Interim Progress Report

Data Mining and Machine Learning I, MSc Data Analytics (2019)

HIMANSHU GUPTA, STUDENT ID: 18203302

1 Motivation:

Crop-Production: We all know in Today's world how important agriculture is for everyone. If we consider the data from FAO (Food and Agriculture organization of United Nations), 60 percent of world population is depending on Agriculture [1]. Directly and indirectly agriculture is linked with everyone through food safety and health. Farming and agriculture have also bestowed employment in every country either in small or big way and contribute to the overall economy especially in developing countries like INDIA where 43% people have employment in Agriculture [2]. With the coming time world population has going to be increase so does the demand for overall crops. Many times, due to some unforeseen reason a good percentage of crop production got wasted, farmers have to bear loss which leads to suicide of farmers across various region in the world. World has limited resources and smart farming can do wonders for conserving resources or we should at least have an idea about crop production in the coming year which will help farmers and agriculture related sectors to plan things in advanced.

Online Retail Store: Maximum Shopping has done online, and it is very important part that how you are selling your products. Your sale is mainly depending on factors like pricing, offers and availability of an item.

Credit card Fraud-Detection: On the different aspect, in the world almost everyone does online payments and we have so many transaction records, there is huge chances that people do fraud transactions to earn money. Sometimes there are loopholes in the system and some people would get advantage of that and civil people and government suffer because of that.

To achieve the above analyzing different datasets using various ML methodologies in depth to create ML models and then evaluate those models.

2 Research Question:

By analyzing different agriculture related dataset and its attributes to predict about crop production prediction based on previous year data?

Predicting Customer behavior by analyzing online retail dataset?

Analyzing the credit card transaction dataset to predict the credit card fraud detection?

3 Initial Lit Review:

3.1 Related work of Machine Learning Prediction on Crop Production

[3] Anil Ramakrishna states that Estimation of crop production can be a very useful tool while planning a nation's agriculture. In India the major portion of population is depend on agriculture and contributing factor in the economy. I will be taking two other crops among 5 different from the ones used in this research.

[4] Chen Zhang et.al (2019) states that in agriculture insightful information can be get by using crop planting map. Firstly, it considered large areas and applied machine learning framework on that. They have used machine learning methods based on Artificial Neural Network (like artificial neurons in human brain in connecting nodes neural network), convolutional neural network (subset of ANN) and recurrent neural network (subset of ANN). For data preprocessing they analyze the certain pattern in crop plantation based on previous year data and create training and test data.

3.2 Previous work Related to Online Retail, Market Basket Analysis

[5] Gabriel Kronberger (2011) states that the target in MBA (Market Basket Analysis) is to find subsets of items that are purchased together frequently. It will help in understanding customer behavior and to efficiently used the space in the store (in case of physical store).Apriori algorithm is used to find the items which buy together frequently where the data is very large and transactional.

[6] Run-Qing Liu (2018) states that it is the world of big data, customer data and data mining analysis is very critical to make CRM strategies. They show how to do customer classification for MBA using k-means clustering technique and to find associations.

[7] Xuan HUANG and Zhijun Song (2014) analysis the E-commerce transaction and observing clustering using K-means Clustering technique.

3.3 Previous work Related to credit card fraud detection

[8] Fabrizio Carcillo ,Yann-Ael Le Borgne, Olivier Caelen ,Yacine Kessaci ,Frederic Oble (2019) and Gianluca Bontempi use combination of supervised and unsupervised learnings. In credit card fraud detection, mainly supervised machine learning approach use in this in which it makes the model learn by identifying patterns using transactions data.

[9] Nuno Carneiro, Gonçalo Figueira and Miguel Costa (2017) gives the detailed explanation about making a data mining-based system using machine learning techniques in e-tail applications. It gives a practical approach to design such system.

4. Data Sources:

1. https://data.gov.in/resources/district-wise-season-wise-crop-production-statistics-1997

Crop statistics about crop production in INDIA during 1997 to 2015 Data is available in csv format.

2. https://archive.ics.uci.edu/ml/datasets/Online%20Retail

This dataset is transactional dataset which have transactions during the period of 01/12/2010 to 09/12/2011 recorded through online retain store.

3. https://data.world/vlad/credit-card-fraud-detection

This dataset has credit card transaction. It has numerical input variables which are the result of a PCA transformation because of confidentiality. There are features V1, V2, ... V28 in which PCA has been applied to get principal component. It is available in csv format.

I will be following KDD methodology for Data mining and to find knowledge in the proposed data. Data can be containing things which are not meaningful so we need to filter out relevant data/meaningful information to extract knowledge which can be further use in taking the decision according to the relevant subject.

5. Identification of Machine Learning Methods:

5.1 For Crop Production:

I am going to use two machine learning modals here "**Simple regularized multivariate regression method**" and "**Regression Trees**".

[10] Using the simple regression method specifically Multivariate regression. Consider multivariate regression with Q response variables z1, ⋯, zq and P prediction variables v1, ⋯, vp". I will create training and test dataset for only 3 to 4 crops. Two attributes yield and area on different scales, so I need to scale them and normalize them.

We have number of machine learning techniques, in one of the such technique the learning from a set of independent instances resulted in like decision trees [11]. A "divide and conquer" methodology will be used to classify an unknown instance. Generally, we create two subsets and validate the new value to check in which subset it is lying. We can predict both categories as well as numerical value. In case of numerical, each leaf represents a numeric value which is average of all the training set values to which the leaf applies. Predicting numerical value using averaged numerical values of decision trees is called regression trees.

I am choosing these methods as there are two categorical attributes named as "district" and "season"

5.2 For Online Retail store:

I am going to use two machine learning modals here **K- Means Clustering** for doing Market basket analysis.

Whenever we have a dataset to do classification and association, we can make use of clustering approach. There are several instances and the similar instances will come in a same cluster. It is generally used whenever there is no class to be predicted. One of the main approach of Clustering is 'K-means' [et al.[11]] in which we have to choose a central reference point called "K" and then using the Euclidean distance metric all instances has been assign to their nearest cluster. Then we calculated mean of clusters and choose the new center points and repeat the process again till it will be stable. I am using K-Means Clustering because it is suitable for unsupervised learning especially when we need to observe the product items frequently bought together to do the Market Basket Analysis.

5.3 For Credit card fraud detection:

I am going to use two machine learning modals here one is **Naïve Bayes Method** and second is **Support Vector Machine**.

[12] Naive Bayes Method is based on the Bayesian statistical methods which is based on probability of events. It is mostly used in the binary classification where the modal would consider all the features and then predict the probability of an event. Suppose you need to predict whether the new instance would fall in which category of the classification. For example, if you are saying it is raining 4 out of 10 days by considering climate factors and the probability of event happening is 40 percent and not happening is 60 percent. These two are the only output responses. I am taking this modal for the credit card fraud detection because it also has classification problem attach with it, like either it would be fraud, or it would not be.

Support Vector Machine is a supervised machine learning method which is used in classification as well as regression problems. In SVM we create a line which separates the two classes[et al.12]. This line is considering as hyperplane, and support vectors can be created by joining the closest data points with the hyperplane based on maximum margin distance. I am considering this data model by doing classification by making two classes of Yes or no and will do the prediction. The main benefit of using SVM is that it will use few features to optimize the model.

Using these models because I am doing the supervised machine learning and have a classification problem. Naïve Bayes would be suitable in fraud detection because it would predict the probability and it does not matter whether it would be 51 percent or 100 percent, it will fall in the yes category (so we are expecting more accuracy in such scenarios).

6. Identification of Evaluation Methods:

Whatever the analysis I have done till on my data, I am expecting following evaluation methods would be suitable for my machine learning models.

6.1 For Simple multiple Regression (multivariate):

I will evaluate this machine learning model using MPE (Mean Percentage Error) as this is simple regression method. [Li et al.[10]] explains that this kind of a course of action is independent of the data pre-processing steps.

6.2 For Regression Trees:

Similar like above I am going to evaluate this model using (Mean Percentage Error). MPE is exactly like Mean Absolute Percentage Error but there is only one difference that MPE will consider both positive and negative values while MAPE will only take absolute value. Sometimes it is beneficial to use MPE over MAPE because it will tell the about your model estimates like it is overestimating (more positives) or underestimating (more negatives).

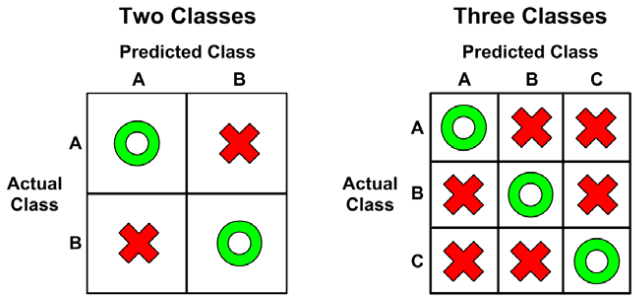
6.3 For K-Means Clustering:

I will evaluate this model using ROC (Receiver operating characteristic) which is used to identify true positives and ignores false positive [et al. [12]]. For this clustering model, we need to calculate type 1 error(E1) and type 2 error(E2) , if I two point are supposed to belong in one cluster but will not come then it's a Type 1 error. For every k value we will calculate E1 and E2 and join these two to get the ROC curve.[13]

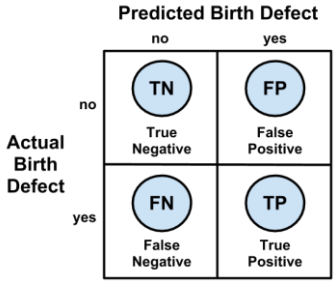
6.4 For Naïve Bayes Method:

I will be evaluating this machine learning method using Confusion Matrix. [et al. [12]] In Confusion Matrix, we predict the variable and match it with the original value in the data. If the predicted value is same as original, then it is "True" otherwise it "false". All the correct predictions show as diagonal. It will be suitable for binary classification model like this.

Confusion matrix will be look like below: We will be using Two classes Confusion matrix for this method.



[et al.[12]]



Accuracy = TP+TN / (TP+TN+FP+FN)

We need to consider adding error rate also.

6.5 For Support Vector Method:

I will be evaluating this model with confusion matrix because it is also binary classification problem like above. I will compare the expected results of accuracy of both the models. Uses of confusion matrix is also depends on what type of prediction you are doing. Here I am predicting whether it will be a fraud detection or not so confusion matrix supposed to work fine. But in case of critical medical disease prediction, it might be risky to evaluate your model based on confusion matrix and it might affect the person's life.

Bibliography:

[1] Agriculture Remains Central to the World Economy"(October 2014).[online].Available: http://www.expo2015.org/magazine/en/economy/agriculture-remains-central-to-the-world-economy.html [Accessed on: Nov 1,2019].

[2] International Labor Organization "Employment in agriculture"(% of total employment , September 2019).[online].Available: https://data.worldbank.org/indicator/sl.agr.empl.zs [Accessed on: Nov 1,2019].

[3] Ramakrishna Anil, A statistical approach to estimate seasonal crop production in India, Department of Computer Science, University of Southern California, Los Angeles, CA.[online].Available: https://pdfs.semanticscholar.org/5c74/91174c1d8b2029a8b273c6e04a95aa77cea4.pdf [Accessed on: Nov 2,2019]

[4] Chen Zhang, Liping Di , Li Lin and Liying Guo (2019), "Machine-learned prediction of annual crop planting in the U.S. Corn Belt based on historical crop planting maps" , Journal of Computers and Electronics in Agriculture, Elsevier, Volume 166 , October 2019 .

[5] Gabriel Kronberger, Michael Affenzeller (2011), "Market Basket Analysis of Retail Data: Supervised Learning Approach", Journal of ResearchGate , [online].Available:https://www.researchgate.net/publication/221431835_Market_Basket_Analysis_of_Retail_Data_Supervised_Learning_Approach

[6] Run-Qing Liu , Young-Chan Lee and Hong-Lei Mu (2018) , "Customer Classification and Market Basket Analysis Using K-Means Clustering and Association Rules: Evidence from Distribution Big Data of Korean Retailing Company" , Journal on ResearchGate.

[7] Xuan HUANG and Zhijun Song (2014) , "Clustering Analysis on E-commerce Transaction Based on K-means Clustering" Journal of Networks ,Vol 9,No 2 [online].Available: https://pdfs.semanticscholar.org/4e76/d4958135ba88dec1d1391eb63af3823a1e06.pdf

[8] Fabrizio Carcillo ,Yann-Ael Le Borgne, Olivier Caelen ,Yacine Kessaci ,Frederic Oble (2019) , "Combining unsupervised and supervised learning in credit card fraud detection", Journals at Science Direct, Elsevier.

[9] Nuno Carneiro, Gonçalo Figueira and Miguel Costa (2017) , "A data mining based system for credit-card fraud detection in e-tail" , Vol 95 , Pages 91-101 , Journal at Science Direct ,Elsevier.

[10] NIH Public Access **"**Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer**"** Mar.2010.[online].Available:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3905690/ [Accessed on: Nov 7,2019].

[11] I.H.Written , E.Frank, M. A. Hall and C.J.Pal,Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Cambridge, MA: Elseview,2017.

[12] Brett Lantz, Machine Learning with R,1st ed. PACKT publications,2013.

[13] Helena Aidos1 , Robert P.W. Duin2 and Ana Fred1 ," THE AREA UNDER THE ROC CURVE AS A CRITERION FOR CLUSTERING EVALUATION" in 2013, Published in ICPRAM , DOI:10.5220/0004265502760280.[online].Available: https://www.semanticscholar.org/paper/The-Area-under-the-ROC-Curve-as-a-Criterion-for-Aidos-Duin/b5c77886b725aa680396a17522d0779d02e8b97f [Accessed on 8 November 2019].