Probabilistic Distillation On Datasets Using Binning Techniques

Siddharth Shyamsunder

Student, University Of Essex, Colchester

Abstract: One of the most compelling ideas that have come up in the area of Machine Learning is the idea of ensemble learning or rather the concept of integrating multiple training methods on the same data to get a more efficient output. However there are some pitfalls in this method like as in here is a lot of latency gap in the training of multiple methods, also there is a potentiality for a presence of redundancy in the data in the training examples which must be taken into account. This has been accounted by ensuring the creation of a hidden layer of datasets that can probabilistically relate the original dataset and the higher refined dataset. In our model we focus on developing this middle layered dataset to range probability into buckets called bins, these bin values decide the class of the new database which can then be tested for accuracy purposes on our old original database using other machine learning techniques. Furthermore to test the accuracy on a performance measure, we are considering making use of 3 random datasets, on E-Coli protein Synthesis, Letter Recognition, and Automobile brand Detection through Insurance Claims.

Index Terms: Random Forest Machine Learning, Data Binning, Ensemble Learning, Histogram Binning, Confusion Matrix.

1. INTRODUCTION

Machine learning primarily focuses on having a system train some data on a dataset, identify certain patterns in the feature to label relationship and finally use the pattern recognition skill that it learnt in predicting the values of the labels for the new dataset. One of the most quantifying facets that must been taken into consideration while carrying out a particular machine learning study is the proper means of evaluating the performance of the techniques used and its accuracy and confidence in identifying the right class label for a dataset with unpredicted class labels.

One can consider Machine learning in a real case scenario as to how the human brain handles the Learning curve. Much of how humans learn is quantifiable with his/her surroundings, upbringing and education. As an analogy between the real world scenario and a computer's Machine Learning, a computer learns through means of different algorithmic techniques, some of which may include a boundary specification in a pattern or finding the salient attributes that gives more definition to the output label or finding the class through the nearest neighbors of the point in an arbitrary space.

But an underlying condition in any of the above methods is in the fact that none of these methods are efficient enough to properly deduce the exact labels of a test dataset. This can be handled by training multiple models on the same dataset thought by Caruana[1] and obtain an optimal output which can then deduce the optimal class that is required for the test set. But training multiple modules can be a cumbersome process as machine learning can be a time expensive process that can be an issue to real time applications that focuses on latency retention. Hence a pragmatic solution that has been considered in creating a new dataset modelled on the cumbersome output table of the integrated machine learning test dataset. This new model can This new model can be used to extract a well-defined structure from the old model, which can then generalize the model[2].

One identifiable issue which was taken into account in such a regularized model was in the identification of a proper way to relate the quantifiable labels in our newly generalized models that could help in developing a generalized class label for the test set.

Another roadblock was in the handling of the redundant features and patterns that were observed in most machine learning datasets. This method of structuring the model in a quantifiable relationship model can be achievable through making use of probabilistic values. However there is a generalistic noise that can be observed through this method, wherein 2 models can be incorrect in classifying the labels but there could be a probabilistic degree in identifying the incorrectness in the data, example a cat may not be exactly a Siamese cat but the probability of the image not being Siamese is more than the probability of the image being say a monkey.

So using this new model of distilling the cumbersome dataset into a new and more generalized model, we create a more powerful approach to generalizing the model. We further create a single more unified model of all the probabmodel by making use of binning techniques like that of a Histogram by categorizing the probabilities into a certain set of ranges, and identifying which class a particular label belongs to. Then using other standard machine Learning techniques on the new dataset, we can identify the accuracy of the model on the earlier training dataset. Using this method, not only will the cumbersome task of integrating Machine Learning techniques be optimized in a more efficient manner, but also a clear cut range can be identified in decision of the classes for our new dataset.

The basic steps involved in our methodology will include:

1. Training a Random Forest Technique on 3 unique
2. Create a multiclass dataset with the random forest technique and setting probabilities on the output values of each of the forests.
3. Create probability bins on each of the probability value ranges to determine the original number of classes from the multi class model.
4. Once the new dataset has been ceated and binning has been adopted, make use of another or other Learning techniques to rest the accuracy of the new dataset on the old dataset.

2. BASIC CONCEPTS

Random Forest Method: Random Forest[3] method is basically an integrated learning methodology wherein this algorithm basic all represents a forest of decision trees, The end goal of the Random forest method is to remove the concept of overfitting, which is majorly attributed to a single decision tree. The Random Forest method is attributed to calculating the mode ormost commonly occurring attribute from each of the outputs from the forest, and give a better judgement to the to the output of the classifier.

Binning Techniques: Binning[4] is a data preprocessing technique to categorize small errors in observations into Bins. This technique will be used in our model to identify ranges in the multiclass probabilistic outputs to determine the classes for our new dataset.

Histogram Binning: In addition to understanding the concept of Binning, the method of binning we will incorporate in our model will be the Histogram Binning Model, A histogram is a method of accurately classifying data though means of a probabilistic distribution bucket called Binning. In a Histogram model, each bin carries with it a certain range and the values belonging in that range are added to the specified bins. Depending on the range, the bin size can be either large, small or medium, in the best case scenario, it is advisable to use medium bin ranges as too large and too small bins can be inadequate is determining accuracy.

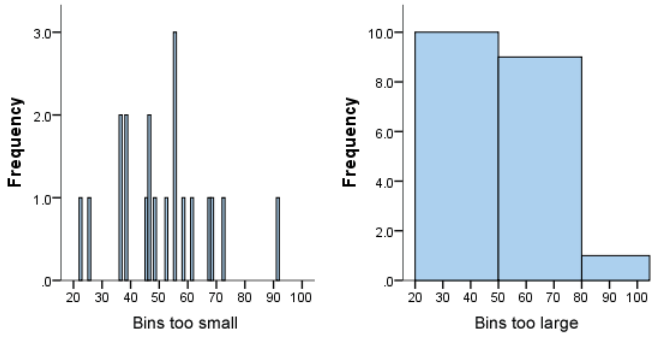

Figure 1: Data Binning example

datasets.

BACKGROUND

The amount of work that has been put into designing an efficient classifier for machine Learning has been tremendous. The introduction to ensemble training methods to integrate learning methods has led to more efficient predictability of outputs in comparison as compared to the earlier used single techniques like SVM, Decision tree techniques and so on.

In Caruana's[1] approach, a much more efficient method was considered to standard machine Learning techniques, where in ensemble Learning methods would have to be incorporated on a training dataset, and this would create a new dataset, unto which a probabilistic measure would be considered on each of the individual training outputs, and then these probabilistic measures would curtail to a more generalized standard approach for identifying the output classes than making use of a cumbersome integrated machine learning matrix of class values.

Furthermore the adoption of the random forest learning method by Tao Shi[5] proves as a way to measure labeled or unlabeled data and act as a dissimilarity metrics between an integrated learning predicted output from multiple trees. Furthermore inclusion of a random forest method proves to be an additional asset in it can be used to identify not only numerical but also categorical data and hence this proves to be an efficient Learning method to correctly predict using integrated learning techniques the predictability of the range of the output class.

3. METHODOLOGY

We aim to introduce the Random Forest Method on 3 random training sets. What we wish to achieve from this random forest is a multi class label predictive list typically a class label list from each of the trees in our forests, in our case about more than 100 trees. Using this list, of predicted class label values we will be able to generate a probabilistic distribution chart on our estimated outputs. We can do so with the below formulation.

$$P(i) = zi/(zi+zj)$$

Where z(i) is the total count in occurrences of a class I and zj is the total number of occurrences of z(j).

Once this measure has been carried out, we will make use of binning techniques to derive a common class by considering putting the probabilistic values into bins of a particular range. After obtaining this new class, we will make use of decision

tree and SVM(Support Vector Machines) to classify the new dataset and identify the accuracy with the old dataset standard.

The dataset [6] that we will be using has the protein localization sites on the Escherichia Coli organism, the most widely studied organism due to its easy DNA Structure. Here the main reason to select this database, is it has a multiclass structure, where different values in the attributes stand for the signal recognition of protein analysis at different points in localization in the Cell Body of the cell.

For this type of data there are less chances of redundancy in the data which makes it ideal to be used in our experiment.

4.2. Letter Recognition Data

The objective of this dataset[7] is to recognize each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on different types of fonts and each letter within these different types of fonts was randomly distorted to produce a file of 20,000 unique training 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..

4.3. Automobile Dataset

This data set[8] consists of three types of entities:
(a) the specifications of a vehicle in terms of various characteristics.
(b)its assigned insurance risk rating.
(c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates.

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.

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

4.1. E-Coli Data:

Having these 3 unique datasets will give us more perspective in types of data that can be fitted in our model.

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.

5. EXPERIMENTS

We will initially perform the Random Forest machine Learning Techniques and identify the modes of the classes on all the different ranges, then we will carry out a probability mapping of the class output by applying some distillation techniques, which would allow for us tp create and normalize the data which will allow us to easily transfer the information into Bins giving us a more efficient reading on the class values of our new dataset.

Once we have obtained the new class values, as these class values are created on an inaccurate platform of assuming the probabilities to lie within a particular bin to belong to a particular class, it is necessary to have an evaluation testing to be carried out on the new dataset.

First and foremost we will employ the Decision Tree technique on the new database so asto get an accuracy of how the new database will compare with the original database, then we will be creating a confusion matrix determining the rate of precision and recall on the values generated in the new dataset in comparison with that of the original dataset.

Once this has been created, we will also have other machine learning techniques carried out to compare which has a better precision and recall on the original dataset.

6. DISCUSSION

Upon generating the new dataset, it is utmost crucial, on how we evaluate our new dataset to meet the accuracy of original dataset. In order to do so, we will first train our modeled

dataset with a Machine Learning technique like that of the decision tree. Alsthough decision tree has a major tendency of causing overfitting during evaluation, decision tree is also a fast and efficient way in identifying the salient attributes that would help to identify the correct class labels ona test dataset.

We have chosen to make use of 3 random datasets as evaluating on only one dataset would begeneric to that type of dataset.

Once we generate the predicted values, we will be able to generate a confusion matrix, which will compare the true positives, the false positives, the true negatives and the false negatives from out original and new datasets. This will allow us to have 3 confusion matrices for all of our individual datasets.

A typical Confusion matrix will look like the below Table

Truth Values	**Predicted Yes**	**Predicted No**
Actual Yes	TP	FN
Actual No	FP	TN

Table1: Confusion Matrix Example

Through making use of these confusion matrices, we will be able to generate the precision and recall measures to identify the performance of our Machine Learning System on the original dataset.

Similarly we will use 2-3 other machine Learning techniques to further work on creating more confusion matrices to give amore optimal performance on which Machine Learning technique satisfies our new model.

7. CONCLUSION

The area of machine Learning is forever expanding, and new techniques of ensemble Learning are helping improve the performance factor of machine Learning. Our method is one of many methods that can give an optimal class label prediction on ensemble learning techniques and there is always more scope to move forward in this area.

We will then use the decision tree to predict the class labels of our original 3 datasets, assuming these 3 original datasets are now test sets.

8. 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 '06, Pages 535-541, New York, NY, USA, 2006, ACM[Accessed: February 15th 2019].

[2] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014[Accessed: February 17th 2019].

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

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

[5] Tao Shi ,Steve Horvath: Unsupervised Learning With Random Forest Predictors. Journal of Computational and Graphical Statistics Pages 118-138 | Published online: 01 Jan 2012[Accessed: February 20th 2019].

[6] Paul Horton & Kenta Nakai. "A Probablistic Classification System for Predicting the Cellular Localization Sites of Proteins".Intelligent Systems in Molecular Biology, 109-115. St. Louis, USA 1996[Accessed: February 20th 2019].

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

[8] 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.

Project Plan- Gannt Chart

Task/Time	Feb13th 2019-Feb21st 2019	Feb 21st 2019-Feb-28th	March 1st 2019-March 7th 2019	March 8th 2019-March 15th 2019	March16th 2019-March 23rd 2019	March 24th 2019-March 31st 2019	April 1st 2019-April 7th 2019	April 8th 2019-April 15th	April 16th 2019-April 22nd 2019
Dataset Collection/Hypothesis Building									
Building Initial Plan									
Literature Review and designing model									
Performing initial Random Forest Measures and generating multi class model									
Creating Probability Distribution and carrying out Binning Techniques									
Evaluating New Dataset with Machine Learning Techniques against original datasets.									
Carrying out any new techniques discovered in the process									
Building Final Deliverables and structuring code									