Analyse en Composantes Significatives

J.C. Risch*

*7 rue Frédéric Clavel - 92287 Suresnes Cedex - France,
jean-charles.a.risch@capgemini.com

1 Introduction

Ces dernières années, le mot clef « Big Data » a beaucoup fait parler de lui. Big Data est synonyme de données volumineuses mais aussi de données avec une très grande dimension. Pour faire face à cette dimension, des méthodes existent : Analyse en Composante Principale (Jolliffe, 2002), Locality Sensitive Hashing(Gionis et al., 1999) et d'autres (Tufféry, 2012). Cependant ces méthodes sont coûteuses en temps de calcul et pas forcément adaptées au Big Data.

La méthode présentée est une méthode de réduction de la dimension dédiée à un environnement Big Data et qui se base sur l'analyse des variables significatives. La méthode est à base de seuil, ainsi elle s'applique aux variables numériques. Cette dernière s'adapte à un regroupement hiérarchique et une visualisation de regroupement à caractéristiques fortes.

2 Réduction de la dimension

2.1 Pré-traitements

Cette méthode utilise un seuil. Ce dernier permet d'affirmer si oui ou non une variable a une valeur significative pour tel ou tel individu. Le seuil utilisé est le même pour l'ensemble des variables. Ainsi, il est nécessaire d'effacer les unités de ces dernières en les centrant et réduisant.

2.2 Méthode par seuil

Le seuil est défini au préalable. Chaque individu est défini par un vecteur V_i à n dimensions numériques. Pour un individu et une variable donnée, si la valeur de la variable est supérieure au seuil ou inférieur à la négation de ce dernier, alors on considère cette variable comme significative pour cet individu. Les variables analysées comme significativement fortes (respectivement faibles) pour un individu sont notées "+NomVar" (respectivement "-NomVar").

Ainsi, pour chaque individu, l'algorithme va construire un nouveau vecteur V_i' composé des noms des variables significatives accompagnés des signes associés. La dimension des nouveaux vecteurs sera variable d'un individu à l'autre.

Le tableau 1 représente une sortie de la méthode de réduction de la dimension.

V'1 :	-Var1	+Var4	
V'2 :	-Var2	+Var4	
V'3 :	+Var1	-Var2	-Var4

TAB. 1 – *Résultat de la réduction de dimension.*

2.3 Construction de la matrice des similarités/dissimilarités

A partir de ces nouveaux vecteurs, nous allons construire la matrice des similarités avec en ligne et en colonne les individus analysés. Nous commençons par initialiser la matrice à 0. Ensuite, pour tous les couples (V'_x, V'_y), à chaque similarité rencontrée (variable en commun ayant le même signe), on ajoute un point au score de similarité correspondant et on soustrait un point pour chaque dissimilarité (variable en commun n'ayant pas le même signe). Ainsi, les valeurs des cellules de la matrice correspondent au score de similarité entre les individus. Plus le score est fort, plus les individus concernés sont proches et inversement.

3 Regroupement et visualisation

Une fois la matrice des similarités construite, il est aisé de construire un arbre hiérarchique et ainsi regrouper les individus en fonction de leurs similarités et dissimilarités.

A chaque étape de l'algorithme de regroupement hiérarchique, il est possible de retrouver quelles sont les variables significatives décrivant chacun des groupes. Ainsi, cette méthode s'adapte aux méthodes de visualisation de regroupement par caractéristiques fortes (Risch et al., 2014).

4 Conclusion

Nous avons proposé une méthode de réduction de la dimension dans le but d'analyser des individus décrits par des vecteurs à très forte dimension. En outre, nous avons montré que cette méthode peut s'appliquer aisément à un regroupement hiérarchique ainsi qu'à une visualisation de regroupement par caractéristiques fortes.

Références

Gionis, A., P. Indyk, et R. Motwani (1999). Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, San Francisco, CA, USA, pp. 518–529. Morgan Kaufmann Publishers Inc.

Jolliffe, I. T. (2002). Principal component analysis and factor analysis. In *Principal Component Analysis*, Springer Series in Statistics, pp. 150–166. Springer New York.

Risch, J.-C., J. Brunet, E. Soulier, et F. Rousseaux (2014). Méthode de visualisation de regroupement statistique à relativement haute dimension. IHM'14, 26e conférence francophone sur l'Interaction Homme-Machine. Poster.

Tufféry, S. (2012). *Data Mining et Statistique Décisionelle. Quatrième édition*. Paris : TECHNIP.

Summary

Significant components analysis is a method for dimension reduction. It is dedicated to very high-dimensional datasets. This method was build to be easily combined to a hierarchical clustering and visualization for significant features clustering.