IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Special Issue-9, April 2019 E-ISSN: 2347-2693

A Comparative Model of Feature Engineering With and Without Domain Knowledge

Rohit Bohra[1], Pankaj Karki[2], Kumudavalli M V[3]

[1,2]Scholar, Bangalore
[3]Dept. of Computer Applications, Dayananda Sagar College of Arts, Science and Commerce

Corresponding Author: rohitbohra23051994@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si9.5659 | Available online at: www.ijcseonline.org

Abstract: One of the key aspects of building a good machine learning model is Feature engineering. Feature engineering is a process where we create new features from existing raw features. To create new features, we require domain experts who have knowledge of the subject. By using their knowledge they create new features which are helpful for a machine to learn better. The time taken by the domain experts to understand the data and then create new features is time-consuming and expensive. This problem is addressed with a neural network which will not require domain experts to engineer new features. Current paper deals with the case study pertaining to the data of Human Action Recognition. Using the data, the machine predicts the various physical actions and appearances of a person like if the person is sitting, standing, walking, walking up stairs, and walking downstairs or lying. We compare the accuracy of the model using data which was feature engineered by experts and the model which was not feature engineered by the domain experts.

Keywords- Machine Learning, Feature Engineering, Domain Knowledge, Human Action Recognition, Neural Networks.

I. INTRODUCTION

Data is the world's most valuable resource, it is very important for an industry to grow and be a helping hand to the human. Data is used to build intelligent machines which can work on complex problems which require human understanding. Data can be used in any domain like in health care to detect disease like cancer at earlier stages, improve business revenue, targeting the right audience in digital marketing, etc. The key factor for the machines to work on relatively complex problems is the availability of rich information. The rich information is not readily available in the form of data but needs to be generated from the existing raw data. Feature Engineering is the process of generating rich information from the raw data. The process of generating new features requires a lot of experience about the particular subject that the problem is related to [1]. This requires a lot of time and is highly expensive. For many real-world problems, we try finding a solution using Artificial Intelligence which needs to be faster and cheaper. But generally, we cannot actually generate features manually and then train a model because of high time complexity. For this reason, we use the Neural Network.

Neural Network uses the raw data and generates features by itself which works well for complex problems. This avoids the requirement for a domain expert to work on finding complex data from raw data. In this paper, we are using Human Action Recognition data to do the comparative study. The dataset is collected through a smartphone by wearing it on to the waist [2], and it is categorized as Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing and Laying.

A. Feature Engineering

Feature engineering is one of the essential steps in the applications of machine learning. For training a machine learning model well, we require pre-processed data. In the process of data pre-processing, other than data cleaning we also need to feature engineer the data in such a way that a model's performance is high. In feature engineering, we require domain expertise of the subject to create new features from the existing raw features. The features used in training a machine learning model is important as it influences the result that would be achieved by the model. This process is time-consuming and expensive because the domain expert needs to understand the problem and data before creating new features.

B. Classical machine Learning

Classical machine learning is a set of algorithms and statistical modeling which takes data as an input and models it using statistics and algorithms to give the desired output. The most common classical machine learning models are Linear Regression, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, etc [3]. Even though the time is taken to train this model is less but the time taken by domain experts to create new features is very high. The domain experts need to try out many methods which might or might not work well to build a model [4]. The time taken by a domain expert to create new feature is a big disadvantage because in the real world, time available to solve a problem is very narrow and the classical machine learning models even though take less time to train a model but the amount of time taken to feature engineer the data is very high.

C. Neural Network model

A neural network is basically inspired by how neurons in our brain work to send signals to each other to do a particular work. A set of neurons are interconnected in such a way that the output of a neuron will be an input to the other neurons. In a neural network, we don't need domain expertise to do feature engineering as the model engineers the new features by itself by taking raw data as input and the accuracy achieved is similar to the output accuracy achieved using classical ML models with feature engineering [5].

The performance of the models are compared using two types of data, one with feature engineering done by domain experts and another without any domain expert but using an only neural network to build a model, and then compare the performance of the models using Confusion matrix and Accuracy as performance matrix. The paper is organized as follows: Section II gives a view of related works pertaining to the subject of study. Data set creation details are given in Section III. Section IV deals with the methodology followed whereas Section V describes the various outputs obtained after the study. SECTION VI gives the conclusion of the study.

II.RELATED WORKS

The field of Artificial Intelligence is gaining popularity as the utilization and implementation factors have increased exponentially. M. B. Holteet.al [6] have discussed the human recognition activity through multi-view video and the recent developments in the domain. K. Charalampous and A. Gasteratos [7] have given insight about online deep learning methods which can be a better aid in action recognition. Since not much emphasis is given on model building with and without feature engineering the current work is carried out.

III. DATASET DESCRIPTION

These experiments were conducted for a group of 30 volunteers between the ages of 19 to 48 years. Everyone wears a smartphone on the waist (Samsung Galaxy S II) for six activities (Walking, Walking upstairs, Walking downstairs, Standing, Sitting, Laying). Using its embedded accelerometer and gyroscope[8], we capture 3-axis acceleration and 3-axis angular velocity at a constant rate of 50 Hz. Data has been manually tagged by video recording experiments [9]. The obtained data sets were randomly divided into two groups, of which 70% of the volunteers were selected to generate training data and 30% of test data.
The sensor signals (accelerometer and gyroscope) were pre-processed by applying a noise filter and then sampled in a fixed width sliding window (128 readings/window) of 2.56 seconds and 50% overlap. The sensor acceleration signal has a gravity and body motion component that is separated into body acceleration and gravity using a Butterworth low pass filter. It is assumed that gravity has only a low-frequency component, so a filter having a cutoff frequency of 0.3 Hz is used. From each window, the feature vector is obtained by calculating variables from the time domain and the frequency domain.

The raw features that were taken:
body_acc_x, body_acc_y, body_acc_z, body_gyro_x,
body_gyro_y, body_gyro_z
total_acc_x, total_acc_y, total_acc_z

These are the engineered features by the domain expert
using raw features:
mean(): Mean value, ***std()***: Standard deviation, ***mad()***: Median absolute deviation, ***max()***: Largest value in array, ***min()***: Smallest value in array, ***sma()***: Signal magnitude area, ***energy()***: Energy measure. Sum of the squares divided by the number of values, ***iqr()***: Interquartile range, ***entropy()***: Signal entropy, ***arCoeff()***: Autorregresion coefficients with Burg order equal to 4, ***correlation()***: correlation coefficient between two signals, ***maxInds()***: index of the frequency component with largest magnitude, ***meanFreq()***: Weighted average of the frequency components to obtain a mean frequency, ***skewness()***: skewness of the frequency domain signal, ***kurtosis()***: kurtosis of the frequency domain signal, ***bandsEnergy()***: Energy of a frequency interval within the 64 bins of the FFT of each window [10],***angle()***: Angle between to vectors.

IV. METHODOLOGY

The model's performance was analyzed on both features and non-featurized data sets.
In the first step, we build classical machine learning models with both raw and featurized data. The results were noted down for each linear, non-linear and tree-based models with both the data sets.
In the second step, we build deep learning model (LSTM) with raw data to see how the model performs in a neural network without taking feature engineered data.

V. RESULTS AND DISCUSSION

The various results obtained after model building and testing its efficiency is presented as below:

Step 1: The Human Action Data is being collected from UCI which provides open source data. The data is in two formats a) Raw data b) Feature engineered data.

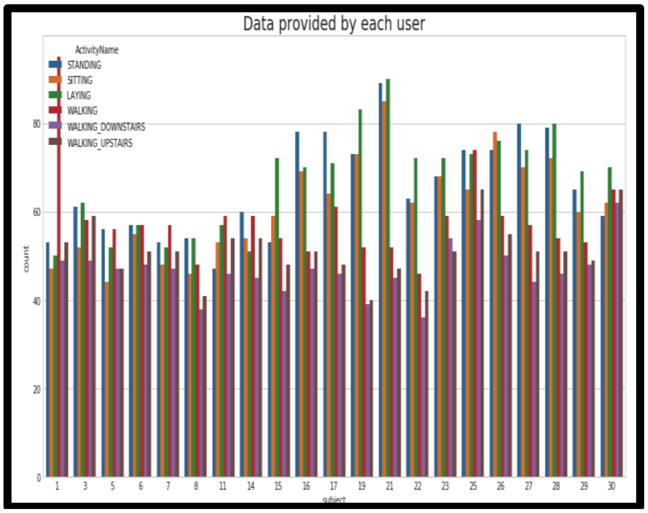


Figure 1: Histogram depicting the raw data from each user

Step 2: Raw data contains only the data which is not featured by a domain expert and thus hard to classify its activity. We have used TSNE(dimensionality reduction technique) to reduce the dimensionality of the data to see whether the data is separable or not and we can clearly interpret that the activities are hard to classify by seeing the TSNE plots and similarly we also applied TSNE on the featured data and saw that we can interpret most of the classes clearly which are separable from each other.

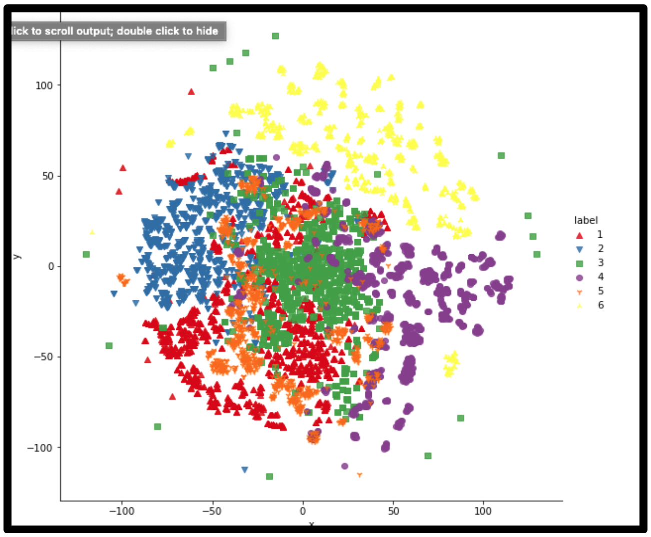


Figure 2: Image showing that there is no good separation of each class in raw data

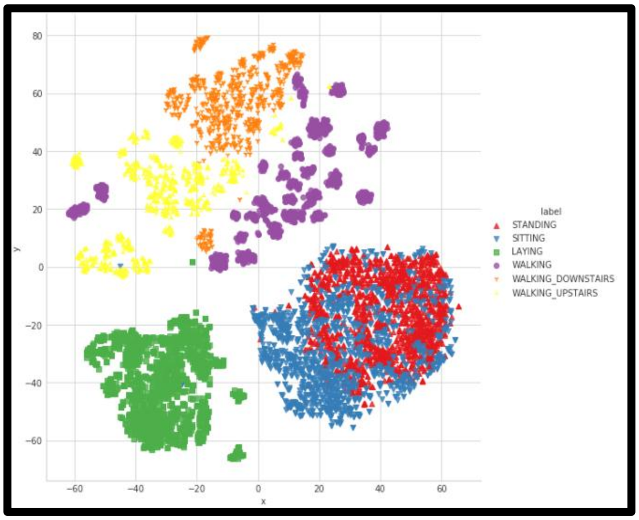
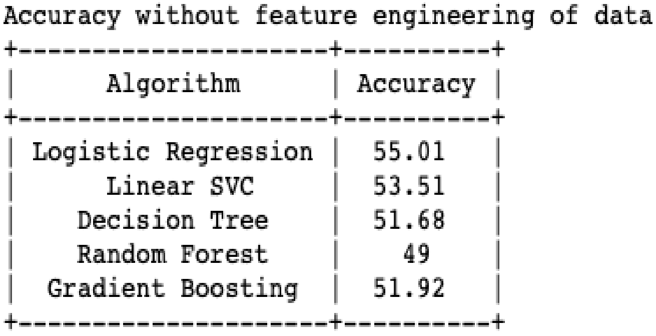


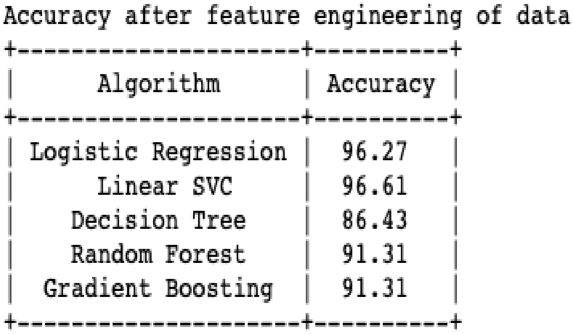
Figure 3: Image showing that there is good separation of each class after feature engineering of raw data

Step 3: We implemented classical machine learning algorithms with raw data to see how well the model is being trained. In the table, we can see that the accuracy is approximately 50% for various classification models algorithms.



```
Accuracy without feature engineering of data
+---------------------+----------+
|      Algorithm      | Accuracy |
+---------------------+----------+
| Logistic Regression |  55.01   |
|      Linear SVC     |  53.51   |
|    Decision Tree    |  51.68   |
|    Random Forest    |    49    |
|  Gradient Boosting  |  51.92   |
+---------------------+----------+
```

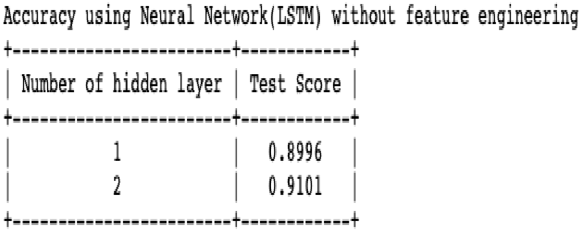
Step 4: To improve the accuracy, we tried using the feature engineered data which was done by the domain experts. The accuracy has been drastically improved as seen in the table. The best accuracy was around 96% for the model with the classical machine learning model.



```
Accuracy after feature engineering of data
+---------------------+----------+
|      Algorithm      | Accuracy |
+---------------------+----------+
| Logistic Regression |  96.27   |
|      Linear SVC     |  96.61   |
|    Decision Tree    |  86.43   |
|    Random Forest    |  91.31   |
|  Gradient Boosting  |  91.31   |
+---------------------+----------+
```

Step5: We tried reaching the accuracy of around

96%without feature engineering and only by using the raw data. As the data is a Time series data, using LSTM technique was the best. We experimented using one hidden layer and two hidden layered neural network with parameter tuning of the hidden layers and dropout rate, and got an accuracy of 91% in two hidden layer LSTM model.



```
Accuracy using Neural Network(LSTM) without feature engineering
+------------------------+------------+
| Number of hidden layer | Test Score |
+------------------------+------------+
|           1            |   0.8996   |
|           2            |   0.9101   |
+------------------------+------------+
```

VI. CONCLUSION

From the above experiment, it is seen that with raw data it is difficult to train a classical machine learning model when the problem is complex and hence domain expert is needed to generate new features which would be useful to build a better performing model. When we need to solve a complex machine learning problem we can use a neural network as a model training technique which would give accuracy similar to the classical machine learning model with feature engineering. Even though training a neural network model is time-consuming but it is comparatively less when a domain expert takes time to understand the data and generate new features which might or might not be useful for training a machine. Training a machine using the neural network is a better option when we don't want to interpret the output predicted and if the output needs to be interpreted then domain knowledge and classical machine learning model is a better option.

ACKNOWLEDGMENT

Rohit Bohra thank and acknowledge the support extended by Applied AI Course team for performing the model building and testing process.

REFERENCES

[1] B. Fish A. Khan N. Chehade C. Chien and G. Pottie, "Feature selection-based on mutual information for human activity recognition," IEEEInternational Conference on Acoustics, Speech and Signal Processing,vol.37, pp.1729-1732, 2012.

[2] D.Anguita A. Ghio L. Oneto X. Parra and J. Reyes-Ortiz, "Human Activity Recognition on Smartphones Using a Multiclass Hardware- Friendly Support Vector Machine." International Conference on Ambient Assisted Living and Home Care, pp.216-223, 2012.

[3] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions:A local SVM approach," in Proc. Int. Conf. Pattern Recognit., vol. 3.2004, pp. 32–36.

[4] Changki Lee, Gary Geunbae Lee, "Information gain and divergence based feature selection for machine learning based textcategorization," Information Processing & Management, vol. 42,Issue 1, pp. 155-165, January 2006.

[5] J. Yang, J. Wang and Y. Chen, "Using acceleration measurements foractivity recognition: An effective learning algorithm for constructingneural classifiers," in Pattern Recognition Letters, vol.29, no.16, pp. 2213-2220, 2008.

[6] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognitionfrom multi-view videos: Comparative explorations ofrecent developments," IEEE Journal of Selected Topicsin Signal Processing, vol. 6, pp. 538–552, 2012.

[7] K. Charalampous and A. Gasteratos, "On-line deep learningmethod for action recognition," Pattern Anal. Appl., pp. 1–18,Aug. 2014.

[8] Q. Li J. Stankovic M. Hanson and A. Barth, "Accurate, Fast Fall Detection Using Gyroscopes and Accelerometer-Derived Posture Information,"Sixth International Workshop on Wearable and Implantable Body SensorNetworks, pp.138-143, 2009.

[9] N. Gkalelis, N. Nikolaidis, and I. Pitas, "View independent human movement recognition from multi-view video exploiting a circular invariant posture representation," in 2009 IEEE International Conference on Multimedia and Expo, June 2009, pp. 394–397.

[10] O. Banos J. Galvez M. Damas H. Pomares and I. Rojas, "Windowsize impact in human activity recognition," Sensors, vol.14, no.4, pp.6474-6499, 2014.

AUTHORS PROFILE

Rohit Bohra, is an aspiring Machine Learning Engineer who like solving real world problme using data. He has done two internships in field of machine learning and currently a freelancer, has few certifications in Data Analytics, Data Science and Machine Learning. He has participated in few hackathons.

Pankaj Karki, is an inspiring Data Scientist who like playing with data. He has done an internship in Data Analytics. A free lancer who likes solving data driven problems. He has various certificates in field of Data science and Machine Learning to his credit and has participated in various hackathons.

Dr. Kumudavalli M.V, is working as an Associate Professor at Dayananda Sagar College of Arts, Science & Commerce, Bangalore, India. Her research interests are Bioinformatics, Data Science, Operations Research, Networks etc. She has 14 years of academic experience. She has many research papers and certifications to her credits.