A large e-commerce data set released to benchmark categorization methods

Bruno Goutorbe[1], Yang Jiao[2], Matthieu Cornec[1],
Christelle Grauer[1] & Jérémie Jakubowicz[2]

[1] *Cdiscount, 33000 Bordeaux,*
{*bruno.goutorbe, matthieu.cornec, christelle.grauer*}*@cdiscount.com*
[2] *SAMOVAR, CNRS, Télécom SudParis, Univ. Paris-Saclay, 91011 Evry,*
{*yang.jiao, jeremie.jakubowicz*}*@telecom-sudparis.eu*

Résumé. En 2015, Cdiscount a mis la communauté au défi de prédire la catégorie correcte de ses produits à partir de certains attributs comme leur description et leur photo. Le concours était basé sur l'intégralité du catalogue, qui contient environ 15.8 millions de produits répartis dans 5789 catégories, hormis une petite partie qui a servi d'ensemble de test. La qualité des données est loin d'être homogène et la répartition des catégories est extrêmement déséquilibrée, ce qui complique la tâche de catégorisation. Les cinq algorithmes gagnants, sélectionnés parmi plus de 3500 contributions, atteignent un taux de prédiction correcte de 66–68 % sur l'ensemble de test. La plupart utilisent des modèles linéaires comme des régressions logistiques, ce qui suggère que les étapes préliminaires telles que le pré-traitement du texte, sa vectorisation et le rééchantillonnage des données sont plus cruciales que le choix de modèles non-linéaires complexes. En particulier, les gagnants corrigent tous le déséquilibre des catégories par des méthodes d'échantillonnage aléatoire ou de pondération en fonction de l'importance des catégories. Les deux meilleurs algorithmes se distinguent par leur aggrégation de grands nombres de modèles entrainés sur des sous-ensembles aléatoires des données. Le catalogue de produits est mis à disposition de la communauté, qui disposera ainsi de données réelles issues du e-commerce pour étalonner et améliorer les algorithmes de classification basés sur le texte et les images.

Mots-clés. Classification, e-commerce, big data, jeu de données public.

Abstract. In 2015 Cdiscount challenged the community to predict the correct category of its products from some attributes such as their description and their image. The challenge was based on the whole catalog of products, which contains about 15.8 millions items distributed over 5789 categories, a subset of which served as testing set. The data suffers from inconsistencies typical of large, real-world databases and the distribution of categories is extremely uneven, thereby complicating the classification task. The five winning algorithms, selected amongst over 3500 contributions, are able to predict the correct category of 66–68 % of the products in the testing set. Most of them are based on linear models such as logistic regressions, which suggests that preliminary steps such as text preprocessing, vectorization and data set rebalancing are more crucial than resorting to complex, non-linear models. In particular, the winning contributions all carefully cope with the strong imbalance of the categories, either through random sampling

or sample weighting. A distinguishing feature of the two highest-scoring algorithms is their blending of large ensemble of models trained on random subsets of the data. The data set is released to the public, as we hope it will prove of valuable help to improve text and image-based classification algorithms.

Keywords. Classification, e-commerce, big data, public dataset.

1 Introduction

E-commerce companies have become major actors of the retail business over the past decade. As the product catalog of the largest companies now routinely exceeds several millions of distinct items, a large part of which from third-party sellers, a salient yet increasingly tough need consists in filling correctly the products' characteristics in order to efficiently guide the customers towards the products they desire.

In 2015 Cdiscount challenged the community through a datascience.net contest (www.datascience.net) on a simple, real-word question: how can one guess the category of a product from its description, its image and other available attributes? The participants had access to Cdiscount's catalog of products, a subset of which had their category hidden to serve as testing set, thus turning the problem into one of supervised classification. The contest was held between May–August 2015 and attracted over 800 participants. It seems that most of them participated in a personal capacity and worked individually. Cdiscount released the data set to the public to be used as a real-world benchmark and encourage improvements over text and image-based classification algorithms.

2 Data set

The data set consists of about 15.8 millions of products, which represents virtually the whole catalog of Cdiscount as of May 2015, and takes about 4 Gb in text format. Each product is associated with a unique identifier, a three-level hierarchical category, a title, a description, a brand, a seller (Cdiscount or third party) and a price. Besides, part of Cdiscount products are accompanied by a representative image in jpeg format. As described hereafter, the data suffer from flaws and inconsistencies typical of large databases involving strong user interaction. Products do not necessarily have a brand, and their description is sometimes cut-off, ending in this case with an ellipsis. The price is set to -1 for out of stock products, and can take unrealistically large values. More importantly, the category filled by third-party sellers is not as reliable as that of Cdiscount's products.

As a consequence, the vast majority of the populated categories are not strongly reliable, third-party sellers accounting for almost 95% of the database (Table 1). As can be expected, the 5789 available categories are strongly unevenly distributed amongst the products: the distribution of the number of products per category approximately follows a power law, which exhibits a long tail of categories containing a large number of

Table 1: Key numbers on the data set.

15,821,950	products
791,453	products sold by Cdiscount
15,030,497	products sold by third-parties
5789	distinct categories
27,982	distinct brands

products (Fig. 1a). As a matter of fact, about 700 categories hold 90% of the products (Fig. 1b). Similar trends are observed for the distribution of the approximately 28,000 brands (Fig. 1c) and of the descriptions' vocabulary (Fig. 1d) amongst the products.

3 Description of the challenge

In 2015, Cdiscount offered a simple challenge based on the data set described in the previous section: given the information described in the previous section, what is the correct category of a product? A subset of 35,065 products, the category of which was hidden, served to evaluate the prediction algorithms proposed by the candidates. As evaluation metric, we simply used the proportion of correct predictions within the testing set. Note that, in order to build up a testing set not too biased towards the most popular categories, no more than 20 products may belong to the same category. The resulting distribution of categories amongst the products consequently strongly differs from that of the whole data set.

The challenge attracted 838 participants who submitted 3533 contributions. The five highest-scoring contributions were sent to a jury, which made the final ranking based on the score, quality and originality of the proposed solutions. The following section briefly describes the winning contributions, which received money prizes between 500€ and 9000€.

4 Analysis of the winning contributions

The winning algorithms are able to predict the correct category of 66–68 % of the products in the testing set (Table 2). The four best algorithms use linear models, mostly with a logistic loss function (Walker and Duncan, 1967). Interestingly, the squared loss function also gives good results (contribution #3), although it is known to lack robustness against outliers. Two other linear classifiers appear in the winning contribution, namely, the passive aggressive classifier (Crammer et al., 2006) and the naive Bayes method with multinomial distribution of the features (Zhang, 2004). The only non-linear model is the convolutional neural network (Johnson and Zhang, 2015), which is used by candidate #5.

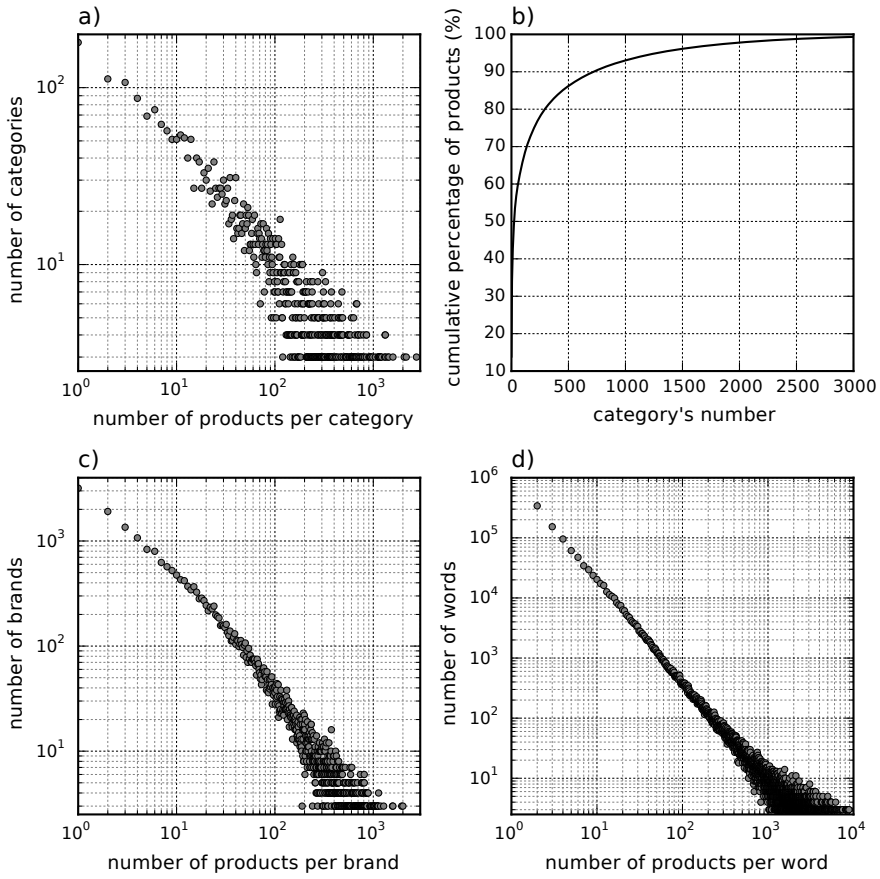

Figure 1: (a) Distribution of the products per category. (b) Cumulative percentage of products held by the categories (categories containing the largest number of products first). (c) Distribution of the products per brand. (d) Distribution of the products per word of the descriptions' vocabulary (given a word, a product is counted once if the word appears in its description).

Table 2: Summary of the winning contributions.

Rank	Score	Language/library	Method(s)
#1	68.3 %	Python/scikit-learn	Logistic regression with stochastic gradient descent + multinomial naive Bayes + passive aggressive classifier
#2	68.0 %	Python/scikit-learn	Two-stage logistic regression
#3	66.9 %	Python, R, Vowpal Wabbit	Linear classifier with squared loss function
#4	66.3 %	Python/PIL, C++, Dataiku	Logistic regression
#5	66.3 %	R/ConText	Three-stage convolutional neural network

All the candidates concatenate at least the title, brand and description of the products to build input features. Candidate #2 applies larger weights to the title and the brand. Because of the large range of values it takes and the errors it contains, the price seems more delicate to include, but it nevertheless appears as input in two contributions (#1 and #3). In order to tackle the above-mentioned issues, the winner assumes that values above 10,000 actually correspond to thousandths of euros and uses as input the interval to which the price belongs, which takes a limited number of values. Only one candidate (#4) fully integrates the images, by associating with each product the category of the three nearest neighbors of its image, weighted by their inverse distance, as an additional input feature. Curiously, candidate #1 finds that simply appending a piece of text describing the image's geometry (rectangular or not rectangular) significantly improves the categorization of books.

As can be expected when dealing with large chunks of text data with potential inconsistencies, the candidates have to apply a variety of preprocessing techniques before the vectorization step. These usually include lower-casing, removal of stop words, conversion to plain ASCII text and word stemming. Some candidates additionally remove numbers, or replace them with generic strings such as |NUMBER| or |DIGIT|. Candidates #1 and #4 apply a special treatment to prepositions such as "for", as they can alter the meaning of the following sentence.

The vectorization step is then most often realized with the tf-idf statistic, wherein the concatenated text associated with a product is converted to a vector whose i^{th} element is proportional to the number of appearances of the i^{th} token of the corpus within the product's text, and offset by the frequency of the token in the corpus. Candidate #5 takes on a different approach that partly preserves the order of the words, wherein the text is split in successive regions of 15–20 contiguous words, and a word count is applied to each region.

In order to cope with the unevenness of the distribution of the categories outlined in section 2, most of the candidates resort to some form of stratified sampling: in other words, subsets of the catalog are randomly selected as training sets, with a limit of a few hundreds products per category and with replacement oversampling for underrepresented categories. Candidates #1 and #2 repeat this subsetting procedure several thousands of times and blend the predictions from the resulting ensemble of models, which we assume to be a key ingredient to their success. The winning algorithm actually goes a step further by (1) random parameterizing several processing steps applied to the subsets (e.g., tf-idf or word count, word stemming or not, unigrams or bigrams...) and (2) including three families of classifiers in the ensemble of models (Table 2). Only candidate #3 chooses not to re-sample the catalog of products, but rather assigns them weights inversely proportional to the frequency of appearance of their categories.

As for the better reliability of the category of Cdiscount products (section 2), it an information only two candidates take advantage of (#1 and #5): the former candidate specifically trains models on Cdiscount or third-party products and assigns them different

weights in the final blend; the latter candidate explicitly gives priority to Cdiscount products in the stratified sampling step described in the previous paragraph.

Finally, candidates #2 and #5 use the three-level hierarchical structure of the categorization to reduce ambiguity between categories belonging to different branches, by performing classifications by stage. The idea is to successively predict the category across the levels of the hierarchical tree from top to bottom.

5 Conclusion

We gave statistical insights into the catalog of products of Cdiscount in order to highlight the kind of pitfalls and difficulties a classification algorithm applied to a real-world data set has to cope with. Specifically, the potential inconsistencies of the products' attributes, the varying reliability of the data, the large number of categories and the extreme imbalance of their distribution obviously complicate the classification task.

The five winning contributions of the datascience.net challenge are able to predict the correct category of 66–68 % of the products in the testing set. Most of the algorithms are based on simple, linear classifiers, which outlines the importance of preliminary processing steps such as text preprocessing, vectorization and rebalancing of the training data.

The whole data set is released to the public. The availability of a large, real-world catalog of products with associated images and text attributes, together with benchmark results from the most accurate models to date, should prove of valuable help to the scientific community in order to improve over existing text and image-based classification algorithms.

References

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Johnson, R. and Zhang, T. (2015). Semi-supervised convolutional neural networks for text categorization via region embedding. In *Advances in Neural Information Processing Systems*, pages 919–927.

Walker, S. and Duncan, D. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1867):167–179.

Zhang, H. (2004). The optimality of naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach. FL: AAAI Press.