Probabilistic Distillation Using Data Binning on Ensemble Learning

Siddharth Shyamsunder

Student, University Of Essex

Abstract: To improve the efficiency and performance of any Machine Learning system, it is always important to try different Machine Learning algorithms, have a look at their performances, their ability to adapt or rather bridge the latency gap that is expected during delivering results in a real world scenario. However, this is merely an ideal situation, and training multiple models on a single dataset can prove to be a cumbersome process and expensive. A solution that we have proposed as an extension to other researcher's work like Caruana's[1] is to first integrate the knowledge from multiple machine learning techniques and average them[2] and then compress this extended model to create a more generalized unimodal that can not only execute the knowledge of multiple techniques in a single iteration but also improve the efficiency of delivering results. For better accuracy, we have considered testing model with 3 different datasets, an ecoli, protein synthesis dataset, a letter recognition dataset and an automobile import dataset, and in all three datasets, we have come to identify that our model works much more efficiently than merely using a single learning technique. Using our model, we have also considered in a different type of integration, by adding probabilistic values of the outcomes to the original dataset, that can aid in the faster retrieval of information to give us a higher performance. Through this probabilistic technique, we are not only able to increase performance, but also provide a better opportunity to have the system judge classes more efficiently.

Index Terms: Ensemble Learning, Random Forest Machine Learning technique, probabilistic distillation, Data Binning, Evaluation Parameters-Precision, recall and F-Score, Decision Tree Model, Supervised Learning Models.

1. INTRODUCTION

When one considers machine learning, one usually is trying to find a way to have the machine find relative patterns in hordes of data, so that the machine will be able to properly judge the class of unseen information. So basically, the machine is made to learn how the output is considered based on a particular pattern of input, and then the technique us tested upon unseen or a related but new dataset, and the machine is supposed to use its prior pattern recognition abilities to identify the classes for the new dataset. In order to this, its not simply enough the pattern recognition skills of the system be effective to judge a given class, but the system should also perform quickly, and maintain a sence of low latency or time inexpensiveness.

The initial motivation of machine learning was to have the computer system replicate the features of the human brain, in terms of pattern recognition and decision making. The way a human brain works is by identifying input though means of the 5 senses, and putting together patterns that is generates from the influx of inputs that it receives and be able to jufge a particular mode of action , that is deemed necessary. In a likely manner, a machine is provided with information, through means of datasets, which could be textual, unstructured, auditory or visual information, and then use a machine learning technique like a supervised(linear classification, decision tree classification, probabilistic classification), unsupervised(Clustering, association), reinforcement learning, and beable to judge and make effective decisions.

As we will be looking more into the Supervised Learning techniques, the system has a means to identify patterns in the dataset, through making decisions at each step, and collaboratively come up with a final decision(decision tree). Through means of dividing classes along a particular boundary, or a hyperplane(Support Vector Machine), by finding nearest classes, a particular data would relate to in space(k nearest neighbour) and so on. However, with rise of technological advancements, individual machine learning techniques are not merely enough to give efficient results. Hence there is a need to perform ensemble learning techniques, or rather integrate, more than machine learning techniques with one another and use the new model to make better optimal judgements[1]. However, although integrating machine learning techniques together can bring about having a system perform better in terms of judgement. The entire process is still cumbersome

and expensive to carry out, moreover such a system is slow to produce results. Hence, we provide a model that acquires the knowledge obtained from ensemble training, and work as a much simpler model, so that the system can use the knowledge learnt from integrating machine learning techniques to form better judgements and use the simplicity of the model to improve the speed and performance of the learning. We handle this by first, having the machine learn to generalize ,using an ensemble learning technique and identify a pattern to judge unseen data and then we make use of this information, and create a probability list, using binning techniques to capture a new set of probability fields into our original dataset. Now having this knowledge acquired from the ensemble training, we can use this new dataset, to learn to judge classes more quickly and efficiently.

The major roadblock that came, associating with general ensemble training, was the ability to bridge the latency gap, to handle real world situations, this wehave learnt to achieve by creating a more simplistic model, and the second roadblock would be the handling of redundant information, that comes in package with ensemble training. By making use of a probabilistic distillation approach, it makes it a lot easier, to create a more generalized model that can be used to provide better results.

The motivation towards this model, was brought about after having a decent study on the human sub conscious behavior. When a child is small, or when a person is initially taking say, a driving class, the brain's role initially is to acquire information, having a child take more time in trying to understand how something works, or having the driving learner understand the rules of the machine. But at the later stage, the information stored in the human brain becomes secondary, having the child make collective decisions or a driving learner feel more comfortable on the road with relative ease. Complementing the initial phase of learning to that of ensemble learning in our case, we later create a generalized model that can make learning of unseen data almost like a secondary process.

The basic steps involved in our methodology will include:

1. First we will take 3 unique datasets and split the datasets into their training and test sets individually.
2. Then we will perform an ensemble learning technique like random forest technique or any other integrated learning technique.
3. Using the skills learnt on this technique, we will use a probabilistic approach to our earlier predicted values on the entire dataset, and generalize the model through means of data binning, for example, if the classes are for a dog and a cat, we would be creating probabilistic values as in dog: 0.1, cat 0.9, dog 0.2, cat 0.8 etc.
4. Then we will add the newly generated probabilities as fields to our entire dataset and now we will use this generalized dataset on any simpler learning technique like Decision Tree or Support Vector Machine and generate our simplistic model.

2. BACKGROUND

There has been an extensive level of research that has been carried out in trying to create an efficient model that can be used to make effective judgements on classes as well as perform Machine Learning is a much more faster and simplistic manner On the contrary to earlier used individualistic machine learning techniques, ensemble training techniques have provided a better platform integrating the results of many machine Learning techniques and averaging them[2] to provide more efficient judgement to classes of unseen data. Caruana[1] has an approach wherein a much more efficient model is considered as an addition to the traditional machine Learning techniques. In his approach, a method like an ensemble Machine Learning method is applied on a dataset, and a probabilistic output is generated which is then concatenated to the entire dataset in correspondence to each training output, which would act as a basis dataset to a more generalized model.

In further addition to this, the random forest method that was adopted by Tao Shi[3] proved as a way to integrate multiple decision tree, which is an individualistic Machine Learning approach, and average the scores[2] to remove the dissimilarity factors that was involved in the different decision trees, so as to create based on a general mean information, the right judgement for classes of unseen data. . Also a random forest machine learning ensemble method has the unique nature to also be able to judge between categorical data and numerical data, hence as compared to lesser traditional approaches like the SVM, or the decision tree, it provides a more holistic approach to making better judgements, without having to worry of further costs incurred in encoding the label information using priority label encoders. Also the use of data binning, has been very crucial in the identification of a more generalized pattern to achieve a simplistic model using the knowledge of an integrated Machine Learning model like the random forest method.

3. METHODOLOGY

When we consider discussing the methodology, that we intend to incorporate in our model, it is necessary to know a few basic terms that will be used ti understand our method more clearly.

Random Forest Method: Random Forest[7] method is

Nothing, but an ensemble training methodology, wherein, his algorithm represents itself, like a forest of decision trees. Making use of random forest methodology, one, can successfully prevent the concept of overfitting or rather the issue wherein, the model can only make a pattern with the training dataset and not the test dataset. Making use of the random forest method, a machine learning consultant can engineer would take the mean or average of all the trees[2] in the forest to properly make the right judgement for the output class values.

A model of a sample Random Forest method can be shown in the below figure.

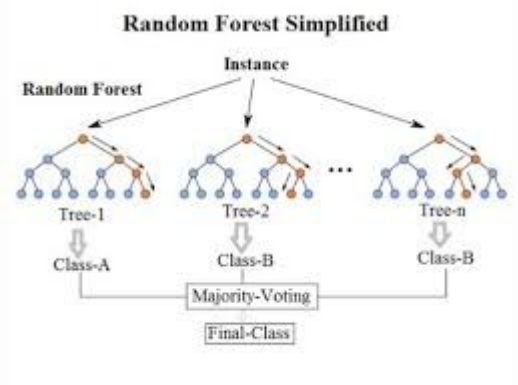

Figure 1: Sample Random Forest Methodology

Binning[8] is a technique used in data preprocessing. Basically binning plays an active role, in identifying values of a similar range, and placing them in the same bucket or bin. In our model, we will make use of the concept of binning to scale the probabilistic values that we predict on the entire dataset, to give a more optimal result.

An example of binning is shown in the below figure.

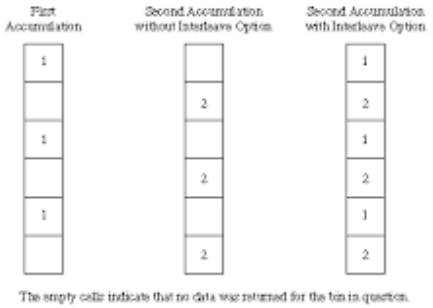

Figure 2: Data Binning

The aim of our model is to incorporate the concept of ensemble training on 3 random datasets. Once the dataset is trained under the random forest method, we will predict the probability rate for each training example for each of our classes. Once that has been achieved, we will make use of binning techniques to scale the predicted probabilities, and incorporate these probabilities on our dataset. In doing so, we generalize the model, and we can use other Machine Learning techniques to work on the new dataset to provide more optimal results. To further enhance the effectiveness of our algorithm, we will perform the random forest list using not few but 100 or more than 100 decision trees in our random forest. Once we generate the predicted values, we can use the same classifier on the entire dataset, to create a probabilistic list of all classes per training record. This can be formulated as below:
$P(i) = zi/(zi+zj)$

Where zj is the total number of occurrences of the class zj, and zi would be the total count of the classes in the dataset.
After the probabilistic measure has been considered, we will be able to use binning techniques to scale these probabilistic measures in order to make it easy to determine the optimal class. Then we will add this scaled probability values to our earlier dataset, and carry out anindividualistic Machine Learning Technique like decision tree or Support Vector Machine on the entire dataset.

The dataset that we intend to use for testing this model are:

3.1. E-Coli Data:

The E- Coli dataset [4] that we are considering as our dataset, gives us an idea on the on the localization of proteins on one of the most widely used microorganism for DNA Replication and genome projects., the Escherichia Coli organism, due to its DNA structure being easy to understand. One of the reasons we chose this structure, is because of its simplicity in organization of the dataset, yet wide number of uses. Each of the dataset attributes here take account of the protein analysis and signal recognition at different localization areas in an E-Coli cell body.
Also we selected this dataset, as each sequence number hs a unique ID, with its own unique features, thereby making redundancy appear less in this dataset.

3.2. Letter Recognition Data

The Letter recognition dataset[5] has its main goal being to recognize the structure, of a number of black and white symbols, and how the black symbols, correspond to appearing as the shape of one of the26 alphabets of the English language. The character

images were based on different handwriting patterns, including legibly or illegibly scripted human handwriting, different fornta on the computer, part letters, where the system;s role is to identify, the letter just by viewing a part of the letter. Akk of these examples corresponded to 20000 different training examples, out of which we selected near around 2200 training and test examples.
Each training example was converted into 16 primitive number attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. We typically train on the training items and then use the resulting mode to predict the letter category for the test dataset.

3.3. Automobile Dataset

This Automobile Import dataset[6] consists of three types of features:
(a) the specifications of a vehicle in terms of various characteristics.
(b)its assigned risk rating on the term of insurance.
(c) The losses incurred as and when comparing the current car in the training model, with respect to other cars after normalization.
The second feature type gives us a degree of how risky the vehicle is in comparison to the price one pays for the vehicle.
Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.
The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.
Having these 3 unique datasets will give us more perspective in types of data that can be fitted in our model.

[9]When we consider making use of the E-Coli Dataset, the values in this dataset are more categorical, patterns work on a unique ID for an ecoli organism hence there is opportunity for finding similarity in the probabilistic values gathered.
As far as the letter recognition dataset is concerned, there are only minor possibility for outliers to change the value of a letter, for example the value of the letter P can be changed to R, hence, these outliers using certain machine learning techniques could result in a probabilistic distribution of classes. When we consider the case of the automobile dataset, similar classes tend to lie with similar attributes. hence probabilistic distribution will tend to move more towards the original class, giving more accurate readings using Binning techniques.

4. EXPERIMENTS AND EVALUATION:

To improve the performance and to validate our model, A number of experiments were carried out. As mentioned in the previous section, 3 datasets were used, the E-Coli Dataset, the Letter Recognition Dataset and the Automobile Import Dataset. First the letter recognition dataset was worked upon. This dataset had 1 class field called ‘lettr’, which consisted of 26different type of class labels for each of the different types of letters, in the English alphabet. To validate the moel, 3 different types of models were compared. For the first model, the letter recognition dataset was read, andall the labels were encoded as an individualistic machine learning algorithm like a decision tree cannot work on categorical data.

For the first step, we split the entire data intro 80% training data, and 20%test data. Once this was done, we trained the training data against the training outputs using a Random Forest Classifier having an integrated number of 100 trees. Then we tested our fitted model with the test data, and got the predicted values. Comparing our predicted values with the actual values in the dataset, we were able to generate a 97.25% accuracy.
Also to validate the difference between using an ensemble technique instead of the individualistic training, we trained the same dataset, with the same splits on a Decision Tree Classifier, and we were able to generate only a 75%accuracy proving that ensemble training techniques give a higher accuracy on terms of judging the classes.
Once we trained the Random Forest Classifier on the dataset, we again predicted, this time in terms of probability using the same fitted Random Forest Classifier on the entire dataset instead of the 80% training dataset. In doing so, we were able to get a number_of_recordsX26 probability distillation matrix, where upon each probability lay in the range of 0-1. Hence, we then we applied Pandas Data binning function, keeping the bins of scale of 0.1, and we were able to scale the probability distillation matrix.
Once this matrix was generated, we concatenated this matrix through columns to our original dataset, now creating our generalized model.
Once the Generalized model, we performed a decision tree classifier on the new dataset to obtsin an accuracy, with the predicted value as 91%.

From this, we can make some points clear.

- Using an Ensemble Machine Learning technique gives a higher performance than using a simple Individual Machine Learning technique, like Decision Tree, which further gavean accuracy of 79% after distillation as opposed to 91% accuracy, when trained initially with the Random Forest technique.
- Although making use of the Decision Tree after distillation proved to reduce the accuracy of the code from 97% to 91%, it not only bridges the gap between the latency, but also performs much better than s dimple Decision Tree Classifier.
- Aside from Decision Tree after distillation, we also compared with Support Vector Machine Technique after Random Forest Didtillation to get an accuracy of 98% giving highest accuracy for optimal judgement.

Based on our individual readings, for the remaining 2 datasets, we were able to obtain the accuracy values as such:

Experiment/Dataset	E-Coli	Auto Import/Part Dataset
RF On Train test	89.70 %	87%
DT after Distillation	75%	35.40%
DT Without Distillation	79.40 %	19.30%
SVM after RF Distillation	69.10 %	22.50%

Table 1: Accuracy Parameters for E Coli and Aauto Iimports dataset

As can be seen, in terms of the E-Coli Database, although Decision Tree without distillation turned out toprovide a higher accuracy thana model after distillation, however, keeping into consideration, Latency gap reduction, this still provides the more optimal result.

Once we tested the training we evaluated the 3 datasets using sklearn's Classification Evaluation Report and Confusion Matrix.
For the letter Recognition dataset, we were able to achieve a weighted average precision, recall and F1 Score of 97% during the initial Random Forest Evaluation, and after implementing a Decision tree after distillation of Random Forest, were were able to achieve a weighted average precision of 97% and weighted average recall of 91% and a weighted average F1 score of 93%, this clearly proves that although compare accuracy was reduced from 97% to 91%, the precision count was still high at97% claiming that more of the right classes were returned, optimally passing our model. Similarly, with a 98% Precision, recall and F1 Score, with an SVM Classifier after distillation, our system was optimally able to return 98% of the results.
However, in the event of simply using a Decision Tree Classifier, the precision, recall and F2-Score were around 80% showing ensemble training makes our system model better with judging classes and distillation helpd improve latency gap.

Similarly, we were able to generate the evaluation for the other 2 datasets.

Ecoli Dataset			
Experiment/Evaluation Metric	P	R	F1
RF On Train test	87.00 %	90%	88%
DT after Distillation	64%	75.00 %	68%
DT Without Distillation	73.00 %	79.00 %	75%
SVM after RF Distillation	67.00 %	69.00 %	67%

Table 2: Evaluation Metrics for E Coli dataset

Auto Imports			
Experiment/Evaluation Metric	P	R	F1
RF On Train test	83.00 %	87%	84%
DT after Distillation	19%	35.00 %	22%
DT Without Distillation	8.00%	19.00 %	10%
SVM after RF Distillation	5.00%	23.00 %	9%

Table 3: Evaluation Metrics for Auto Imports dataset

5. CONCLUSION

As far as machine Learning is concerned, the field is always ever growing. The methodology, that we applied, still has a variance on the type of dataset. This is one Scope of area, where our model can be further progressed upon. Furthermore, the aim of any Mmachine Learning experiment isto achieve near 100% accuracy, hence, there is always scope ofnew models to come up to optimize performance, as well as reduce latency gap.

REFERENCES

[1] C. Buciluˇa, R. Caruana, and A. Niculescu-Mizil. “Model compression,” In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, New York, NY, USA ’06, Pages 535-541, ACM[Accessed: April 10th 2019].

[2] T. G. Dietterich. “Ensemble methods in machine learning.” In *Multiple classifier systems*, Springer, 2000, pages 1–15[Accessed: April 12th 2019].

[3] Tao Shi ,Steve Horvath. “Unsupervised Learning With Random Forest Predictors.” In Journal of Computational and Graphical Statistics Pages 118-138 | Published online: 01 Jan 2012[Accessed: April 12th 2019].

[4]For a reference of thre E-Coli Dataset Please go to: https://archive.ics.uci.edu/ml/datasets/ecoli [Accessed: February 17th 2019].

[5]Dua, D. and Karra Taniskidou, E. (2017). UCI. “Machine Learning Repository” [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science.Irvine, CA: [Accessed February 20th 2019].

[6]Geraldine E. Rosario and Elke A. Rundensteiner and David C. Brown and Matthew O. Ward.. “Mapping Nominal Values to Numbers for Effective Visualization.” INFOVIS. 2003[Accessed February 20th 2019].

[7]Random Forest Wiki.[Online]. Available: https://en.wikipedia.org/wiki/Random_forest[Accessed:February 17th 2019].

[8] Data Binning Wiki.[Online]. Available: https://en.wikipedia.org/wiki/Data_binning[Accessed: February 19th 2019].