Prévision de la validation d'un brevet

Sandra Fourcade & Ketsia Guichard & Marion Vichery

École nationale de la statistique et de l'analyse de l'information (Ensai), Rennes
sandra.fourcade@live.fr, ketsia.guichard@gmail.com, marionvichery@yahoo.fr

Résumé. Notre objectif est de prévoir la validation en France de brevets délivrés par l'Office Européen des Brevets (OEB). Cette étude, proposée par la Caisse des Dépôts et Consignation Propriété Intellectuelle et Bluestone, s'inscrit dans le cadre plus général de la notation des brevets, dont la validation nationale (en France et dans d'autres états membres de l'OEB) est une composante importante.
Nous disposons de données décrivant l'ensemble des brevets du champ technologique « IT methods for management » délivrés par l'OEB avec une date de dépôt postérieure à 1988.
Une analyse descriptive préalable ne permet de révéler que peu de variables discriminantes de la validation en France des brevets, mais elle révèle toutefois de nombreuses corrélations entre les variables potentiellement explicatives - corrélations prises en compte par la suite. Les modèles de régressions logistiques ensuite mis en œuvre avec différents schémas d'échantillonnage montrent un pouvoir prédictif très relatif ; en revanche, la modélisation par forêt aléatoire révèle finalement une meilleure capacité prédictive.
Nos modélisations permettent également de mettre en lumière un certain nombre de caractéristiques des brevets apparemment liées à leur validation en France, parmi lesquelles : le nombre de déposants de la demande de brevet, le nombre de revendications du brevet ou encore le pays prioritaire.

Mots-clés. Régression logistique, arbres de décision, forêts aléatoires, sélection de variables, brevet

Abstract. Our goal is to predict the national validation in France of patents granted by the European Patent Office (EPO). This study, proposed by the Caisse des Dépôts et Consignation Propriété Intellectuelle and Bluestone, is part of a more general analysis on patent ratings for which national validation (in France and within other members of the EPO) is a major component.
Our dataset includes all patents from the technological field "IT methods for management," delivered by the EPO with a registration date posterior to 1988.
The preliminary descriptive analysis only keeps a few variables that discriminate the patent validation in France. However, it also reveals many correlations between potential explanatory variables, which are subsequently considered. Logistic regression models implemented with different sampling schemes show a very weak predictive power. On the other hand, random forest models display a better predictive capacity.
Our models highlight several patent characteristics linked to their national validation in

France, among which are the number of patent owners, the number of claims, and the priority country.
Keywords. Logistic regression, decision trees, random forests, variable selection, patent

1 Introduction

Depuis les années 1980, le brevet a élargi ses fonctions traditionnelles de protection d'une invention. Ce titre de propriété est actuellement considéré comme un quasi-actif financier, c'est-à-dire transmissible et susceptible de produire des revenus à son détenteur. Ainsi vu comme un placement, les portefeuilles de brevets sont davantage valorisés lors des rachats d'entreprises, en ce qu'ils évaluent le capital intellectuel des firmes. La Caisse des Dépôts et Consignations Propriété Intellectuelle souhaite proposer, avec l'assistance de Bluestone, un service d'évaluation de la valeur des brevets.

Les procédures européennes de brevet sont longues et cheminées d'étapes. Une fois la demande déposée, un examen du dépôt et des éléments formels de la demande – dessins, désignation de l'inventeur, traductions et taxes – a lieu. Un rapport de recherche est ensuite établi afin d'énumérer l'ensemble des documents disponibles pouvant être pertinents pour apprécier l'activité inventive liée au brevet. Dix-huit mois après le début de la procédure, la demande de brevet est publiée, et le fond du brevet peut alors être examiné. La division d'examen publie ensuite sa décision : le brevet européen est alors ou non délivré. La dernière étape consiste en la validation du brevet au sein de chaque Etat. Nous nous intéressons ici au passage de la délivrance à la validation, en nous centrant sur le cas de la France.

Quels sont les critères encourageant le plus la validation d'un brevet ? Après une première phase d'analyse descriptive de la base, nous mettrons en place des modèles de régression logistique, dont les performances seront confrontées à d'autres méthodes d'apprentissage statistique telles que les arbres de décision et les forêts aléatoires.

2 Spécificités de la base de données

Notre base de données concerne les demandes de brevets, du champ technologique « IT methods for management », déposées auprès de l'Office européen des brevets (OEB) entre 1988 et 2009. Cette base comportait initialement un millier de brevets caractérisés par un identifiant et plusieurs centaines de variables. Ces variables nous renseignent entre autres sur les principales caractéristiques des demandes de brevets (langue et date de dépôt, nombre d'inventeurs, de citations, de mots et de figures utilisés) à l'instant de la publication, mais aussi lors de la délivrance du brevet. Nous disposions également de variables préalablement recodées, créées en prévision de la mise en place de modèles de

régressions logistiques. Elles rassemblent des variables déjà présentes dans notre base de données, découpées en classes.

Nous avons décidé de mettre en place en première approche une régression logistique, étant donné que notre variable cible, la validation ou non d'un brevet, est binaire. Pour réaliser cette modélisation, il est primordial de s'intéresser à la colinéarité entre les différentes variables explicatives, qui entraîne des problèmes d'identifiabilité du modèle.

Par exemple, certaines variables de notre base de données renseignent les mêmes caractéristiques d'une demande de brevet à deux étapes de son processus, la publication et la délivrance. Peu de changement étaient notables entre ces étapes, les variables de ce type étaient fortement colinéaires. Notre objectif étant de prédire la validation d'un brevet à partir du moment où celui-ci est délivré nous décidons de ne garder que les variables concernant la délivrance, ce qui permettra de s'affranchir des dépendances pouvant exister avec les caractéristiques du brevet à la délivrance. Cette épuration, étape clé de la modélisation, nous permet à présent de mettre en place des régressions logistiques et de modéliser la validation d'un brevet.

3 Méthodes et résultats

3.1 Régression logistique

La variable à expliquer, la validation ou non d'un brevet, étant binaire, une des méthodes statistiques possibles dans ce contexte est la régression logistique.

Notre base de données présente plusieurs variables véhiculant la même information (variables recodées). Trois démarches sont donc menées : construction d'un modèle ne prenant pas en compte les variables quantitatives recodées puis d'un autre ne contenant que ces variables pour finalement en exposer un dernier comprenant les variables ressorties dans les modèles précédents. En revanche, il n'y a pas de différenciation pour les variables qualitatives – seules celles recodées sont intégrées aux modèles antérieurs. Ces dernières présentent moins de modalités et seront ainsi plus facilement interprétables.

Pour chacune de ces approches, plusieurs échantillons d'apprentissage/test ont été créés. Cependant, les variables significatives, déterminées par une méthode « pas à pas », changent d'un échantillon à l'autre : les modèles manquent de robustesse. Nous avons donc regardé, sur une quinzaine d'échantillons d'apprentissage, les variables explicatives qui impactaient le plus la validation d'un brevet.

Ces trois méthodes sont ensuite comparées à l'aide de critères basés sur la prévision (taux de mal classés, AUC,...), notre objectif étant principalement de prédire la validation d'un brevet. Une approche est ressortie en moyenne meilleure que les autres. Il s'agit de la troisième méthode utilisant les variables les plus fréquentes et influentes pour expliquer la validation d'un brevet ressorties dans les deux approches précédentes.

À l'issue de cette phase de modélisation, la difficulté rencontrée pour présenter un modèle robuste capable de prédire la validation des brevets en France nous conduit à utiliser d'autres modèles statistiques. Qui plus est, ces modèles permettent de pallier certains inconvénients de la régression logistique tels que sa sensibilité aux valeurs aberrantes.

3.2 Des arbres de décision aux forêts aléatoires

La lisibilité, la facilité d'interprétation et la possibilité de détecter des liaisons non linéaires avec la variable cible ont entre-autres motivé notre choix vers la mise en œuvre d'un arbre de segmentation. C'est la méthode CART (Classification And Regression Trees) qui a ici été retenue. Par divisions successives et binaires de l'ensemble de nos individus statistiques, cette méthodologie nous a fourni une partition la plus homogène possible du point de vue de la variable cible. L'arbre CART maximal obtenu a ensuite été élagué afin de conserver le pouvoir discriminant du modèle sur de nouvelles observations, c'est-à-dire dans le souci d'éviter le sur-apprentissage. Nous avons alors retenu le sous-arbre qui présentait le plus faible taux d'erreur estimé par validation croisée (méthode du leave-one-out). Ce sous-arbre, optimal dans le sens du critère précédent, s'est révélé être l'arbre stump composé uniquement de deux feuilles. Il s'avère que le pays prioritaire est la variable la plus discriminante du point de vue de ce modèle.

Dans l'optique d'obtenir des prédictions toujours plus robustes, nous nous sommes finalement intéressées aux méthodes d'agrégation d'arbres (Breiman, 2001 [1]). Intuitivement l'intérêt de ces méthodes se comprend très bien : en agrégeant les prédictions d'un grand nombre d'arbres, nous avons en moyenne plus de chance de prédire des phénomènes complexes.

Parmi celles-ci, la forêt aléatoire a retenu notre attention. Une fois avoir tiré un grand nombre d'échantillons bootstrap de nos observations, un arbre CART a été élaboré sur chacun d'entre eux, puis nous avons agrégé ces différents arbres pour construire notre forêt. L'agrégation a été effectuée par vote : pour chaque individu, la forêt prédit la classe qui lui a été majoritairement attribuée par les arbres qui la composent.

En plus de la randomisation des individus, présente également dans le bagging, la forêt aléatoire a l'avantage d'offrir une autre randomisation : celle des variables. En effet, pour chaque arbre de la forêt, chaque scission de nœud est effectuée non pas à partir de l'ensemble des variables explicatives mais d'un sous ensemble de celles-ci tirées aléatoirement. Cette double randomisation est justement la force de la forêt aléatoire. Elle permet d'obtenir une grande variété d'arbres qui sont donc potentiellement peu corrélés les uns aux autres : nous obtenons un modèle agrégé de faible variance c'est à dire plus robuste.

Concrètement, la forêt aléatoire présente quelques difficultés pour prédire les brevets non validés. En revanche, elle détecte très bien les brevets validés et est le modèle le plus robuste que nous ayons réalisé. Cette approche a également l'avantage de fournir un critère mesurant l'importance d'une variable explicative sur la validation d'un brevet.

Nous avons donc pu identifier les principales caractéristiques qui influencent la validation d'un brevet. Parmi elles se trouvent le pays prioritaire, l'année de début de la vie du brevet, le nombre de déposants, le nombre de concepts dans la description à la délivrance, autant de variables qui avaient été mises en évidence dans les modélisations précédentes.

4 Conclusion

Si les modélisations effectuées prédisent correctement la validation d'un brevet, sa non-validation demeure plus complexe à détecter. Néanmoins, ces travaux ont souligné l'influence de certaines variables sur la décision de validation d'un brevet, notamment :

- des variables classiquement utilisées dans les études sur les brevets, à l'instar des travaux de Sampat (2005) [2], Squicciarini (2005) [3] ou encore Van Zeebroeck (2013) [4], comme le nombre de citations au sein du brevet, son nombre de classes IPC (ou domaines technologiques), ainsi que le nombre de revendications, c'est-à-dire de contributions distinctes à une invention
- d'indicateurs plus rarement référencés, à l'instar du nombre de mots ou du nombre de déposants.

Bibliographie

[1] L. BREIMAN (2001), Random Forests, *Machine Learning* 45 : 5–32
[2] B. N. SAMPAT (2005), Determinants of Patent Quality : An Empirical Analysis, School of International and Public Affairs, Columbia University
[3] M. SQUICCIARINI and H. DERNIS and C. CRISCUOLO (2013), Measuring Patent Quality : Indicators of Technological and Economic Value, *OECD Science, Technology and Industry Working Papers*
[4] N. VAN ZEEBROECK and B. VAN POTTELSBERGHE DE LA POTTERIE (2007), Filing strategies and patent value, *Economics of Innovation and New Technology*