Random Forests and Text Mining

Armando Segnini
Institut Mines-Télécom - Télécom Bretagne
Brest, France
armando.segnini@telecom-bretagne.eu

Juanita Joyce Tayou Motchoffo
Institut Mines-Télécom - Télécom Bretagne
Brest, France
juanita.tayoumotchoffo@telecom-bretagne.eu

***Abstract*—Data come from varying sources and given their increasing importance and relevance, they are continuously being collected and stored in databases across the world. While traditional data mining methods assume that data to be mined are represented in structured relational databases, text mining is concerned with the extraction of information from unstructured textual data. Random forest have emerged as a powerful learning technique in text classification applications within the text mining domain. The extensive literature review on random forests and text mining provided in this paper makes clear the link and relevance that exists between these two fields, and shows how academia and industry are doing an increasing number of studies on these subjects, keeping them an interesting issue for further development.**

***Index Terms*—Random forest, data mining, text mining, text classification, machine learning**

I. Introduction

The proliferation of the use of digital technology in everyday's life has resulted in the generation of massive amounts of data. Data come from varying sources, ranging from governments, to business entities and individuals. Given their increasing importance and relevance, data generated are continuously being collected and stored in databases across the world. Different types of data exist however. In general terms we have qualitative and quantitative structured, or unstructured data. Because the huge amounts of data available conceal a wealth of invaluable information and knowledge several methods have been developed (and still are) to help process and make sense out of all data types.

Enterprises are growing and with them the amount of data generated by each transaction. By the need to see patterns in users and generate highers incomes, there are different types of methods that had been created to generate better results. Data mining has been one of the most popular methods, and there had been a lot of studies in order to improve the tools under the hood. One of this tools are random forests (RF). This paper attempts to show the main concepts of this technique, the latest studies done around them. We will show that it is still an interesting approach and has not reached its limits because it has not been replaced by other methods.

This paper focuses on the application of random forests (and its derivatives) in the classification of text data and is organised as follows. We begin a formal definition of the concepts of classification using decision trees (random forests) in Section 2. Describe in detail the original random forest algorithm as defined by Breiman [1]. Sections 3 and 4 describe the main ideas underlying the data and text mining respectively. Our discussion of random forests applied to text mining starts in Section 5, where we list a few applications of RF in this domain. In section 6 we briefly discuss modified and improved versions of RF (derivatives of random forests) used in text mining applications. Finally, section 7 concludes by supporting the relevance of RF in text mining in other fields and suggests novel RF classifiers whose application in the text mining domain could be explored.

II. Random Forests

Hereafter referred to as RF, random forests are a supervised learning algorithm based on machine learning theory that belongs to the family of ensemble methods. It employs the supervised learning methodology whereby information from labeled data sets (training set) is captured, used to derive predictions and to build a model. The derived model from the labeled dataset can then be used to classify unlabeled data [2].

Introduced by Leo Breiman and Adele Cutler in the 2000s [1], RF build prediction ensembles using decision trees that are generated randomly in randomly selected subspace of data [3]. The decision trees generated are used to represent the rules that result from modeling. Training datasets contain several attributes, in the case of RF, randomness is also applied to choose the best attribute to split on at different levels of the decision tree. The average is calculated from random classifiers.

Random forests (RF) are an increasingly popular technique for a variety of tasks in classification, prediction, the study of variable importance, variable selection, and outlier detection [4].

A. Classification process in random forests

To uncover patterns that exist in data, the decision trees in RF are grown through the combining of several appropriately selected machine learning algorithms. Below is a description of this process based on [1]:

1) Model Construction: Given a training data set with N number of cases, a training sample set of size N is selected with replacement, from the original training set. The sampled set contains a number of input variables or features M. A number of input variables m is specified and kept constant throughout the tree building process, (where $m \ll M$) such that, at each node m variables are randomly selected and the best split on these m is used to split the node. The trees are grown in this manner (random feature selection) to maximum size without pruning. This process is repeated several times and a collection of predictors, refered to as a weak learners, is created at the end of this process. The inclusion of randomness ensures that the weak learners in RF have low bias, low correlation and high variance [1]. Making them ideal predictors as low bias and low correlation are essential for accuracy.

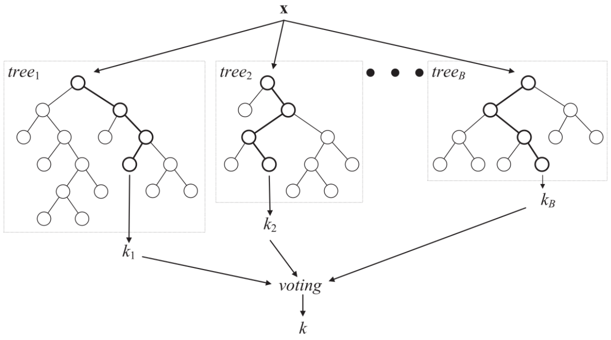

Fig. 1. A general architecture of a random forest [4].

Figure 1 gives a general overview of the RF process summarising the model construction process. The illustration shows that a RF consists of a set of B decision trees ($\mathrm{tree}_1, \mathrm{tree}_2, \ldots \mathrm{tree}_B$), with each tree having a class label (k_1, $k_1, \ldots, k_B$) that is used to get a *k* voted class label [4].

2) Use of model in prediction: Prediction is made by aggregating the predictions of the weak learners. To classify a new object from an input vector, it passes the sample vector to each of the trees in the forest. Each tree will classify the new object and the classification results of each tree combined. The chosen classification for the new object is that which has the most votes over all the classification trees in the forest.

As the number of trees grown in RF increases, the test set error rates converge to a limit, meaning that there is no over-fitting in large RFs [1]. Additionally, RF presents a good solution for classification of sparse data sets, datasets with errors and missing values [5].

III. Data Mining

Data mining has emerged as a field dedicated for the analysis of data. Simply put, it aims at uncovering or extracting of useful and meaningful information from data. Data mining is essentially concerned with information extraction from structured databases. Knowledge extraction from data can be approached in several ways and for the different data types that exist. Efficient data mining methods for large databases have been studied and thus many interesting methods applicable to the plethora of data types available have been developed.

IV. Text Mining

While traditional data mining methods assume that the data to be mined are represented in structured relational databases [6]. Text mining differentiates itself in that; it is concerned with the extraction of information from unstructured textual data. Inherently connected to the field of machine learning (ML), it seeks the retrieval of useful information from unstructured textual data through the identification and analysis of patterns within data [7]. Text mining is a multidisciplinary area that integrates methods from data mining (DM) and knowledge discovery in databases (KDD). In order to discover patterns in data and to make predictions from them, text mining makes use of different techniques such as classification, clustering and association rule techniques.

For text classification, a set of training data is analysed and a model for each class based on the features in the data is constructed. The classification process results in the generation of classification rules that can later be used to classify future data [8].

V. Random Forests and Text Mining

Random forests are predominantly used in text classification applications within the text mining domain. Because textual data are typically characterised by highly dimensional feature spaces [9], it is essential that the chosen classifier performs well within this context. Thus the primary concern when choosing a text classification method is about the prediction accuracy of the classifier. Because the dimensionality of textual data increases the risk of excessive detailing when building the decision trees, an equally important aspect to consider is the classifier's feature selection. That is, the tendency the classifier has to overfit the training data during classification [9].

Despite being relatively new, RF have gained considerable attention. In text classification studies, RF is being chosen over the popular Support Vector Machine (SVM) and Naive Bayes (NB) classifiers. This is due to the theoretical guarantees for optimal classification performance RF provides (as specified by Breiman in [1]). In relation to the factors to consider when choosing a text classification method (specified above), RF are increasingly seen to outperform other classifiers in text mining applications such as text categorisation and text filtering. Advantages of the RF methods are discussed in [3][10][11].

The conventional approach followed in text mining generally is the Bag Of Word approach (BOW). Here, each word from a text is treated as a term and is commonly represented using a Vector Space Model (VSM) [12]. The BOW approach is also used for small phrases, treating the text as a collection of phrases and being usually represented in n-gram models as bigram (2 words) or trigram (3 words), etc. In order to transform the text in one of these models, different techniques such as term frequency-inverse document frequency (tfidf) and others ranking models procedures are regularly used [13]. However, sometimes text cannot be seen only as a set of words or phrases because the interconnection that exist between these elements in a text is also meaningful. Considering this limitation, there is another approach based on Natural Language Processing that focuses its model in semantic relations with minimal loss of this type of information [14].

Although RF is a proven accurate classification method for highly dimensional datasets, its prediction accuracy can be compromised when processing certain text data corpus. When applied to training sets that contain large number of features with the percentage of truly informative features being small the performance of RF in speed and prediction accuracy declines significantly. This results in the construction of trees whose nodes are populated by non-informative and redundant features, and the generation of a wrong predictors as a consequence.[15]

A. *Applications of Random Forests in Text Mining*

Below is a (not comprehensive) review of the use of RF classifiers to main applications of the field of text mining. (The reader should note that due to the plethora of relevant academic work available on RF application to text mining this review is understandably not a comprehensive one).

1) Email filtering: According to [16] text filtering is a way of classifying a dynamic collection of texts, that is a stream of incoming documents that are dispatched in an asynchronous way by an information producer to an information consumer. Following Rios & Zha 2004 [17], RF are comparable to SVM in classification accuracy when applying RF for spam detection on time indexed data. Five years after Breiman first introduced RF and at a time where Bayesian approaches were the most widely used in text categorization and e-mail classification, Koprinska et al. [18] studied the application of RF for the classification of emails. Their investigation focused on the ability of random forest to accurately file emails into folders and filter spam emails. Comparing RF against the following state-of-the-art algorithms: Support Vector Machines (SVM), Naive Bayes (NB) and Decision Trees (DT), they showed that RF outperformed them in both supervised and semi-supervised settings for e-mail classification, asserting that RF is a better choice in terms of its high prediction accuracy, running and classification time (when working on large and high dimensional databases), and RF's simplicity to tune.

2) Marketing: In 2013 Agrawal et al. [19] designed an algorithm to automatically suggest phrases from Internet sites to advertisements companies in order to get a better commercial impact. This method is commonly referred to as bid phrase recommendation. Based on investigations carried out in 2006 [13] and 2010 [20] on the current methods used in this area that suggest techniques based on ranking and Natural Language Processing (NLP), they approached bid phrase recommendation as a multi-label learning problem and evaluated their tool on a corpus of 10 million labels. Using a derivative of standard RF, called Multi-Label Random Forest (MLRF), it was possible to make at least 100 suggestions significantly superior than the ones generated by the current methods on relatively all ad landing pages in a few milliseconds, breaking the limits concerning the number of millions labels that can be predicted.

3) Finance: Mori, H. & Umezawa, Y. [21] in 2007 successfully applied a method based on random forests to evaluate the financial data of energy utilities in the market. The purpose of their study was to present a new intelligent method for credit risk evaluation in the power market at a time where artificial neural networks (ANNs) where the most conventional machine learning models used for this task .

4) Bioinformatics: With research on the application random forests in the text mining of biological data carried out as early as 2003 by [22] just two years after Breiman's [1] description of RF, they have emerged as good performers in the field of bioinformatics. The performance of RF in high-dimensional data classification makes them successful in the classification of gene expression microarray data and Mass spectrometry (MS)-based proteomics studies as demonstrated in [22] and [23] It should however be noted that in 2008, Amaratunga et al. [24] proposed a very successful modification to the original RF called ”enriched random forests” in order to increase the prediction accuracy of RF in bioinformatics.

We performed a search in three well known and respected academic research libraries namely: IEEE Xplore, Springerlink and Science Direct in order to observe the amount of research on random forests associated to text mining published till date. The results of this is summarised in Figure 2, which shows the aggregated frequency distribution of documents containing the terms random forests AND text mining in the aforementioned libraries. An increasing trend is clearly apparent from the figure over time from 2004 when the first paper containing those two key words is observed. The highest value is observed in the year 2014 with 99 documents containing both keywords with some documents already scheduled to be published in 2015. Showing the relevance of random forests in the frame of text mining applications. (Note: as stated, this search was only performed for articles that contained the keywords ”text mining” and ”random forest”. Which does not guarantee the inclusion of all articles pertaining to RF applied to text mining).

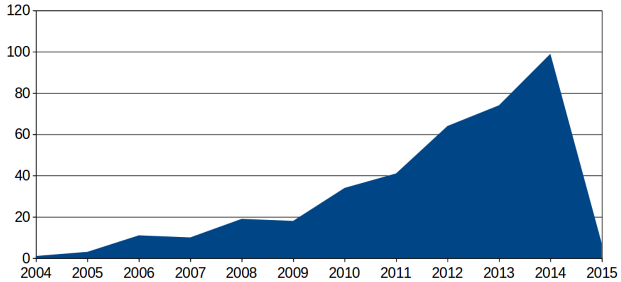

Fig. 2. Frequency distribution of documents containing the terms ”random forests” AND ”text mining” in different research libraries

VI. Related Works

Improved RF can handle the limitations faced by standard RF in areas such as spam filtering, user profiling analysis and for the case of complex imbalanced data text (the number of instances in one class greatly outnumbers the number of instances in the other class). Inspired by studies published in 2008 by [24] and 2011 by [25], Wu, Q. in 2014 proposed a novel stratified sampling method for feature subspace selection and built a splitting classifier for data partition at each node of the tree models in the random forest ensemble [26]. His approach uses a supervised learning technique to learn the splits at the nodes of the tree with the help of Support Vector Machine (SVM) to create a more robust model that fits class imbalanced data even better. When tested, the modified method's classification performance was superior to that of standard RF and different variants of the SVM algorithm.

Having identified an inadequacy in Breiman's standard random forest for the modelling of text data that contains many uninformative features to a class. In 2012 Xu et al. used Amaratunga et al.'s feature weighted method [24] and the out-of- bag accuracy proposed in [25] to develop a novel feature weighting method and tree selection method for the random forest process. The improved RF was tested on six text data sets with diverse characteristics and the results of the study revealed that the improved RF outperformed SVM, NB, k-nearest neighbor (KNN) and decision trees.

Figure 3 shows the relations and dependencies that exist between the papers documented on the use of the random forest algorithm and its derivatives in the frame of text mining applications in this paper. The nodes represent the papers and the edges represent the papers that were referenced by the various authors. With the stronger colour indicating the papers that were used at a higher frequency. The figure shows Breiman [1] as the most common influence, followed by A. Liaw [10] and G. Biau 2008 [11].

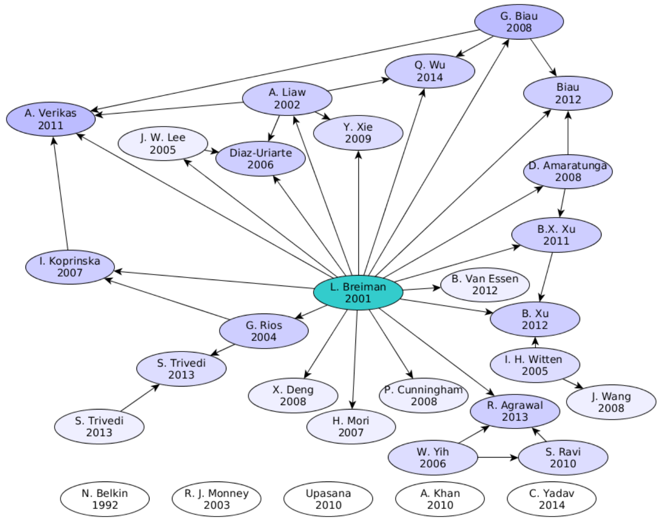

Fig. 3. Reference Dependencies

VII. Discussion

Our study makes clear the link and relevance that exists between RF and text mining. Figure 2 shows that there is an increasing trend in research on the use of these two subjects together. It is seen that standard random forests have evolved since they were first introduced by Breiman in 2001 [1] with these random forests derivatives consistently outperforming conventional text mining machine learning techniques, hence keeping random forests relevant in the area of text mining.

Other impovements of random forests applied to areas other than text mining could also be studied within the framework of text mining. An example of which is compact random forests (CRF, not to be confused with conditional random fields) that can generate decision trees more suitable for acceleration than traditional decision trees, as it was proved by Essen in 2012 [28]. It may be interesting to observe and evaluate the performance of CRF in text mining.

Another improved RF that could be applied to text mining is the improved balanced random forests (IBRF) proposed by Xie et al. [27] in 2009.

Finally, in a study carried out in 2013 on text categorization detecting unsolicited emails (spams), Trivedi & Dey [29] built an ensemble of several weak Genetic Programming (GP) classifiers based on same concepts from previous studies. That is, subsets randomly selected, a fitness function and genetic operators like selection, crossover and mutation. They conclude that their Enriched GP was better than standard RF in terms of performance accuracy as well as false positive rate [30]. It could be a good review repeat those test with the RF derivation created by Qingyao Wu and others, since the difference between the EGP and the standard RF was not considerable [26].

References

[1] L. Breiman, *Random forests*. Machine Learning, 45(1):5-32, 2001.

[2] P. Cunningham, M. Cord and S. J. Delany, *Supervised Learning*. In Machine Learning Techniques for Multimedia Cognitive Technologies, 2008, pp. 21-49.

[3] G. Biau, *Analysis of a Random Forests Model*. Journal of Machine Learning Research 13. 2012, pp. 1063-1095.

[4] A. Verikas, A. Gelzinis, M. Bacauskiene, *Mining data with random forests: A survey and results of new tests*. In Pattern Recognition, Volume 44, Issue 2, 2011, pp. 330-349.

[5] X. Deng, Y. Ye, H. Li, J. Z. Huang, *An improved random forest approach for detection of hidden Web search interfaces*. In Machine Learning and Cybernetics, 2008 International Conference, Vol. 3, 2008, pp. 1586-1591.

[6] R. J. Mooney and U. Y. Nahm, *Text Mining with Information Extraction*. In Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, 2003, pp. 141-160.

[7] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Second Edition (Morgan Kaufmann Series in Data Management Systems), Morgan Kaufmann Publishers Inc., 2005

[8] J. Wang, *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*. Information Science Reference, 2008.

[9] Upasana and Chakravarty, *A Survey of Text Classification Techniques for E-mail Filtering*. In Second International Conference on Machine Learning and Computing, 2010. pp. 60-76.

[10] A. Liaw and M. Wiener, *Classification and Regression by randomForest*. In R News, Vol. 2/3, 2002. pp. 18-22.

[11] G. Biau, L. Devroye and G. Lugosi, *Consistency of random forests and other averaging classifiers*. In The Journal of Machine Learning Research, 2008. pp. 2015-2033.

[12] A. Khan, B. Baharudin, K. Khan, *Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification*. In Computer Engineering and Applications (ICCEA), Vol. 2, 2010, pp. 398-403.

[13] W. Yih, J. Goodman, V. R. Carvalho, *Finding Advertising Keywords on Web Pages*. In WWW '06 Proceedings of the 15th international conference on World Wide Web, 2006, pp. 213-222.

[14] C. Yadav, A. Sharan, M. Joshi, *Semantic Graph Based Approach for Text Mining*. In Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014, pp. 596-601.

[15] B. Xu, X. Guo, Y. Ye and J. Cheng, *An Improved Random Forest Classifier for Text Categorization*. In Journal of Computers, Academy Publisher, Vol. 7, No. 12, 2012, pp. 2913-2920.

[16] N. Belkin and W. B. Croft, *Information filtering and information retrieval: two sides of the same coin?*. In Magazine Communications of the ACM - Special issue on information filtering, Vol. 35, Issue 12, 1992. pp. 29-38.

[17] G. Rios, H. Zha, *Exploring support vector machines and random forests for spam detection*. In Proc. First International Conference on Email and Anti Spam (CEAS), 2004, pp. 398-403.

[18] I. Koprinska, J. Poon, J. Clark and J. Chan, *Learning to classify e-mail*. In Journal Information Sciences: an International Journal, Vol. 177, Issue 10, 2007. pp. 2167-2187.

[19] R. Agrawal, A. Gupta, Y. Prabhu and M. Varma, *Multi-Label Learning with Millions of Labels: Recommending Advertiser Bid Phrases for Web Pages*. In Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee. 2013, pp. 13-24

[20] S. Ravi, A. Broder, E. Gabrilovich, V. Josifovski, S. Pandey, B. Pang, *Automatic generation of bid phrases for online advertising*. In WSDM '10 Proceedings of the third ACM international conference on Web search and data mining, 2010, pp. 341-350.

[21] H. Mori, Univ. Meiji, Kawasaki and Y. Umezawa, *Credit Risk Evaluation in Power Market with Random Forest*. In IEEE International Conference on Systems, Man and Cybernetics, ISIC, 2007, pp. 3737-3742.

[22] J. W. Lee, J. B. Lee, M. Park, *An extensive comparison of recent classification tools applied to microarray data*. In Computational Statistics and Data Analysis. Vol. 48, Issue 4, 2005, pp. 869885.

[23] R. Díaz-Uriarte and S. Alvarez de Andrés, *Gene selection and classification of microarray data using random forest*. BMC Bioinformatics, 7:3, 2006, pp. 869885.

[24] D. Amaratunga, J. Cabrera and Y. Lee,*Enriched random forests*. In Bioinformatics, Vol. 24, No. 18, 2008, pp. 2010-2014.

[25] B.X. Xu, J.J. Li, Q. Wang and X.J. Chen, *A Tree Selection Model for Improved Random Forest*. In Proc. of the International Conference on Knowledge Discovery, 2011, pp. 382-386.

[26] Q. Wu, Y. Ye, H. Zhang, M. K. Ng and S. Ho, *FORES TEXTER: An efficient random forest algorithm for imbalanced text categorization*. Elsevier. Knowledge-Based Systems 67. 2014, pp. 105-116.

[27] Y. Xie, X. Li, E.W.T. Ngai and W. Yingc, *Customer churn prediction using improved baanced random forests*. Expert Systems with Applications, Vol. 36, Issue 3, Part 1, 2009, pp. 54455449.

[28] B. Van Essen, C. Macaraeg, M. Gokhale and R. Prenger, *Accelerating a Random Forest Classifier: Multi-Core, GP-GPU, or FPGA?*. In Field-Programmable Custom Computing Machines (FCCM), 2012 IEEE 20th Annual International Symposium on, 2012, pp. 232-239.

[29] S. Trivedi and S. Dey, *Effect of feature selection methods on machine learning classifiers for detecting email spams*. In the Proceedings of the 2013 ACM Research in Applied Computation Symposium. 2013, pp. 35-40.

[30] S. Trivedi and S. Dey, *An Enhanced Genetic Programming Approach for Detecting Unsolicited Emails*. In IEEE 16th International Conference on Computational Science and Engineering. 2013, pp. 1153-1160.