Magical Thinking in Data Mining : Lessons From CoIL Challenge 2000 par Charles Elkan

Présenté par : Mirija RAKOTONANAHARY

Résumé

Le data mining a pour objet l'extraction d'un *savoir* à partir de quantités de données, par des méthodes. Cet article discute des résultats et des méthodes lors de la compétition CoIL (Computational Intelligence and Learning) challenge 2000. C'est une compétition d'apprentissage surpevisé pour le fouille de données (data mining).

Contexte

On donne un ensemble de données sur des clients d'une compagnie d'assurance pour prédire les clients qui sont susceptibles d'avoir une assurance caravane et de dire pourquoi?

Discussion

L'auteur commence sur les algorithmes utilisés lors de la compétition – le bayésien naïf, la retropropagation des réseaux de neurones, les cartes autoorganisatrices... Il souligne que les deux premiers lors de la compétition, le sien qui est premier, ont utilisé un classifieur bayésien naïf.

Un classifieur bayésien naïf est plus performant si on ajoute de nouveau attribut, qui est dérivé de la combinaison des attributs existants. Les valeurs des attributs ont été discrétisé par les organisateurs du concours en classe.

Il introduit de nouveaux attributs (Elkan et al. 1997) dans les données qui sont le produit croisé d'attributs et qui apporte plus d'information.

Il met en évidence les corrélations entre la question et l'un des faits suivants :

- avoir 2 assurances voiture
- avoir une assurance voiture et n'est pas dans une classe inférieure à 5
- haut niveau d'achat (niveau 5)
- avoir une troisième assurance
- avoir une assurance bateau
- avoir une sécurité sociale et avoir un haut niveau assurance feu (niveau 4)

Ensuite il parle du problème au cours de l'apprentissage, l'overfitting – surapprentissage – qui peut être dû à la formule de la variance (Silvey et al. 1975).

Il n'y a pas de méthode clair pour la sélection d'attribut sur lesquels on va baser les algorithmes de classification pour tous les participants du concours. Ils utilisent tous un heuristique.

Il analyse les résultats des méthodes de classification en se basant sur « null hypothesis ». Il aborde aussi ce point avec l'hypothèse de McNemar (Dietterich et al. 1998).

Il propose de reformuler le problème en le regardant sur le coût de ceux qui refusent l'offre et ceux qui en voulaient mais ne l'a pas reçu.

Conclusion

Il y a déjà les bonnes méthodes disponibles pour les problèmes d'apprentissage qui donne une bonne prédiction et des modèles interprétables.

On remarque qu'il n'y a pas suffisamment de données pour appliquer des algorithmes sophistiqués pendant cette compétition.

Il fallait bien poser les règles pour qu'on puisse bien évaluer les concurrents, il parle en particulier de la méthode CRISP-DM (Wirth et al.2000).

Mon avis

Les méthodes simples peuvent toujours apporter de bons résultats. On peut indéniablement dire que il y a des classifieurs qui sont plus performants que d'autres dans certain type de problème.

L'article est bon pour son sens critique des algorithmes et manières de raisonner de certains surtout qu'on ne trouve pas souvent des articles comme ça.