CE888 Assignment 1: Knowledge Distillation from Random Rorests

David Barragn Alcntar, *Data Science and Desicion Making, University of Essex,*

***Abstract*—Data Science plays an important role in dealing with big amounts of information. Different method have been developed to obtain information from data sets. An important field in Machine Learning is data classification in order to make intelligent decisions based on class predictors. This report describes the most employed classifier methods making emphasis on random forests, which will be used to learn from three different data sets. Besides, a procedure known as knowledge distillation will be implemented upon the random forests to measure its impact on the classification accuracy. Finally, the model developed in this report will be compared to state-of-the-art methods.**

***Keywords*—Data Science, Machine Learning, Random Forest, Classifiers, Knowledge distillation, Classification methods, Decision tree.**

I. Introduction

Data science is a multi-disciplinary field of computer science combining statistics with the analysis of considerable amounts of data and machine learning techniques. The final outcome of these procedures is to generate knowledge and understanding from data sets [1].

As the technology as advanced in the last decades, the quantity of information carried out by institutions and organizations has grown exponentially, leading to consider conclusions that can be generated from the information in the data collected through time.

One of the main objectives of data science is to improve decision making capabilities in government, business and medicine, just to name a few, thanks to the extraction of information.

A common use of the techniques of data science that are applied in real world scenarios is in business marketing, such as online advertising, targeted marketing advertising and selling recommendations. Another recurrent application of data science is customer analysis, where financial industries assign a credit score to a client in order to predict whether a person will be able to pay a credit or not [2]

The present report includes a data science and machine learning technique used to classification and regression. The name of this procedure is ”Knowledge distillation from random forest” that basically uses a random forest method to extract and compress information into a single comprehensive and interpretable decision making treelike model (decision tree). This approach will be explained in more detail in the Methodology section of the present work.

The method implemented will be compared with other actual methods for similar classification/regression algorithms applied to three different sets of data obtained from [3], an online repository containing a wide variety of data sets suitable for machine learning techniques.

The three data sets are focused on classification approaches rather than regression procedures. The reason of this is because the majority of the data sets found in [3] are adequate for this specific task. The data sets will be fully explained in the Methodology section of this report.

After the implementation of the random forest method to the data sets, an extraction of information is expected from the sets of data. This new and previously hidden at first appearance information will be useful for making decisions and future predictions on unseen instances.

II. Background

Due to the present work deals with a classification problem, following is a description of some classification techniques that have been used in the past, along with a discussion in previous work on the topic.

Classification algorithms belong to the category of supervised learning. This division refers to the methods used to correlate a number of input parameter into an output and generally categorical value [4].

In order to learn, a classifier needs to somehow generalize the information extracted from a set of example data, where the output class is assigned to a specific set of input values. This example data is generally called ”training data” [5]. After the model has been trained, another, different and unseen set of data is applied to the algorithm to compare the predicted values of the classifier with the real values. This second set of data is called ”test data”. This is useful to measure the performance of the classifier, the more similarity of the model results with the real outputs, the more accurate the algorithm is and the more generalization ability it has.

According to [6] the most prevalent machine learning algorithms for classification are linear classifiers, support vector machines, decision trees, boosted trees, random forest, neural networks and nearest neighbors. In all methods listed above, the model learns from the data to classify new observations into a two or more classes.

A. Linear Classifiers

A linear classifier is one of the simplest ways of group different classes. The mechanics of a linear classifier is to make a decision based on a linear combination of its features (hence its name) [7].

A common problem of linear classifiers arises when the classes are not linearly separable, meaning that the simplicity of the method leads to inaccurate predictions. Thus, more

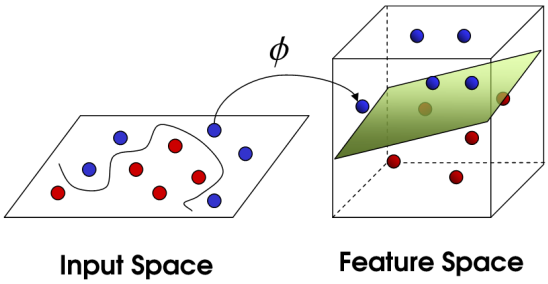

Fig. 1: Support vector machine

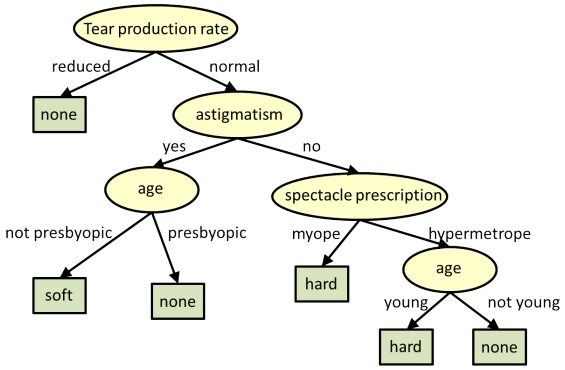

Fig. 2: Decision tree

sophisticated algorithms have been developed in order to deal with this and similar situations.

Some examples of commonly used linear classifiers are naive Bayes classifier, logistic regression and the linear perceptron [8].

B. Support Vector Machine

Support vector machine (SVM) is a classifier and regression procedure that usually separates instances belonging to one of two classes (binary classifier). SVMs are capable of performing linear and no linear classification with the correct use of the kernel, a mathematical tool that allows a SVM to map the data into higher dimensions where the instances will be separable with linear hyper-planes (Fig. 1) [9].

One characteristic of SVMs is that is a deterministic algorithm, meaning that with the same training data and parameters, the classifier will perform in the same manner. Another property of this method is that it only uses the most prominent data, ignoring and saving resources for no relevant information [10].

C. Decision Tree

A decision tree is a method generally used to classify instances using a branching approach to separate the different classes according to their key features. A decision tree classifier is a tool that uses a tree-like model consisting on nodes, branches and leaf nodes. Inside each node, a feature is measured and branched into two or more paths, until there is only one class in the leaf node or a stop criteria has been reached (Fig. 2) [11].

The important step in building a decision tree is the feature selection for each node. A good feature selection will lead to smaller trees, whereas a bad selection of these features can provide no information valuable to classify.

An advantage of a decision tree is that it can be easily understandable, and can be represented graphically as well. Each node represents explicitly a decision making process [12].

D. Boosted tree

Gradient boosting is a technique used to classification and regression in supervised learning. This method makes use of weak decision tree predictors, combining and converting them into a stronger decision tree through ensembling them [13].

The term ”boosted” refers to the method of which the weak predictors become stronger through several iterations, In each iteration, the last decision tree predictions are used to generate the next classifier [14].

E. Random Forest

Random forest is a similar approach to a boosted tree classifier. Both model share the dependency on weak decision trees to combine them and construct a stronger one. The difference with random forest is the method to generate the weaker decision trees and the procedure to merge them into a better tree classifier.

Random forest is also an ensemble learning method. It first produce a set of random generated decision trees with random selected features during the training period, hence the name. Then, for constricting the final decision tree classifier, the method uses the mode of the classes of all individual trees (Fig. 3) [15].

An advantage of this method over single decision trees is that a random forest prevents over-fitting on the training set, achieving more generalization and accuracy on the test set [16].

Another quality of random forest is that it allows to measure the feature importance of the overall random forest by examining the same feature across all single decision trees and how that selection led to a decrease in the impurity in the forest [17].

Apart from the advantages mentioned above, random forest has the drawback of the need of a big number of trees, causing slowness in the model and converting it an ineffective real time classifier. Generally, because of the random generation of trees in the training session, the algorithm is faster to train than to actually predict on the test set. Nevertheless, for real world applications, random forest are fast enough to handle the majority of classification tasks [17].

F. Neural Network

A neural network, also know as artificial neural network (ANN) to differentiate it from the human nerve system, is a

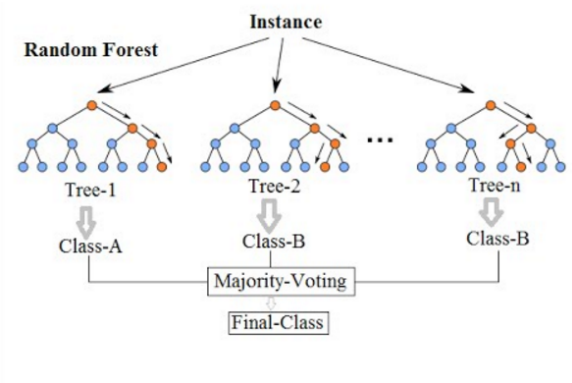

Fig. 3: Random forest

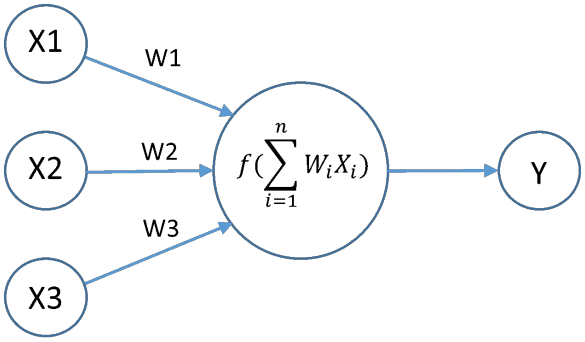

Fig. 4: Perceptron of a neural network

model based on the interconnection and flow of information in biological neurons.

ANNs learn from examples and adjust their parameters to generate the same or a near output as the training data. They are used for classification and regression as well. An ANN is composed of interconnected nodes called ”perceptrons” that perform mathematical operations with the outputs they receive and produce an output passed to subsequent perceptrons (Fig. 4) [18].

Generally, the more perceptrons a neural network has, the more power of generalization it achieves. Furthermore, there are several classification of neural networks depending of their interconnection. ANNs are widely used in image classification, speech recognition, medical diagnosis and game playing [19].

All methods described above are commonly used for classification purposes. For the intention of this report, the random forest algorithm will be selected to classify three different sets of data obtained from [3] and the results will be compared with previous work on the same data sets. The following section describes in depth the selected method as well as the information in the data sets.

III. Methodology

The goal of the present work is a supervised learning classification method of three different sets of data. The outcome will be a random forest classifier that will be able to extract or distil knowledge ”hidden” in the data sets.

Firstly, the three data sets will be described, these information was obtained from [3], a machine learning repository with more than 400 different data sets for machine learning, data science and data mining purposes. The first data set selected corresponds to the ”Abalone Data Set” to predict the age of Abalone shellfishes. The second data set is the ”Human Activity Recognition Using Smart-phones Data Set” which classifies six different behaviours. Finally, the third data set is called ”Adult Data Set” and is used to predict the income of a certain person.

A. Abalone Data Set

The Abalone data set is used to predict the age of abalone shellfishes depending on physical measurements. The age of an abalone is measured through cutting its shell and counting the number of rings with the help of a microscope. The data comes from a population study of abalone described in [20].

The data set contains information of 4,177 instances, each one with 8 different features that can be categorical, composed of natural numbers or real numbers. The following list shows the attributes of this data set.

- Sex (M / F / Infant)
- Length (mm)
- Diameter (mm)
- Height (mm)
- Whole weight (grams)
- Shucked weight (grams)
- Viscera weight (grams)
- Shell weight (grams)

- Output: Rings (+1.5 gives the age in years)

B. Human Activity Recognition Using Smart-phones Data Set

Human Activity Recognition data set is used to classify a person’s behaviour using an accelerometer on a smart-phone located on the waist. The data was generated with a group of 30 volunteers from 19 to 48 years old. Each person performed six actions (walking, walking upstairs, walking downstairs, sitting, standing and laying) [21]

The data set contains information of 10,299 instances, each one with 561 attributes. The following list shows the main features of this data set.

- Triaxial acceleration from accelerometer
- Triaxial angular velocity from gyroscope
- 561-feature vector with time and frequency domain variables.
- Identifier of the subject

- Output: Activity label

C. Adult Data Set

The Adult data set is used to predict whether an income of a person exceeds $50,000 per year based on census data. The data was extracted by Barry Becker from the 1994 census database [22].

The data set contains information of 48,842 instances, each one with 14 different features that can be categorical or composed of natural numbers. The following list shows the attributes of this data set.

- Age (continuous)
- Work-class (categorical)
- Fnlwgt: (continuous)
- Education (categorical)
- Education-num (continuous)
- Marital-status (categorical)
- Occupation (categorical)
- Relationship (categorical)
- Race (categorical)
- Sex (categorical)
- Capital-gain (continuous)
- Capital-loss (continuous)
- Hours-per-week (continuous)
- Native-country (categorical)

- Output: $>$\$50K/yr, $\leq$\$50K/yr,

These data sets were selected because of their instances size, the association task (classification) and the different types of data (continuous, categorical, time series).

D. Tools for the analysis of data

Regarding the tools for analyzing the data, the model that will be used to classify and predict the outcomes of the three different data sets will be a random forest.

For the development of the random forest model, the scikit-leark python toolkit will be employed [23]. This toolkit has a built-in function called "RandomForestClassifier" that performs the method described previously. The classifier is provided with various parameters that can be modified to produce random forest classifiers with different characteristics, for example, the number of estimators, the maximum depth, among other parameters.

IV. Experiments

In order to measure the accuracy of the random forest model, a series of experiments have to be implemented.

In the first place, a reasonable high number of trees is recommended, usually above 100. The experiment will take place increasing the number of estimators (aka. trees) from 100 to 1,000, increasing 100 trees between each iteration.

Then, the method for measuring the quality of the split on the trees will be evaluated among the "gini" method and the "entropy" method in order to decide if there is a significant difference of using a specific criteria.

For parameter optimization, a grid search algorithm will be used in order to find the parameters that best suits a determined data set.

A distillation procedure will be implemented on the random forest model. In order to implement this, the probabilities for each class will be calculated with the calculation of a previous random forest. Then a new data set will be generated adding the probability information to it in the form of multi-class classification encoding. Decision tree classifiers will be selected to learn from the new data set (the distilled data). The final product of this distillation procedure is to obtain a single and interpretable decision tree compressing the knowledge generated by random forests.

Furthermore, in the case that the random forest model would take long time to conclude with a solution, a maximum depth parameter will be added to prevent significant time-consuming evaluations.

To assess the performance of the random forest, a ten k-fold cross-validation method will be performed, thus giving an estimate of the model performance. This overall cross-validation accuracy will be compared with a test set of unseen examples, the test set.

Finally, other algorithms applied to the same sets [20],[21],[22] will be compared with the random forest model to notice if there are advantages or disadvantages with the current classifier.

V. Discussion

To finalize the part of experimentation it is necessary to have a way of evaluating the performance of each one of the different random forest models with different parameter selection.

In the first place, the accuracy percentage against the a test data will be calculated and compared among all models. For a more quantitative approach, a confusion matrix will be calculated for each different model. The confusion matrix will help to determine the level of specificity and the level of sensibility of the model. Also, the power of the model can be calculated from this analysis.

Then, a graphical representation of feature importance will be generated for each model. Along with the confusion matrix, these evaluation methods will make easier to perceive the behaviour between different parameters of the model, and they will reveal the reaction on the random forest for individual parameter modification.

Finally, the performance of same size and shape random forest with and without the distillation procedure will be compared to decide if there is an increase in the performance with the inclusion of knowledge distillation.

VI. Conclusion

The reasoning presented here is a conclusion in advance to the experiments, it can be counted as the expected results from the experiments and discussion of results obtained.

It is expected that the knowledge distillation will produce better results for classifying the data sets selected. This results can be addressed to the process itself, because it uses actual

random forests to generate new information based on probabilities that help to produce better prediction for even simpler decision tree methods.

Also, a set of parameters will be optimized for each specific set of data through the grid search process, leading to a more optimized knowledge extraction/distillation classifier based on a decision tree model.

REFERENCES

[1] V. Dhar, "Data Science and Prediction", *Communications of the ACM*, December 2013, Vol. 56, No. 12, Pages 64-73

[2] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making", *Big Data*, February 2013, Vol. 1 No. 1

[3] D. Dua and E. Karra, "UCI Machine Learning Repository", *University of California, Irvine, School of Information and Computer Sciences*, 2017, url: http://archive.ics.uci.edu/ml

[4] S.J. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach", *Prentice Hall*, Third edition, 2009

[5] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "Foundations of Machine Learning", *MIT Press*, Second edition, 2018

[6] M. Sidana, "Types of classification algorithms in Machine Learning", *Sifium Technologies*, February 2017

[7] G. Yuan, C. Ho, and C. Lin, "Recent Advances of Large-Scale Linear Classification", *Proc. IEEE*, 2012, Vol. 100, No. 9

[8] T. Mitchell, "Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression", *Draft version*, 2005

[9] C. Cortes, V.N. Vapnik, "Support-vector networks", *Machine Learning*, 1995, Vol. 20, No. 3, Pages 273-297

[10] Scikit-learn, "Support Vector Machines", *Journal of Machine Learning Research*, 2011, Vol. 12, Pages 2825-2830

[11] L. Rokach, and M.O. Lior, "Data mining with decision trees: theory and applications", *World Scientific Pub Co Inc.*, 2008

[12] J.R. Quinlan, "Induction of decision trees", *Machine Learning*, 1986, Vol. 1, Pages 81-106

[13] J.H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", *IMS 1999 Reitz Lecture*, February 1999

[14] (Electronic Version): StatSoft, Inc. (2013). "Electronic Statistics Textbook", Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/

[15] T.K. Ho, "Random Decision Forests", *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, August 1995, Pages 278-282

[16] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning", *Springer*, Second edition 2008

[17] D. Nikla, "The Random Forest Algorithm", *Towards Data Science*, February 2018

[18] M. van Gerven, S. Bohte, "Artificial Neural Networks as Models of Neural Information Processing", *Frontiers in Computational Neuroscience*, January 2018

[19] D. Silver, et al. "Mastering the game of Go with deep neural networks and tree search", *Macmillan Publishers Limited*, 2016

[20] W.J. Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn and W.B. Ford, "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait", *Sea Fisheries Division*, 1994, Technical report No. 48

[21] J.L. Reyes, D. Anguita, A. Ghio, L. Oneto and X. Parra, "Smartlab - Non-Linear Complex Systems Laboratory", *Universit degli Studi di Genova*, Genoa, Italy

[22] R. Kohavi and B. Becker , "Data Mining and Visualization", *Silicon Graphics*

[23] Pedregosa et al., "Scikit-learn, Machine Learning in Python", *JMLR*, 2011, Vol. 12, Pages 2825-2830