A Hybrid Embedded-Filter Method for Improving Feature Selection Stability of Random Forests

Conference Paper *in* Advances in Intelligent Systems and Computing · February 2017

DOI: 10.1007/978-3-319-52941-7_37

3 authors, including:

Afef Ben Brahim

19 PUBLICATIONS **137** CITATIONS

SEE PROFILE

Essoussi Nadia

Université de Tunis

60 PUBLICATIONS **232** CITATIONS

SEE PROFILE

A Hybrid Embedded-Filter Method for Improving Feature Selection Stability of Random Forests

Wassila Jerbi[1(✉)], Afef Ben Brahim[2(✉)], and Nadia Essoussi[1(✉)]

[1] LARODEC, Institut Supérieur de Gestion, Université de Tunis, Avenue de la Liberté, 2000 Le Bardo, Tunisie
jerbi_wassila@hotmail.fr, nadia.essoussi@isg.rnu.tn
[2] LARODEC, Tunis Business School, Université de Tunis, El Mourouj 2074, Tunisie
afef.benbrahim@yahoo.fr

Abstract. Many domains deal with high dimensional data that are described with few observations compared to the large number of features. Feature selection is frequently used as a pre-processing step to make mining such data more efficient. Actually, the issue of feature selection concerns the stability which consists on the study of the sensibility of selected features to variations in the training set. Random forests are one of the classification algorithms that are also considered as embedded feature selection methods thanks to the selection that occurs in the learning algorithm. However, this method suffers from instability of selection. The purpose of our work is to investigate the classification and feature selection properties of Random Forests. We will have a particular focus on enhancing stability of this algorithm as an embedded feature selection method. A hybrid filter-embedded version of this algorithm is proposed and results show its efficiency.

Keywords: Stability · Feature selection · Classification · High dimensional data · Random forests

1 Introduction

With the evolution of technology that keeps skyrocketing, we are devastated by tremendous amount of data. Therefore, new challenges are imposed in the machine learning field, which have to deal with a large amount of data in their different forms. Besides, recently we are talking more and more about two phenomenons, big data and high dimensional data. Which have drawn several researches. In fact, the increased volume of those type of data makes the learning process painful and less efficient. Henceforth, the pre-processing procedure is crucial to get more adapted data for learning algorithms. Evidently, the more numerous are the features, the higher is the risk of noisiness. For that cause, feature selection is an important step to pre-process data.

A. Abraham et al. (eds.), *Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016)*, Advances in Intelligent Systems and Computing 552,
DOI 10.1007/978-3-319-52941-7_37

Feature selection consists on minimizing as possible the number of features while keeping the original properties of data. It is the process of removing redundant and irrelevant features to make the learning algorithm more efficient, so that we can get better results.

Various researches have been interested on feature selection. However, these researches have focused on enhancing the predictive performance while neglecting the stability issue.

Stability is about preserving at most the same result concerning the selected feature subset, avoiding the variation in the selected set of features, even with small changes in the Data set.

In the present work, we look to get a robust feature selection. For that aim, we will investigate one of the ensemble feature selection techniques, namely Random Forest (RF) [4]. Indeed, we will provide some insights about RF and how it behaves as a feature selection technique. We will underline the instability problem which is caused by the intrinsic randomness in the algorithm design.

We will propose a solution to deal with that issue to get more robust feature selection.

2 Dimensionnality Reduction by Feature Selection

Interested in real world data, information industry collects huge data that are often complex and impractical for use. In several domains experts are faced to high dimensional data sets which have huge number of features and few number of observations. Therefore, mining in such forms of data is becoming a challenging task.

For better mining results we have to improve data quality by pre-processing data [8]. Feature selection is an important step in this process. Feature selection reduces the number of features under consideration by removing redundant, irrelevant or weakly relevant features which do not contribute on the classification process [8].

The crucial challenge with feature selection is about how to retain the minimum number of parameters which present the pertinent properties of the data [13], and how to preserve the original meaning of the data making interpretation more feasible. Besides, how can we uncover unlike features? The attribute selection techniques are various. They essentially divide into wrappers, filters, and embedded.

2.1 Filters

Filters assign a score for each attribute to obtain a feature ranking and then select the best subset of features. As they do not depend on a specific type of predictive model, they only take into consideration the intrinsic characteristics of the data [7].

Known to be not time consuming, they are faster than wrappers, but compared to embedded methods they have shown to be competitive in that

respect [7]. A commonly used filter method is the t-test. Which is a filter technique that assigns a score for each attribute. It is generally used to compare two normally distributed samples of population. The t-test is a statistical method that works better with features which have a maximal difference of mean value between groups and a minimal variability within each group [7].

2.2 Wrappers

Wrapper methods consist on assigning a score relative to the usefulness of the subsets of features. Wrappers consider a learning machine algorithm as a black box, it works as follow: firstly, a search algorithm gives a set of features that will be evaluated later by measuring the performance given by the learning algorithm. Thereafter, the set of selected features will be returned back for a next search iteration if it does not reach the required quality. However, if the classifier gives a good predictive performance the subset of features will be returned as the selected set of features. Iterative models like wrappers require massive amounts of computation. They are criticized to be time consuming. In addition, they suffer from the lack of generality since the resulted set of features depends on a specific classifier [7].

2.3 Embedded

As for wrappers, embedded methods depend on a specific learning algorithm. Besides, while the search and evaluation procedures are separated in the wrappers, the embedded method performs feature selection into the classifier construction using its internal parameters. Therefore, they are faster than wrappers and they are more efficient as they avoid the use of all the available data by not needing to divide it into a training set and a validation set [7]. Decision trees such as CART are famous example of embedded methods.

Decision tree induction for feature selection: Decision tree algorithms are often applied for classification task but they are also used as an embedded feature selection method. At each step, an attribute is chosen following an evaluation measure, in order to be represented in one of the tree nodes. Iteratively, the best feature is selected according to its discriminative power of separating different classes. This procedure is repeated until reaching some stopping criterion. Consequently, the obtained model is a tree that uses a specific subset of features. To put it differently, the reduced set of features can be determined from features appearing in nodes. Thus, feature selection has been performed implicitly into the algorithm [8]. Perceived as an ensemble form of decision trees, thus RF are also considered as an embedded feature selection.

3 Random Forests for Feature Selection

Aiming to get better results, the principle of ensemble learning has been proposed to enhance the prediction performance in the machine learning field. Ensemble

learning [5] relies on constructing a set of classifiers dealing with the same problem. Thus, to classify a new object, predictions made by individual classifiers, are combined by voting or averaging to get a final consensus [5].

In that context, looking to enhance classification performance, RF have been proposed by Breiman [4]. In fact, RF revolves around the generic principle of classifiers combination. It consists basically on generating a large number of trees, to let them later vote for the most popular class. In order to grow these ensembles, iteratively, we take a bootstrap sample from the available Data set to govern the construction of each tree in the ensemble. Then, to determine the split at each node, a random selection of features is applied [4].

RF gained a great interest as an ensemble method in the machine learning field [4], because they have significantly improved the classification performance and have remarkably proved its efficiency compared to single tree classifier [1]. For that reason the idea has been extended and has been adopted in the feature selection domain.

In addition, we can easily remark that the RF algorithm proceeds by selecting features which improve the most the predictive performance to put them in the tree nodes [4]. Under those circumstances, RF is considered as an embedded feature selection method which yields to high prediction accuracy.

3.1 Stability Issue of Random Forests

In several domains databases are constantly updated. Hence, many changes occur regularly whether with adding more observations or new features. Consequently, those small adjustments in the data set cause remarkable variation in the selected set of features and that is not practical for use. This occurs specially with data with small instances compared to the number of features [9]. Stability of feature selection is defined as the sensitivity of a given process to variations in the training set [9] i.e., it is obtained by measuring the similarity between different resulted set of features. In fact, it is about keeping at most the same resulted feature subset, even with small changes in the Data set. Many researches have been interested on stability of feature selection for high dimensional data, using the ensemble learning concept [11] or hybrid methods [3].

As described before, RF introduced by Leo Breiman [4] is an ensemble learning method that intends to enhance the predictive performance.

Here, we talk about its performance as an embedded feature selection due to selection that already occurs at each node of the tree. To be considered as an efficient feature selection method, it must satisfy the two criteria of a robust feature selection method, which are the classification performance and the stability.

In addition to the small sample size, RF use random components to generate diversity in the ensemble of trees and this is also a source of instability. Applied experiments on high dimensional data [10], have proved that the selected best features with RF changes dramatically even with little change on data. Therefore, used as a feature selection method RF have shown good results in prediction accuracy, whereas still the problem of stability which have not been resolved.

4 Hybrid Embedded-Filter Feature Selection

To speak about robust feature selection we have to assess the prediction performance and the stability of the selected features. So it is about finding a consensus between the stability of feature selection and the prediction accuracy, as they are both important for classification task. For that goal, we choose an algorithm that have already shown its performance in terms of prediction accuracy, and we will try to improve its stability index. In fact, the idea of our proposal is to take advantage of the standard RF algorithm to ensure a good classification accuracy, then working on improving its stability on selecting features, while preserving at most the high prediction accuracy.

The instability of RF is caused by random components involved in the learning algorithm. At first, the bagging method is used to select a bootstrap sample for each tree growth then the randomization is used to choose features for each split. Those two steps developed in the algorithm make harder to reproduce the same feature selection. However, the randomization is required to achieve diversity of trees in order to assure later, efficient results. Still, as we work on high dimensional data, even a little variation causes a significant change in the selection of a feature. Thus, bagging is very sufficient to get diverse trees. Hence, in order to improve the feature selection stability of RF we have to focus on the randomization and try to work on reducing its effects.

In our proposed method, we make the preference to eliminate one of the two random components and act through the choice of the split node. To make it clearer, in the basic RF algorithm, each time to find the split node, a number of features are selected randomly. In our approach, as argumented we look to

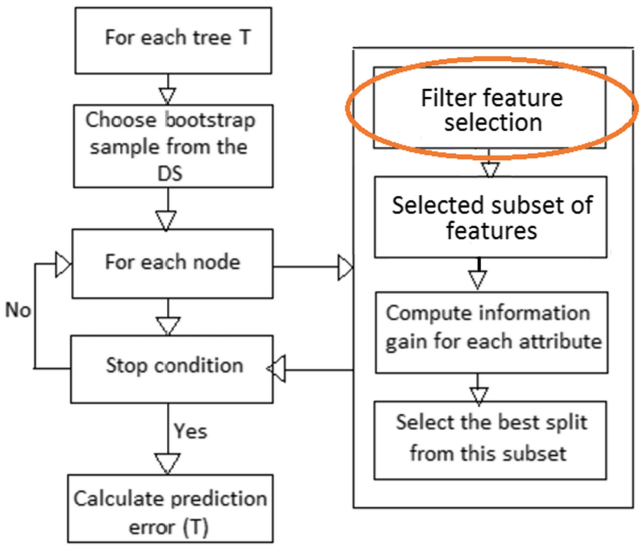

Fig. 1. Flowchart: hybrid random forests.

eliminate randomness at this level, so we proceed by replacing the random choice of the subset of features by introducing a filter method. Consequently, these features will be selected using a filter method instead of randomness. Figure 1 describes the process of these hybrid approach. Notably, we look to minimize the randomness to get more stable selection.

Furthermore, we will also propose different combination methods trying to more evaluate the stability. Accordingly, at first, features are ranked according to their importance on the classification process and we preserve only 1% of the top ranked features to test the stability. Then, we will proceed by diversifying the number of features with different combination strategies; top ranked features and weight aggregation.

5 Experimental Study

As described above, our proposal consists on replacing the random selection which occurs for the split node by a filter selection method. For this purpose, in our hybrid feature selection experiments we introduce the t-test, one of the filter selection methods known to be fast and efficient [7]. To better analyze the performance of our proposal, we evaluate the resulted hybrid RF model for different settings where we variate the number of trees in the forest (nTrees) from 5 to 500.

To assess our approach we use 10 fold cross validation consisting on using 90% of the data as a training set. This protocol allows us simulating small changes in the data, so we get diverse training sets to generate different RF. Then we perform a comparison between results obtained by the basic RF algorithm and the proposed hybrid RF algorithm.

In a second experimental setting, to evaluate our approach we apply two different methods that will serve to combine the feature subsets obtained by each forest. For that, we use the top rank method and the rank aggregation method. Our aim is to see whether the combination technique affects results with the focus on obtaining the best stability.

5.1 Data Sets

Three data sets are used in this experimental process; Lymphoma, Bladder and DLBCL. Characterized by thousands of features and tens of instances, we employ this specific type of data sets known as 'high dimensional data' because they present great challenges for classification and feature selection algorithms. Table 1 contains brief description of the used databases.

5.2 Evaluation Metrics

To evaluate our approach we use the stability measure (Kuncheva). This index ranges within [−1,1], the higher is this index the more are common features across the different sets. As we can not work on improving the stability measure while neglecting the prediction accuracy, we adopt the (F-measure) which is a performance evaluator that describes the harmonic mean of precision and recall.

Table 1. Data set characteristics.

Name	#instances	#Features	References
Lymphoma	45	4026	[2]
Bladder	31	3036	[6]
DLBCL	77	7029	[12]

5.3 Results

To estimate the robustness of the embedded-filter feature selection we proceed as explained above. We run both algorithms (basic RF and hybrid RF) across different data sets. Only 1% of features are chosen to test stability i.e., for lymphoma data set we keep 40 features from the top ranked features, for bladder data base the number of retained features will be 30 and for DLBCL we perserve 70 features.

Table 2 shows the stability and prediction results while running the basic RF algorithm and then, the hybrid algorithm (RF with t-test). We can easily see the increase of stability for all data sets with the hybrid version using the t-test. However, we remark that F-measure is slightly better when running the basic RF algorithm. At first sight, some can argue that the increase in stability, comes at cost of lower accuracies. But once focusing more in the results we can deduce that the tiny decrease of F-measure can be neglected via the significant increase of stability. So, we can deduce that the predictive performance is at most the same while the stability has remarkably improved.

It can be observed that the hybrid feature selection algorithm (RF with t-test) provides more robust feature selection comparing to the basic algorithm (RF). However, the difference on the stability is dependent of the data set and other parameters like the number of trees in the forest (nTrees). Besides, it is

Table 2. 10 fold CV F-measure and Kuncheva index for stability measure are evaluated with data variation.

Data set			nTrees=5	nTrees=50	nTrees=100	nTrees=200	nTrees=500
Lymph	RF (basic Algo)	Fm	0.7826	0.9111	0.9583	0.9362	0.9565
		Stab	0.6482	0.1566	0.1746	0.1723	0.1712
	RF(with t-test)	Fm	0.7755	0.9167	0.9362	0.9131	0.9235
		Stab	0.8367	0.4041	0.5264	0.4810	0.4822
Bladder	RF (basic Algo)	Fm	0.8182	0.8696	0.8696	0.8696	0.9091
		Stab	0.7598	0.1187	0.1269	0.1157	0.1441
	RF(with t-test)	Fm	0.8000	0.8571	0.8571	0.8571	0.8901
		Stab	0.8526	0.3379	0.3237	0.3663	0.3973
DLBCL	RF (basic Algo)	Fm	0.9106	0.9106	0.9421	0.9180	0.9470
		Stab	0.7246	0.0958	0.1610	0.1916	0.1998
	RF(with t-test)	Fm	0.9500	0.9076	0.9412	0.9178	0.9468
		Stab	0.7486	0.4071	0.4521	0.4989	0.5049

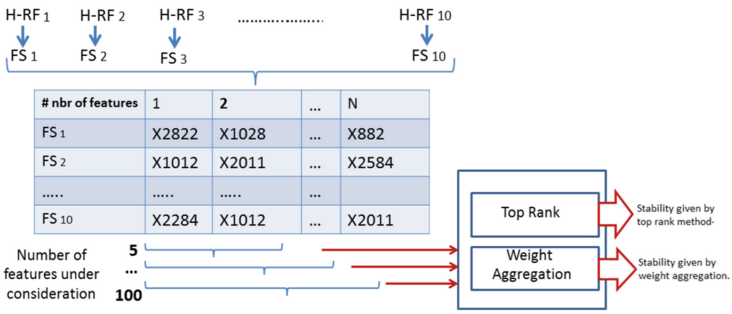

Fig. 2. Testing different feature combination techniques while variating the number of features under consideration.

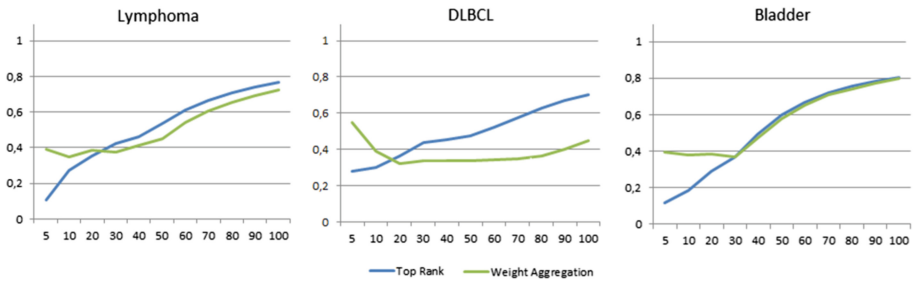

Fig. 3. Stability results of Hybrid RF using the top rank and weight aggregation methods with various feature cardinality.

important to underline that for all data sets the stability measure increases while running the new approach but with various expansion rates.

Aiming to more evaluate the stability of our proposal, we investigate other strategy. As described in Fig. 2, The weight aggregation and the top rank methods are tested with various number of features.

Looking to get a higher Kuncheva index. The results given by applying these proceeds at the three data sets are illustrated in Fig. 3. Let's denote that FS_n refers to the Feature Selection obtained from the hybrid RF $H-RF_n$. In the first part of the experiments, to evaluate stability we proceed by considering only 1% of the top ranked features while diversifying the ensemble size. In this part, as described in Fig. 2, we have fixed the ensemble size at 50 trees, and we tested the stability results with various feature cardinality [5...100] using top ranked and weight aggregating methods.

As it can be observed Fig. 3, the weight aggregation leads to better results than top rank methods with a number of features lower than 20. However, when the number of features under consideration increases, the stability with top rank process exceeds the one obtained by weight aggregation. Moreover, the stability of our hybrid random forests measured with kuncheva index is improved while

increasing the number of retained features, and reaches 0.8 using 100 features for Bladder data set (Fig. 3).

So that, we can conclude that our embedded-filter feature selection provides high stability measure while considering a substantial number of features. Furthermore, it is important to underline that in addition to these high stability results we have obtained a good classification performance that varies between 0.8 and 0.9.

Under those circumstances, we get through hybrid method to deliver a modified random forests that provide a stable feature selection while maintaining a good classification performance.

6 Conclusion

In this work, our objective was to obtain a robust feature selection. We have proposed an hybrid approach, based on joining both embedded and filter feature selection methods. The main idea of our proposal is to take advantage of the standard random forests algorithm to ensure a good classification accuracy, then working on improving its stability on selecting features. For this purpose, we used to reduce the effect of randomization involved in the standard algorithm. So, instead of proceeding by choosing randomly a number of features we proceed by applying a filter feature selection (t-test). Our experimental study is a combined analyze of predictive performance and robustness of feature selection. The experimentation shows that our hybrid method, based on joining filter method (t-test) to embedded method (RF), finds a trade-off between these two important criteria. Thus, our initial objective is satisfied proving the efficiency of our proposal.

References

1. Ali, J., Khan, R., Ahmad, N., Maqsood, I.: Random forests and decision trees. Int. J. Comput. Sci. Issues (IJCSI) **9**(5), 1–7 (2012)
2. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature **403**(6769), 503–511 (2000)
3. Ben Brahim, A., Limam, M.: A hybrid feature selection method based on instance learning and cooperative subset search. Pattern Recogn. Lett. **69**(C), 28–34 (2016)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000). doi:10.1007/3-540-45014-9_1
6. Dyrskjøt, L., Thykjaer, T., Kruhøffer, M., Jensen, J.L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., Ørntoft, T.F.: Identifying distinct classes of bladder carcinoma using microarrays. Nat. Genet. **33**(1), 90–96 (2003)
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)

8. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Data Management Systems. Morgan Kaufmann, San Francisco (2000)
9. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. **12**(1), 95–116 (2007)
10. Li, S., Harner, E.J., Adjeroh, D.A.: Random KNN feature selection-a fast and stable alternative to random forests. BMC Bioinformatics **12**(1), 1 (2011)
11. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008). doi:10.1007/978-3-540-87481-2_21
12. Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., et al.: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. **8**(1), 68–74 (2002)
13. van der Maaten, L.J.P., van den Herik, H.J.: Dimensionality reduction: A comparative review. Technical report. Tilburg Centre for Creative Computing, Tilburg University, Tilburg, Netherlands Technical Report: 2009–005 (2009)