Étudiants : *Gautier DRAN*
Clément Spies
Encadrants : *Armelle Brun*
Benjamin Gras

Les moutons noirs : étude de la différence

Présentation générale

Les systèmes de recommandations fournissent des recommandations qui ne sont pas adaptées aux goûts des utilisateurs dit «moutons noirs». Le but de ce projet était donc de définir des hypothèses et de les valider afin d'améliorer les recommandations par rapport à ces utilisateurs atypiques.

Déroulement du projet

Dans un premier temps, il a fallu chercher des méthodes de clustering que nous allions développer.
Nous avons donc utilisé ces algorithmes (k-means, k-médoïdes) afin de regrouper les films.
Une fois ces films regroupés nous avons tenté de répondre à deux hypothèses :

- Les films avec une forte erreur de recommandation sont regroupés dans un ou deux clusters.
- Les films avec une faible erreur de recommandation sont regroupés dans un ou deux clusters.

Nous avons trouvé plusieurs résultats. Dans un premier temps nous avons observé des effets d'absorptions lors de la classification de nos films (un groupe absorbait tous les autres groupes). Ensuite il a fallu fixer un seuil afin de décider quand un film a un fort ou un faible taux d'erreur de recommandation. Pour cela nous avons observé nos données : 25% de nos données se trouvent entre un score de 1.13 et 3.64. Nous avons donc dans un premier temps choisi de prendre un seuil égal à la moyenne entre les deux bornes citées précédemment. Mais seulement 2% de nos données correspondaient à ce critère. Nous avons donc choisi un seuil de 1.5 qui correspondait à 10% de nos données.
Pour le second seuil, 25% de nos données étaient entre 0 et 0.75 nous avons donc placé le seuil à 0.38 qui correspondait à 6% de nos données.
Les essais sur ces données nous ont conduit à valider nos deux hypothèses, nos films se retrouvaient en majorité dans le même cluster.
Pour finir, nous pouvons conclure que nos hypothèses sont valides. Mais si l'on regarde plus précisément nos cluster, on peut voir qu'il y a toujours un cluster très gros et d'autres très petits. Si l'on regarde aussi la répartition des films par rapport aux critères citées précédemment, on a pu observer qu'ils étaient tous dans le même cluster.
Nous avons donc conclu que, d'une part il ne faut pas traiter ces deux «types» de films séparément mais bien voir s'ils sont répartis sur des clusters différents.