International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Comparative Study of SVM and KNN for Tumor Prediction

J. Sivapriya, Nishanth Prem, G. Venkatesh Prasad, Balasubramanian.C.L

Abstract: ***Classification algorithms have played a vital role in the field of machine learning and data science. They cannot be downplayed. There are several variants of classification algorithms. In this paper, we compare KNN (K- nearest neighbors) and SVM (Support Vector Machine) algorithms. The attributes of both the algorithms are conferred. The benefits and drawbacks of each of these algorithms are assessed and finally arrive at a conclusion on which one has higher efficiency. We shall examine the efficiency of each algorithm based on their learning curve, comparing their accuracy on tumor prediction.***

Index Terms: ***Big Data, Healthcare, SVM Algorithm, Tumor detection.***

I. INTRODUCTION

Statistics have shown that tumors are the cause of cancer related deaths in teens and males in the age group of 24-40. Tumors are also one of the major factors that causes cancer in females of the similar age group. These facts increase the significance of the researches on the tumor identification and detection and this will present a chance for doctors to help save lives by identifying the disease prematurely and take necessary actions. Various image processing techniques are available at our disposal to be applied on various imaging processes for tumor detection that will diagnose certain features of the tumors such as the structure, outline, calcification and texture. These features help make the detection procedure more accurate and easier as there are some standard attribute of each features for a specific tumor. Tumors start off small and expand with time. As they expand, they will become more visible and it increases the probability of showing their characteristic attributes. A person with tumor usually has certain symptoms and this discomfort causes him to consult a doctor. From this, the doctor will be able to detect the smallest possible symptomatic malignant tumor that is in earliest stage possible and the other benign tumors in the screening process. There are many elements that can impact the appearance of tumors in different kind of processed images despite some quotidian features of malignancies because of variation in the type of tissue and tumor. For slightly bigger tumors, characteristic features can be easily found, whereas in smaller tumors these features are hard to detect many and some of them might exhibit themselves by other side effects such as distortion in its structure. The doctor can order several screening tests, including computed tomography scan (CT) or magnetic resonance imaging scan (MRI) , to locate the precise location of the tumor and show its size.

Manuscript published on 30 April 2019.
* Correspondence Author (s)
J. Sivapriya*, Assistant Professor (O.G.) at Department of Computer Science Engineering, SRM Institute of Science and Technology, Chennai
Nishanth Prem, B.Tech Computer Science Student at SRM Institute of Science and Technology, Chennai.
G. Venkatesh Prasad, B.Tech Computer Science Student at SRM Institute of Science and Technology, Chennai
Balasubramanian.C.L, B.Tech Computer Science Student at SRM Institute of Science and Technology, Chennai

II. CLASSIFICATION

In machine learning and data science, a mainstream technique is used, by which one can identify the category or set of categories to which the new data belongs i.e. it can be categorized. The basis of a training set of data containing observations (or instances) have a known category membership. This technique is called classification. Examples are used to assign a given email instances to either non spam or spam category, and assigning an interpretation to the given patient based on observed characteristics of the patient. Classification may be visualized as an example of pattern recognition. In machine learning nomenclature, supervised learning, where data with the output is first fed to the machine, which can help train the machine to predict future outcomes. Unsupervised learning involves grouping data into categories based on some similarity or distance. It is known as clustering. In machine learning, there is a supervised learning technique that involves the use of information it learnt to classify new data on its own, after it learns from the data input given to it. This is called classification. This data set obtained may be multi class or it may simply be bi class (like classifying whether a person is male or female or that the mail is spam or non-spam). Some examples of classification problems include speech recognition, Handwriting recognition, Bio Metric ID and document categorization etc.

A. K-Nearest Neighbor (KNN):

The k-nearest neighbors (KNN) algorithm is a supervised machine learning algorithm which is simple, easy-to-implement. Both classification and regression problems can be solved with the help of KNN. A non-parametric approach is used in k-nearest neighbor algorithm (KNN). KNN is generally used for both for classification and regression models. In both classification and regression model it consists of k as the input, which is the number of closest neighbour's in a training space. The output of the KNN algorithm is based on the type of learning model, whether it is classification or regression.

The class membership in KNN classification algorithm is the output. The object is classified based on the plurality vote, the most common object in its k-nearest neighbour is assigned to the class. The value of k is a small positive integer. The class of the single nearest neighbour is allotted when the value of k =1 in KNN algorithm.

For KNN regression, property value for the object is the output. This value is made out to be the average of the values of its *k* nearest neighbours. KNN algorithm is shown in Fig 1. KNN algorithm may be referred as instance-based learning or lazy learning algorithm, where the computation is set aside until classification and the function is used only locally. KNN is one of the simplest and easy to use algorithm in machine learning.

The nearer neighbours contribute more to the average and the distant ones contribute less to the average. So, there should a useful method required to assign weight for regression and classification for the contribution of the neighbours.

For example, a weight of 1/d is given to all neighbours in a common weighting scheme, distance to the neighbour is measured as d.

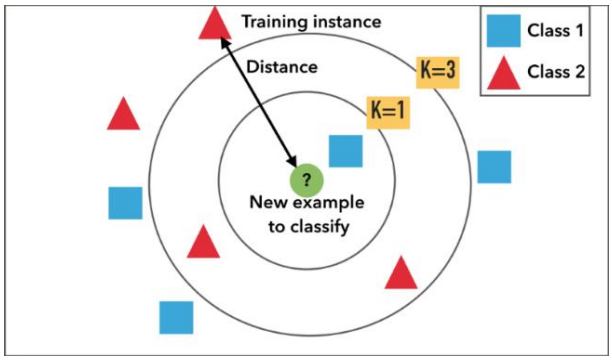

Fig 1 KNN Algorithm

B. Support Vector Machine (SVM)

SVM stands for Support Vector Machine. SVM is a supervised type of machine learning model. SVM can be used for both classification and regression problems.
SVM is used mostly for classification of data. Data points are plotted on a n-dimensional space in this technique. Then, data is classified by finding the hyper-plane that differentiate the two classes very well (refer Fig 2).

In machine learning, supervised learning models of SVM and its associated learning algorithms that are used to analyze the data, used for classification and regression solutions as well as for analysis and then tend to learn on their own. Given a set of training data, with each one segregated as belonging to either of the two categories, an SVM training algorithm then builds a model that assigns new examples to one of the categories, which generalizes it as a non-probabilistic binary linear classifier An SVM model is a depiction of data points in space, which are mapped in such a way that the form a huge split as wide as possible as shown in figure 2. New data points may then be mapped into that same space and can be used to predict which category they belong to, based on the side of the split they fall.

In addition to linear classification, SVM uses kernel trick to perform non-linear classification efficiently, using the high-dimensional feature spaces into which their inputs are mapped.

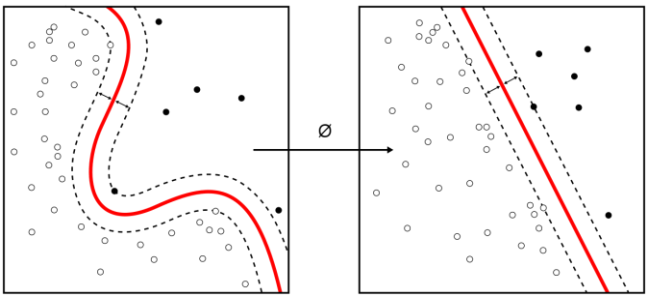

Fig 2 SVM Algorithm

III. SYSTEM ARCHITECTURE

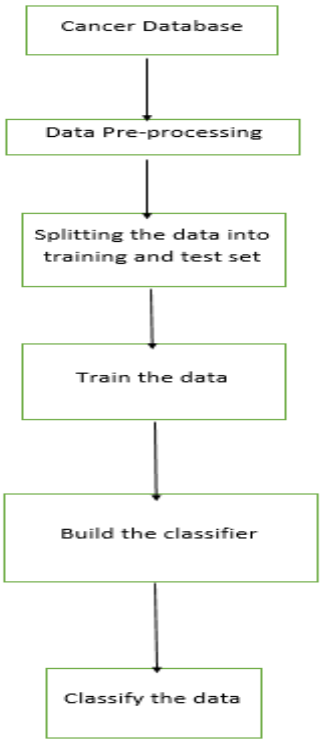

Fig 3 SVM System Architecture

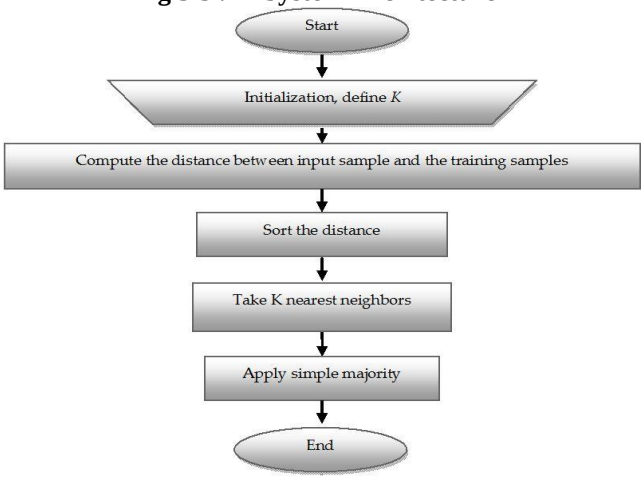

Fig 4 KNN System Architecture

IV. COMPARATIVE STUDY OF SVM AND KNN ALGORITHMS

The data in the KNN algorithm is generally classified based on length and distance metric, while a proper training phase is required by the SVM algorithm.

Typically, A multi-class classifier is used in KNN algorithm, while the standard SVM algorithm separates the data belonging to either one of the classes. The multi-class SVM uses one-vs-all and one-vs-one approach.
In one-vs-one approach, A unknown pattern is fed to the entity and then, the final decision of SVM is based on the majority result out of all the results of the data. This is generally used in multi-class classification.

In one-vs-all approach, number of SVMs to be trained should be the same as the available classes of unlabeled data. Although SVM appears to be more intensive on the computation part, once the training is finished, classes can be predicted when new and unlabeled data is encountered. On the counter-part, in KNN, whenever new unlabeled data is encountered, the distance metric is calculated. So, in KNN, the distance metric needs to be defined. In SVM, classes may be separated in either a linear way or a non-linear way.

Therefore, in KNN, the distance metric suitable for classification needs to be selected after the K parameter is set while the regularization factor is set for the SVM algorithm.

V. DATASET

The dataset is obtained from Kaggle dataset. The dataset is Breast cancer(diagnostic) data set. The dataset contains following attributes. The “diagnosis” attribute is dependent variable to determine malignant or benign and all other are independent variable. The sample dataset is represented in (table-1) and result of the accuracy of KNN and SVM is represented in table (2).

1. id- ID number
2. radius_stder-It is the measurement of standard error of mean of distances from center.
3. texture_stder-It is the measurement of standard error for standard deviation.
4. perimeter_stder- standard error for the perimeter
5. area_stder- standard error for the area
6. smoothness_stder-It is the measurement of standard error in radius lengths.
7.compactness_stder-It is the measurement of standard error for the value of (perimeter^2 / area - 1.0).
8.diagnosis-It classifies the breast tissue as either M=malignant or B=benign

VI. IMPLEMENTATION AND RESULT

A. Implementation

Here we have used python as our computer language. Here dataset contains breast cancer data from different women. The results are mainly shown in confusion matrix which is useful for comparison of classifiers performance. Results of KNN and SVM (in table2) is explains in result section. Performance analysis is conduct under accuracy of result, search time for data set, memory used to process.

Table1- Sample Data (Breast Cancer Kaggle Dataset)

id	radius_stder	texture_stder	perimeter_stder	area_stder	smoothness_stder	compactness_stder	diagnosis
842302	1.095	0.9053	8.589	153.4	0.006399	0.04904	M
842517	0.5435	0.7339	3.398	74.08	0.005225	0.01308	M
84300903	0.7456	0.7869	4.585	94.03	0.00615	0.04006	M
84348301	0.4956	1.156	3.445	27.23	0.00911	0.07458	B
84358402	0.7572	0.7813	5.438	94.44	0.01149	0.02461	M
843786	0.3345	0.8902	2.217	27.19	0.00751	0.03345	M
844359	0.4467	0.7732	3.18	53.91	0.004314	0.01382	B
84458202	0.5835	1.377	3.856	50.96	0.008805	0.03029	M
844981	0.3063	1.002	2.406	24.32	0.005731	0.03502	M
84501001	0.2976	1.599	2.039	23.94	0.007149	0.07217	M
845636	0.3795	1.187	2.466	40.51	0.004029	0.009269	M
84610002	0.5058	0.9849	3.564	54.16	0.005771	0.04061	M
846226	0.9555	3.568	11.07	116.2	0.003139	0.08297	M
846381	0.4033	1.078	2.903	36.58	0.009769	0.03126	B
84667401	0.2121	1.169	2.061	19.21	0.006429	0.05936	M
84799002	0.37	1.033	2.879	32.55	0.005607	0.0424	M
848406	0.4727	1.24	3.195	45.4	0.005718	0.01162	M
84862001	0.5692	1.073	3.854	54.18	0.007026	0.02501	B
849014	0.7582	1.017	5.865	112.4	0.006494	0.01893	M

B. Result

Table2- Result (Accuracy Rate of SVM and KNN)

Algorithm	Training Data(no)	Test Data(no)	Accuracy Measurement rate (in %)
SVM	369	200	98.9
KNN	369	200	96.47

Thus, from the above table, it is observed that the accuracy rate of SVM classifier is much higher than that of the KNN classifier. SVM classifier has an accuracy rate of 98.9% whereas the KNN classifier has an accuracy rate of 96.4%. This doesn’t imply that the KNN classifier is inefficient. Classifiers are dependent on the dataset. For the given dataset, SVM classifier is the most efficient one.

VII. CONCLUSION

The results and observations show that the SVM algorithm is more reliable than KNN. Though, it remains that the SVM algorithm is more computationally intensive. The multi-class data classification should be done using KNN as it is the easier one to implement. The data which is encountered decides which algorithm to be used hat would present reliable detection in unpredictable situations. When the data points are heterogeneously distributed, both of them should function well. But when the data is homogeneous, non-linear classification of data through SVM should be preferred. In case f larger data set, KNN would be a bad choice as computation would just take too much time.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

VIII. FUTURE WORK

Future work can be suggested for highly accurate results with the help of topic modelling and hybrid models can be used with more effective structure so that results can further improved upon. In future various other genetic algorithms can be combined to make the pre-existing classifiers more efficient. With the accuracy and efficiency of SVM and the constantly evolving KNN algorithms, their combinations would provide various applications in the field of machine learning.

REFERENCES

1. www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.
2. wikipedia.org/wiki/Knearest_neighbors_algorithm H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
3. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/
4. https://www.researchgate.net/post/The_accuracy_of_k-NN_is_greater_than_SVM_What_would_be_the_main_reason.
5. https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn.

AUTHORS PROFILE

J.Sivapriya, Assistant Professor (O.G.) at Department of Computer Science Engineering, SRM Institute of Science and Technology, Chennai. Areas of interest include DBMS, Big Data and Machine Learning.

Nishanth Prem, 3rd year B.Tech Computer Science Student at SRM Institute of Science and Technology, Chennai. Areas of interest are Data Science, Python Programming and Machine Learning.

G. Venkatesh Prasad, 3rd year B.Tech Computer Science Student at SRM Institute of Science and Technology, Chennai. Areas of interest are Big Data, Business Analytics and Financial Analytics.

Balasubramanian.C.L, 3rd year B.Tech Computer Science Student at SRM Institute of Science and Technology, Chennai. Areas of interest are Machine Learning and Software Development.