EasyChair Preprint

№ 5704

Cancer Prediction Using Machine Learning

Sairaj Nanaware, Yogita Garje, Priyanka Salunke and
Vaishnavi Jamdar

June 4, 2021

Cancer Prediction Using Machine Learning

Sairaj Nanaware, Yogita Garje, Priyanka Salunke, Vaishnavi Jamdar
Department of Computer Engineering
Vishwakarma Institute of Technology, Bibwewadi, Pune, India.

sairaj.nanaware18@vit.edu
yogita.garje18@vit.edu
priyanka.salunke18@vit.edu
vaishnavi.jamdar18@vit.edu

ABSTRACT
Machine learning is an application of artificial intelligence (AI). Machine learning focuses on the development of computer programs that can access data. machine learning is frequently used in cancer diagnosis and detection. Among the better designed and validated studies, it is clear that machine learning methods can be used to improve the accuracy of predicting cancer susceptibility, recurrence and mortality.

Keywords
KNN algorithm, machine-learning, prognosis, Brest cancer, classification tool.

1. INTRODUCTION

The kNN algorithm is a non-parametric instance based algorithm that can be used for regression and classification. In k-nearest neighbors' algorithm (k-NN) is a non-parametric method used for classification and regression. In. It classifies a target object based on the number of members of the class nearest to it. 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. calculating the relative position between license plate and vehicle. Few people calculate the yaw angle of vehicle to the lane using an image; most of the researchers use a sensor.

scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

2. MOTIVATION

An algorithm by blending shapes and colors characteristic of lane is proposed and it enables quick and accurate identification of lane and calculate the vehicle

Usually we measure the distance of all the objects, used for classification, to the target object and assign to it the class more common within its k nearest neighbors. An example is shown in Figure 1 where the target object is shown as the black point with an interrogation mark. The example shows nine objects with two classes. The table in the figure shows the Euclidean distances of some neighbors to the target object sorted in ascending order. If we use 3 of the neighbors for classification the class assigned to the target object will be in this case"a".

3. SOLUTION AND IMPLEMENTATION

We implemented the kNN algorithm and run simulations using the breast cancer prognosis data. The original data has 289 records of breast cancer patients, but four of these records lack the number of lymph nodes and we excluded them in this study. The breast cancer has 2 classes are magninent and begnin. Algorithm decides whether the tumor is magninent or not. Each record has 32 features. The first one is the ID of the patient, the second one indicates with one letter the recurrence (R) or not recurrence of cancer (N).The KNN algorithm is as follows:

1.Load the data in instance space form.

2.Select K value where k is no of nearest neighbor included in majority voting process .Consider 2 cases

* If k lowest: noisy, less stable prediction

*If K high :forever to process ,more stable prediction with increasing errors

Select k value is very important task

*Hence select the value of k in by :

a)square root of n where n is total number of data instances

b)odd value of k to avoid confusion between two classes

3. For each instance in the data

a)Calculate the Euclidean distance with formula

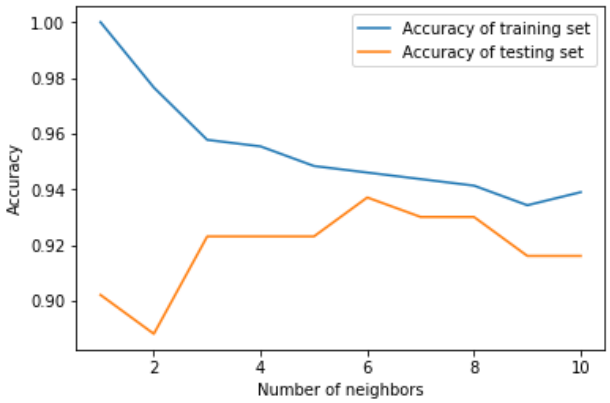
Distance(d)= $\sqrt{(x - a)^2 + (y - b)^2}$

b) Add the Euclidean distance to dataset

4.Pick the first K entries which have shortest Euclidean distance

5.Get the labels of the selected K entries

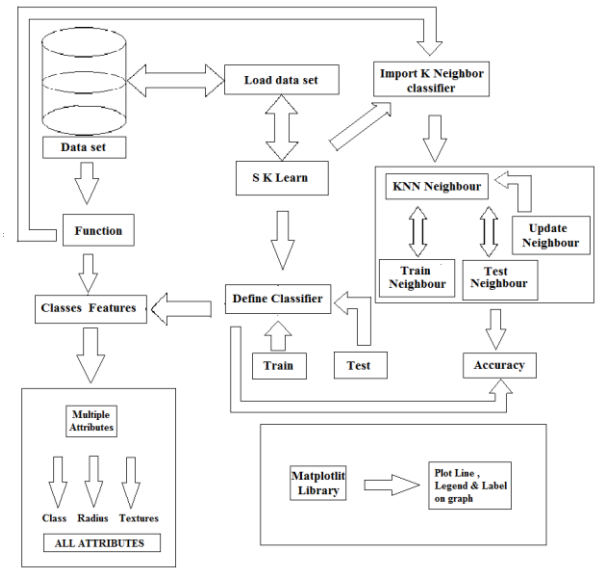
6.New instance belong to class with majority voting



This graph shows accuracy on shape and size of tumors. So algorithm uses different types of libraries such as numpy, scipy, matplotlib to help in prediction. Here K-nearest algorithm has 5 neighbor points and from that it decides class of new data from dataset.

Algorithm works very well and we got accuracy of 93%.

4 . ARCHITECTURE



5.APPLICATION

1. This KNN algorithm increases the accuracy of diagnosis such as breast cancer and other diseases.

2. The algorithm can be used to enhance the automated diagnosis which include diagnosis of multiple diseases showing similar symptoms.

Future Scope

1. This work can be extended to surveillance application and monitoring the patient health.

2. Extensive Research can be carried out on this application using this techniques.

3. In addition, a more detailed dataset can be developed which covers additional attribute related to patient medical history.

6.CONCLUSION

We have implemented and run several simulations with the kNN algorithm to evaluate its accuracy when using it for breast cancer recurrence prediction. The implementation of the algorithm was done in Python . The running time is of almost two hours in a 2.7 GHz PC for the 940,500 simulations run when generating average, maximum, and minimum values for nineteen settings of k as shown in figures 5 to 9. Our results show that the kNN algorithm is simple and powerful, but it also could give very poor results. Our implementation is the simplest one and uses all the 32 values of the record of a patient. We plan to study it when applying to it principal component analysis to determine the most important features in a record. We also plan to study the effect of using other distance definitions as a measure of similarity . There are a lot of variants for the kNN algorithm .We plan to study and evaluate them for cancer prognosis and expect to develop and suggest new approaches to improve it

7.REFERENCES

[1]http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/kakutei1/ dl/11_h7.pdf

[2] Arihito Endo, Takeo Shibata and Hiroshi Tanaka, “Comparison of Seven Algorithms to Predict Breast Cancer Survival,” Biomedical Soft Computing and Human Sciences, Vol. 13, No.2, pp. 11-16, 2008.

[3] Jini R. Marsilin and G. Wiselin Jiji, “An Efficient CBIR Approach for Diagnosing the Stages of Breast Cancer Using KNN Classifier,” Bonfring International

Journal of Advances in Image Processing, Vol. 2, No. 1, pp.1-5, March 2012

[4] Shomona G. Jacob and R. Geetha Ramani, "Efficient Classifier for Classification of Prognosis Breast Cancer Data Through Data Mining Techniques," Proceedings of the World Congress on Engineering and Computer Science 2012, Vol. I, October 2012.

[5] E. Fix and J. L. Hodges, "Discriminatory analysis, nonparametric discrimination: Consistency properties," Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[6] http://archive.ics.uci.edu/ml/datasets.html

[7] Seyyid A. Medjahed, Tamazouzt A. Saadi, and Abdelkader Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules," International Journal of Computer Applications, Vol. 62, No.1, pp.1-5, January 2013.

[8] Sung-Hyuk Cha, "Comprehensive Survey of Distance/Similarity Measures between Probability Density Functions," International Journal of Mathematical Models and Methods in Applied Sciences, Vol. 1, Issue 4, pp.300-307, 2007.

[9] Nitin Bhatia and Ashev Vandana, "Survey of Nearest Neighbor Techniques," International Journal of Computer Science and Information Security, Vol. 8, No. 2, pp.302-305, 2010.