Input Data Characterization Using Machine Learning and Deep Learning

Rati Shukla¹, Vinod Kumar², Vikash Yadav³ and Mayur Rahul⁴

¹GIS Cell, MNNIT Prayagraj, Allahabad, India
²ABES Engineering College, Ghaziabad, India
³Department of Computer Application, UIET, CSJM Kanpur, India
mca.rati@gmail.com¹, vinod.kumar@abes.ac.in², vikash.yadav@abes.ac.in³, mayurrahul209@gmail.com⁴

ABSTRACT- Seeing the current trends in Information technology, a large volume of heterogeneous data is produced widely across the world by means of social media sites such as Facebook, Instagram, Google plus, etc. and electronic gadgets used by humans for instance sensors. The generated data act as garbage and makes no sense until they are categorized. A strong need of data analytics or data filtering came into existence to filter the data into concrete categories. The generated data contain enormous features with huge dimensions. The high dimension data must be reduced to low dimension data to avoid curse of dimensionality and to build a better machine learning (ML) model.

A ML model is built to perform the classification task that result into labeling of the data. In the first phase of work, a framework has proposed to process the data having features belonging to text and images using ML and DL algorithms. In the second phase of work, we have shown data related behavior using ML and DL framework. The proposed work is generic in sense of FS techniques and FE techniques using ML classifiers and DL classifiers. After carefully implementing the algorithms on the datasets we have evaluated the results taking accuracy as our performance metric. After carefully analyzing the accuracy it can be concluded that DL algorithms give better results than ML algorithms on both text and image dataset.

Keywords: Machine Learning, Deep Learning, Feature Selection, Feature Extraction, Naïve Bayes, Natural Language Processing.

1. Introduction

A large volume of data is generated wide across the world through social media sites for instance Google, Facebook, Twitter, etc. and electronic gadgets for instance sensors. These social media sites serve as a top level platform in market research inspection and decision making by the humans and the sensors are used to collect the real world and transmit it to the computer. There are lot of challenges faced by an analyst or data scientist in day to day life such as capturing, storing, sharing, transferring, analyzing, filtering and visualizing the data. To overcome all the above mentioned challenges is not that easy task so we focused upon the filtration of data. Taking filtration as a core challenge we thought to work on classification of data around us.

The size of the data is so huge to manage. It contains enormous features with huge dimensions and acts as garbage and makes no sense until they are categorized. This high dimensional data must be reduced to low dimensional to make the processing of the data easier and build a better machine learning
model to classify the data. The low dimension data can easily train the model of high accuracy in lesser time.

In the real world we daily come across the data and the objects that need to be classified to provide correct label. To provide correct label to the data and object is very significant task. Suppose you forget the name of the faculty who teaches you machine learning. It is shameful if somebody ask you the teacher name, although she teaches you daily. Then how would describe the teacher to the person. It is the set of features that the teacher possesses with the help of which you can correlate the teacher to the person. So classification is most important part of our life. To make our life easier and convenient, we select the data features and perform the classification task.

The main objective is to propose a common platform to evaluate the data related behavior of heterogeneous data. By heterogeneous data we mean data belonging to different formats such as texts, images, audio, video, etc. these heterogeneous data contain features. Every object possesses unique features to recognize it. The features of the object are in high dimensions due to which it becomes difficult to process the data and get high accurate results. The high dimension data must be reduced to low dimensions through FS techniques in order to avoid curse of dimensionality (COD). The researches show that the performances of the classifiers becomes saturated or declined with the increases in count of the features. Another goal of the project is to understand the role of the FS techniques such as wrappers, filter and embedded. The figure 1 below describes the effect on classifiers performance with the increases in count of features.

Figure 1. Effect on classifiers performance with increase features

2. Related Work

A lot of research has been done in this area. In [3] the authors have proposed the automated text classification process using ML techniques. The study shows extensive information related to text representation and Feature Selection (tf-idf, term frequency, etc.). In [4] the authors have performed an extensive analysis of Feature Selection techniques. A vivid description of these techniques provides the greater insights to Feature Selection. In [5] the authors have worked upon the idea of feature extraction and criteria of Feature Selection methods and few challenges. In [6] the authors have proposed the ML framework for image classification and image recognition. They showed great interest in feature extraction technique; instead of global color FE they have used Speed up Robust Feature. They have used Bag of Words technique. Accuracy of model is measured on ‘Caltech 101’ images.

In [7] the authors have proposed a hybrid MLP-CNN predictor to classify remotely sensed image. The
efficiency of ensemble MLP-CNN predictor was calculated in both rural and urban areas. It was calculated using collection of aerial photography with extra satellite data. MLP-CNN predictor gain significant enhancement. In [8] authors have proposed a conventional FS technique namely Information Gain to increase the efficiency for twitter sentiment analysis. The analysis is performed on movie reviews using Naïve Bayes classifier. In [9] the authors have proposed a classification technique for power network icing detection (PNID) image on the basis of CNN classifier to efficiently classify and recognize PNID image. In addition, keeping the shortcomings of CNN algorithm into consideration authors proposed the hybrid classification system combining CNN and SVM. On very first, CNN was used to extract features, and then SVM was used to replace the soft max layer of the CNN to classify PNID images.

2.1. Conventional Feature Selection Techniques

In ML, conventional FS techniques such as PCA, LDA, IG, etc. are generally used in the classification of text related data. These conventional techniques are majorly used by the researchers. So, a survey is performed to find out the contributions made by the researchers in the field of FS in last few years. This survey helps us to figure out the feature role on different datasets in a single glance.

In [10] the author has proposed a spam detection technique using wrapper method as the Feature Selection technique. The author compared MBPSO with traditional Feature Selection techniques like SFS and SBS. The obtained results show that MBPSO shows superiority over SBS and SFS. In [11] the authors have suggested a conventional FS technique namely IG is used to increase the efficiency for twitter sentiment analysis. The analysis is performed on movie reviews using Naïve Bayes classifier. In [12] the author has proposed two new FS metrics: (a) Relevance Frequency FS (RFFS) and (b) Accuracy for text classification and performance are measured. The result is then compared to chi-square and IG. The new proposed methods outperform and can be used in the area of image processing and malware analysis.

In [13] the author has proposed an effective FE through advance use of PCA techniques such as Folded-PCA (FPCA), spectrally-Segmented-PCA (SSPCA), Segmented-PCA (SPCA) and Super pixel wise PCA (SuperPCA). The result shows that these advance techniques outperforms. In [14] the authors have proposed a comparison of supervised and unsupervised FE techniques such as PCA, LDA and the combination of both PCA and LDA for hyper spectral image classification.

2.2 Dissimilar Feature Selection Techniques

In ML, conventional FS techniques such as PCA, LDA, IG, etc. does not outperforms. To achieve the results of high accuracy the researchers have worked upon some exotic methods in the classification of text and image related data. So, a survey is performed to find out the contributions made by the researchers in the field of FS in last few years. This survey helps us to figure out the feature role on different datasets in a single glance.

In [15] the author has proposed an extensive analysis of twelve FS methods such as IG metrics for text classification. After analyzing the results, a new FS method called Bi-Normal Separation (BNS), is proposed. In [16] the author has proposed a novel approach for FS technique namely, feature unionization. The evaluated result showed that the proposed approach worked effectively.
In [17] author has proposed a strategy to detect the brain areas and subjects allied to Alzheimer’s disease (AD) based on 3Dimensional MRI scans with the help of ML and Eigen brain. In [18] author has proposed a novel and improved strategy of feature extraction for classifying image with the help of new ranking strategy.

A wide-ranging survey is performed by us on FS techniques and various framework of ML and DL framework in text classification and image classification. In our work we have cited 20+ research papers and journals published by the researchers and scholars in some great publications such as IEEE, Springer, Science Direct, etc. After thoroughly reviewing more than 75+ papers we have concluded that there is no common FS approach to select the features related to different formats (text and image) of data. Every heterogeneous data has its own features i.e. text, images, videos have different, etc. So we found nothing in common. But we get ideas to classify the data using some common classifiers.

3. System Design and Methodology

ML and DL frameworks are most commonly used to provide intelligence to the machine. These frameworks are used for solving complex problems. The algorithms are highly dependent upon the input data. Data plays a very crucial part in the performances of the algorithm. Initially, data is separated into training and testing in both ML and DL. There are some standard criteria for splitting data for training and testing, so it is highly desired to select the dataset for training and testing carefully. ML algorithms works better for small amount of data but DL algorithms give better results with larger datasets. When the dataset is small the DL algorithms doesn’t perform that well. The figure 2 below clearly depicts the performance of ML and DL with increase in the size of the data.

![Figure 2. Performance evaluation of ML and DL model with increase in size of data.](image)

3.1 Machine Learning Framework

Machine Learning, a vast interdisciplinary area of AI which borrows and builds upon ideas from various domains like computer science, statistics, mathematics and cognitive science. Machine Learning can be classified into two major areas including supervised and unsupervised. The aim of supervised Machine Learning is to locate the association between input variables (independent variables) and output variables (dependent variables). The goal of unsupervised ML is to drive the conclusions from datasets with no labeled outputs. Supervised learning can be further clubbed into classification and regression problems. In ML we use various statistical techniques which provides, computer the capability to understand and act without being explicitly programmed by the user. It helps the computer systems in acquiring new knowledge and in combining different type of information for betterment of decision
making of the agents. The behavior of machine learning techniques depends upon various reasons. One of them is related to dataset we have taken into consideration. ML algorithm works well with the smaller size dataset. After performing a lot of research we conclude that it is better to work with some commonly used algorithms. We have used the following algorithms under ML frameworks for result analysis. The algorithms used for ML framework are:

- Naïve Bayes (NB)
- Support Vector Machine (SVM)
- Multilayered Perceptron (MLP)

### 3.1.1. Naïve Bayes

This classifier uses Bayes’ theorem and it is categorized into supervised Machine Learning algorithm. The variables that are used for the purpose of generating the model are independent to each other.

\[
P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)
\]

The process of classification is performed by deriving utmost posterior probability which is said to be maximal probability \(P(X|C_i)\) using above supposition applied to Bayes theorem. The supposition minimizes the processing cost by including the class distribution. Even though the supposition is not valid in most cases since the attributes are dependent, surprisingly NB performed impressively. NB is a very simple algorithm to implement and it provides good result in most cases.

### 3.1.2. Support Vector Machines (SVM)

SVM is a supervised Machine Learning technique used for both regression and classification. In maximum cases it is used for the purpose of classification. It depends on the inspiration of finding the decision boundaries called hyper-planes that help ML model to classify data points. Data points falling on the different side of the decision boundaries can be classified into distinct classes. Also, quantity of features represents dimension of the hyper-plane. When input features are two, then the hyper-plane is a line. When it is three, then hyper-plane becomes a 2-D plane. When the number of features exceeds three, then it will become very difficult to imagine. The classification of data point using hyper-planes is depicted in the figure 3.

![Figure 3. Depiction of hyper-plane dividing the dataset into two classes](image-url)
3.1.3. Multilayer Perceptron

MLP are generally applied to supervised learning problems. MLP model consist three or more layers, single input, single output and more than one hidden layer. The number of hidden layer depends upon the complexity of a problem. The model trains itself on input-output pairs and learns to find the dependencies between those inputs and outputs. The main focus is on minimizing the error by adjusting the parameters like weights and biases, while activation function is mainly used to evaluate the output. Back-Propagation technique plays a major role in adjusting the weights and biases relative to the error. The error can be calculated in various ways. But the most commonly used technique is root mean squared error (RSME). A MLP model is shown in figure 4.

![Multilayer Perceptron Model](image)

Figure 4. Multilayer Perceptron Model

3.2. Deep Learning Framework

A Deep Learning model structure is like how a human brain works or how a human draw conclusion, it is created to persistently examine data using logic structure related to human methodology to represent conclusions. To attain this, layered architecture of algorithms in deep learning is used called an ANN. The structure of ANN is motivated by the biological neurons of human brain. This makes for machine insight that is undeniably more competent than that of standard ML models. DL algorithms work in such a way that the learning models itself extracts the features on which classification of dataset will be performed. That’s why the DL model is very complex in comparison of ML model. After performing a lot of research we conclude that it is better to work with some commonly used algorithms. We have used the following algorithms under ML frameworks for result analysis. The algorithms used for DL framework are:

- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)

3.2.1. Convolutional Neural Network (CNN)

A Convolution Neural Network is a type of Deep Learning algorithm. It is majorly used in analyzing visual imagery. It is most popular DL architecture. It is also applied to NLP, recommender systems, etc. A
very less pre-processing is requisite in Convolutional Network (ConvNet) as compared to other
classification algorithms. It can take image as input, allocate biases and weights to different
objects/aspects in image and be able to discriminate one from the other. A sequence of pooling and
convolution operations are processed which is followed by a various fully connected layers. If we are
performing multiclass classification the output is softmax (activation function). While in primitive
methods filters are implemented with enough training, ConvNets have the ability to learn these
characteristics. The model can be seen in figure 5.

3.2.2 Recurrent Neural Network (RNN)

RNN are the techniques used for sequential data and used by Google’s Voice Search and Apples Siri.
Because it is first technique that remembers the input, because of an internal memory, so this algorithm is
perfect suited for ML problems that move around sequential data. It is most popular used algorithms in
DL in the past few years. RNN is a controlling and robust type of neural networks. Due to internal
Memory, the ability of RNN’s to remember all the useful features about the input that is received by
algorithm or RNN model, which allows them to be very precise in predicting the next outcome. The
working model of RNN is shown in the figure 6.
3.3 Data related description

We performed the experimental evaluation on:
- Text data (SMS Spam Collection dataset) [19]
- Image data (MNIST dataset) [20]

3.3.1. SMS dataset

The SMS dataset is a collection of text messages written in English language. It consists of 5572 samples each sample is labeled with spam or ham as shown in the figure. There is only one message per line. Out of total messages 87% of the messages are ham messages and 13% of the messages are spam messages as shown in the figure. The sample of the dataset is in the below shown table 1.

| Labels | Messages |
|--------|----------|
| Ham    | Go until jurong point, crazy... Available only in bugis n great world la e buffet… Cine there got amore wat… |
| Ham    | Ok lar… Joking wif u oni… |
| Spam   | Free entry in 2 a wkly comp to win FA cup final tkts 21st May 2005. Text FA to 87121 to receive entry question (std txt rate ) T&C’s apply 08452810075over18’s |
| Ham    | U dun say so early hor… U c already then say… |

3.3.2 MNIST dataset

MNIST stands for “Modified National Institute of Standards and Technology”. The datasets contain 70000 images of handwritten digits. Among those 70000 samples, 60000 samples are of training images and 10000 samples are of testing images. It is subset dataset of NSIT dataset. The testing samples has 5000 images of original NSIT training dataset and 5000 images of original test dataset. These first 5000 datasets are much cleaner than last 5000 datasets. Each digit is normalized and is centered in a gray-level image with 28 X 18 i.e. 784 in total as features. The sample of dataset is shown in the figure 7.
3.4 Proposed Model

Figure 8. Proposed Model

The model is generated by using classification algorithms of either ML or DL also known as classifiers. These algorithms are part of the process of classification as they define the method of classification. Firstly, data pre-processing consisting of data cleaning, feature extraction and feature selection is performed. Secondly data is separated into two different set known as training and testing set. Data splitting plays a vital role in the final result, if the training data is less than testing data the results are not much accurate. So the train-test splitting should be done in such a way that training data is more than testing to get more accurate results. Thirdly, apply the classifier on the training data to train the model and after training, model will ready for validation test to classify data in to different categories. After classification of testing data, performance of the model is measured. Some of the performance measures metrics are accuracy, precision, recall, etc. The model proposed is shown in the figure 8.

4. Results

4.1. Results with text dataset

The behavior of various ML algorithms used in our prediction with SMS dataset is shown in the table 2. The performance of the algorithms is examined using accuracy as the performance metrics.

| S No. | Classifier | Accuracy  |
|-------|------------|-----------|
| 1     | NB         | 0.98061   |
| 2     | SVM        | 0.99066   |
| 3     | MLP        | 0.98707   |

The behavior of various DL algorithms used in our prediction with SMS dataset is shown in table 3.
Table 3. DL behavior with text data

| S NO. | Classifier | Accuracy |
|-------|------------|----------|
| 1     | CNN        | 0.9924   |
| 2     | RNN        | 0.9985   |

From the above results we can infer that DL algorithms show better results than ML algorithms with the text dataset.

4.2. Results with image dataset

The behavior of various ML algorithms used in our prediction with image dataset (MNIST dataset) is shown in table 4.

Table 4. ML behavior with image data

| S No. | Classifier | Accuracy |
|-------|------------|----------|
| 1     | NB         | 0.8365   |
| 2     | SVM        | 0.9787   |
| 3     | MLP        | 0.9713   |

The behavior of various DL algorithms used in our prediction with image dataset (MNIST dataset) is shown in table 5.

Table 5. DL behavior with image data

| S No. | Classifier | Accuracy |
|-------|------------|----------|
| 1     | CNN        | 0.97801  |
| 2     | RNN        | 0.9911   |

From the above results we can infer that DL algorithms are showing better results with image dataset. After carefully analyzing the accuracy we conclude that DL framework gives better results than ML framework on both text and image dataset. The pictorial representation of results is shown in the figure 9 and figure 10.

Figure 9. Result comparisons of text data
5. Conclusion and Future Work

After carefully analyzing the accuracy we conclude that DL framework gives better results than ML framework on both text and image dataset. We are able to extract the data related behavior of several of various formats of data using FS techniques. We can easily provide labels to the data generated through various sources and objects in real life. An approach to FS in the scope of classification problems related to various data formats (text, audio, image, etc.) is also discussed. The literature work done is quite descriptive and is in full support of the FS. The social media sites for instance Facebook, Instagram and Twitter are generating data in huge volume. So our project can be expanded to provide the labels to heterogeneous data. We can easily implement the ML and DL algorithms to provide the labels to data. Along with that the data miners are working passionately for a very long time in the field of ML and DL to give some contributors to the community. Our proposed model can provide some glimpse to the researchers and data miners in the same domain.

References

[1] Misra Rajiv, Live internet stats, Big Data Computing, NPTEL. Retrieved from https://drive.google.com/open?id=1VUDWqSSBr5AXbMWAH0tCW-cv7veCPgm.

[2] Cowley Benjamin, The spiral model of design. Retrieved from https://www.researchgate.net/profile/Benjamin_Cowley/publication/287268554/figure/fig1/AS:614209679261713@1523450407826/The-spiral-model-of-design-This-consists-of-four-phases-1-Determine-objectives.png.

[3] Ikonomakis, Emmanouil & Kotsiantis, Sotiris & Tampakas, V. (2005). Text Classification Using Machine Learning Techniques, WSEAS transactions on computers, 4, pp. 966-974.

[4] Girish Chandra shekar, Ferat Sahin, A survey on feature selection methods, Computers & Electrical Engineering, 40(1), 2014, pp. 16-28, ISSN 0045-7906.

[5] S. Visalakshi and V. Radha, A literature review of feature selection techniques and applications: Review of feature selection in data mining, 2014 IEEE International conference on Computational Intelligence and Computing Research, Coimbatore, 2014, pp. 1-6.

[6] Loussaief, Sehla & Abdelkrim, Afe. (2016), Machine Learning framework for image classification, pp. 58-61.
[7] Ce Zhang, Xin Pan, Huapeng Li, Andy Gardiner, Isabel Sargent, Jonathon Hare, Peter M. Atkinson (2017), A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification.

[8] S. Widya Sihwi, I. PrasetyaJati and R. Anggrainingsih, “Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier,” 2018 International Seminar on Application for Technology of Information and Communication, Semarang, 2018, pp. 190-195.

[9] Jiazheng Lu, Yu ye, XunjianXu and Qinpu Li (2019), Application research of convolutional neural network in image classification of icing monitoring in power grid.

[10] Yudong Zhang, Shuihua Wang, Preetha Phillips, GenlinJi, Binary PSO with mutation operator for future selection using decision tree applied to spam detection, Knowledge-Based Systems, 64, 2014, pp. 22-31, ISSN 0950-7051.

[11] S. WidyaSihwi, I. PrasetyaJati and R. Anggrainingsih,” Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier,” 2018 International Seminar on Application for Technology of Information and Communication, Semarang, 2018, pp. 190-195.

[12] Durmuş Özkan Şahin & Erdal Kılıç (2019) Two new feature selection metrics for text classification, Automatika, 60:2, pp. 162-171.

[13] Md. Palash Uddin, Md. Al Mamun & Md. Ali Hossain (2019) Effective feature extraction through segmentation-based folded-PCA for hyperspectral image classification, International Journal of Remote Sensing.

[14] A. A. Joy, M.A.M. Hasan and M.A. Hossain,” A comparison of Supervised and Unsupervised Dimension Reduction Methods for Hyperspectral Image Classification,” 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox’s Bazar, Bangladesh, 2019, pp. 1-6.

[15] Forman, G. (2003), An Extensive Empirical Study of Feature Selection Metrics for Text classification, Journal of Machine Learning Research, 3, pp. 1289-1305.

[16] Abbas Jalilv and, Naomie Salim, Feature unionization: A novel approach for dimension reduction, Applied Soft Computing, 52, 2017, pp. 1253-1261, ISSN 1568-4946.

[17] Zhang Y, Dong Z, Phillips P, Wang S, Ji G, Yang J, Yuan TF, Detection of subjects and brain regions related to Alzheimer’s disease using 3D MRI scans based on eigen brain and machine learning, Front Comput Neurosci. 2015 Jun 2; 9:66.

[18] Xuan Zhou, Jiajun Wang, J. (2015), Feature Selection for Image Classification Based on a New Ranking Criterion, Journal of Computer and Communications, 3, pp. 74-79.

[19] Kaggle, SMS dataset. Derived from https://www.kaggle.com/ishansoni/sms-spam-collection-dataset.

[20] Kaggle, SMS dataset. Derived from https://www.kaggle.com/ngbolin/mnist-dataset-digit-recognizer/data.