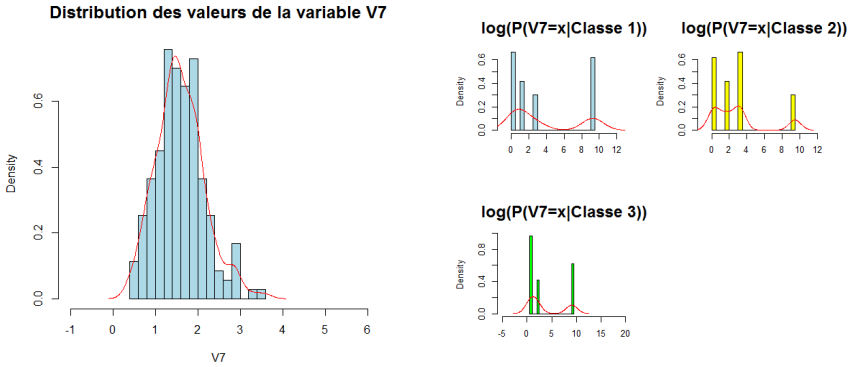
1. Dans cette étude, on fixe le nombre de replicates à R= 50

2.2 Représentation supervisée des données

Une représentation supervisée des données est utilisée. Elle recode les données brutes grâce à une technique de groupage supervisée ou de discrétisation supervisée qui utilise la variable *cible* contenant la liste des classes à prédire.

Les variables continues sont discrétisées (Boullé, 2004), c'est à dire découpées en intervalles, tandis qu'une méthode de groupage est appliquée sur les variables catégorielles (Boullé, 2005). Le prétraitement des données est réalisé à l'aide de l'approche MODL. Cette approche consiste à trouver la partition des valeurs de la variable continue (respectivement catégorielle) qui donne le maximum d'information sur la répartition des classes à prédire connaissant l'intervalle de discrétisation (respectivement le groupe de modalités).

A la fin du processus de prétraitement, les variables numériques et catégorielles sont donc recodées : chaque variable m est recodée en une variable qualitative contenant I_m valeurs de recodage. Chaque objet de données est alors recodé sous forme d'un vecteur de modalités discrètes $D = D_{1i_1}, D_{2i_2}, \ldots, D_{Mi_M}$. D_{mi_m} représente la valeur de recodage de D_m sur la variable m, avec la modalité discrète d'indice i_m. Ainsi les variables de départ sont alors toutes représentées sous une forme numérique. Le vecteur initial contenant M composantes de variables numériques et catégorielles devient un vecteur de $M * J$ composantes numériques : $log(P(D_{mi_m}|C_j))$.



(a) avant le prétraitement de la variable (b) après le prétraitement de la variable

FIG. 1 – *la distribution des valeurs de la variable V7 avant et après le prétraitement*

A titre illustratif, la figure 1 présente la discrétisation d'une variable numérique de la base UCI (Blake et Merz, 1998) "Wine" qui contient 13 variables explicatives et une variable cible à 3 classes ($Y \in \{1, 2, 3\}$). Après le prétraitement, on remarque que la distribution des variables prétraitées est multimodale et non gaussienne.

2.3 Initialisation des centres

L'initialisation des algorithmes de clustering basés sur le partitionnement influence la qualité de la solution trouvée et le temps d'exécution. C'est pourquoi le choix de la méthode d'initialisation de l'algorithme est un choix important lors de l'implémentation d'un algorithme de

clustering. Cependant, il n'y a pas une méthode d'initialisation meilleure que toutes les autres dans la littérature (Meila et Heckerman, 1998) mais plusieurs bonnes méthodes. Parmi ces dernières la méthode nommée K means++ a été utilisée (Arthur et Vassilvitskii, 2007). Cet algorithme est défini comme suit :

1. Choisir un centre uniformément au hasard parmi l'ensemble des points de données E.
2. Pour chaque point D, calculer $S(D)$: la distance entre D et le centre le plus proche qui a déjà été choisi.
3. Choisir le centre prochain $c_i = D' \in E$ suivant la probabilité $\frac{S(D')^2}{\sum_{D \in E} S(D)^2}$.
4. Répéter les étapes 2 et 3 jusqu'à ce que l'on ait placé tous les centres.

2.4 Choix de la meilleure replicate

Afin de prémunir contre le problème lié à l'initialisation et au fait que l'algorithme ne garantit pas d'avoir un minimum global, on exécute l'algorithme de clustering plusieurs fois. On obtient donc un certain nombre de partitionnements différents, dont on souhaite garder uniquement le meilleur. Pour se faire, et puisqu'on est dans le cadre de clustering supervisé, on utilise une mesure de qualité nommée *EVA* qui mesure la qualité d'un clustering supervisé en prenant en considération la variable 'cible'.

EVA mesure le gain qu'une partition établissant un compromis entre le nombre de groupe et la répartition des étiquettes peut apporter par rapport à la partition ayant un seul groupe. Plus formellement, EVA est une description scalaire comprise entre 0 et 1, décrite par la formule suivante : $EVA = 1 - \left(\frac{c(K)}{c(1)}\right)$, où

$$c(K) = log(N) + log\left(\begin{array}{c} N+K-1 \\ K \end{array}\right) + \sum_{k=1}^{K} log\left(\begin{array}{c} N_k+J-1 \\ J-1 \end{array}\right) + \sum_{k=1}^{K} log\left(\frac{N_k!}{N_{k1}! \ldots N_{kJ}!}\right) \tag{1}$$

et où K est le nombre de cluster, N_{kj} est le nombre d'individus du cluster k et de classe j et N_k le nombre d'individus dans le cluster k.

$c(K)$ mesure d'une manière supervisée l'intérêt d'une partition de Voronoi relative à un échantillon. Il quantifie le compromis entre le nombre de groupes de la partition et la distribution de la variable cible, ce qui correspond à un compromis entre complexité du modèle et ajustement du modèle aux données de l'échantillon. D'une manière générale, on cherche à maximiser cette mesure. Cette mesure est détaillée dans(Ferrandiz et Boullé, 2010).

2.5 Présentation des résultats du clustering

A la fin de la convergence de la méthode de clustering, on présente les résultats à l'aide des groupes de modalités et des intervalles créés lors de l'étape de prétraitement, en calculant les effectifs des individus dans chaque groupe de modalités ou intervalle pour chaque cluster. A titre d'exemple, le tableau 1 présente les effectifs des individus dans l'ensemble des intervalles de la variable V7 de la base Wine pour les trois clusters.

		id-cluster			
	Intervalle / Groupe de modalités	Cluster 1	Cluster 2	cluster 3	Total
V1	...	...	...	...	...
	...	...	...	...	...
	...	...	...	...	...
...					
V7	$]-\infty ; 0.975]$	0	1	38	39
	$]0.975 ; 1.575]$	0	13	10	23
	$]1.575 ; 2.31]$	1	38	0	39
	$]2.31 ; +\infty]$	58	19	0	77
...					
V13	...	...	...	...	...
	...	...	...	...	...
	...	...	...	...	...
	...	...	...	...	...

TAB. 1 – *Discrétisation de la variable V7*

3 Choix d'une méthode de tri adaptée au contexte

3.1 Contribution d'une variable

Dans la littérature, plusieurs indices de qualité de clustering ont été développés afin de mesurer la contribution d'une variable au résultat d'un clustering. Cette problématique de mesure de contribution, de mesure d'importance, dans un clustering peut être divisée en deux sous-problèmes que l'on peut respectivement caractériser de *global* ou *local*. L'importance *globale* a pour but de mesurer l'impact que la variable a eu sur la structure entière du partitionnement et non pas l'impact qu'elle a eu sur un cluster en particulier. Par contre, l'importance *locale* a pour objectif de savoir quelle variable a été déterminante dans la formation d'un cluster en particulier. Nous nous intéressons dans cet article uniquement à l'importance globale.

Parmi les méthodes de l'état de l'art permettant de mesurer cette importance on trouvera de nombreux indices tels que : (i) l'indice de Dunn (Dunn, 1974) ; (ii) l'indice de Davies-Bouldin (DB) (Davies et Bouldin, 1979) ; (iii) l'indice Silhouette (Rousseeuw, 1987) ; l'indice SD (Halkidi et al., 2000) ; l'indice S_Dbw(Halkidi et Vazirgiannis, 2001) ...

La plupart de ces méthodes utilisent le théorème de Huygens et la décomposition de l'inertie totale en la somme de l'inertie intra cluster et de l'inertie inter cluster. La contribution d'une variable est alors, par exemple, calculée en mesurant la valeur de l'inertie inter calculée uniquement avec cette variable vis-à-vis de la somme des inerties inter calculée sur toutes les variables ((Benzécri, 1983), (Celeux et al., 1989) section 2.10 p154-164).

3.2 Notre proposition

Notre but est l'ordonnancement du tableau 1 selon la contribution des variables à l'affection des clusters. Nous pensons que dans le cadre de notre contexte et de nos prétraitements les critères classiques tel que ceux présentés ci-dessus ne sont pas totalement adaptés. La figure 1 montre par exemple que pour la base de données Wine la distribution de départ de la variable V7 (partie gauche de la figure) devient après prétraitements « multimodale » (partie droite de la figure).

Nous décidons alors de poser le problème comme un problème de classification supervisée. Le but sera d'essayer d'apprendre à prédire le cluster d'appartenance d'un individu (l'id-cluster du tableau 1) en utilisant une seule variable (classification univariée). Puis de trier les variables selon leur pouvoir prédictif vis-à-vis de l'id-cluster.

Comme on désire trier les variables selon le résultat de clustering initialement obtenu on s'interdira les classifieurs qui créent une nouvelle représentation des données. En effet on ne souhaite pas mesurer l'importance des variables dans un nouvel espace mais l'importance des variables avec la représentation supervisée obtenue juste avant la création des clusters. Le but est l'aide à l'interprétation du clustering de manière à permettre à l'analyste de concentrer son attention sur les variables les plus importantes vis-à-vis du clustering obtenu.

Parmi les méthodes capables d'utiliser la représentation issue de nos prétraitements supervisés et le tableau d'effectifs qui sert à présenter les résultats on choisit d'utiliser la méthode MODL qui mesure le pouvoir prédictif (appelé "level") d'une variable numérique dans (Boullé, 2004) et le pouvoir prédictif d'une variable catégorielle dans (Boullé, 2005).

Dans le cas d'une variable numérique [respectivement catégorielle] si les intervalles de discrétisation [les groupes de modalités] sont fixés, alors le critère se calcul à l'aide des effectifs observés dans les intervalles [groupes de modalités]. Nos prétraitements supervisés nous donnent les intervalles [les groupes de modalités] et la projection des individus sur les clusters (tableau 1) nous permettent d'avoir en notre possession les effectifs. L'ensemble des éléments nécessaire au calcul du level par variable est donc disponible pour toutes les variables explicatives.

4 Expérimentations

4.1 Jeu de données utilisé

Pour évaluer le comportement de notre nouvelle approche en termes de tri des variables selon leur importance, des tests préliminaires ont été effectués sur les bases de données suivantes (Blake et Merz, 1998) :

- Wine : Cette base contient les résultats d'une analyse chimique des vins produits dans la même région en Italie, mais provenant de trois cultivateurs différents (trois classes à prédire). Elle est constituée de 178 données caractérisées par 13 attributs continus.
- Letters : Cette base est constituée de 20000 données caractérisées par 16 attributs et 26 classe à prédire.
- Iris : Cette base est constituée de 150 données caractérisées par 4 attributs continus et trois classe à prédire.

4.2 Algorithme utilisé pour comparer les mesures d'importance

Une bonne mesure d'importance doit permettre de trier les variables en fonction de leur importance. Les moins bonnes de ces variables ne contiennent pas, ou peu d'information utile à la formation des clusters. Le résultat d'un clustering sur le jeu de données privé de cette variable, et donc sa qualité, devrait rester sensiblement identique, ou même être légèrement meilleur (moins de bruit). Inversement, le retrait d'une variable importante, priverait l'algo-

rithme d'une information importante pour former les clusters produisant alors un clustering de moins bonne qualité.

On définit alors un algorithme simple qui nous permet de recueillir les informations pour comparer les différentes mesures d'importance :

1. exécuter l'algorithme de clustering afin d'obtenir un premier partitionnement.
2. trier les variables selon leur importance, à l'aide de la méthode de tri que l'on souhaite tester.
3. exécuter l'algorithme de clustering afin d'obtenir un nouveau partitionnement.
4. estimer la qualité de ce partitionnement à l'aide des critères EVA et AUC.
5. retirer du jeu de donnée la variable la moins importante, d'après le tri effectué en 2.
6. réitérer à partir de l'étape 3, jusqu'à un critère d'arrêt (par exemple, toutes les variables ont été retirées).

On peut alors tracer la courbe des valeurs d'EVA (respectivement AUC (Fawcett, 2004)) en fonction du nombre de variables. L'examen des résultats peut alors être fait visuellement en observant l'évolution de la courbe des valeurs d'EVA (respectivement AUC) et/ou en calculant l'aire sous la courbe des valeurs d'EVA (respectivement AUC) (ALC = Area Under Learning Curve (Salperwyck et Lemaire, 2011)). Plus l'ALC est élevée plus la méthode de tri est de bonne qualité.

4.3 Les méthodes de tri implémentées

Nous avons listé dans la section 3.1 plusieurs indices permettant de trier les variables en fonction de leur importance dans un clustering. Pour des raisons de temps et de coût d'implémentation, à ce jour deux indices ont été implémentés à savoir Davies-Bouldin (BD) (Davies et Bouldin, 1979) et SD (Halkidi et al., 2000). Dans cette section ces deux indices seront comparés à notre approche présentée dans la section 3.2.

4.4 Résultats

Level	V3	V8	V5	V4	V2	V9	V11	V1	V6	V13	V10	V12	V7
Indice-DB	V4	V2	V13	V11	V3	V1	V10	V8	V9	V5	V12	V6	V7
Indice-SD	V4	V5	V3	V2	V13	V8	V11	V9	V1	V6	V10	V7	V12

TAB. 2 – *Tri des variables (de la moins importante à la plus importante) à l'aide des trois méthodes*

Le tableau 2 présente à titre illustratif l'ordonnancement des variables en fonction de leur importance dans le clustering obtenu, pour la base Wine, à l'aide des trois méthodes de tri implémentées. A partir de nos prétraitement [2], les deux dernières méthodes (Davies-Bouldin et SD) calculent pour chacune de ces variables trois valeurs de contribution conditionnellement à la classe à prédire. Cela veut dire qu'une seule variable peut avoir une forte contribution à la

2. Dans le cas du jeu de données wine (cas de 3 classe à prédire C_1, C_2 et C_3), chaque variable est prétraitée de la manière suivante : $log(P(X_i = x|C_1), log(P(X_i = x|C_2), log(P(X_i = x|C_3)$ avec $i \in [\![1, 13]\!]$.

construction des clusters conditionnellement à une classe à prédire et en même temps une faible contribution conditionnellement à une autre classe. Dans ce cas, on définit une contribution d'une variable comme étant la somme des trois valeurs [3]. La méthode proposée (level), est une méthode capable d'utiliser la représentation issue de prétraitement et le tableau des effectifs pour fournir une valeur de contribution par variable.

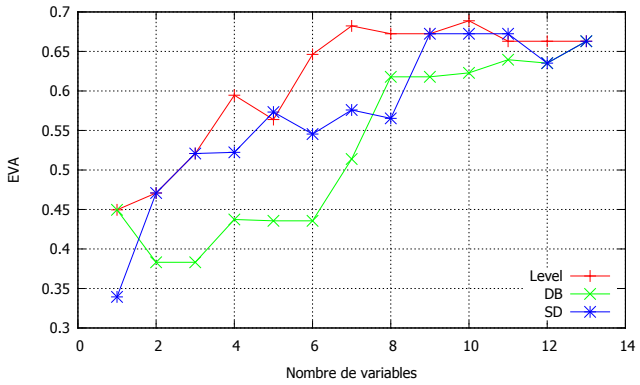


FIG. 2 – *Evolution du critère EVA pour les trois méthodes (K=3)*

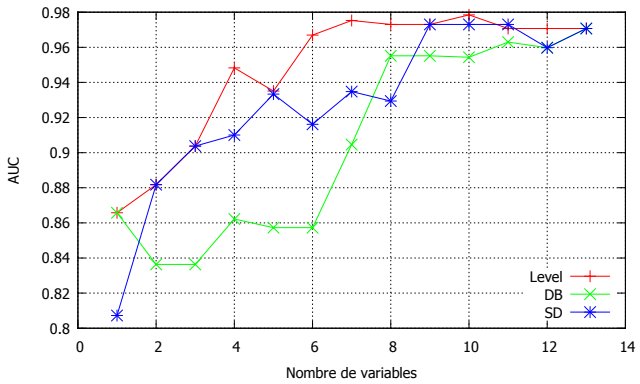


FIG. 3 – *Evolution du critère AUC pour les trois méthodes (K=3)*

La figure 2 (respectivement la figure 3) présente les trois courbes d'EVA (respectivement AUC [4]) à titre illustratif sur la base Wine en fonction de la méthode utilisée.

Le tableau 3 présente quand à lui les valeurs d'ALC pour EVA, l'AUC et l'ACC (le taux de bonne classification) selon le critère utilisé.

3. Une variable considérée comme moins contributrice pour un clustering, doit être retirée entièrement du jeu de données

4. Le critère AUC est donné par la formule suivante : $\sum_{j\in[\![1,J]\!]} P(C_j)AUC(C_j)$ avec J est le nombre de classe. Il permet de mesurer la qualité de la classification en traitant chaque cluster individuellement. Notons que la classe prédite d'un cluster est définit comme étant la classe majoritaire de celui-ci.

L'évolution du critère EVA (respectivement AUC) à l'aide de la méthode proposée est meilleure vis-à-vis de l'évolution des deux autres critères à mesure que l'on retire les variables jugées les moins contributrices pour le clustering obtenu.

		DB	SD	level
Wine	ALC(EVA)	0,5257	0,5714	0,6116
	ALC(AUC)	0,9060	0,9281	0,9472
	ALC(ACC)	0,8574	0,8863	0,9123
Letters	ALC(EVA)	0,3628	0,2871	0,3749
	ALC(AUC)	0,8930	0,8558	0,8952
	ALC(ACC)	0,3475	0,2813	0,3555
Iris	ALC(EVA)	0,6304	0,4571	0,6304
	ALC(AUC)	0,9675	0,9078	0,9675
	ALC(ACC)	0,9350	0,8267	0,9350

TAB. 3 – *Les valeurs d'ALC pour les trois méthodes selon le critère utilisé.*

On remarque également qu'il possible de trouver un nombre restreint de variables produisant la même valeur d'EVA que l'ensemble complet des variables de départ. Par exemple sur la base de données Wine, et à l'aide de la méthode proposée, on aurait pu déterminer un jeu de 7 variables qui auraient produit un clustering supervisé presque de même qualité que celui obtenu à l'aide de 13 variables.

5 Conclusion

Cette contribution a présenté une nouvelle méthode de tri des variables en cours d'élaboration dans notre contexte industriel particulier. Cette méthode trie les variables en fonction de leur importance à la fin de la convergence de notre clustering qui est "supervisé" en partie. Les résultats préliminaires qui ont été obtenus sont encourageants et semblent montrer l'intérêt de la méthode. Néanmoins ces résultats devront être confirmés sur d'avantage de base de données et comparés à un jeu de critère de qualité de la littérature de plus grande taille.

Références

Arthur, D. et S. Vassilvitskii (2007). K-means++ : The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pp. 1027–1035.

Benzécri, J. P. (1983). Analyse de l'inertie intraclasse par l'analyse d'un tableau de correspondance. pp. 351 – 358.

Blake, C. L. et C. J. Merz (1998). Uci repository of machine learning databases. last visited : 01/12/2013, http ://archive.ics.uci.edu/ml/.

Boullé, M. (2004). A Bayesian approach for supervised discretization. In Zanasi, Ebecken, et Brebbia (Eds.), *Data Mining V*, pp. 199–208. WIT Press.

Boullé, M. (2005). A grouping method for categorical attributes having very large number of values. In P. Perner et A. Imiya (Eds.), *Proceedings of the Fourth International Conference*

on Machine Learning and Data Mining in Pattern Recognition, Volume 3587 of *LNAI*, pp. 228–242. Springer verlag.

Celeux, G., E. Diday, G. Govaert, Y. Lechevallier, et H. Ralambondrainy (1989). *Classification automatique des données*. Dunod.

Davies, D. L. et D. W. Bouldin (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-1*(2), 224–227.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics 4*(1), 95–104.

Fawcett, T. (2004). Roc graphs : Notes and practical considerations for researchers. *Machine learning 31*(7), 1–38.

Ferrandiz, S. et M. Boullé (2010). Bayesian instance selection for the nearest neighbor rule. *Machine Learning 81*(3), 229–256.

Halkidi, M. et M. Vazirgiannis (2001). Clustering validity assessment : finding the optimal partitioning of a data set. In *Proceedings IEEE International Conference on ICDM 2001*, pp. 187–194.

Halkidi, M., M. Vazirgiannis, et Y. Batistakis (2000). Quality scheme assessment in the clustering process. In D. A. Zighed, J. Komorowski, et J. ?ytkow (Eds.), *Principles of Data Mining and Knowledge Discovery*, Volume 1910 of *Lecture Notes in Computer Science*, pp. 265–276. Springer Berlin Heidelberg.

Jajuga, K. (1987). A clustering method based on the l1-norm. *Computational Statistics & Data Analysis 5*(4), 357–371.

Kashima, H., J. Hu, B. Ray, et M. Singh (2008). K-means clustering of proportional data using l1 distance. In *19th International Conference on ICPR*.

Lemaire, V., F. Clérot, et N. Creff (2012). K-means clustering on a classifier-induced representation space : application to customer contact personalization. In *Annals of Information Systems, Springer, Special Issue on Real-World Data Mining Applications*.

Meila, M. et D. Heckerman (1998). An experimental comparison of several clustering and initialization methods. *Machine Learning*.

Rousseeuw, P. J. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*(0), 53 – 65.

Salperwyck, C. et V. Lemaire (2011). Learning with few examples : An empirical study on leading classifiers. In *IJCNN*, pp. 1010–1019.

Summary

This article presents a preliminary study made in an industrial context. We have a typology of customers that the marketing service want to contact. This typology is a segmentation of customers into groups, whose profils will be used to propose differentied marketing campaigns. The constitution of groups is realised by using a clustering technique which does not currently allow the importance of the variables. This article proposes to solve this problem by using a methodology which gives in our industrial context the importance of variables. This method will be compared with some others methods from the literature.