La Revue de Modulad

L'ANALYSE DES DONNEES SUR TABLEAU DE DISTANCE
ou
"QUI SE RESSEMBLE S'ASSEMBLE"

Gérard Thauront
Ingénieur de recherche à l'INRETS
Institut National de Recherche sur les Transports et leur Sécurité
2, avenue du Général Malleret-Joinville, BP 34
94114 Arcueil Cédex
Téléphone (1) 49 86 12 12
EARN : THAURONT@FRIRTS71

Résumé

Dans les programmes classiques d'analyse des données, la "formule" qui permet de calculer la dissemblance entre éléments est écrite à l'intérieur des programmes. C'est généralement une distance euclidienne. Il est cependant possible d'utiliser en entrée de certains programmes d'Analyse des Données un tableau de distances, voire même un tableau de dissimilarités.

L'Analyse des Données sur tableau de distances
ou
"Qui se ressemble s'assemble"

Gérard Thauront
Ingénieur de recherche à l'INRETS
Institut National de Recherche sur les Transports et leur Sécurité
2, avenue du Général Malleret-Joinville, BP 34
94114 Arcueil Cédex
Téléphone (1) 49 86 12 12

Résumé

Dans les programmes classiques d'analyse des données, la "formule" qui permet de calculer la dissemblance entre éléments est écrite à l'intérieur des programmes. C'est généralement une distance euclidienne. Il est cependant possible d'utiliser en entrée de certains programmes d'Analyse des Données un tableau de distances, voire même un tableau de dissimilarités

La notion de distance en Analyse des Données

Lorsqu'un ensemble est classifié[1], on doit trouver dans une même classe des éléments qui se ressemblent, alors que deux éléments de deux classes différentes seront dissemblables.

Ce concept de ressemblance/dissemblance est distinct du concept spatial proche/éloigné, mais lorsqu'on veut modéliser, pour pouvoir "mesurer avec des nombres" la distance de deux points de l'espace, ou la dissemblance de deux individus à classifier, on est amené à utiliser la même structure ou presque la même, à un axiome près (Cf annexe). Ces structures sont appelées distance et dissimilarité par les mathématiciens.

Mais, alors que la distance spatiale possède un étalon universel, et que l'on comprend une phrase telle que : "Ces deux moutons sont éloignés de dix mètres", on ne sait pas exprimer à l'aide d'un nombre la ressemblance de ces deux moutons qui, tous les deux sont blancs avec la tête noire mais l'un des deux a une patte noire.

Remarquons que, dans les cartes factorielles, on identifie ces deux concepts en représentant une similitude par une distance spatiale que l'on peut mesurer en centimètres sur le papier, d'où l'adage utilisé en sous-titre.

---

[1]Nous dirons que l'on fait un classement ou que l'on classe des éléments si on les affecte à des classes déjà existantes, par exemple : classer une lettre dans un dossier, et nous dirons que l'on fait une classification ou que l'on classifie des éléments s'il y a simultanément création du système de classes et affectation des éléments à ces classes

Le choix de la distance

Chaque analyse typologique comporte son propre étalon de mesure, sa propre formule pour chiffrer la dissemblance entre deux éléments. Le choix de cette formule est soit implicite, soit explicite.

Dans les programmes usuels, la formule qui calcule la dissemblance est implicitement définie au travers des choix suivants :

- Le choix des variables actives. C'est l'élément fondamental car de lui dépendra la bonne adéquation de la typologie à la question posée.

- Le choix d'une pondération de ces variables. Cela permet d'affiner le choix précédent, mais peu de programmes proposent cette option.

- Le choix de la "métrique". Ce n'est souvent qu'un choix technique lié au type des variables utilisées. Par exemple : métrique du Chi^2 pour les variables qualitatives, métrique euclidienne simple sur variables centrées réduites pour les variables quantitatives, métrique de Mahalanobis pour les variables bouléennes...

Remarquons que dans tous ces cas, il s'agit de distances euclidiennes.

D'autres programmes travaillent directement sur des tableaux de dissimilarités. On les utilise principalement dans les deux cas suivants :

- Le protocole de recueil des données fournit directement un tableau de dissimilarités, ou bien il fournit des données qui s'agrègent naturellement sous forme d'un tableau de dissimilarités.

- Les distances implicitement fournies dans les programmes usuels ne satisfont pas le statisticien. Celui ci se construit alors explicitement une formule de dissimilarité qui représente mieux la notion de dissemblance qu'il veut utiliser.

Parmi les méthodes pouvant fonctionner directement sur un tableau de dissimilarités, trois existent ou existeront prochainement dans Modulad : la CAH, les Nuées Dynamiques et l'Analyse Factorielle ou Analyse du Triple.

La classification ascendante hiérarchique sur tableau de dissimilarités

Le cas de la CAH est traité ailleurs. Notons simplement que la méthode classique --du moins dans sa version naïve qui n'intégre pas toutes les améliorations qui ont permis de rendre cet algorithme très performant en supprimant un certain nombre de calculs de distances non utilisées, et en limitant l'étendue des tris-- commence par calculer le tableau de distances. Il suffit donc de remplacer ce calcul par la lecture d'un tableau de dissimilarités.

L'analyse factorielle sur tableau de dissimilarités ou analyse du triple

L'analyse du triple sur tableau de dissimilarités pose le problème théorique suivant dans le cas non euclidien. Le calcul d'une analyse factorielle commence

par transformer le tableau de distances en un tableau le produits scalaires. Ce calcul utilise la notion d'angle (plus exactement de cosinus) calculé à l'aide de la formule de résolution du triangle, qui n'est valable que dans le cas euclidien.

Dans le cas de distances non euclidiennes, le concept d'angle passe par celui de géodésiques, qui nécessiterait bien plus qu'un tableau de distances, à savoir la connaissance d'un espace muni d'un champ de courbure.

Exemple simple sur la sphère (géométrie rimanienne) : Tout triangle équilatéral de côté $\pi r/2$ a trois angles droits, alors que si on lui applique formellement la formule de résolution du triangle, on oblient bien sûr des angles de 60°.

Benzecri in [BEN 73] §6.5 pp 86-89 propose d'appliquer l'analyse du triple à des distances euclidiennes entachées d'erreurs.

Or la pratique montre que, même dans le cas de dissimilarités fortement non euclidiennes, l'analyse du triple donne des résultats très satisfaisants et que l'on arrive à interpréter des axes liés à des valeurs propres positives inférieures au module de la plus "forte" valeur propre négative. Par contre, on ne sait pas interpréter les coordonnées imaginaires liées aux axes possèdant des valeurs propres négatives.

Mais ce n'est pas la première fois, dans l'histoire de l'Analyse des Données, que l'application fructueuse précède la justification théorique.

Les Nuées Dynamiques sur tableau de dissimilarité

L'algorithme de Nuées Dynamiques nécessite :

- une notion de dissimilarité pour regrouper les éléments semblables,
- une notion de centre pour calculer le centre des classes,
- une notion d'inertie pour calculer la qualité de la classification obtenue.

Calcul du centre d'une classe

Dans les programmes classiques de nuées dynamiques, le calcul du barycentre des classes se fait au travers d'un calcul de moyenne des différentes coordonnées.

La moyenne, mais également d'autres notions centrales, peuvent être définies directement à partir des dissimilarités. Trois formules sont classiques et correspondent respectivement aux normes l1, l2 et l^{∞} des espaces vectoriels sur R. Je les appelle médiane, moyenne et milieu par analogie au cas mono-dimensionnel.

Cela fournit un point G de l'espace (pas nécessairement unique) auquel ne correspond généralement pas une observation.

	MÉDIANE	MOYENNE	MILIEU
non pondéré	G minimise $\sum_{i\in I} d(Xi,G)$	G minimise $\sum_{i\in I} d^2(Xi,G)$	G minimise $\sup_{i\in I} d(Xi,G)$
pondéré	G minimise $\sum_{i\in I} \mu i\, d(Xi,G)$	G minimise $\sum_{i\in I} \mu i\, d^2(Xi,G)$	Le poids n'intervient pas en ℓ^∞

La définition de l'espace de référence de G n'est pas toujours facile, et dans les cas d'une dissimilarité quelconque qui ne dérive pas directement de coordonnées, il n'est pas possible de calculer l'ensemble des points qui minimise la formule. Par contre il est toujours possible de choisir un élément central parmi les éléments d'une classe.

Ce peut même être plus satisfaisant. Le français moyen est-il un monstre hermaphrodite, ou une certaine femme de taille moyenne ... ?

Cette notion d'élément central existe dans les premiers programmes de nuées dynamiques sous le nom d'étalon.

Le critères de qualité de la partition

On aprécie la qualité d'une partition grâce à la part d'inertie expliquée (inertie interclasse/inertie totale). Mais la fameuse équation :

> Inertie totale = Inertie interclasse (ou expliquée)[2]
> + Inertie intraclasse (ou résiduelle)

n'est vraie (ainsi que le théorème de Huygens utilisé pour sa démonstration) que dans le cas d'espace euclidien. De plus il existe pour un ensemble de points, deux formules d'inerties qui ne sont équivalentes que dans le cas euclidien.

J'appelle respectivement inertie centrée, la première qui se réfère explicitement à un centre, et inertie générale la seconde qui utilise toute les paires d'observations.

	Inertie centrée en G	Inertie générale
non pondéré	$\sum_{i\in I} d^2(Xi,G)$	$1/2N \sum_{i\in I, j\in I} d^2(Xi,Xj)$
pondéré	$\sum_{i\in I} \mu i\, d^2(Xi,G)$	$1/2M \sum_{i\in I, j\in I} \mu i \mu j\, d^2(Xi,Xj)$

Avec N nombre de points et M poids total du nuage.

---

[2] L'inertie intraclasse est la somme des inerties des classes, alors que l'inertie interclasse est l'inertie des centres de classes pondérée par la masse des classes.

Cette situation inhabituelle pose un certain nombre de problèmes dont le plus immédiat est : Par quelle(s) formule(s) permettant de choisir le meilleur résultat, remplacer l'habituel critère du pourcentage d'inertie expliquée par la partition ?

Le programme NUDIS

Le programme NUDIS effectue une classification par la méthode des nuées dynamiques sur des éléments (observations) décrits par un tableau de dissimilarités (euclidiennes ou non), ces observations pouvant être ou non munies d'une pondération.

Initialisation de l'agorithme

La méthode est initialisée soit :

- par le tirage aléatoire d'étalons initiaux,

- par le choix d'étalons initiaux désignés par
  -- leur rang,
  -- leur nom.

Dans ces cas, il faut également fournir le nombre de classes. La phase d'itération commence par la création d'une partition par rattachement des observations à l'étalon le plus proche.

L'initialisation peut également être la lecture d'une partition initiale, auquel cas, le nombre de classes est celui de cette partition, et la phase d'itération commence par le calcul des centres des classes.

L'itération

Le coeur de l'algorithme consiste à boucler jusqu'à la convergence sur les deux phases de calcul suivantes :

- mise en classe des éléments en les rattachant au centre le plus proche,

- calcul du nouvel élément central de chaque classe.

Le programme NUDIS utilise la moyenne pondérée.

Le rattachement des observations au centre le plus proche

Puisque le centre est un étalon, les "calculs" des dissimilarités se réduisent à de simples lectures.

- Remarque : "le centre le plus proche" est une notion invariante par transformation monotone croissante et le rattachement est fait sur les carrés des dissimilarités

Les conditions d'arrêt

Les itérations s'arrêtent soit :

- quand on obtient un point fixe (aucune observation ne change de classe),
- quand on atteint le nombre maximum d'itérations prédéfini.

Bien que ce soit classique dans les programmes, nous avons éliminé l'arrêt sur faible variation relative du critère, car un palier dans la croissance du critère peut cacher une bonne solution.

Le programme NUDIS utilise les formules d'inertie générale pondérée et d'inertie centrée pondérée pour calculer l'inertie totale du nuage ainsi que celles des groupes (et donc l'inertie intraclasse par sommation); dans ces deux cas l'inertie expliquée par la partition est évaluée de façon formelle par différence et non par le calcul habituel.

Les sorties du programme

NUDIS fournit le listage de la classification résultante et les pourcentages d'inertie expliquée par cette partition, ainsi qu'un fichier au standard Modulad pouvant être utilisé dans d'autres programmes de Modulad.

Références bibliographiques

NUDIS est une mise aux normes Modulad et Fortran 77 d'un programme décrit dans la thèse de 3ème cycle de Gérard Thauront du 18 juin 1979.

Pour une présentation générale de la méthode des nuées dynamiques, voir par exemple : E. DIDAY - La méthode des nuées dynamiques (Rev. stat. app., vol. 19, n° 2, pp 19-34. 1971)

[BEN 73] L'Analyse du triple est décrite dans "l'Analyse des Données" de J-P Benzécri et collaborateurs, Tome 2 : L'analyse des correspondances [Repr. Eucl.] & [Repr. Eucl. ex.] pp 64-132.

Voir également l'article "Etude des puissances d'une distance" de S. Joly et G. Le Calvé dans "Statistique et Analyse des Données" 1986 - vol. 11 n° 3 pp. 30-50.

Annexe - Les axiomes

<table>
<tr><td>Dissimilarité</td><td>Distance</td><td>Distance euclidienne</td></tr>
<tr><td colspan="3">Axiome de positivité
$\forall i, \forall j : d(i,j) \geq 0$</td></tr>
<tr><td colspan="3">Axiome de séparabilité
$\forall i, \forall j : d(i,j)=0 \Leftrightarrow i=j$</td></tr>
<tr><td colspan="3">Axiome de symétrie
$\forall i, \forall j : d(i,j) = d(j,i)$</td></tr>
<tr><td></td><td colspan="2">Inégalité triangulaire
$\forall i, \forall j, \forall k : d(i,k) \leq d(i,j + d(i,k)$</td></tr>
<tr><td></td><td></td><td>Caractère euclidien
cf infra</td></tr>
</table>

L'axiome de séparabilité

Dans le cas d'une distance, il est commode de dire que deux individus sont à distance nulle, s'ils ont exactement les mêmes valeurs pour tous les caractères, alors qu'il faudrait dire pour respecter l'axiome de séparabilité, qu'ils sont représentés par le même point d'un espace métrique.

Par contre, dans le cas général d'une dissimilarité, on ne peut se dispenser de l'axiome de séparabilité qu'à la condition d'y substituer la formule suivante (qui, dans le cas des distances est une conséquence de l'inégalité triangulaire) :

$$\forall i, \forall j, \forall k : d(i,j) = 0 \Rightarrow d(i,k) = d(j,k)$$

Le caractère euclidien

Il n'existe pas de formule simple pour qualifier le caractère euclidien d'une distance.

Une dissimilarité entre N éléments est euclidienne si et seulement si il existe N points d'un espace RP (en général P = N-1) tel que :

$$d2(i,j) = \sum_{k=1,P} (x(i,k) - x(j,k))2$$

Une dissimilarité est euclidienne si et seulement si l'analyse du triple ne présente que des valeurs propres positives ou nulles.