Machine learning in Spark Challenge in Categorical Data Modelling

Mehdi Tadayoni[1], Jingwei Min[1], Elham Babaee[1]

Abstract -- **Apache Spark as an open-source distributed processing has gained growing attention in the past few years. It supports execution of a variety of workloads, including SQL queries and machine learning applications. Recently, many enterprises use Spark to use its fast in-memory processing of big volumes of data. Machine Learning has plenty of useful algorithms which can deal with categorical problems. Using the DT, FT, and Neural Network to handle the data set. Prediction Primary crime types with 34 categorical data and Arrest data with only 2 categorical data were the main challenges in this study. By using Spark.ml library different algorithms were tested and ANN by 0.74 accuracy and Random Forest by 1 accuracy for Arrest and Crime type data respectively had been the best approaches. However, Logistic regression reaches the best result in comparison to other algorithms.**

Key words: Spark, Algorithm, MLlib, Crime, Machine Learning, Arrest, ANN

1.INTRODUCTION

In recent years, the amount of data collected, saved, and analyzed has skyrocketed, notably in the areas of Web and mobile device usage, as well as data collected from the physical world via sensor networks. Human-powered systems quickly become infeasible when confronted with this volume of data. As a result, the use of big data and machine learning systems has increased.

Many open source technologies by distributing data storage and computation across a cluster of machines, successfully handles massive data volume. Apache Hadoop [1] (through Hadoop MapReduce, a framework for doing computation tasks in parallel across multiple nodes in a computer cluster) is the most widely used of these technologies. Moreover, MapReduce has some significant flaws, including high start-up costs for each task and a dependency on saving intermediate data and calculation results to disc, both of which make Hadoop inappropriate for iterative or low-latency use cases.

Machine learning has become more and more popular in recent years. Because it provides a ton of algorithms to handle the big data problems. Machine learning is used in almost every field of the industry.

The paper aims to answer “how the spark is used in machine learning to deal with big volumes of categorical data '' and “what’s the benefit about using spark”. provides a brief introduction of big data and machine learning systems. Section 2 describes Spark ‘s core technologies. In Section 3, this paper will introduce what we prepare before doing the data analysis and what the problem statement is about this research. Section 4 will analyze the sample data by using Spark to generate the results about how Spark can deal with the sample data. Section 5 summarizes the work of this paper and gets to the conclusion.

2.BACKGROUND

In the previous research, Spark and Machine Learning are the most important things. And they will be used in the following research. This section will briefly give the background information about Spark and Machine Learning Algorithms.

A. Spark

Spark [2] is a general-purpose cluster computing platform for analyzing massive amounts of data which is in memory, fast, fault-tolerant, and scalable. It allows you to create distributed applications using the programming languages Java, Python, Scala, and R. Spark introduces Resilient Distributed Dataset (RDD), as a new data structure that can be persevered in memory to enhance the performance of interactive and iterative applications. It allows users to store data in disk and memory explicitly. We can utilize RDD to implement several new features which most of current cluster programming does not support. RDD has a large number of operations for manipulating data. Furthermore, Spark provides different kinds of data analysis with its own modules which are built on top of Spark Code engine.

The modules include Spark SQL, Spark Streaming, GraphX and MLlib[3]. Spark is generally utilised by applications that need to process data in real time. As a result, it's critical to keep improving Spark's execution performance for diverse sorts of applications. This can be achieved in a variety of forms, including the ones listed below:

- enhancing a Spark job's physical execution plan;
- effective spark scheduling strategy for heterogeneous cluster
- selecting the appropriate cluster setup, such as the number of machines and available resources of each machine.

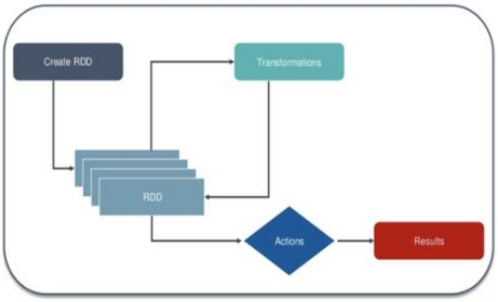


Figure_1: RDD Building block ()

Apache Spark machine learning library consists of popular learning algorithms and utilities which take advantage of both data and process parallelization [4]. The library provides a set of machine learning algorithms, such as classification, regression, clustering, dimension reduction, and rule extraction, making large-scale machine learning applications simple and quick to construct in practice. Apache Spark MLlib also includes a range of multi-language APIs for evaluating machine learning algorithms, including optimization, latent Dirichlet allocation, linear algebra, and feature engineering pipelines [5]. Many aspects of data science approaches have been improved by such a library in recent years [6], and many machine learning scientists and engineers have contributed to the big data analytics community throughout the world by establishing novel Apache Spark MLlib components.

B. Machine Learning Algorithms

The logistic regression provides the probability of the presence/absence of each land use at each location based on their drivers [7]. It is often used to handle the classification problems, and there are 2 outcomes of the logistic regression -- 0 or 1. Logistic regression also used a logistic function for example sigmoid function[9].

Decision Tree which is supervised learning is based on the if-else statement. And it can be used to solve both the classification and regression problems. It will predict the class or label by learning from training data.

An RF is an ensemble of separately trained binary decision trees [8]. And RF is included in bagging. RF is using a set of 'weak' learn like DT to build 'strong learner'. And it can be used for high dimensional data. RF also can deal with the overfitting which is a hard problem in DT. And all the basic DT will run at the same time in RF.

Gradient tree boosting is used to solve the regression problem and classification problem. It always optimizes the value of the model. GBT is one of the most effective machine learning models for predictive analytics [10].

Neural network is generated by a series of algorithms to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.[11] Neural networks can handle the changes very quickly. And it has more than one input, which means even one input has some noise data, and neural networks can figure out easily. Neural networks consist of 3 layers -- input layer, hidden layer and output layer. The structure of the neural network is similar to human's brain.

3.RELATED WORK

We are using the Chicago crime data set which has a big volume of data. And because it had plenty of features in the dataset which can fit our topic perfectly. This dataset reflects reported incidents of crime that occurred in the City of Chicago from 2001 to 2017.

Before the data analysis, it is necessary to do the data visualization to get the general information about our data.

In the Chicago crime data, it is easy to see we have over 7 millions data and over 20 features. Each feature has multiple labels which we will generate later. For instance, the 'Arrest' is a boolean with 2 different labels which are arrest and unarrest. Compared with 'Arrest', 'Location description' has many labels like 'Street', 'Resident', and 'Alley'. The following figure gives the parts of our data visualization.

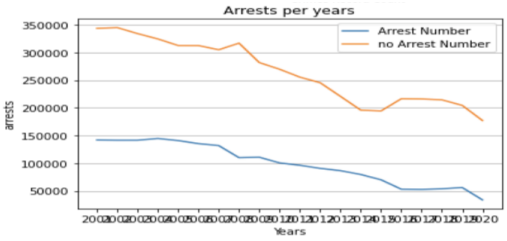


Figure_2: Arrest Data Visualization

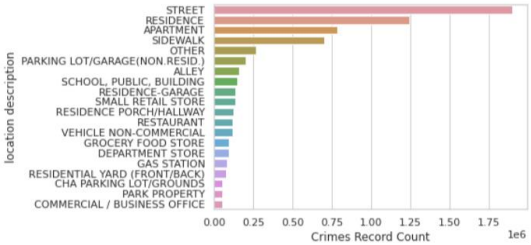


Figure_3: Location Description Data Visualization

4.DATA ANALYSIS

In this section two label data; Primary crime type with 34 categorical data and Arrest data with two categorical data were modelled. Different algorithms such as ensembles and Deep Learning were investigated. Table_1 and 2 indicate different features which were used in this study. The main important goal is exploring high and low categorical target modelling by various approaches. PySpark as a one of the Spark API was used to investigate Crime type and Arrest data.

Type	Data	Category
Feature data	Month day	31
	FBI Code	26
	Hour	24
	Year_month	12
	Week_day	7
	Arrest	2
Target	Crime Type	34

Table_1: Six Features used in Crime Type Modelling

Type	Data	Category
Feature data	Month day	31
	FBI Code	26
	Hour	24
	Year_month	12
	Week_day	7
Target	Arrest	2

Table_2: 5 features used in Crime type modelling

In this study, all algorithms such as Decision Tree, Random Forest, GBT, Logistic Regression and Neural Network were applied from the pyspark.ml library. For optimization of dataset and evaluation of result, total data was divided into train sets and test sets with 80 and 20 percent, respectively. In prediction of crime data by using the first dataset, only ensemble approaches were run and among them RF and DT were the best algorithm. However, based on table_3, ANN was the best algorithm to predict Arrest data.

Algorithm	Crime Type	Arrest
Decision Tree	1	0.69
Random F.	1	0.67
GBT	-	0.72
Logistic Reg.	0.97	0.5
ANN	-	0.73

Table_3: Evaluation Result in Crime and Arrest Data Modelling

5.CONCLUSION AND FUTURE WORK

A. Conclusion

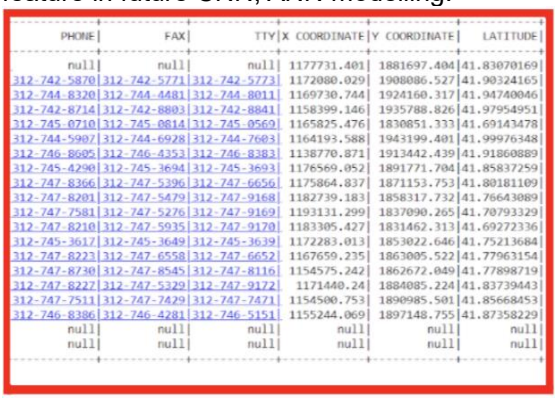
The main goal of this research was investigating in big categorical data by Spark that the important findings are as below items:

- Dealing with big data (7.3m records) and large categorical Crime data in Chicago.
- Prediction Primary crime types as a categorical dataset includes 34 categories
- The best accuracy on primary dataset was 1 for RF and DT by second features
- Arrest dada as a second target includes 5.3 million cases not arrested (False) and about 2 million of them are arrested (True).
- LR, DT, RF, GBT and ANN were run in the third dataset. Logistic regression with accuracy 0.50 and ANN with accuracy of 0.74 were the worst and the best accuracy.
- Future issue: adding CNN and RNN by joining new data about police station information

B. Future Work

For prediction of social data such as Crime type data, there are many important factors which must be considered as well. Poverty rate, social welfare, financial items, education and police characterization are

other data that can be used to increase the accuracy of the models. Police station data is one of the supplementary data that will be appended to main crime data to increase the Arrest prediction. The distance from police station to crime place can be an effective feature in future CNN, ANN modelling.



PHONE	FAX	TTY	X COORDINATE	Y COORDINATE	LATITUDE
null	null	null	1177731.401	1881697.404	41.83070169
312-742-5870	312-742-5771	312-742-5773	1172080.029	1908086.527	41.90324165
312-744-8320	312-744-4481	312-744-8011	1169730.744	1924160.317	41.94740046
312-742-8714	312-742-8803	312-742-8841	1158399.146	1935788.826	41.97954951
312-745-0710	312-745-0814	312-745-0569	1165825.476	1830851.333	41.69143478
312-744-5907	312-744-6928	312-744-7603	1164193.588	1943199.401	41.99976348
312-746-8605	312-746-4353	312-746-8383	1138770.871	1913442.439	41.91860889
312-745-4290	312-745-3694	312-745-3693	1176569.052	1891771.704	41.85837259
312-747-8366	312-747-5396	312-747-6656	1175864.837	1871153.753	41.80181109
312-747-8201	312-747-5479	312-747-9168	1182739.183	1858317.732	41.76643089
312-747-7581	312-747-5276	312-747-9169	1193131.299	1837090.265	41.70793329
312-747-8210	312-747-5935	312-747-9170	1183305.427	1831462.313	41.69272336
312-745-3617	312-745-3649	312-745-3639	1172283.013	1853022.646	41.75213684
312-747-8223	312-747-6558	312-747-6652	1167659.235	1863005.522	41.77963154
312-747-8730	312-747-8545	312-747-8116	1154575.242	1862672.049	41.77898719
312-747-8227	312-747-5329	312-747-9172	1171440.24	1884085.224	41.83739443
312-747-7511	312-747-7429	312-747-7471	1154500.753	1890985.501	41.85668453
312-746-8386	312-746-4281	312-746-5151	1155244.069	1897148.755	41.87358229
null	null	null	null	null	null
null	null	null	null	null	null

Figure 1: Future Work

6.REFERENCE

[1] J. Dean, S. Ghemawat, MapReduce: simplified data processing on large clusters, in: Proc. USENIX Conf. on Operating Systems Design and Implementation (OSDI), 2004, pp. 137–150.

[2] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, I. Stoica, Resilient Distributed Datasets: a fault-tolerant abstraction for in-memory cluster computing, in: USENIX Conf. on Networked Systems Design and Implementation (NSDI), 2012.

[4] M. Armbrust, R.S. Xin, C. Lian, Y. Huai, D. Liu, J.K. Bradley, X. Meng, T. Kaftan, M.J. Franklin, A. Ghodsi, et al., Spark SQL: Relational data processing in Spark, in: Proc. ACM Int. Conf. on Management of Data (SIGMOD), 2015, pp. 1383–1394.

[4] X. Meng et al., MLlib: Machine Learning in Apache Spark, J. Machine Learning Res., 17 (1) (2016), pp. 1235-1241

[5] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, et al., "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing", Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation USENIX Association, pp. 2-2, 2012.

[6] E. R. Sparks, A. Talwalkar, D. Haas, M. J. Franklin, M. I. Jordan and T. Kraska, "Automating model search for large scale machine learning", Proceedings of the Sixth ACM Symposium on Cloud Computing, pp. 368-380, 2015.

[7] Ravì, D., Bober, M., Farinella, G.M., Guarnera, M., Battiato, S.,Semantic segmentation of images exploiting DCT based features and random forest. Pattern Recogn. 52, 260–273, 2016.

[8]Verburg, P.H., Overmars, K.P., and Witte, N. Accessibility and land-use patterns at the forest fringe in the northeastern part of the Philippines.https://web-b-ebscohost-com.ezproxy.auckland.ac.nz/ehost/pdfviewer/pdfviewer?vid=1&sid=c171d215-fa7f-422f-bb43-00c77a4c9edb%40sessionmgr101 Geographical Journal, 170, 238–255, 2004..

[9]En.wikipedia.org. 2021. *Sigmoid function - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Sigmoid_function> [Accessed 1 June 2021].

[10] Semanjski, I. and Gautama, S., 2015. Smart City Mobility Application—Gradient Boosting Trees for Mobility Prediction and Analysis Based on Crowdsourced Data. *Sensors*, 15(7), pp.15974-15987.

[11] Investopedia. 2021. *Neural Network Definition*. [online] Available at: <https://www.investopedia.com/terms/n/neuralnetwork.asp> [Accessed 1 June 2021].