Performance Comparision of Machine Learning Algorithms for Tissue microarray (TMA)

Abstract — In this paper Compare the performance of two classification algorithm. It is useful to differentiate algorithms based on computational performance rather than classification accuracy alone. As although classification accuracy between the algorithms is similar, computational performance can differ significantly and it can affect to the final results. So the objective of this paper is to perform a comparative analysis of two machine learning algorithms namely, K Nearest neighbor, classification and Logistic Regression. In this paper it was considered a large dataset of 7981 data points and 112 features. Then the performance of the above mentioned machine learning algorithms are examined. In this paper the processing time and accuracy of the different machine learning techniques are being estimated by considering the collected data set, over a 60% for train and remaining 40% for testing. The paper is organized as follows. In Section I, introduction and background analysis of the research is included and in section II, problem statement. In Section III, our application and data analyze Process, the testing environment, and the Methodology of our analysis are being described briefly. Section IV comprises the results of two algorithms. Finally, the paper concludes with a discussion of future directions for research by eliminating the problems existing with the current research methodology.

I. INTRODUCTION

Tissue microarray (TMA) is a recent innovation in the field of pathology. A TMA contains many small representative tissue samples from hundreds of different cases assembled on a single histologic slide, and therefore allows high throughput analysis of multiple specimens at the same time. The classifier identifies epithelial and stromal regions from images in large patient cohorts, allowing of quantification of the interaction between cancer cells and normal cells.

Machine learning is a field of artificial intelligence dealing with algorithms that improve performance over time with experience. Supervised learning algorithms for regression are trained on data with the correct value given along with each variable This allows the learner to build a model, based on the attributes that is best fit the correct value. By giving more data to the algorithm the model can be improved.

Learning can be described in this way as improving performance. The measure of performance is how well the algorithm predicts the regression value given a set of variables or attributes. Machine learning algorithms provide excellent solutions for building models that generalize well given large amounts of data with many attributes by discovering patterns and trends in the data. Machine learning algorithms are a natural solution for sifting through these large datasets and determining the important pieces of information for prediction.

Machine learning algorithms may also provide a fast and efficient method to predict data, which will be often more useful to applications than static.

In this study, a performance analysis of a wide range of machine learning algorithms using real-world data for predict is performed.

II. PROBLEM STATEMENT

Machine learning algorithms are an advanced and efficient solution for determining the accurate models to predict patient survival. But the most suitable machine learning algorithm with maximum performance have to be decided, as our intended purpose is to increase the accuracy. Since in predict patient survival if the processing time of algorithm is high, the can become inaccurate. This problem can be overcome by implementing these kind of research works.

Prediction estimation has garnered a good deal of interest from both academia and industry, with numerous systems being proposed using a variety of technologies. The studies have shown that a number of different algorithms are able to achieve high classification accuracy. The effect of using different sets of statistical features on the same dataset has seen little investigation.

Further, these algorithms are limited by the size of the dataset since a large dataset will require a substantial amount of time to detect pattern, hindering real-world deployment. Systems that build signal propagation maps for a building have achieved similar accuracy.

III. METHODOLOGY

In this paper there are several technical tools were used to implements machine learning algorithm. There are scikit-learn and pandas. Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. This package focuses on bringing machine learning to non-specialists using a general-purpose high-level language. Emphasis is put on ease of use, performance, documentation, and API consistency. Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. pandas is a Numfocus sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project.

A. Working Preinciple of Two Algorithms

1) K–Nearest Neighbors Classifiction

In pattern recognition, the *k*-nearest neighbors algorithm (*k*-NN) is parametric method use for classification and regression In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression.

- In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, *k* is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point.

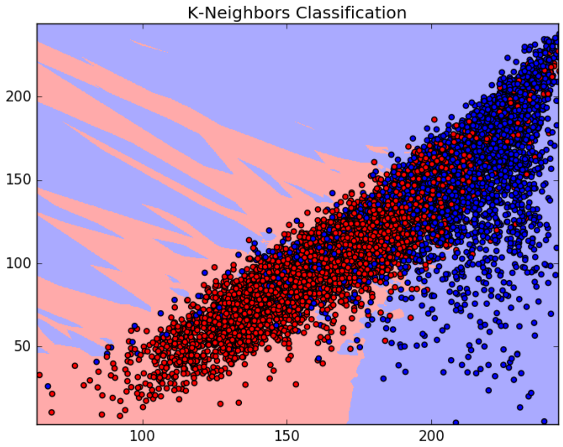

Fig.01. K-Nearest neighbor Classify train data set.

2 Logistic Regression

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. Cases where the dependent variable has more than two outcome categories may be analyzed in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression. In the terminology of economics, logistic regression is an example of a qualitative response discrete choice model.

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression.

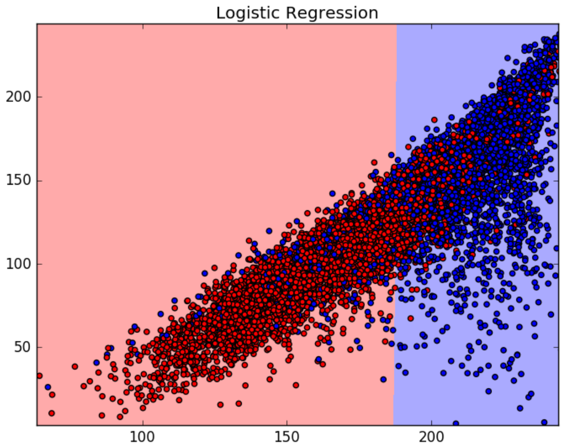

Fig.02. Logistic Regression classify train data set.

B. Analysis

The tsv file was converted to ".csv" file format and was used that “.csv” file as an input to the python programme by using the “pandas” library. Then the all the sensor values were stored as variable arrays by using “numpy” library. The machine learning algorithms were applied to those stored data sets and the data acquisition time of each and every algorithm was obtained. The data set was divided into several sets and obtained the execution time and a graph was obtained for the easiness of comparison.

IV. RESULTS

For performing comparative analysis, this paper principally focuses on the time taken to form classification and accuracy of both algorithms.

The performance evaluation results as shown in the bellow table. In this evaluation same size of train and test data were used for both model. Also this execution time was calculated by only considering model training period.

TABLE 1: THE TABLE OF EXECUTION TIME FOR DIFFERENT ALGORITHMS

Name of the Algorithm	**Execution Time of the Algorithm (s)**	**Accuracy**
K-Nearest neighbor	0.068	0.580
Logistic Regression	0.288	1.000

V. CONCLUSION

Much of this existing research focuses on the achievable accuracy of different machine learning algorithms. The accuracy is the most important thing when compare machine learning algorithms. The accuracy will change according to input data set. In this paper we have used two kind different algorithms. Logistic regression has linear classification functionality but K-Nearest neighbor algorithm is the one of non-linear pattern detection algorithm.

When we consider about execute time for both algorithm it is clear that K-Nearest neighbor has more processing power to analyze the given data set than Logistic regression. Therefore K-Nearest neighbor was best in analyzing large number of data than Logistic regression.

But when we create machine learning model high priority should be added to the accuracy of the particular model. By refer the table 01 we can see that the high accuracy algorithm was Logistic regression with 100% accuracy. It is clear that Logistic regression algorithm has 100% prediction capability than K-Nearest neighbor.

In this paper we are going to finalize performance of two different algorithm as we discuss earlier. Therefore the accuracy is the most important feature to select best algorithm Logistic regression is the best algorithm to create prediction model for Tissue microarray (TMA) data set.