Computing (2020) 102:1187–1198
https://doi.org/10.1007/s00607-019-00785-6

Resource-efficient fast prediction in healthcare data analytics: A pruned Random Forest regression approach

Khaled Fawagreh[1] · Mohamed Medhat Gaber[2]

Received: 16 April 2019 / Accepted: 23 December 2019 / Published online: 9 January 2020

Abstract
In predictive healthcare data analytics, high accuracy is both vital and paramount as low accuracy can lead to misdiagnosis, which is known to cause serious health consequences or death. Fast prediction is also considered an important desideratum particularly for machines and mobile devices with limited memory and processing power. For real-time health care analytics applications, particularly the ones that run on mobile devices, such traits (high accuracy and fast prediction) are highly desirable. In this paper, we propose to use an ensemble regression technique based on *CLUB-DRF*, which is a pruned Random Forest that possesses these features. The speed and accuracy of the method have been demonstrated by an experimental study on three medical data sets of three different diseases.

Keywords Random Forest · Healthcare data analytics · CLUB-DRF · Ensemble classification · Ensemble pruning

Mathematics Subject Classification 68T10 · 62H30

1 Introduction

Random Forest (RF) has proven its effectiveness as a classification and a regression method in a variety of applications [10]. In [11], a new method termed *CLUB-DRF* was introduced to select diverse decision trees drawn from groups of similar trees (i.e. clusters of trees), to form a pruned Random Forest ensemble that is much smaller than the initial and traditional RF ensemble [7], and yet, performs at least as good as

✉ Mohamed Medhat Gaber
mohamed.gaber@bcu.ac.uk

Khaled Fawagreh
kfawagreh@pmu.edu.sa

[1] Prince Mohammad Bin Fahd University, Dhahran, Saudi Arabia

[2] Birmingham City University, Birmingham, UK

the original ensemble. Since a well established principle in ensemble classification and regression is that ensembles tend to perform better when the individual classifiers in the ensemble exhibit a high level of diversity [1,8,19,31], we have adopted data clustering [17] to inject more diversity into an RF, and to prune the RF, leading to a faster inference. The premise is that grouping of similar classifiers in clusters according to their classification patterns, and then choosing a representative classifier (or more) of each cluster can result in a pruned and more diversified ensemble.

Since diversity can lead to better performance as previously described, in a nutshell, there are two levels of diversity already applied in an RF. The first level is when each decision tree is constructed using sampling with replacement from the training data. The samples are likely to have some diversity among each other as they were drawn at random. The second level is achieved by randomisation which is applied when selecting the best node to split on. The ultimate objective of our new method is to add a third level of diversity by injecting more diversity in an RF using clustering as described in Sect. 3.3.

In [11], *CLUB-DRF* has proved its effectiveness. With at least 92% or above pruning level, while retaining or outperforming the RF accuracy (before pruning). Since in predictive healthcare data analytics low accuracy can lead to improper diagnosis, which in turn can be both fatal and devastating, in this paper, we aim to target one domain only, namely, the health domain. Hence, we apply our proposed method *CLUB-DRF* on three medical data sets of three different diseases. It is worth noting that in this paper, it is the first time to apply *CLUB-DRF* for regression. Thus, the method has been modified by replacing a clustering technique (k-modes) that is used for categorical data by another that is tailored to operate on numerical data (k-means). This is due to the replacement of the type of output between the two approaches (i.e. categorical for classification, and numerical for regression).

This paper is organised as follows. Section 2 presents related work in the domain of healthcare data analytics. Overview and detailed description of our proposed new method are given in Sect. 3. Experiments and results are presented in Sect. 4. The paper is finally concluded with a summary and pointers to future work in Sect. 6.

2 Related work

Healthcare data analytics, also known as clinical data analytics, is the gathering and interpretation of data from a variety of sources (e.g. the electronic health record, billing claims, cost reports, and patient satisfaction surveys) to help organisations improve the quality of care, lower the cost of care, and enhance the patient experience. According to [26], several researchers have conducted studies which proved that, by utilising healthcare data analytics technologies, they were able to reduce mortality rates, healthcare costs, and medical complications at various hospitals. Moreover, healthcare analytics has the potential to reduce costs of treatment, avoid preventable diseases, predict outbreaks of epidemics, avoid preventable deaths, and improve the quality of life in general [28]. According to a recent Research and Markets report [27], healthcare analytics is expected to rise and reach a $34.27 billion industry by the end of 2022. To improve healthcare as described above, healthcare analytics became a hot area of interest and attracted attention from diverse disciplines such as databases, data mining, information retrieval, image processing, medical researchers, and healthcare practitioners [26]. In this paper, we will exploit one of data mining's tasks known as regression to improve the performance of the traditional RF on medical data sets, both in terms of accuracy and regression speed.

The prediction of chronic kidney diseases by using Decision Tree (C4.5) algorithms was investigated by [5]. In terms of accuracy and execution time, the classifier used proved its performance. A new method that is based on linear regression to correct Partial Volume Effects (PVE) in Arterial Spin Labeling (ASL) MRI was developed by [3]. An online Stochastic Gradient Descent (SGD) algorithm with logistic regression is implemented by [23] using Apache Mahout to develop the best scalable diagnosis model. The proposed prediction model achieved 81.99% accuracy for training sample and 81.52% accuracy for testing sample.

The prediction of three diseases, namely, leukemia, lung cancer, and heart disease was investigated by [24]. To do this, the researchers used three different classifiers: Naive Bayes, C4.5, and Random Forest. According to the researchers, the proposed method has better accuracy, precision, recall and Fmeasure.

Using stepwise logistic regression models, Higdon et al. [13] developed a model based on three different counts of medications: outpatient, inpatient drug classes, and individual inpatient drug names. The model was used to rank the patient population for medical complexity. Thanks to this model, based on the number and type of medications, simple admission screens for predicting the complexity of patients were implemented.

The approach offered in this paper aims at applying a pruned RF regression approach on medical data sets of different diseases. In the pruned RF ensemble produced by our approach, clustering is used as the main diversity approach for selecting diverse trees that form the pruned RF ensemble. The following section provides greater insight about the proposed approach.

3 CLUB-DRF for regression

In this section, we present our *CLUB-DRF* approach starting with a general overview of how it works. After this, a more nuts-and-bolts description of the method is presented including the algorithm and other supporting details.

3.1 Random Forest: an overview

Random Forest (RF) is an ensemble learning method used for classification and regression. Developed by Breiman [7] over a decade ago, the method combines Breiman's bagging sampling approach [6], and the random selection of features, introduced independently by [14,15] and Amit and Geman [2], in order to construct a collection of decision trees with controlled variation. Using bagging, each decision tree in the ensemble is constructed using a sample with replacement from the training data. Statistically, the sample is likely to have about 64% of instances appearing at least once in the sample. Instances in the sample are referred to as in-bag-instances, and the remaining instances (about 36%), are referred to as out-of-bag (OOB) instances.

To enhance diversity, at each node in all trees, a best split feature is selected using a goodness measure (e.g. Gini index) from a set of randomly selected features (typically $\sqrt{n}$, where n is the total number of features). Each tree is grown to the largest extent possible and is unpruned. A maximum depth is usually allowed to prevent trees from growing out of memory in high dimensional data sets.

3.2 CLUB-DRF

RF algorithms tend to build between 100 and 500 trees [12]. Some empirical and theoretical studies have also clearly demonstrated that adding more trees to an RF beyond a certain number (i.e. 500) won't necessarily improve the RF accuracy [4]. Our research aims at pruning RF ensembles by producing subsets of the original ones that are significantly smaller in size and yet, have accuracy performance that is at least as good as that of the original RF from which they were derived. In other words, we aim at finding the optimal or near-optimal ensemble of trees that will be used to generate an accurate RF.

As mentioned earlier, to create groups of similar trees, clustering will be used. This novel technique has been used extensively as a diversity technique in many applications [9,18,20,21,29,30]. Unlike classification, clustering is an unsupervised learning technique that attempts to organise objects into clusters (groups) where the members in one cluster are more similar to each other than those members in other clusters. Each group is referred to as a cluster, hence, a cluster is a group of similar objects which are dissimilar to other objects belonging to other clusters. Clustering is considered a data exploration method as it helps to unveil the natural grouping in a data set without a prior knowledge of the groups to be produced. One of the earliest and most popular clustering algorithms is called K-means. It was developed by MacQueen et al. [22] in the late sixties and despite its seniority, it is still considered as one of the most widely used algorithms, mainly due to its simplicity, efficiency, and empirical success [16]. We have used clustering in [11] to produce a pruned RF classification technique for a variety of data sets from a variety of domains.

3.3 CLUB-DRF: an adapted method for regression

In this section, we propose an enhancement of RF called *CLUB-DRF* that spawns a child RF that is (1) much smaller in size than the parent RF and (2) has an accuracy that is at least as good as that of the parent RF. In the remainder of this article, we will refer to the parent/original traditional RF as simply *parentRF*, and refer to the resulted child RF based on our method as *CLUB-DRF*.

Figure 1 shows the *CLUB-DRF* approach and the corresponding algorithm is displayed in Algorithm 1 where T refers to the training data set, and S refers to the size of the *parentRF* to be created . The constant k refers to the number of clusters to be created which we define as a multiple of 5 in the range 5–50. This way and as we shall see in the experimental section, we can compare the performance of *CLUB-DRF* of

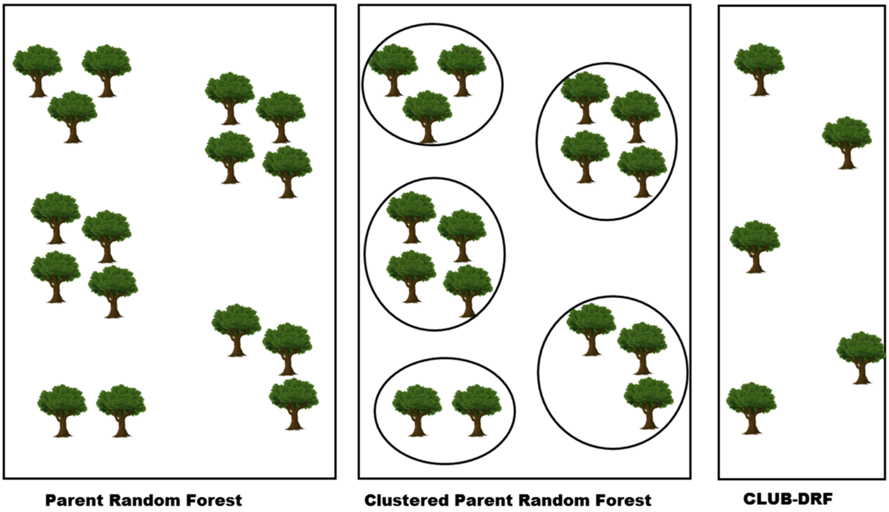

Fig. 1 CLUB-DRF approach

different sizes with that of *parentRF*. As outlined in the experimental section, the size of the *parentRF* to be created is 500 trees. When the number of clusters is multiple of 5 in the range 5–50, this means the pruning levels will be in the range 99–90% respectively, which we consider a reasonable range for extreme pruning.

Algorithm 1 CLUB-DRF Algorithm

```
{User Settings}
input $T$, $S$, $k$
{Process}
Create an empty super ordered list $AllPredictions$
Create an empty ordered list $T_{rf}$ to represent parentRF
Create an empty ordered list $T_{clubdrf}$ to represent CLUB-DRF
Using the traditional Random Forests Algorithm, create $T_{rf}$ of size $S$
For each tree in $T_{rf}$, find its predictions on T and add it to AllPredictions
for $i = 1 \rightarrow S$ do
    $AllPredictions = AllPredictions \cup R(T_{rf}.tree(i), T)$
end for
Using K-means, cluster $AllPredictions$ into a set of $k$ clusters: $cluster_1 \ldots cluster_k$ $clusters \leftarrow MakeClusetrs(AllPredictions)$
From each cluster, find a representative tree and add it to $T_{clubdrf}$
for $i = 1 \rightarrow k$ do
    $repTree \leftarrow FindRep(clusters(i))$
    Add repTree to $T_{clubdrf}$
end for
{Output}
An ordered list of trees $T_{clubdrf}$
```

As shown in Algorithm 1, a clustering-based technique is applied to produce diverse groups of trees in the *parentRF*. Assuming that the trees in the *parentRF* are denoted by the ordered list $T_{rf} = \langle t_1, t_2, \ldots, t_n \rangle$ (where n is number of trees in the *parentRF*), and the training set is denoted by $T = \{r_1, r_2, \ldots, r_m\}$. Each tree in T_{rf} is used to regress each record in the training set to determine the class label c. We use $R(t_i, T)$ (where $t_i \in T_{rf}$) to denote an ordered list of continuous values obtained after having t_i regresses the training set T. That is, $R(t_i, T) = \langle c_{i1}, c_{i2}, \ldots, c_{im} \rangle$. The result obtained of having each tree regress the training records will therefore be a super ordered list *AllPredictions* containing ordered lists of continuous values produced by each tree. That is,

$$AllPredictions = R(t_1, T) \cup R(t_2, T) \cup \cdots R(t_n, T)$$

This super ordered list is fed as input to a K-means clustering algorithm as shown in Algorithm 1. When clustering is completed, we should have a set of clusters where each cluster contains ordered lists that are similar and likely to have the least number of discrepancies.

It is important to remember that the number of trees of the resulted *CLUB-DRF* is determined by the number of clusters used. For example, if the number of the clusters is 5, then the resulted *CLUB-DRF* will have 5 trees, and so on.

The final step in the algorithm is to select a representative from each cluster. To do this, from each cluster, we pick the tree that has achieved the highest accuracy on the training data. It is worth mentioning that this is not the only way to select a representative. One other way is to randomly select a tree from each cluster without using accuracy as the main selection criterion in the selection process. Yet another method is to pick the tree that has achieved the highest performance on the out-of-bag (OOB) instances. These are the instances that were not included in the sample with replacement that was used to build the tree, and they account for about 36% of the total number of instances. Using the OOB samples to evaluate a tree gives an unbiased estimate of its predictive accuracy since, unlike training data that was seen by the tree when it was built, OOB data was not seen and therefore, it is a more accurate measure of the tree's predictive accuracy. We are considering these new methods for selecting representative selection in future research as covering these methods is beyond the scope of this paper.

4 Experimental study and results

In this section, an experimental study of our approach on three medical regression data sets of three different diseases will be presented. These diseases are parkinson's, diabetes, and breast-cancer. Basic information about these data sets are given in Table 1. The first data set was obtained from the University of Waikato Repository, and the last two from the University of California, Irvine (UCI) Machine Learning Repository. The experiments were conducted on a laptop running Windows 7 Enterprise, with a dual processor of 2.60 Ghz and 8 GB of RAM.

The diabetes data set contains the necessary attributes to predict the dependence of the level of serum C-peptide on the various other factors in order to understand the patterns of residual insulin secretion. In a scale of 1–4, the attributes in the the breast-

Table 1 Experiments data sets

Data set name	Number of features	Number of instances
Diabetes	3	43
Breast-cancer	11	700
Parkinson’s	26	5875

Table 2 Experiments results: diabetes data set

(CLUB-DRF)(RF)	MAE	MSE	RMSE	R^2
ParentRF (500 trees)	0.64	0.62	0.78	−0.46
5	(**0.42**)(0.64)	(**0.45**)(0.67)	(**0.66**)(0.8)	(−**0.29**)(−0.8)
10	(**0.39**)(0.63)	(**0.42**)(0.64)	(**0.64**)(0.78)	(−**0.2**)(−0.59)
15	(**0.37**)(0.63)	(**0.4**)(0.62)	(**0.62**)(0.77)	(−**0.15**)(−0.52)
20	(**0.34**)(0.63)	(**0.37**)(0.61)	(**0.59**)(0.77)	(−**0.05**)(−0.48)
25	(**0.33**)(0.62)	(**0.35**)(0.6)	(**0.58**)(0.76)	(−**0.01**)(−0.44)
30	(**0.32**)(0.62)	(**0.34**)(0.6)	(**0.57**)(0.77)	(**0.02**)(−0.49)
35	(**0.32**)(0.62)	(**0.34**)(0.6)	(**0.56**)(0.76)	(**0.03**)(−0.51)
40	(**0.31**)(0.62)	(**0.33**)(0.6)	(**0.56**)(0.76)	(**0.06**)(−0.48)
45	(**0.31**)(0.63)	(**0.34**)(0.61)	(**0.56**)(0.77)	(**0.04**)(−0.48)
50	(**0.31**)(0.63)	(**0.33**)(0.62)	(**0.56**)(0.77)	(**0.05**)(−0.5)

cancer data set are used to predict the type of the breast-cancer. Using a scale range of bio-medical voice measurements from 42 people with early-stage Parkinson’s disease, the attributes in the parkinson’s data set are used to predict the clinician’s Parkinson’s disease symptom score on the Unified Parkinson’s Disease Rating Scale (UPDRS) scale.

To use the holdout testing method, which is the simplest type of cross validation, each data set was divided into sets: training and testing. Two thirds (66%) were reserved for training and the rest (34%) for testing. The selection of the two sets was done randomly using uniform distribution, where each instance has the same probability of being selected. The size of the Parent Random Forest (refer back to Fig. 1), which we called *parentRF*, was 500 trees; a typical setting for Random Forest [12]. This setting was chosen for two main reasons. First, the more trees we have, the more diverse ones we can get. Secondly, the more trees there are, the more unlikely the problem of empty clusters [25] does not surface.

The *CLUB-DRF* algorithm outlined in Algorithm 1 was implemented using the Python programming language utilising its machine learning library Scikit-learn. It was run 10 times on each data set where a new *parentRF* was created in each run. The average of the 10 runs for each resulted *CLUB-DRF* was calculated to produce the average for a variety of metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R Squared.

Table 3 Experiments results: breast-cancer data set

((CLUB-DRF)(RF)	MAE	MSE	RMSE	R^2
ParentRF (500 trees)	0.13	0.26	0.51	0.71
5	(**0.04**)(0.13)	(**0.08**)(0.26)	(**0.28**)(0.51)	(**0.91**)(0.72)
10	(**0.04**)(0.13)	(**0.08**)(0.25)	(**0.28**)(0.5)	(**0.9**)(0.72)
15	(**0.04**)(0.13)	(**0.08**)(0.25)	(**0.28**)(0.5)	(**0.91**)(0.72)
20	(**0.04**)(0.13)	(**0.09**)(0.26)	(**0.29**)(0.5)	(**0.9**)(0.72)
25	(**0.04**)(0.13)	(**0.09**)(0.26)	(**0.30**)(0.51)	(**0.9**)(0.72)
30	(**0.04**)(0.13)	(**0.09**)(0.26)	(**0.29**)(0.51)	(**0.9**)(0.72)
35	(**0.04**)(0.13)	(**0.09**)(0.26)	(**0.29**)(0.5)	(**0.9**)(0.72)
40	(**0.04**)(0.13)	(**0.09**)(0.26)	(**0.29**)(0.5)	(**0.9**)(0.72)
45	(**0.04**)(0.13)	(**0.09**)(0.26)	(**0.3**)(0.5)	(**0.9**)(0.72)
50	(**0.05**)(0.13)	(**0.09**)(0.26)	(**0.3**)(0.51)	(**0.9**)(0.71)

Table 4 Experiments results: parkinson's data set

(CLUB-DRF)(RF)	MAE	MSE	RMSE	R^2
ParentRF (500 trees)	0.04	0.0	0.05	0.69
5	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.89**)(0.69)
10	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.9**)(0.69)
15	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.89**)(0.69)
20	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.90**)(0.68)
25	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.89**)(0.68)
30	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.89**)(0.68)
35	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.89**)(0.68)
40	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.89**)(0.68)
45	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.89**)(0.69)
50	(**0.01**)(0.04)	(0.0)(0.0)	(**0.03**)(0.05)	(**0.89**)(0.69)

Results of our experiments on the three data sets **diabetes**, **breast-cancer**, and **parkinson's** are displayed in the Tables 2, 3 and 4 respectively. In each of these tables, the second row displays performance metrics of the *parentRF*. Following that, performance metrics of *CLUB-DRFs* of sizes in the range 5–50, reflecting extreme pruning levels from 99 to 90% respectively, as previously described in Sect. 3.3, are listed. To show the effectiveness and efficiency of *CLUB-DRF* over the traditional RF, we also listed performance metrics for traditional RFs in the same range that were derived from the *parentRF*. Each performance metric is represented as a couple of numbers each enclosed between parentheses, where the first number is for *CLUB-DRF*, and the second for a traditional RF. If the first number for *CLUB-DRF* is highlighted in boldface, it means that *CLUB-DRF* outperformed its counterpart traditional RF of the same size.

Table 5 Inference time per instance (in μs)

(CLUB-DRF)(RF)	Diabetes	Breast-cancer	Parkinson’s
ParentRF (500 trees)	2030.05	173.32	182.27
5	(**16.67**)(23.33)	(2.10)(1.50)	(**1.50**)(1.75)
10	(**50.00**)(80.00)	(**1.05**)(2.10)	(**3.68**)(3.83)
15	(**33.33**)(130.00)	(4.20)(3.15)	(**4.55**)(5.61)
20	(**100.00**)(163.33)	(**2.1**)(6.30)	(**6.71**)(7.18)
25	(**66.67**)(163.33)	(14.71)(9.45)	(**8.41**)(8.66)
30	(**133.33**)(180.00)	(**6.30**)(13.66)	(**9.51**)(10.79)
35	(**150.00**)(196.67)	(**11.55**)(14.71)	(**11.01**)(12.66)
40	(**166.67**)(230.01)	(**9.45**)(16.81)	(15.52)(14.67)
45	(**100.00**)(246.67)	(**14.71**)(16.81)	(**14.37**)(16.12)
50	(**183.34**)(246.67)	(**12.61**)(19.96)	(**16.42**)(17.39)

Table 5 shows the inference time per instance (ITPI) in microseconds for the three data sets diabetes, breast-cancer, and parkinson’s. This refers to the time needed to predict all the instances in the testing data set divided by the number of instances. As in the previous tables, we have highlighted in boldface the first entry when *CLUB-DRF* achieved less ITPI than its counterpart traditional RF.

5 Discussion

To test our *CLUB-DRF* method, the training and testing data sets should be prepared, as aforementioned. The training data set is fed as input to *CLUB-DRF* which constructs an initial RF of 500 trees as was outlined in Sect. 3.2. Pruned RF of sizes 5–50 are then produced by *CLUB-DRF* which are then used to predict the value of the target feature for each instance in the data set. Since *CLUB-DRF* generates several pruned forests (with different sizes), and reports the performance of each forest, it is recommended to pick the forest that has achieved the highest performance in terms of accuracy for deployment.

As depicted in Tables 2, 3 and 4, it is obvious from the key performance indicator MAE (Mean Absolute Error) that *CLUB-DRF*, regardless of its size, outperformed not only the *parentRF*, but also a traditional RF of the same size. Furthermore, our approach seems to be insensitive to the dimension and size of the data sets. As outlined in Table 1, the 3 data sets have different dimensions and sizes and yet, *CLUB-DRF* performed consistently well as was demonstrated in Tables 2, 3 and 4. An interesting observation in these tables is that *CLUB-DRF* performs even better as the number of dimensions and the data set size increase.

As for the he inference time per instance (ITPI) in Table 5, it is easy to see that in almost all cases, *CLUB-DRF* has outperformed the traditional RF. The interesting thing is that, not only *CLUB-DRF* performed better accuracy-wise as was demonstrated in Tables 2, 3 and 4, but it also achieved better ITPI as was demonstrated in Table 5.

The results show that R Squared reached 0.9 for CLUB-DRF in both the breast-cancer and parkinson's data sets. This implies that 90% of the variance in the data has been explained. Contrasting with traditional RF, the best performing RF for the breast-cancer measured in R Squared is 0.72, and for the parkinson's data set is 0.69. This clearly shows the superiority of CLUB-DRF in explaining the variance in the data. We argue that the diversity created through the clustering process is the main reason for this performance boost up. However, for the diabetes data set, RF seems not to be able to explain the variance in the data. For CLUB-DRF with the number of trees greater than or equal to 30, the model is able to explain some of the variance (6% when the number of trees is 40). It is worth noting that there are two observations in this case: (1) the diabetes data set is small (only 43 instances); and (2) CLUB-DRF has shown positive values for R Squared when the traditional RF was not able to show any case of a positive value.

6 Conclusion and future work

In predictive healthcare data analytics applications, it is imperative for such applications to be as accurate as possible to minimise misdiagnosis which can be fatal sometimes. To ensure proper diagnosis of diseases, we presented in this paper a Random Forest regression-based prediction approach that not only is more accurate than the traditional Random Forest, but also runs faster due to the small size of the ensembles it produces. To achieve both accuracy and speed, we empirically validated the principle that diversity in ensembles can lead to better performance. To do this, we have used clustering as a diversity technique to further diversify the traditional RF, resulting in *CLUB-DRF*. Since trees in the original ensemble (*parentRF*) were clustered into groups of similar trees, and a representative was selected from each group, many redundant trees were eliminated. Hence, *CLUB-DRF* ensembles with extreme pruning levels reaching as high as 99% were produced. As was demonstrated in the results obtained in Sect. 4, these *CLUB-DRF* ensembles not only run faster due to their small size, but also perform at least as good as the *parentRF* mainly due to the high level of diversity in their constituent trees.

As future work, the application of the proposed method on other healthcare data sets will be applied for both regression and classification problems. Also the deployment of the pruned models on handheld devices like smartphones will be experimented. Due to the small size and fast prediction, personalised healthcare through deployment of these models can be an interesting prospect.

References

1. Adeva JJG, Beresi U, Calvo R (2005) Accuracy and diversity in ensembles of text categorisers. CLEI Electron J 9(1):1–2
2. Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. Neural Comput 9(7):1545–1588
3. Asllani I, Borogovac A, Brown TR (2008) Regression algorithm correcting for partial volume effects in arterial spin labeling MRI. Magn Reson Med 60(6):1362–1371
4. Bernard S, Heutte L, Adam S (2009) On the selection of decision trees in random forests. In: International joint conference on neural networks, 2009. IJCNN 2009. pp 302–307
5. Boukenze B, Mousannif H, Haqiq A (2016) Predictive analytics in healthcare system using data mining techniques. Comput Sci Inf Technol 1:1–9
6. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
7. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
8. Brown G, Wyatt J, Harris R, Yao X (2005) Diversity creation methods: a survey and categorisation. Inf Fusion 6(1):5–20
9. Brown RD, Martin YC (1998) An evaluation of structural descriptors and clustering methods for use in diversity selection. SAR QSAR Environ Res 8(1–2):23–39
10. Fawagreh K, Gaber MM, Elyan E (2014) Random forests: from early developments to recent advancements. Syst Sci Control Eng Open Access J 2(1):602–609
11. Fawagreh K, Gaber MM, Elyan E (2015) CLUB-DRF: A clustering approach to extreme pruning of random forests. In: International conference on innovative techniques and applications of artificial intelligence. Springer, pp 59–73
12. Graham W (2011) Use R: data mining with rattle and R: the art of excavating data for knowledge discovery. Springer, Berlin
13. Higdon R, Stewart E, Roach JC, Dombrowski C, Stanberry L, Clifton H, Kolker N, van Belle G, Del Beccaro MA, Kolker E (2013) Predictive analytics in healthcare: medications as a predictor of medical complexity. Big Data 1(4):237–244
14. Ho TK (1995) Random decision forests. In: Proceedings of the third international conference on document analysis and recognition, 1995, volume 1. IEEE, pp 278–282
15. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844
16. Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recognit Lett 31(8):651–666
17. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv (CSUR) 31(3):264–323
18. Kuncheva LI, Hadjitodorov ST (2004) Using diversity in cluster ensembles. In: IEEE international conference on systems, man and cybernetics, 2004, volume 2. IEEE, pp 1214–1219
19. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51(2):181–207
20. Lee J, Sun Y, Nabar R, Lou H-L (2008) Cluster-based transmit diversity scheme for mimo ofdm systems. In: IEEE 68th vehicular technology conference, 2008, VTC 2008-Fall. IEEE, pp 1–5
21. Li J, Yi K, Zhang Q (2010) Clustering with diversity. In: Automata, languages and programming. Springer, pp 188–200
22. MacQueen J, et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, p 14. California, USA
23. Manogaran G, Lopez D (2018) Health data analytics using scalable logistic regression with stochastic gradient descent. Int J Adv Intell Paradig 10(1–2):118–132
24. Nagarajan VR, Kumar V (2018) An optimized sub group partition based healthcare data mining in big data. Int J Innov Res Sci Technol 4(10):79–85
25. Pakhira MK (2009) A modified k-means algorithm to avoid empty clusters. Int J Recent Trends Eng 1(1):1
26. Reddy C, Aggarwal C (eds) (2015) Healthcare data analytics. Chapman and Hall/CRC, New York. ISBN: 9780429183447. https://doi.org/10.1201/b18588
27. Research and Markets (2016) Global big data in healthcare: focus on hardware, software type, deployment model, analytic service type, analytic service applications, and geography—estimates and forecast, 2015–2022

28. Sarada J, Lakshmi M (2017) An introduction to data analytics in healthcare industry. Int J Adv Sci Technol Eng Manag Sci 3(1):169–173
29. Sharpton T, Jospin G, Dongying W, Langille M, Pollard K, Eisen J (2012) Sifting through genomes with iterative-sequence clustering produces a large, phylogenetically diverse protein-family resource. BMC Bioinform 13(1):264
30. Shemetulskis NE, Dunbar JB Jr, Dunbar BW, Moreland DW, Humblet C (1995) Enhancing the diversity of a corporate database using chemical database clustering and analysis. J Comput Aided Mol Des 9(5):407–416
31. Tang EK, Suganthan PN, Yao X (2006) An analysis of diversity measures. Mach Learn 65(1):247–271

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.