INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES
ISSN: 2322-0821(O) ISSN: 2394-9910(P)
VOLUME 5 ISSUE 2
Apr 2017 - Jun 2017
www.irjeas.org

MACHINE LEARNING TECHNIQUE FOR PRODUCT CLASSIFICATION IN E-COMMERCE DATA USING MICROSOFT AZURE CLOUD

Anurag Shrivastava[1], Jyoti Sondhi[2], Bharat Kumar[3]

[1]*Asso. Professor &Head, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*
[2]*Assistant Professor, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*
[3] *M.Tech Scholar, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*

***Abstract*— Rapid growth of online shopping and marketing increase the field of e-commerce, which got boosted rapidly with the application of machine learning. Machine Learning (ML) has empowered businesses by finding useful patterns from customers' search patterns and buying behaviours on web. Predictive analytics based on machine learning can enhance sale probability and find customer churn by analyzing customers past click-through behaviour, purchases, and history in real time. Applying popular the traditional ML libraries do not support well processing of large datasets, so new approaches and platforms are needed. Cloud computing paradigm turned out to be valuable alternatives to speed-up machine learning platforms. The paper, first discusses the importance of machine learning in predictive analytics. The need of cloud platforms to analyze ever growing E-commerce data is briefly surveyed then. Finally, the work proposes a framework to predict the Product Category in a large E-commerce dataset having 9 categories and 93 features of products (like electronics, fashion, etc.). The dataset is released by a famous e-commerce company for a competition. The classifier is build which is based on 'Multiclass Decision Forest' Machine Learning Algorithm and is deployed on Microsoft's Azure Machine Learning (Azure ML) platform. Azure ML is public cloud platform. The results obtained by proposed model are evaluated in terms of accuracy and the comparison is done with benchmarks provided by competition administrators. The results obtained are promising and the paper also directs the future research work in the field.**

***Keywords*— E-commerce, Classification, Big Data, Machine Learning, Microsoft Azure Cloud, Cloud Computing**

1. INTRODUCTION

1.1 **Predicting Product Category in E-Commerce Data:** For doing business in this communication era, web is the best medium. E-commerce has allowed businesses to offer choices to consumers. .To address this data and information explosion, e-commerce stores are applying machine learning to customization principles to their presentation in the on-line store [1]. Machine learning has empowered businesses to analyze all queries, whether searched or abandoned from all the users..Machine learning can be defined as an intelligent way to find secret patterns or information even in large datasets or databases. Machine learning often included in the category of predictive analytics as it helps to predict the future analysis.

1.2 Microsoft Azure Cloud Computing Environment for Machine learning [3]: Microsoft's Azure Machine Learning (Azure ML) [4] is a cloud service that enables execution of machine learning process. Microsoft Azure is a public cloud platform. The benefits of using public cloud computing platform (Azure ML) includes: handling big data and access from anywhere in the world. The process of Azure ML is shown in Figure – 1, which is same as that of basic process of ML. Azure ML provides a graphical tool for managing the ML process, a set of data pre-processing modules, a set of machine learning algorithms, and an API to launch a model to applications. ML Studio is a graphical tool that is used to control the process from beginning to end i.e. from data pre-processing to run experiments using a machine learning algorithm, and test the resulting model. ML Studio also helps its users deploy that model on real cloud.

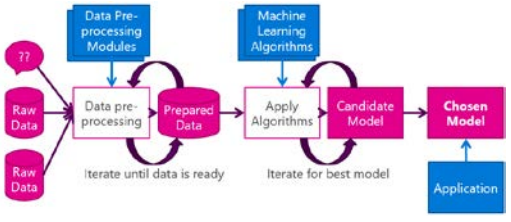

Figure 1: Machine Learning Process

International Research Journal of Engineering & Applied Sciences, IRJEAS
www.irjeas.org, ISSN(O): 2322-0821, ISSN(P): 2394-9910, Volume 5 Issue 2, Apr 2017- Jun 2017, Page 11-13

the need of cloud platforms to analyze E-commerce data is established in next section. The rest of the paper organized as: Section 2 briefly surveys the need of cloud platforms to analyze ever growing E-commerce data. The work proposed is presented in section 3.Experimental setup and result analysis is shown in section 4 and paper is concluded in section 5.

2. Literature Review (Need of Cloud Platforms to Analyze E-commerce Data)

Multiple choices of cloud computing models are available for different work load management, performance and computational requirements. The popular statistical tools and environments like Octave, R and Python are now embedded in the cloud as well [5].

A. Fast Analysis: The important findings of work [6] indicate the area of customer retention received most research attention.

B. Machine Learning on Cloud environment for Fast Prediction in Big Data: As the data is growing at faster rate and becoming "Big Data", the computation speed for prediction and other operations is inevitable. This paper [7] focused on the specific problem of classification of network intrusion traffic which is a Big Data.

3. Proposed Framework for Product Categorization:

The Proposed Framework which employs simple ML model with little change. The input dataset is suitably processed and converted into a suitable format. The machine learning algorithms are iteratively applied in the next step, and candidate model is determined. These ML algorithms typically apply some statistical analysis like regression or more complex approaches like decision forest to the data. Here in the proposed framework, the ensemble methods [12] are also applied to the model for better predictive accuracy. At last the model is deployed and tested on test data the snapshot of actual model build using specified steps, at Microsoft Azure ML platform, is shown in Figure – 2.

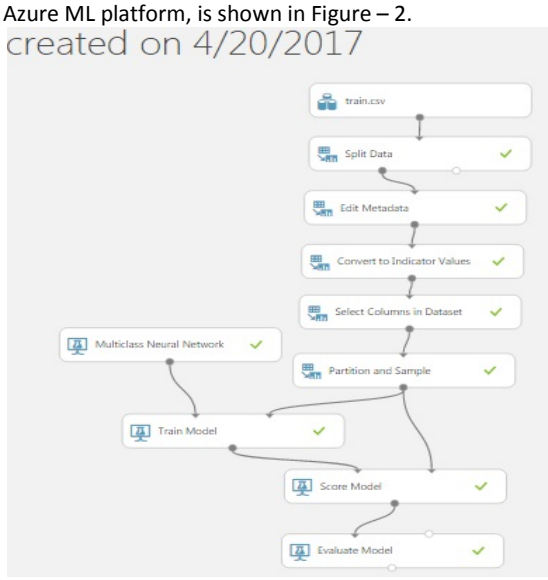

Figure 2: Model built using Azure ML

4. Simulation Environment Setup and Result Analysis:

Azure ML provides ML studio, a graphical tool that can be used to control the process from beginning to end. It includes: a set of data pre-processing modules; a set of machine learning algorithms; An Azure ML API to access model deployed on Azure. ML Studio allows a user to import datasets and data pre-processing methods.

4.1 Dataset Description: The dataset is provided by the Otto Group [8] which is a large e-commerce company. This dataset includes 61,878 instances Each product having 1 or more features out of 93 features provided for datasets.

4.2 Execution of Implemented Work (Experiment Steps): The experimental steps that are and represented in Figure–2, are explained below:

1. Create New Resource: Machine Learning Analytics solution.
2. Import/Upload the dataset.
3. Pre-process the dataset. Data pre-processing can also be done using modules written in R or Python.
4. Randomly split and partition the data into 70% training and 30% testing, using the 'Split Data' module.
5. Identify categorical attributes and cast them into categorical features using the 'Edit Metadata' module.
6. Convert to Indicator Values module to convert columns that contain categorical values which can more easily be used as features.
7. Select Columns in Dataset those are relevant
8. Apply Ensemble Method
9. Apply Machine Learning Algorithm to Train the model.
10. Now Score and Evaluate the Model. The 'Evaluate model' also visualize the results through confusion matrix .

4.3 Experimental Results: Analysis and Discussion: The experiment is evaluated on a simple multi-class classification accuracy parameter. Accuracy is defined as the number of correctly classified instances divided by the total number of instances:

$$\text{Accuracy} = \frac{\text{Number of correct Predictions}}{\text{Number of Instances}}$$

The results obtained using the benchmark code by setting the neural network [9] model with 100 trees got the accuracy of 0.9302 in [8], while the benchmark results given by competition administrators with 10 trees, is 0.50241. Here we have performed experiment at cloud platform with Multicast Neural network ML [10] method with 10 trees and an ensemble method. The evaluation results are inferred from confusion matrix shown in Figure – 3. A confusion matrix also known as error matrix and is used to describe the performance of a classifier (classification model). The overall accuracy obtained with our simulation is 0.6859, which is higher than the benchmark provided. The comparison of proposed model is done with benchmark provided by administrators and

International Research Journal of Engineering & Applied Sciences, IRJEAS
www.irjeas.org, ISSN(O): 2322-0821, ISSN(P): 2394-9910, Volume 5 Issue 2, Apr 2017- Jun 2017, Page 11-13

competition's winning results.

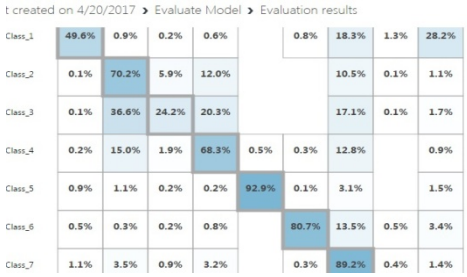

Figure 3: Confusion Matrix with Multicast Neural network

The comparison for accuracy obtained, is shown in Figure–4.

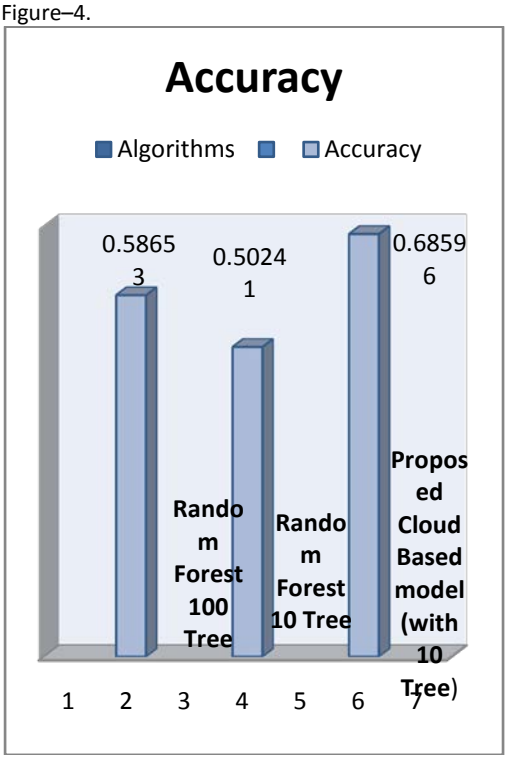

Figure 4: comparison for Accuracy

5. Conclusion & Future Work

The companies doing online business wants to utilize machine learning potential .In this paper, we proposed an Azure ML based model for E-commerce product categorization. The model used Multicast neural network algorithm to train the classifier. The evaluation results show that the proposed classifier performs better in terms of accuracy. We have performed experiment with 10 trees and an ensemble method. Our experiments showed that feature bagging lead to the better accuracy value (i.e. 0.55339) than benchmark .The proposed research can provide potential approach for training and testing of big data for addressing multi-class classification problems. So, further research will evaluate the framework with different ML algorithms, optimization parameters, ensemble methods and e-commerce databases. In future the model can be optimized to handle imbalanced datasets from various sources and domains. Also, the model can be modified for applying on Hadoop MapReduce [11] platform.

References

[1] Pine II, B.J. and Gilmore, J.H. 1999. The Experience Economy. Boston: Harvard Business School Press.

[2] j. Ben schafer, joseph a. Konstan, john riedl e-commerce recommendation applications", data mining and knowledge discovery, 5, 115–153, 2001, kluwer academic publishers, netherlands.

[3] David Chappell, "introducing azure machine learning: a guide for technical professionals", Sponsored by Microsoft Corporation, 2015 Chappell & Associates

[4] https://portal.azure.com

[5] Daniel Pop, "Machine Learning and Cloud Computing: Survey of Distributed and SaaS Solutions", https://www.researchgate.net/publication/257068169.

[6] E.W.T. Ngai ,, Li Xiu, D.C.K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications 36 (2009) 2592–2602, Elsevier

[7] Suthaharan, S., "Big data classification: Problems and challenges in network intrusion prediction with machine learning" Performance Evaluation Review, 41(4), 70-73, ACM 2014.

[8] https://http://www.kaggle.com/c/otto-group-product-classification-challenge

[9] Andy Liaw and Matthew Wiener, "Classification and Regression by randomForest", R News, ISSN 1609-3631, Vol. 2/3, December 2002.

[10]https://www.MulticlassDecisionForest.html

[11]Apache Hadoop Website http://hadoop.apache.org/

[12] J. a. H. Friedman, Trevor and Tibshirani, Robert, The elements of statistical learning vol.1: Springer series in statistics Springer, Berlin, 2001.