Détection de groupes outliers en classification non supervisée

Amine Chaibi[*,**], Mustapha Lebbah[*], Hanane Azzag[*],

[*] {prenom.nom}@lipn.univ-paris13.fr
[*]LIPN-UMR 7030
Université Paris 13 - CNRS
[**] Anticipeo
4 bis, impasse Courteline 94800 Villejuif, France

Résumé. Nous proposons dans ce papier une nouvelle méthode de détection de groupes outliers. Notre mesure nommée GOF (Group Outlier Factor) est estimée par l'apprentissage non-supervisé. Nous l'avons intégré dans l'apprentissage des cartes topologiques. Notre approche est basée sur la densité relative de chaque groupe de données, et fournit simultanément un partitionnement des données et un indicateur quantitatif (GOF) sur "la particularité" de chaque cluster ou groupe. Les résultats obtenus sont très encourageants et prometteurs pour continuer dans cette optique.

1 Introduction

Un outlier est un petit ensemble de données, une observation ou un point qui est différent du reste des données. Souvent, les outliers contiennent des informations précieuses sur le processus d'enregistrement et de la collecte de données. Ils peuvent aussi être problématiques, car ils risquent de biaiser les résultats, notamment pour les méthodes basées sur les distances entre individus, c'est pour cela que ces points devraient être étudiés attentivement.

L'analyse des outliers dans la littérature scientifique remonte à longtemps. En effet, les statisticiens se sont intéressés à ce genre de problématique afin de rendre les modèles mieux adaptés à leurs besoins. La détection des outliers dans les séries chronologiques représente un axe important dans ce genre de problématique, c'est ainsi que (Box et Tiao, 1965) ont étudié le changement de niveau dans les séries chronologiques. Le changement de la variance du premier ordre des modèles auto-régressifs a été examiné par (Wichern et al., 1976). L'approche Local Outlier Factor (LOF) apparue dans les années 2000 par (Breunig et al., 2000) reste la plus utilisée dans les modèles qui se basent sur la densité. L'avantage de cette méthode est qu'elle ne fait aucune hypothèse sur la distribution des données. Les auteurs de (Alhasan et al., 2009) ont donné une définition simplifiée de l'approche LOF. En effet, cette méthode consiste à comparer la densité locale d'une observation avec la densité moyenne de ses k-plus proches voisins (*k-ppv*). (Zengyou et al., 2003) ont utilisé LOF au niveau des clusters pour donner de l'importance aux données au niveau local. Le modèle utilisé permet d'affecter une mesure pour chaque cluster (CBLOF) afin d'identifier les outliers.

Les cartes auto-organisatrices aussi appelées carte de Kohonen ou carte SOM sont souvent

utilisées pour la classification et la visualisation dans le but d'analyser des données (Kohonen, 1995). C'est dans cette optique que (Cai et al., 2009) ont proposé une méthode pour l'étude du comportement des outliers dans les cartes auto-organisatrices. Les outliers sont des points anormaux qui ont des valeurs d'attributs significativement distinctes par rapport à leurs voisins. La particularité de la méthode est l'utilisation de la distance Mahalanolis.

Dans cet article, nous introduisons une nouvelle mesure pour qualifier la «particularité» de chaque groupe/cluster. Cette mesure est intégrée et estimée dans un processus d'apprentissage non supervisé. Nous l'appelons par la suite GOF. Pour la validation, nous avons choisi d'intégrer cette mesure aux cartes topologiques. Ceci permet l'apprentissage de la structure des données tout en fournissant un nouveau paramètre GOF. Ce paramètre est basé sur la densité et quantifie la particularité d'une cellule de la carte : plus la valeur est grande, plus le groupe est susceptible d'être un groupe outlier.

2 Modèle proposé : GOF intégré aux cartes topologiques

Soit $\mathcal{D}$ l'ensemble des données x_i d'apprentissage, de taille N, où chaque observation $x_i = (x_i^1, x_i^2, ..., x_i^j, ..., x_i^d) \in \Re^d$. Notre approche propose de faire un apprentissage avec les cartes topologiques tout en détectant les groupes outliers. Nous rappelons qu'un groupe outlier n'est pas nécessairement un groupe aberrant ; il peut être un groupe d'intérêt, de nouveauté...etc. En fait, c'est un groupe qui est dissemblable et qui peut avoir un comportement largement différent du reste des données. Ce type de groupe peut biaiser les résultats comme il peut constituer un échantillon exhaustif.

Le modèle classique des cartes auto-organisatrices se présente sous forme d'une grille possédant un ordre topologique de C cellules. Les cellules sont réparties sur les nœuds d'un maillage. La prise en compte dans la carte de la notion de proximité impose de définir une relation de voisinage topologique. L'influence mutuelle entre deux cellules c et r est donc définie par la fonction $\mathcal{K}(\delta(c,r))$ où $\delta(c,r)$ constitue la distance de graphe entre les deux cellules c et r. Dans notre approche, chaque cellule c de la grille $\mathcal{C}$ est associée à la fois à deux paramètres : un vecteur référent $w_c = (w_c^1, w_c^2..., w_c^j, ..., w_c^d)$ de dimension d et une nouvelle valeur que nous proposons d'appeler GOF (Group outlier Factor). On note par la suite $\mathcal{W} = \{w_c, w_c \in \mathcal{R}^d\}_{c=1}^C$ l'ensemble des référents et par $GOF_c \in \Re$ l'indicateur outlier associé à chaque cellule c. Chaque référent est associé à un sous ensemble de données affectées à la cellule c qui sera noté P_c. L'ensemble des sous ensembles forment la partition de l'ensemble des données $\mathcal{D}$, $\mathcal{P} = \{P_1, ...P_c, ..., P_C\}$. Dans le cas particulier des cartes topologiques, nous proposons de minimiser la fonction de coût suivante :

$$\mathcal{R}(\mathcal{W}, GOF) = \mathcal{R}(\mathcal{W}) + \mathcal{R}(GOF)$$

où

$$\mathcal{R}(\mathcal{W}) = \sum_{i=1}^{N}\sum_{c=1}^{C} K(\delta(\phi(x_i), c))\|w_c - x_i\|)$$

et

$$\mathcal{R}(GOF) = \sum_{i=1}^{N}\sum_{c=1}^{C} K(\delta(\phi(\mathbf{x}_i), c)) \left(GOF_c - \frac{\sum_{\mathbf{x}_j \in P_c} \frac{1}{f_c(\mathbf{x}_j)}}{\frac{1}{f_c(\mathbf{x}_i)}} \right)^2$$

Où ϕ affecte chaque observation x_i à une cellule unique de la carte. La fonction $f_c(x)$ est notre proposition pour l'estimation de la densité des données au niveau de chaque cellule c, qui est définie comme suite :

$$f_c(x_i) = \exp^{-\frac{\|w_c - x_i\|^2}{2\sigma^2}}$$

Cette densité est définie dans notre cas par une fonction de type gaussienne. Le paramètre σ est l'écart type standard entre les données. Le premier terme $\mathcal{R}(\mathcal{W})$ dépend des paramètres $\mathcal{W}$ et estime les référents. Le deuxième terme est le coût $\mathcal{R}(GOF)$ lié à l'estimation des valeurs GOF associées à chaque cellule. L'algorithme d'apprentissage suivant propose une solution pour la minimisation de la fonction coût en utilisant la méthode de la descente du gradient.

Algorithme

Entrées (1) Les données $\mathcal{D} = \{x_i\}_{i=1..N}$. (2) La carte SOM avec C référents initialisés $\{w_c, c = 1..C\}$. (3) t_{max} : nombre maximum d'itérations. (4) Initialisation des valeurs GOF.

Sorties (1) Une partition $P = \{P_c\}_{c=1..C}$. (2) Les valeurs de GOF= $\{GOF_c, c = 1..C\}$ L'algorithme d'apprentissage est constitué de deux phases :

1. Phase de compétition : Affecter une donnée x_i en utilisant la fonction
$$\phi(x_i) = \arg \min_{1 \leq j \leq C} \| x_i - w_j \|^2$$
2. Phase d'adaptation :
   - Mettre à jour les référents w_c de chaque cellule c
   $$w_c(t) = w_c(t-1) - \varepsilon(t) K(\delta(\phi(x_i, c)))(w_c(t-1) - x_i)$$
   - Mettre à jour les valeurs de GOF_c associées à chaque cellule c :
   $$GOF_c(t) = GOF_c(t-1) - \varepsilon(t) K(\delta(\phi(\mathbf{x}_i, c))) \left(GOF_c(t-1) - \frac{\sum_{\mathbf{x}_j \in P_c} \frac{1}{f_c(\mathbf{x}_j)}}{\frac{1}{f_c(\mathbf{x}_i)}} \right)$$
   où $\varepsilon(t)$ est le pas d'apprentissage
3. Répéter les deux phases jusqu'à un nombre d'itérations fixé, $t = t_{max}$.

3 Expérimentation

3.1 Critère de sélection des groupes outliers : "*Scree Acceleration Test*"

Pour détecter le nombre de groupes outliers (GOF), nous avons utilisé un test statistique proposé par (Cattell, 1966) appelé "Scree Test". Ce test permettra une sélection des valeurs GOF d'une manière automatique. L'utilisation de ce test avec notre vecteur de paramètre, consiste à détecter, par exemple, le changement brutal dans le vecteur $GOF = (GOF_1, GOF_2 ..., GOF_j, ..., GOF_C)$. Ainsi, il faudrait détecter la plus forte décélérations. La procédure de sélection est ainsi composée des étapes suivantes :

1. Ordonner le vecteur des $GOF = (GOF_1, GOF_2..., GOF_j, ..., GOF_C)$ en suivant un ordre décroissant. Le nouveau vecteur ordonné est noté $GOF = (GOF_{..}^1, GOF_{..}^2..., GOF_{..}^j, ..., GOF_{..}^C)$; où l'exposant i de GOG^i indique l'ordre.

2. Calculer les premières différences $df_i = GOF^i_{\cdot\cdot} - GOF^{i+1}_{\cdot\cdot}$ et les deuxièmes différences (l'accélération) $acc_i = df_i - df_{i+1}$
3. Chercher le changement brutal 'scree' avec : $\max_i (abs(acc_i) + abs(acc_{i+1}))$

Ce processus permet de sélectionner les composantes se trouvant avant le changement brutal.

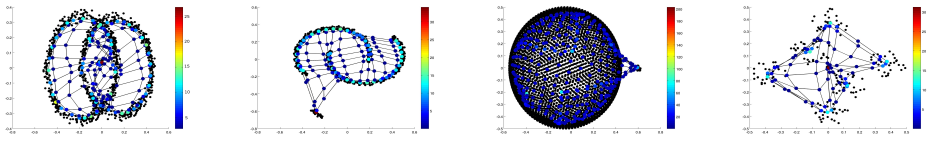
3.2 Validation

Afin de valider notre approche, nous avons utilisé deux types de bases de données, des bases provenant du répertoire UCI Frank et Asuncion (2010) modifiées en créant des difficultés variables et des bases totalement simulées de manière à créer des groupes isolés. Le tableau 1 représente la description des différentes bases ainsi que la taille des cartes utilisées. Le regroupement des données outliers est obtenu quelque soit la méthode de clustering utilisée.

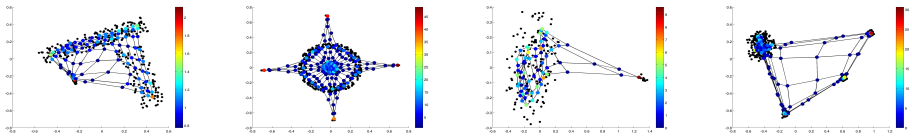
Nom de la base	Nombre de données	Taille de la carte	Nom de la base	Nombre de données	Taille de la carte
anneauxModif	1072	14×12	demicercleModif	638	13×10
HeptaModif	212	9×8	LsunModif	400	11×9
TargetModif	951	13×12	GolfBallModif	4343	19×17
base simulée 1	160	5×13	base simulée 2	234	3×26
base simulée 3	569	8×15	base simulée 4	402	8×13

TAB. 1 – *Description des bases de données.*

Cependant, la nouveauté que nous proposons consiste à estimer une nouvelle mesure au cours de l'apprentissage qui permet de quantifier la "outlierness" d'un groupe (cluster). Cette mesure peut s'intégrer à n'importe quel algorithme de clustering . On a choisi particulièrement le SOM car il permet de faire simultanément du clustering et de la visualisation. Afin de vérifier visuellement les résultats, nous avons projeté les données dans un espace 2D avec les référents de la carte. Plus la couleur est rouge, plus le groupe a une forte valeur GOF.



(a) anneauxModif (b) demicercleModif (c) GolfBallModif (d) HeptaModif



(e) LsunModif (f) TargetModif (g) Simule1 (h) Simule3

FIG. 1 – *Cartes GOF-SOM*

La figure 1 présente les projections de la carte avec la valeur GOF estimée. Nous constatons que les groupes outliers sont clairement visibles et sont associés à des valeurs GOF fortement

indiquées par la couleur rouge. Les référents et le paramètre GOF s'adaptent parfaitement et simultanément avec les groupes isolés. Le tableau 2 présente les résultats obtenus après l'application du Scree Test.

Nom de la base	Nombre de groupes outliers réels	Nombre de groupes sélectionnés avec "Scree Test"	Nombre de groupes outliers sans répétition	Nom de la base	Nombre de groupes outliers réels	Nombre de groupes sélectionnés avec "Scree Test"	Nombre de groupes outliers sans répétition
anneauxModif	1	1	1	demicercleModif	1	1	1
HeptaModif	1	1	1	LsunModif	1	1	1
TargetModif	4	4	4	GolfBallModif	1	1	1
base simulée 1	1	1	1	base simulée 2	2	2	2
base simulée 3	3	5	3	base simulée 4	4	6	4

TAB. 2 – *Détection automatique des groupes outliers.*

Chaque valeur GOF sélectionné représente une cellule outlier. Il existe des cas où plusieurs référents sélectionnés décrivent ensemble le même cluster. Par exemple, dans le cas de la base simulées 3, "Scree Acceleration Test" a sélectionné 5 groupes outliers dont 2 groupes sont des sous ensembles du cluster outlier simulé, les 2 autres appartiennent à un autre cluster et le dernier groupe outlier représente le 3ème cluster. Finalement, les groupes outliers sélectionnés ne détectent que 3 clusters. Nous remarquons que le nombre de groupes outliers réels est toujours égal au nombre de groupes outliers sélectionnés sans répétition, cela confirme nos affirmations de la validation visuelle.

3.3 Evaluation du clustering

Le critère d'évaluation des résultats d'un clustering consiste à comparer la partition calculée avec une partition "correcte". Pour l'évaluation du clustering, nous avons utilisé deux indices classiques : l'indice de pureté et l'indice de Rand. Le tableau 3 représente l'indice de pureté et l'indice de rand calculés sur différentes bases avec notre approche GOF-SOM et l'approche SOM classique dans les mêmes conditions.

Nom de la base	Indice de pureté		Indice de Rand	
	GOF-SOM	SOM	GOF-SOM	SOM
demicercleModif	1	1	0.570	0.570
anneauxModif	1	1	0.574	0.573
LsunModif	1	1	0.643	0.642
TargetModif	1	1	0.674	0.674
Hepta	1	1	0.899	0.904
GolfBallModif	1	1	0.150	0.150
base simulée 1	1	1	0.139	0.137
base simulée 2	1	1	0.262	0.256
base simulée 3	1	1	0.675	0.668
base simulée 4	1	1	0.783	0.768

TAB. 3 – *Indice de pureté et indice de Rand sur GOF-SOM et SOM.*

Nous souhaitons vérifier si le clustering n'est pas perturbé par l'introduction du paramètre GOF au cours de l'apprentissage. Les résultats sur les indices de pureté et rand restent similaires à ceux de SOM, ceci prouve que la qualité du clustering n'a pas été perturbée par GOF.

4 Conclusion et perspectives

Nous nous sommes intéressés dans ce travail au problème de détection de groupes outliers. Nous avons présenté un nouveau paramètre GOF qui se base sur les densités locales des clusters. Ce paramètre a été intégré aux cartes auto-organisatrices. Une série d'expériences ont été réalisées pour valider la méthode proposée. Ceci nous a permis de mieux évaluer notre approche qui s'est avérée prometteuse comme solution au problème de détection de groupes outliers. Les perceptives de ce travail touchent un grand nombre d'étapes de la fouille de données. Nous considérons que la méthode peut être un moyen pour la détéction de nouveautés.

Remerciement Nous remercions Mr. Richard Domps, président directeur général de la société Anticipeo http ://www.anticipeo.fr/ pour le financement de ce travail.

Références

Alhasan, M., V. Chaoji, S. Salem, et J. Zaki (2009). Robust partitional clustering by outlier and density insensitive seeding. *Preprint submitted to Elsevier*.

Box, G. et C. Tiao (1965). *A change in level of a nonstationary t.s*, Volume 52. Biometrika.

Breunig, M., H. Kriege, R. Ng, et J. Sander (2000). Lof: Identifying density-based local outliers. *ACM SIGMOD 2000 International congerence on Management of Data*.

Cai, Q., H.He, et H.Man (2009). Somso: A self-organizing map appoach for spacial outlier detection with multiple attributes. *occedings of International Joint Conference on N.N*.

Cattell, R. (1966). The scree test for the number of factors. *M.B.R 1*, 245–276.

Frank, A. et A. Asuncion (2010). Uci machine learning repository. *Technical report, School of Information and Computer Sciences, available at :http ://archive.ics.uci.edu/ml*.

Kohonen (1995). *Self-Organizing Maps*. Berlin: Springer Verlag.

Wichern, D. W., R. B. Miller, et D. A. Hsu (1976). Changes of variance in first-order autoregressive time series models with application. *Applied statistics 25*, 248–256.

Zengyou, H., X. Xu, et S. Deng (2003). Discovering cluster-based local outliers. *Journal Pattern Recognition Letter 24*, 9–10.

Summary

We propose in this paper a new method to detect groups outliers in unsupervised learning process. A new measure called GOF (Group Outlier Factor) is defined and integrated in Self-organizing Map . Our approach is based on relative density of each group of data and provides simultaneously a partitioning and a quantitative indicator of outlierness (GOF). The obtained results are very encouraging to continue in this direction.