International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Classification and Enrichment of Unlabeled Feedback Data using Machine Learning

B. Kranthi Kiran, Padmaja Pulicherla

Abstract*: *These days' data gathered is unstructured. It is becoming very hard to have labelled data gathered, due to the volume of the data being generated every second. It is almost impossible to train a model on the unstructured/unlabelled data. The unlabelled data will be divided into groups using the ML techniques and CNN/Deep learning/Machine Learning techniques will be trained using the grouped data generated. The model will be enhanced over time by the feedback given by the users and with addition of new data as well. Existing models can be trained over labelled data only. Without labelled data models cannot be used for prediction and reinforcement learning. In this approach though the data is unlabelled if a feature column is specified we will be able to train the model with the help of SME. This will be helpful in many areas of classification and prediction of the trends and patterns. Machine learning, Deep learning techniques (Supervised) will be used to implement the data. Tools used will be Python, PyTorch and TensorFlow. Input can be any data (Audio/Video/pictographic/text). Labelled data and a model file which could be used for further predictions, and which will be improved over feedback.

Keywords* : *Classification, Unstrctured data, Machine Learning, CNN and Prediction.

I. INTRODUCTION

Data Analysis and Analytics are important in the market for any company to foresee its growth; data plays a major role now-a-days. Without proper understanding of data, reliable conclusions cannot be drawn. If drawn, it will have an adverse effect on business. Hence understanding data is vital and key for decision-making, but there is a problem with data being gathered. It cannot be used straight away for making decisions. These days' data gathered is unstructured. It is becoming very hard to have labelled data gathered, due to the volume of the data being generated every second.

There are many ways for labelling the gathered data. In this, we will discuss about labelling the data using machine-learning techniques. It is hard to divide data into qualitative groups. The unlabelled data will be divided into groups using the clustering techniques and CNN/Deep learning/Machine learning techniques will be trained using the grouped data generated. The model will be enhanced over time by the feedback given by the users and with addition of new data as well. By doing this we will be able to qualitatively differentiate the data which can be further used for data analysis and decision-making. As the data is evolutionary, we can reduce the dependency on SME (Subject Matter Expert) for classifying the data. In fact, it will be a very helpful tool for cleansing the data qualitatively for an SME.

In this approach, we will try to group the data given the features using clustering techniques. Descriptions about data manually are taken as inputs, we will try to match those descriptions on the groups clustered, and we will qualitatively classify the data based on the descriptions. Only one-time process. Once the labelling of data is done, then we will pass this labelled data to SME (optional Step). SME will further churn the data and then this input will be given to train an incremental ML algorithm and further data classification will be done by that. It also capable of feedback learning.

Labelling the groups based on the various categories of the data, matching the groups with the category, improving the model over feedback, scaling the model to handle near real time data and improving it on the fly are some serious challenges in this approach.

II. LITERATURE SURVEY

A significant traffic jam in machine learning is data collection and in manifold group of people it is a dynamic investigation issue. From a data administration perspective in this study an author execute a complete study of data collection. For data collection the incorporation of machine learning and data managing is fragment of a greater tendency of big data and AI incorporation and for novel investigation it opens a lot of breaks [1]. Machine learning techniques of current tendencies for the instinctive cataloging of RS pictures are lectured in this paper.Mainly the author concentrated on two novel paradigms such as active learning and semi supervised. In view of SVM based procedures, above-mentioned tactics are hypothetically and experimentally scrutinized.[2] For interference discovery in what way Semisupervised machine learning method can be recycled for both labeled and unlabeled data are shown by an author in this paper. To produce a model an author take a less quantity of labeled data and the prophecy of unlabeled traffic by this model is shown by an author in the suggested method. Future added effort can be allocating with coursing of actual world traffic [3]. Major demolition and defeat of life and assets all over the place have been done by Earthquakes. Moreover to guess the P and S wave appearance periods.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
B. Kranthi Kiran*, Associate Professor, JNTUHCEH. Kukatpally, Hyderabad, Mail:kranthikiran9@gmail.com.
Padmaja Pulicherla , Professor, Department of CSE, Teegala Krishna Reddy Engineering College, Hyderabad. Email: padmaja.j2ee@gmail.com

To assistance in categorizing earthquakes by adding novel features the guessing outcomes can be further enhanced. The author surveyed in this field the completed work and based on this survey the implementation of machine learning, din sifting and epicenter calculation algorithms have been done [4]. An automated classification procedure approach is offered in this paper. To categorize soil based the investigations were conducted. First to section the measured indications, dissection algorithm is applied and then by boundary vigor technique sections topographies are extracted. Specialists' cataloging techniques are successful impersonators by the suggested cataloging systems and the cataloging mission is automated [5]. For cataloging complications the algorithms extensively recycled are machine learning. A relative investigation of diary data considered by recognition algorithms is presented in this paper. Three algorithms are preferred for this research. Out of those the NN algorithm identified better compared to others [6].

For knowledge detection a standard knowledge gaining technique is Data mining. The mapping of information into the predefined class and sets is done by one of the data mining technique known as classification. For data case to envisage group membership this method is recycled. Numerous classification methods and investigation of classification algorithms are offered in this paper. Based on necessary presentation situations One of overhead methods can be selected [7]. By dissimilar approaches an investigation of numerous knowledge detection estimation approaches performance is presented in this paper. For outlier investigation the paper designates a thorough study on numerous methods and distinct operative issues are also defined. In arrears numerous progresses rudimentary theories are summarized by suggested system. In numerous data mining systems the exploitation of outlier investigation are dealt in this paper [8]. In cooperation the diffusion and the dissemination schemes with the deliberation of microgrids for error classification based on co-training of two classifiers a semi-supervised machine learning method offered by this paper. In allocating with many system structures with high accurateness the suggested method delivers suppleness and adaptableness are displayed by the results of this paper. The error classification accurateness can be enhanced by semi-supervised method of suggested system compared to that obtained by other machine learning methodologies are also shown by the results [9]. Designed for identifying electricity fraud in the progressive metering substructure, i.e. MFEFD, a deep-learning-based archetypal is settled on this paper. For the utilization of labeled and unlabeled data MFEFD is skilled in a semi supervised way which overwhelms deception discovery and complications of deficiency of data resources on power-driven system[10].

To develop the din filtering act, the operative use of unlabeled data is sightseen by an author in this paper. To conclude this the author sightseen ESMVU algorithm which is a new din filtering algorithm. By equating with other approaches the efficiency of suggested system is cosidered by the tests.In further work the author deliberate performance of din correction and compares it with din filtering. The big data and heterogeneous data will also deliberated for further work [11]. Since two eons time series classification has become a much enticed classification. From labeled and unlabeled time series data an author study the problem of education discriminative segments. These segments are stated as shape lets. In this paper the author offered SSSL model. In excess of current approaches real-world data trial outcomes exhibit the preeminence of this approach. It has several restrictions which require further improvement. An author provided some future guidelines to further research.[12]

In emergency sections revising radiology reports is crucial, but arduous mission. To speed up the procedure and identify the cases that plea immediate follow up Machine learning methods have been devised. For radiology report classification transversely hospitals this paper probes a semi-supervised assignment learning outline. To control massively access clinical unlabeled data is the main goal of this paper and to develop a knowledge ideal where partial labeled data is accessible. Likened to conservative overseen learning methods CNNs attain knowingly upper efficiency is shown by investigation results. Further operational sample miscellany procedures are added in future work.[13]

III. MACHINE LEARNING

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Deep learning, a subset of machine learning, utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning. The artificial neural networks are built like the human brain, with neuron nodes connected together like a web. While traditional programs build analysis with data in a linear way, the hierarchical function of deep learning systems enables machines to process data with a nonlinear approach.

A traditional approach to detecting fraud or money laundering might rely on the amount of transaction that ensues, while a deep learning nonlinear technique would include time, geographic location, IP address, type of retailer and any other feature that is likely to point to fraudulent activity.

The first layer of the neural network processes a raw data input like the amount of the transaction and passes it on to the next layer as output. The second layer processes the previous layer's information by including additional information like the user's IP address and passes on its result.

The next layer takes the second layer's information and includes raw data like geographic location and makes the machine's pattern even better. This continues across all levels of the neuron network.

A. Traditional Approach

Fig. 1: Traditional Approach

The traditional approach of classification is shown in Figure 1. No algorithmic way classifying the data further, if the data is unlabeled. More time is being spent on classifying the data. Cost is directly proportional to time being spent on classifying data.

IV. OPTIMISING THE DATA CLASSIFICATION USIN G ML AND AI TECHNIQUES FOR PREDICTION.

In this approach we will try to cluster the data. Once clustering is done, we will try to label the clustered data based on the various categories of input given about the data. Third step will be clustering the data into specific labelled groups. This is an optional step where the labelled data will be revised by SME and any corrections are made. Finally, this data is used for training a classification model and further for prediction. Feedback learning will be enabled for this model, corrections and suggestions in real time can be done.

V. DESIGN OF THE PROPOSED SYSTEM

The design of the proposed systems is shown in Figure 2.

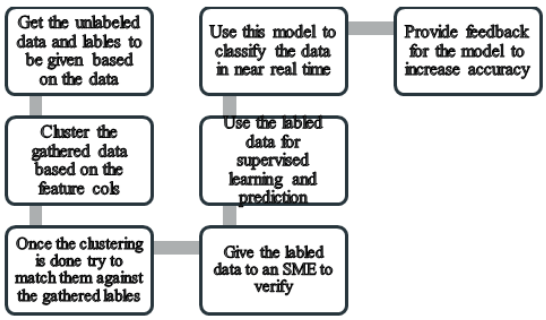

Fig. 22: Design Of The Proposed System

VI. IMPLEMENTATION OF SYSTEM

In this, the discussion is about the actual implementation of the system precisely how about poc and observations made during poc. the python is used as programming language, as it offers many libraries for data science and ml. a large unlabelled file along with key words and respective labels are taken as input and will produce a labelled data file as output for the unlabelled data provided. sme support is optional step, will help in achieving more accuracy. once the labelled file is generated, we will input this labelled file to model for training, this particular model can be enhanced over feedback via re training mechanism. for labelling the data, we have used fuzzy-wuzzy package in python. it has three methods that are best fit for the data. they are token_set_ratio, partial ratio, and ratio. it is proved that the token set ratio will be the best fit for any data as it will ignore the duplicates and does stemming automatically so that we will get the best match for the label.

VII. RESULTS AND DISCUSSION

For clustering the data, we use k means algorithm, as it provides more flexibility and almost fits for every data. before clustering, we need to do stemming and tokenization of data for that we use nltk package in python. entire sentence cannot be stemmed/lemmatised, so we need to first tokenize the data, spilt the sentence into words and lemmatise and regroup it again. for training, we use random forest algorithm on labelled data. we kept some sample data for testing and only 70 percent of the data is being used for training.

For REST service exposure package named flask is being used. It offers great variety of templates to be integrated with UI. It has its own template known as jinja2. It converts objects to displayable text and everything is handled internally. To store the data, we use mongo DB a NoSQL document oriented database. This database will be used to store the data from UI and rest API. Reads are fast in this database. For deployment we used Docker to minimize the installations at the client side, also Docker with kubernetes can scale both vertically and horizontally. Figure 3 shows the APCcall for data cleaning.

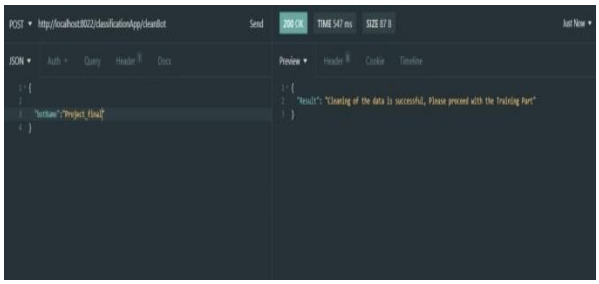

Fig. 3: API call for data clean

Fig.4 : Prediction

Figure 4 shows he screen shot captured in the step of prediction and Figure 5 shows the Predict API call in detail.

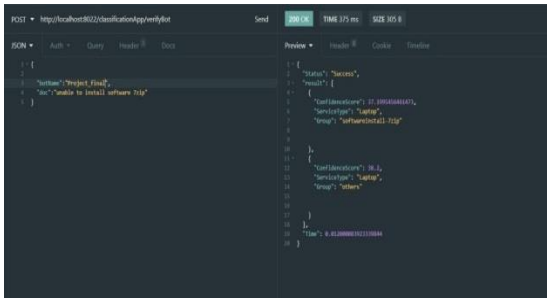

Fig. 5: Predict API Call

VIII. CONCLUSION

This idea of this project is to classify the data without having pre classified data for training purposes, as part of future work we would try to build our own algorithm for clustering the data and classifying the data. Currently it is only limited to text data, we will extend it for all types of data input such as music, video, images, sensor etc.

REFERENCES

1. Roh, Y., Heo, G., & Whang, S. E. (2018). A survey on data collection for machine learning: a big data-ai integration perspective. arXiv preprint arXiv:1811.03402.
2. Bruzzone, L., & Persello, C. (2010, July). Recent trends in classification of remote sensing data: Active and semisupervised machine learning paradigms. In 2010 IEEE International Geoscience and Remote Sensing Symposium (pp. 3720-3723). IEEE.
3. Jaiswal, A., Manjunatha, A. S., Madhu, B. R., & Murthy, P. C. (2016, December). Predicting unlabeled traffic for intrusion detection using semi-supervised machine learning. In 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT) (pp. 218-222). IEEE.
4. Li, W., Narvekar, N., Nakshatra, N., Raut, N., Sirkeci, B., & Gao, J. (2018, March). Seismic Data Classification Using Machine Learning. In 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 56-63). IEEE.
5. Bhattacharya, B., & Solomatine, D. P. (2006). Machine learning in soil classification. Neural networks, 19(2), 186-195.
6. Muhamedyev, R., Yakunin, K., Iskakov, S., Sainova, S., Abdilmanova, A., & Kuchin, Y. (2015, October). Comparative analysis of classification algorithms. In 2015 9th International Conference on Application of Information and Communication Technologies (AICT) (pp. 96-101). IEEE.
7. Sharma, S., Agrawal, J., Agarwal, S., & Sharma, S. (2013, December). Machine learning techniques for data mining: A survey. In 2013 IEEE International Conference on Computational Intelligence and Computing Research (pp. 1-6). IEEE.
8. Swapna, C., & Shaji, R. S. (2015, December). A survey on evolutionary machine learning algorithms for multi-dimensional data classification. In 2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 781-785). IEEE.
9. Abdelgayed, T. S., Morsi, W. G., & Sidhu, T. S. (2017). Fault detection and classification based on co-training of semisupervised machine learning. IEEE Transactions on Industrial Electronics, 65(2), 1595-1605.
10. Hu, T., Guo, Q., Shen, X., Sun, H., Wu, R., & Xi, H. (2019). Utilizing Unlabeled Data to Detect Electricity Fraud in AMI: A Semisupervised Deep Learning Approach. IEEE transactions on neural networks and learning systems.
11. Guan, D., Wei, H., Yuan, W., Han, G., Tian, Y., Al-Dhelaan, M., & Al-Dhelaan, A. (2018). Improving label noise filtering by exploiting unlabeled data. IEEE Access, 6, 11154-11165.
12. Wang, H., Zhang, Q., Wu, J., Pan, S., & Chen, Y. (2019). Time series feature learning with labeled and unlabeled data. Pattern Recognition, 89, 55-66.
13. Hassanzadeh, H., Kholghi, M., Nguyen, A., & Chu, K. (2018). Clinical Document Classification Using Labeled and Unlabeled Data Across Hospitals. In AMIA Annual Symposium Proceedings (Vol. 2018, p. 545). American Medical Informatics Association.