La détection de spams : un jeu d'enfant ?

Tristan Mary-Huard, *chargé de recherche INRA à INRA-AgroParisTech*

Comment distinguer automatiquement un spam d'un message normal? Les filtres anti-spams analysent le texte des messages en utilisant des algorithmes de classification en forme d'arbres. Ceux-ci comportent un nombre optimal de nœuds correspondant à autant de questions pertinentes permettant de déterminer la nature d'un message.

Il y a quelques années, un célèbre fabriquant de jouet proposait le jeu « *Qui est-ce ?* ». Le principe du jeu était simple : chaque joueur devait retrouver, parmi une collection de personnages, celui choisi par son adversaire. Pour cela, les joueurs posaient tour à tour une question de la forme : « Le personnage a-t-il des moustaches ? » ou « Le personnage porte-t-il des lunettes ? ». Le premier joueur à trouver le personnage de l'adversaire gagnait la partie. La victoire appartenait donc au joueur posant successivement les questions les plus pertinentes pour identifier le plus rapidement possible le personnage mystère.

L'histoire ne dit pas si Leo Breiman, mathématicien et chercheur américain, et ses collaborateurs de l'Université de Berkeley étaient des joueurs passionnés de « *Qui est-ce ?* » Toujours est-il que la méthode qu'ils proposèrent en 1984 reprend scrupuleusement la logique de ce jeu et l'utilise pour la résolution de problèmes de classification. Cette méthode a été ensuite régulièrement reprise, et sert aujourd'hui - entre autres - à la détection de spam.

Parmi l'ensemble des messages reçus par un individu sur sa messagerie électronique, on distingue deux types : les messages « réguliers » (envoyés par des amis ou par des sites internet auquel l'individu est abonné) et les messages « spams » (communication électronique non sollicitée). Ces spams sont généralement envoyés à des milliers, voire des millions d'individus sur internet, et les conséquences de tels envois

en masse ne sont pas négligeables : en France, on estime que 95 % des courriers reçus en 2009 étaient des spams.

Une possibilité est de classer les messages à partir de l'étude des fréquences d'apparition de certains mots.

Afin d'éviter aux utilisateurs de voir leur messagerie surchargée, les fournisseurs d'accès à internet cherchent à élaborer des méthodes de filtrage anti-spam, capables de distinguer automatiquement un spam d'un message régulier. De telles méthodes sont appelées algorithmes de classification.

« Gagner », « loterie » : les mots typiques des spams

Sur quelles informations l'algorithme de classification peut-il se baser ? Tout message contient un texte. C'est donc à partir de l'analyse du contenu de ce texte qu'il faut décider de classer un message en spam ou régulier. Bien souvent, un spam capte l'attention du destinataire par la promesse d'un gain quelconque (généralement d'ordre financier), et propose de récupérer ce gain en se connectant sur un site internet qui sera ainsi fréquemment visité. Ces courriers se caractérisent donc par la présence de mots comme « gagner », « loterie », ou encore « cliquer » dans le corps du texte. Il existe donc des mots révélateurs de la nature du message. De ce fait, une possibilité est de classer les messages à partir de l'étude des fréquences d'apparition de certains mots.

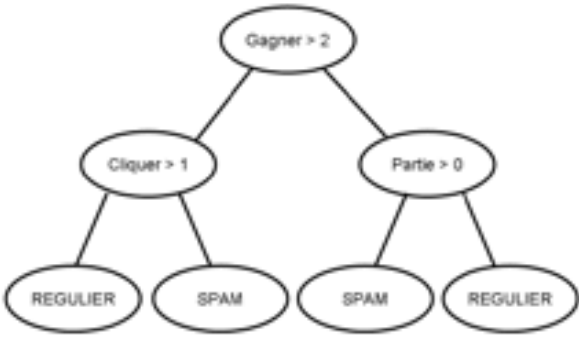

Figure 1. Exemple d'arbre de classification pour le classement des messages électroniques.

L'algorithme proposé par Breiman et ses collègues est illustré en figure 1. Il représente le parcours que doit accomplir un message pour être classé en spam/régulier. Le parcours commence par le haut, en lisant la question inscrite dans le nœud racine (les ellipses sont appelées *nœuds*, les traits entre deux ellipses *branches*, et le premier nœud en haut est appelé *nœud racine*). Si la réponse à la question est affirmative, c'est-à-dire si la fréquence d'apparition du mot « gagner » est strictement supérieure à 2, le message poursuit son chemin en empruntant la branche droite. Sinon, le message emprunte la branche gauche. Le message atteint alors un autre nœud, et comme précédemment poursuit son chemin à droite ou à gauche suivant la réponse faite à la nouvelle question. Ceci se répète jusqu'à ce que le message atteigne une feuille, c'est-à-dire un nœud d'où ne part aucune branche. Dans cette feuille se lit le classement qui doit être attribué au message. Remarquons que le vocabulaire employé (racine, nœuds, branches et feuilles) n'est pas dû au hasard : l'algorithme de Breiman porte en effet le nom d'arbre de classification.

Considérons les messages suivants :

Message 1

Bravo,
Vous venez de gagner à notre grand tirage internet.
Cliquez ici pour recevoir 1.000.000 de dollars !!!
Et pour gagner d'autres cadeaux, cliquez ici.

Message 2

Salut Benoît,
Demain je joue contre Bruno.
Si je gagne la partie je me qualifie directement.
Si je ne gagne pas demain mais que je gagne la suivante,
je monte en pool 3 l'année prochaine !
Bises, Sandra

Dans le premier message, le mot « gagner » n'apparaît que deux fois. La réponse à la question du nœud racine est donc négative, et le message suit la branche gauche. Arrivé au nœud suivant, on vérifie que le mot « cliquer » apparaît deux fois. Le message prend donc la branche droite, pour finalement atteindre une feuille SPAM. Le message sera donc classé en spam. Dans le deuxième message, le mot « gagner » apparaît trois fois, mais il est ici employé au sens de « gagner une partie » plutôt que « gagner de l'argent ». Cette utilisation alternative du mot « gagner » est prise en compte (par le nœud situé en dessous à droite du nœud racine), et le message sera bien classé en REGULIER par l'arbre de la figure 1.

Bien évidemment, l'arbre et les messages considérés ici ne sont qu'une illustration très simplifiée du problème général. Les arbres de classification utilisés en pratique sont souvent plus grands (en termes de nombre de nœuds) et les questions associées à chaque nœud portent sur les fréquences d'apparition de dizaines de mots différents.

Evaluer les différents arbres

Reste la question du choix de l'arbre de classification. Comment choisit-on le nombre de questions à poser ? Ou sur quels mots doivent porter les différentes questions ? Ces choix peuvent s'opérer de la manière suivante : chacun peut proposer l'arbre de classification qui lui semble plus pertinent, puis tous ces arbres sont évalués sur un échantillon test. Un échantillon test est une collection de messages dont la nature est connue parce qu'ils ont déjà été lus et classés en spam/régulier par un individu.

Chaque arbre de classification est tour à tour appliqué à cet échantillon test, et pour chaque message le classement par l'algorithme est comparé au vrai classement. Un arbre classant trop souvent un message régulier en spam ou l'inverse est disqualifié. Parmi les arbres restants, on choisira celui possédant le moins de nœuds, c'est-à-dire celui parvenant à prédire la nature du message en un minimum de questions. L'arbre gagnant est donc celui qui pose les questions les plus pertinentes pour déterminer le plus rapidement possible la nature des messages. On retrouve ici la stratégie du jeu *Qui est-ce ?* !

Améliorer le filtrage

La longévité des arbres de classification, utilisés depuis maintenant 25 ans, ainsi que la diversité des applications qui leur ont été trouvées (dans les domaines de la biologie, de l'écologie ou de la médecine par exemple, voir encadré) ne s'expliquent pas uniquement par le fait que les performances obtenues en pratique par cet algorithme sont satisfaisantes. En effet, du point de vue mathématique, l'algorithme pose bien des questions : peut-on l'utiliser pour tout problème de classification, ou existe-t-il des problèmes où l'algorithme se révélerait inefficace ? Peut-on démontrer que les arbres de classification sont meilleurs (en termes de performance de classement) que d'autres algorithmes couramment utilisés ? Gagnerait-on à utiliser plusieurs arbres de classification plutôt qu'un, comme lorsque l'on convoque un panel d'experts plutôt qu'un seul ? De nombreux articles ont été consacrés à ces questions, et plus généralement à déterminer les propriétés des arbres de classification, et ont permis de justifier théoriquement leur utilisation. Mais ces recherches mathématiques sur les arbres de classification, loin de ne servir qu'à valider la méthode initiale, ont aussi permis aux chercheurs de suggérer des pistes pour faire évoluer l'algorithme et en améliorer les performances. Ainsi, l'application des arbres de classification au filtrage de spams s'est considérablement raffinée au fil des ans. D'un unique arbre de classification, on est passé à une forêt (méthode des *random forests*, cf. encadré), et les arbres sont devenus dynamiques : au fur et à mesure que le propriétaire de la messagerie reçoit des messages, chaque arbre continue de collecter les informations sur les fréquences de mots pour améliorer ses performances de classification. Les arbres de classification, et plus généralement l'ensemble des algo-

rithmes de classification constituent donc un sujet de recherche très actif, auquel de nombreuses conférences mathématiques internationales et prestigieuses sont dédiées.

Au fur et à mesure que le propriétaire de la messagerie reçoit des messages, chaque arbre continue de collecter les informations sur les fréquences de mots pour améliorer ses performances de classification.

Du coté des émetteurs de spams, on n'est pas en reste, et les spams sont aujourd'hui rédigés de manière à contourner les filtrages. Les mots susceptibles de trahir la nature du message sont soigneusement évités, ou volontairement mal orthographiés pour rendre leur identification par l'algorithme de classification plus difficile. La course entre spammeurs et filtreurs ne fait donc que commencer, et continuera à l'avenir de poser aux futurs mathématiciens des problèmes passionnants… et critiques pour le développement informatique et économique de nos sociétés !

La méthode des random forests

Par Robin Genuer, *maître de conférences à l'Université Bordeaux Segalen*

La méthode des *random forests* (forêts aléatoires) peut s'avérer très utile dans des applications en médecine. Par exemple, lorsqu'on étudie les facteurs responsables d'une maladie et qu'on dispose d'information à propos des gènes de sujets atteints ou non de cette maladie. Le but est alors de rechercher quels sont les gènes qui permettent de distinguer au mieux les sujets malades des sujets sains.

Cependant, la plupart du temps, nous disposons d'informations sur des dizaines de milliers de gènes, et donc d'autant de questions potentielles pour découper un nœud de l'arbre. Il est alors très difficile de proposer un bon arbre « en un coup » (faire une partie de *Qui est-ce ?* avec 10000 questions possibles semble en effet assez ardu !).

L'idée des random forests est de construire une collection d'arbres où pour chaque arbre on se restreint à un petit paquet de questions. Après avoir mis en commun tous ces arbres, on peut évaluer le classement donné par la forêt (sur un échantillon test).

De nombreuses études montrent qu'une forêt semble mieux se comporter qu'un seul arbre pour ce type de problème (même si les raisons de ce phénomène restent encore assez méconnues).

Pour aller plus loin

Breiman L., Friedman J., Olshen R., Stone C., (1984) *Classification And Regression Trees*, Chapman & Hall.

Article Spam de Wikipédia, (http://fr.wikipedia.org/wiki/Spam).

European Network and information Security Agency (ENISA), *Spam Survey*, 16 décembre 2009

Surhone L.M., Tennoe M.T., Henssonow S.F., (2010) *Random Forest*, Verlag.