Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Generalized K-Nearest Neighbour Algorithm- A Predicting Tool

**Minakshi Sharma***
*Lecturer*
*Department of Computer Science and IT*
*MIET, Jammu, India*

**Suresh Kumar Sharma**
*Research Scholar*
*Department of Computer Science and IT*
*University of Jammu, India*

***Abstract— k-nearest neighbour algorithm is a non-parametric machine learning algorithm generally used for classification. It is also known as instance based learning or lazy learning. K-NN algorithm can also be adapted for regression that is for estimating continuous variables. In this research paper the researcher endow with a generalized K-nearest algorithm used for predicting a continuous value. In order to better understand the proposed general method is used to forecast the value of maximum humidity. The experimental setup was developed using MATLAB. The interesting facts so obtained are presented in this research paper.***

***Keywords— KNN, Prediction, Mean Square Error, Humidity***

## I. INTRODUCTION

Trying to use machines to solve mathematical problems can be traced to the early 17th century. The computer remained allied with human activity until about the middle of the 20th century when it becomes applied to a programmable electronic device that can store retrieve and process data. One of the currently most active research areas within artificial intelligence is the field of machine learning, which involves the study and development of computational models of learning processes. A major goal of research in this field is to build computer programs or software capable of improving the performance with practice and of acquiring knowledge on their own. Compared with human learning, machine learning learns faster, the accumulation of knowledge is more facilitate the results of learning spread easier. So any progress of human in the field of machine learning, will enhance the capability of computers, thus have an impact on human society [1]

## II. REVIEW OF LITERATURE

K-Nearest Neighbour (KNN) classification divides data into a test set and a training set. For each row of the test set, the K nearest (in Euclidean distance) training set objects are found, and the classification is determined by majority vote with ties broken at random. If there are ties for the Kth nearest vector, all candidates are included in the vote [2]. In the paper "Time Series Prediction Using KNN Algorithms via Euclidian Distance Function: A Case of Foreign Exchange Rate Prediction" researchers discussed that International transactions are usually settled in the near future. Exchange rate forecasts are necessary to evaluate the foreign denominated cash flows involved in international transactions. Thus, exchange rate forecasting is very important to evaluate the benefits and risks attached to the international business environment. A forecast represents an expectation about a future value or values of a variable. In this research paper, Researchers use technique known as KNN which belongs to Machine learning subfield of Artificial intelligence. Here researchers develop an experimental setup to predict the value of USD in term of INR for the next day [3]. In another paper Air Quality Index Prediction using K-Nearest Neighbour Technique researchers said that one of the classical data mining techniques is k-nearest neighbour. This method uses the class of the k nearest neighbour to classify a new instance. The distance is calculated with one of the multiple mathematical distance metrics. In this paper, the technique is used in the air quality forecast domain in order to predict the value of the air quality index. This index is used to categorize the pollution level and to inform the population about some possible episodes of pollution [4]

The research paper "Real-Time Highway Traffic Accident Prediction Based on the k-nearest neighbour Method" researchers talked about the occurrence of a highway traffic accident is associated with the short-term turbulence of traffic flow. In this paper, we investigate how to identify the traffic accident potential by using the k-nearest neighbour method with real-time traffic data. This is the first time the k-nearest neighbour method is applied in real-time highway traffic accident prediction. Traffic accident precursors and their calculation time slice duration are determined before classifying traffic patterns. The experimental results show the k-nearest neighbour method outperforming the conventional C-means clustering method [5]

## III. ABOUT K-NEAREST-NEIGHBOUR

K-Nearest Neighbour (KNN) is a most simple machine learning method. The choice of KNN is motivated by its simplicity and flexibility to incorporate different data types. The main idea of KNN is to base estimation on a fixed number of observations, say k, which are closest to the desired output. It is considered as a lazy learning algorithm because it does not build a model or function previously, but yields the closest k records of the training data set that have the highest similarity to the test. KNN can be used both in discrete and continuous decision making known as classification and regression respectively. For classification select most frequent neighbour, and for regression calculate the average of k neighbour. KNN is a supervised learning algorithm i.e. a training set is given consisting of n pair ($x_i$ , $y_i$) and the problem is to estimate y(x) from a new input x. In order to apply this technique, it is necessary to have a training set and a test sample, to know the value of k (how many neighbours are used for prediction) and the mathematical formula of the distance calculated between the instances.

**DISTANCE USED IN KNN**

The three famous distance functions used with KNN are [6]

(i) Euclidean Diatance: $\sqrt{\sum_{i=1}^{k}(x_i -y_i)^2}$

(ii) Manhattan Distance: $\sum_{i=1}^{k} | xi - yi |$

(iii) Minkowski Distance: $( \sqrt{\sum_{i=1}^{k}(x_i -y_i)^q})^{1/q}$

**FEATURES OF KNN**

The various important features of KNN are:
(i) It is non parametric in nature.
(ii) It is simple to implement.
(iii) It is considered as a lazy learning.
(iv) It is robust with small error ratio.
(v)It is an instance based learning
(vi) It is a local learner
(vii) It has uniform feature weighting.
(viii) It can work with relatively little information.

**CHOICE OF K IN KNN**

In order to avoid ties k is preferably odd. If the value of k is smaller then there is higher variance. On the other hand if the value of k is larger, then there is higher bias. Thus the proper choice of k depends on the data.

**PROBLEMS ASSOCIATED WITH KNN**

The standard KNN method suffer from the curse of dimensionality i.e. The neighbourhood of a given point become very sparse in a high dimensional space, resulting in high variance. Thus in high dimensional 'nearest' become meaningless.

Another problem is over-fitting. It occurs when a learning algorithm performs too good on the training set, compared to its true performance on unseen test data. The most popular method to overcome over-fitting is known as hold-set method. According to hold-set method:
(i) Randomly choose 30% of the training data set, set it aside.
(ii) Train a classifier with the remaining 70% training data.
(iii) Test the classifier's accuracy on the 30%

**ALGORITHM**

```
for all test example x do
     for all training example ($x_i$,$y_i$), do
     compute distance(x,$x_i$ );
     end for
select the k-nearest neighbour of x;
return the average output value among neighbours i.e. $1/k \sum_{i=1}^{k} y_i$ ;
end for;
```

## IV. BENEFITS OF PREDICTING MAXIMUM HUMIDITY

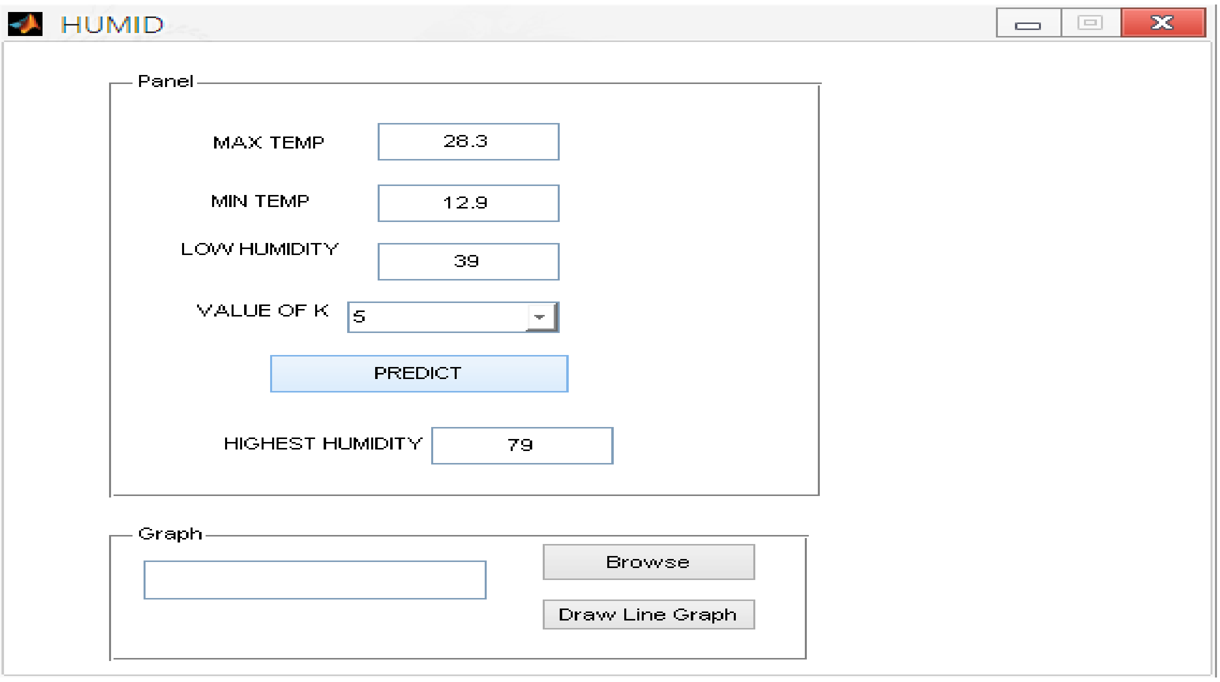
Humidity is the amount of water vapours in air. Humidity affects animals and plants, Human comfort, electronic devices and building constructions. A little knowledge of amount of humidity can prevent the aforesaid adverse effects for instances; moisture may increase the conductivity of permeable insulator of electronic devices which leads to malfunctioning. Climate change and agriculture are interrelated process, both of which take place on a global scale. Humidity is one of the main parameter that effect agriculture. In addition to above, Humidity also affect environment and also used for rain prediction. Humidity is an important metric used in forecasting weather. Humidity indicated the livelihood of prediction, dew/fog which in turn affects the growth of some crops.

**DATA SET**

We use three main parameters that affect highest humidity. These include minimum temperature, maximum temperature and lower humidity. We use previous day value of these parameters in order to predict the highest humidity of next day. This data is obtained from the weather section of the local newspaper. The KNN algorithm is trained using previous 10 days data and next 7 days data are used for testing purpose.

## V. EXPERIMENTAL RESULTS

The following two figures show the experimental result. In figure (a) the running software is predicting the value of highest Humidity. In figure (b) there is a graph between the actual and predicted value of highest humidity. It is found that at the value k = 5, the prediction is very nearer to the actual result as shown in figure (a) and figure(b).



HUMID

Panel

MAX TEMP 28.3

MIN TEMP 12.9

LOW HUMIDITY 39

VALUE OF K 5

PREDICT

HIGHEST HUMIDITY 79

Graph

Browse

Draw Line Graph

**Figure (a)Predicting highest humidity**

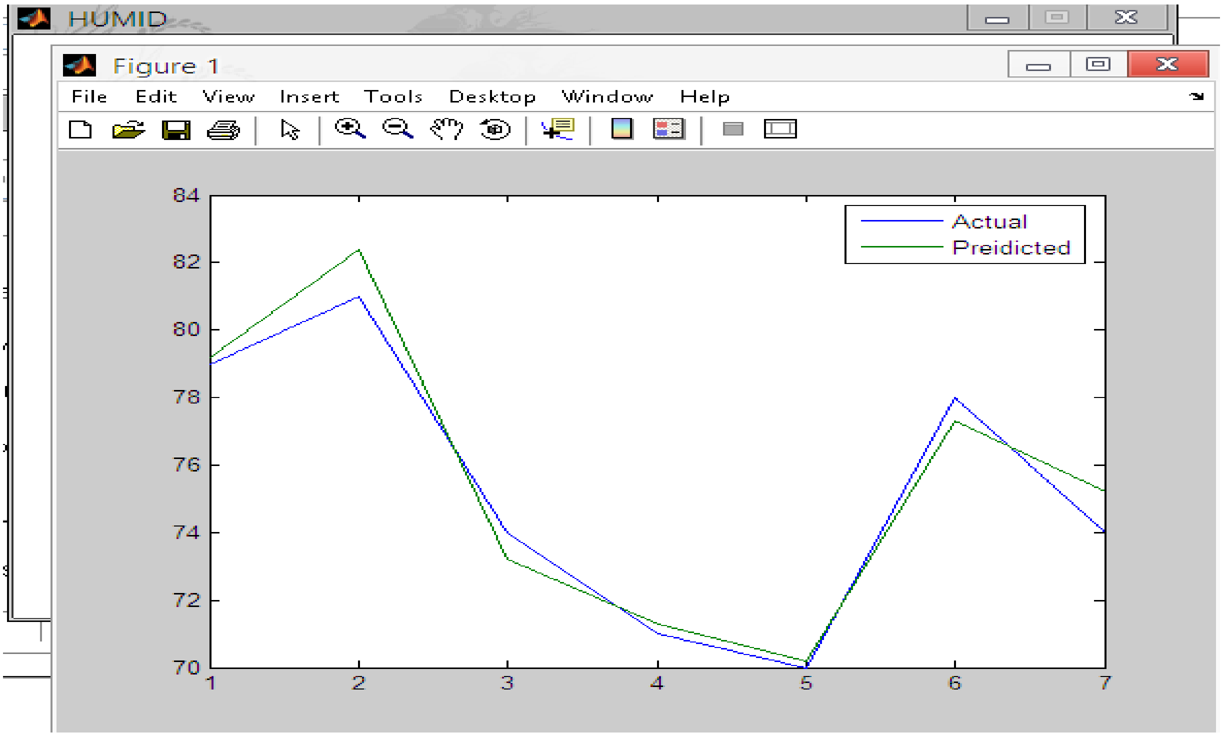


**Figure (b) Graph showing actual and predicted value**

## VI. CONCLUSION

The Mean Square Error (MSE) is the arithmetic mean of the sum of the square of prediction error. This error measure is popular and corrects the cancelling out effects.

$ei \ = |\ (fi - yi\ )|$

$f_i$ : Prediction

$y_i$ : True Value

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(ei)^{2}$$

The MSE index ranges from 0 to ∞, with 0 corresponding to the ideal. Lower MSE is better.
In this case MSE is 0.671429.

## REFERENCES

[1] Wee Keong Ng., Advances in Knowledge Discovery and Data Mining, *Advances in Knowledge Discovery and Data Mining: 10th Pacific-Asia Conference,* PAKDD 2006, Singapore, April 9-12, 2006.

[2] Bhavani Thuraisingham, Latifur Khan, Mamoun Awad, Lei, *Design and Implementation of Data Mining* Tools, 1st ed., CRC Press, 2010

[3] Suresh Kumar Sharma, Vinod Sharma, *Time Series Prediction Using Knn Algorithms Via Euclidian Distance Function: A Case Of Foreign Exchange Rate Prediction*, Asian Journal Of Computer Science & Information Technology,2012.

[4] Elia Georgiana Dragomir, *Air Quality Index Prediction using K-Nearest Neighbor Technique*, Petroleum-Gas University of Ploiesti, Informatics Department, Ploieşti, Romania,2010

**[5**] Yisheng Lv, Shuming Tang ; Hongxia Zhao ,*Real-Time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method* , International Conference on Measuring Technology and Mechatronics Automation, 2009.

[6] Weinberger, K.Q., Blitzer, J., Sau l, L. K, *Distance metric learning for large margin nearest neighbor classification* , NIPS, 2005