Combining Akaike's Information Criterion (AIC) and the Golden-Section Search Technique to find Optimal Numbers of K-Nearest Neighbors

Asha Gowda Karegowda
Dept. of Master of Computer Applications
Siddaganga Institute of Technology
Tumkur, Karnataka,India
9844327268

M.A.Jayaram
Dept. of Master of Computer Applications
Siddaganga Institute of Technology
Tumkur, Karnataka,India
9845843534

A.S. Manjunath
Dept. of Computer Science & Engineering
Siddaganga Institute of Technology
Tumkur, Karnataka, India
9845141040

ABSTRACT
K-nearest neighbor (KNN) is one of the accepted classification tool . Classfication is one of the foremost machine-learning tools used in field of medical data mining. However, one of the most complicated tasks in developing a KNN is determining the optimal number of nearest neighbors, which is usually obtained by repeated experiments for different values of K, till the minimum error rate is achieved. This paper describes the novel approach of finding optimal number of nearest neighbors for KNN classifier by combining Akaike's information criterion (AIC) and the golden-section search technique. The optimal model so developed was used for categorization of a variety of medical data garnered from UC Irvine Machine Learning Repository.

Keywords
Medical Data mining, K-nearest neighbor (KNN), Akaike's information criterion (AIC) and Golden-selection Ratio.

1. INTRODUCTION
With the computerization in hospitals, a massive amount of data is collected. Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified. The use of machine learning tools in medical diagnosis is increasing gradually. Medical data mining has great potential for exploring hidden patterns in the data sets of medical domain. These patterns can be used for clinical diagnosis. Classfication is one the major machine learning tools for medical data mining. It has been used to detect lung cancer, breast cancer, to predict survival of kidney dialysis patients, analyze blood and urine samples, track glucose levels in diabetic, classify diabetic retinopathy and many other medical applications.

This paper investigates the combination of the Akaike's information criterion (AIC) with the golden-section optimization technique to find the optimal number of nearest neighbors for KNN classifier. KNN is one of the most popular and simple classfication tools. Selection of relevant attributes is done using WEKA Genetic algorithm (GA) and Correlation based feature selection (CFS) in cascaded fashion. Section 2 gives a brief introduction of Akaike's information criterion (AIC) and the Golden-selection Ratio. The algorithm which combines the Akaike's information criterion (AIC) with the golden-section optimization technique to find the optimal number of nearest neighbors for KNN classifier has been discussed in section 3. For the sake of entirety KNN has been discussed in section 4. Section 5 gives the facts of data preprocessing and the medical data sets used. Section 6 explains the experimental results and concluding remarks are covered in section 7.

2. AKAIKE'S INFORMATION CRITERION (AIC) AND GOLDEN-SELECTION RATIO

Ren and Zaho[8] derived equations for the selection of the optimal neural network models based on Akaike's information criterion (AIC)[1,2]. The criterion formula for neural network is

$$AIC = n \log(\sigma\hat{}^2) + 2(m+1) \quad (1)$$

$$AICc = n\log(\sigma\hat{}^2) + 2(m+1)(n/(n-m-2)) \quad (2)$$

where n is the number of training data; $\sigma\hat{}^2 = \Sigma\sigma^2/n$ is a mean squared error (MSE) between the target output and actual output; X= m+1 is the number of total parameters, which is equal to the number of parameters in the network model plus one for σ^2 and m is the number of weights and biases used in the neural network.

The AIC defined in equation 1 is for n/ (m+1) >= 40 (large number of training data) and AICc defined in equation 2 is for

the case of n/(m+1) < 40. The AIC consists of two terms. The first term of AIC depends on the MSE of a model. The second term depends on the number of parameters employed in the network model and is used to penalize the over fitting. For a given problem, the number of training data n is fixed, m is the number of weights and biases used in the neural network. The network that achieves the least AIC is the optimal choice.

The authors have attempted to use the Akaike's information criterion combined with golden-selection search to find the optimal number of nearest neighbors for KNN classifier. Equation 2 has been used to compute AIC , m is assumed to be optimal number of nearest neighbors for KNN classifier, while remaining parameters are same as mentioned above.

2.1 Golden-selection Ratio

The golden section search [5] is a technique for finding the extremum (minimum or maximum) of a unimodal function by successively narrowing the range of values inside which the extremum is known to exist. The technique derives its name from the fact that the algorithm maintains the function values for triples of points whose distances form a golden ratio.

3. PROPOSED ALGORITHM TO FIND OPTIMAL NUMBER OF K NEIGHBORS FOR KNN

The lower boundary for the number of nearest neighbors is defined as 1 and the upper boundary for the number of nearest neighbors is defined as large number say 99. AIC is used as a cost function to find the best optimal number of nearest neighbor. To minimize AIC, the golden–section method has been applied. Equation 2 is used to find the AICc of KNN, where authors assume n to be number of training data and m to be the number of nearest neighbors for KNN.

The procedure to find the optimal number of neighbors for k nearest neighbor classification is given as follows.

Step 1. Set the possible minimum number of nearest neighbor as $N_0 = 1$ and the possible maximum number of nearest neighbor as N_1(usually a larger value say 99).

Step 2: Choose the golden-section point

$$N_2 = \lfloor N_0 + .382(N_1 - N_0) \rfloor$$
$$N_3 = \lfloor N_0 + 0.618(N_1 - N_0) \rfloor$$

Step 3: Find the mean square error for KNN with N_2 number of nearest neighbor.

Step 4: Find the mean square error for KNN with N_3 number of nearest neighbor.

Step 5: Calculate the AIC value of the for the two KNN namely,

AIC (N_2) and AIC (N_3).

If ((N1 - N0) <=3) then go to *step* 6.

Else if AIC (N_2) <=AIC (N_3), then let $N_0 = N_0$, $N_1 = N_3$, go to *Step 2.*

Else if AIC (N_2) > AIC (N_3), then let $N_0 = N_2$, $N_1 = N_1$, go to *Step 2.*

In step 5, if AIC (N_2) < AIC (N_3), the minimum AIC point is between N_0 and N_3,

hence $N_0 = N_0$, $N_1 = N_3$, otherwise the minimum AIC point is between N_2 and N_1, hence $N_0 = N_2$, $N_1 = N_1$.

Step 6: K = N0

For(I ← N_0 + 1 ; I<=N_1; I←I+1) /* find K with min AIC*/

If(AICc(I) < AICc(K)

K ← 1

KNN is very sensitive to number of nearest neighbors selected, sometimes with even number of nearest neighbor it results in tie for binary classfication problem. Hence instead of finding one optimal value of nearest value of number of nearest neighbors, the terminating condition ((N_1 - N_0) <=3) is used. Once the terminating condition is reached we compute the AIC for the values between N_1 and N_0 (both inclusive). Finally the optimal number of nearest neighbor is the one with least AIC.

(a) If ((N_1 - N_0) =1) then according to step 2 N_2 and N_3 will have value N_0 as shown below.

$$N_2 = \lfloor N_0 + .382(1) \rfloor = N_0$$

$$N_3 = \lfloor N_0 + 0.618(1 \rfloor = N_0$$

Since $N_2 = N_3$, AICc(N_2) = AICc(N_3), hence N1 = N3 =N0, irrespective of AICc of N1.

(b) If ((N_1 - N_0) =3) then according to step 2 N_2 and N_3 will have same value N_0 +1 as shown below.

$$N_2 = \lfloor N_0 + .382(3) \rfloor = \lfloor N_0 + 1.146 \rfloor = N_0 + 1$$

$$N_3 = \lfloor N_0 + 0.618(3) \rfloor = \lfloor N_0 + 1.854 \rfloor = N_0 + 1$$

Since $N_2 = N_3$, AICc(N_2) = AICc(N_3), hence N1 = N3 ,irrespective of AICc of N1.

Hence the terminating condition (If ((N_1 - N_0) <=3) is used .If the condition is true then in step 6 we compute AIC for values between N_0 and N_1 (both inclusive), and select the one with least AIC for determine the optimal number of nearest neighbor for KNN classifier.

4. K-NEAREST NEIGHBOR ALGORITHM

KNN classifier is one of the most simple classfication methods, which employs the strategy of lazy learners. When there is little or no prior information about the structure of the training set, KNN is preferred over other classification tools since the former is a nonparametric classfication technique.

When a test sample is to be categorized, KNN finds the K training samples that are nearest to the given test sample. One of common measure of closeness is defined using Euclidean distance. The Euclidean distance between two tuples say $A_1 = (a_{11}, a_{12}, a_{13}, \ldots a_{1n})$ and $A_2 = (a_{21}, a_{22}, a_{23}, \ldots a_{2n})$ is given by

$$d(A1_1, A_2) = \sqrt{\sum_{i=1}^{n} (A_{1i} - A_{2i})^2}$$

The test sample is assigned to the most common class among its K nearest neighbors. The optimal value of K is usually determined experimentally starting with K = 1, and then repeating each time by incrementing K by a constant till that K value gives the minimum error rate [5]. By using the proposed approach in section 3, the optimal value of K can be sought which ensures the minimum AICc.

5. DATA PREPROCESSING AND THE DATASET USED

Data preprocessing: Real world data tend to be noisy, incomplete and inconsistent data. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data cleaning, data integration, data transformation and data reduction are some of the data preprocessing steps prior to data mining. One of the data cleaning method involves handling the missing values. Tuples with missing values have been eliminated. Data reduction obtains a reduced representation of the data set. One of the strategies for data reduction is attribute subset selection. Mining on the reduced set of attributes reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand, reduces the time to build the model and finally enhances classification accuracy [5]. For supervised learning the feature selection can be classified into two types, filter methods and wrapper methods. Filter method assesses the relevance of the attributes based on data's intrinsic properties. Filter methods are independent of learning algorithm; hence once the significant features are identified by the filter can be provided as input to different learning algorithm. Wrapper method in supervised learning uses the method of classification itself to measure the importance of feature set, hence the features selected depends on the classifier model used i.e. the feature subset search algorithm is wrapped around the learning model [3]. Relevant attributes have been identified by applying a filter approach by cascading Genetic algorithm(GA)[4] with with Correlation based feature selection(CFS). The weka tool GA rendered global search of attributes with fitness evaluation effected by CFS. The feature selected by GA-CFS when given as input to ANN, improved classification accuracy of ANN [3]. Since filter approach is independent of the classifier, we have used the relevant attributes rendered by GA_CFS filter for KNN classifier. The proposed algorithm has been applied to following four medical dataset garnered from UC Irvine Machine Learning Repository [7]. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. All the four dataset are split randomly into a training set and test set using 70-30 ratio.

Pima Indian Diabetes Database (PIDD) [7] includes the nine attributes (1-8 attributes as input and last attribute as target variable). A total of 768 records are available in PIDD. All patients were females at least 21 years old of Pima Indian heritage. After deleting the records with missing values, there were 392 cases with 130 tested positive cases and 262-tested negative. GA-CFS resulted in the four significant attributes. Wisconsin Breast Cancer dataset [7] has 688 samples each with 7 input attribute and one target attribute. This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. Each sample is either benign or malignant class. The data set contains 16 samples with missing values. After eliminating such samples the remaining 683 samples with 339 malignant and 444 benign samples. GA_CFS produced the same set of attributes, i.e no attributed was eliminated.

The Dermatology database [7] contains 34 attributes and 366 samples. The diseases in this group are of six types: psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. There was no missing data. GA-CFS filter resulted in the 20 significant attributes.

The Heart statlog dataset [7] contains 13 attributes and 270 samples. There were no missing values. GA-CFS filter resulted in the 7 significant attributes. Class attribute represents the absence or presence of heart disease.

6. RESULTS AND DISCUSSION

The preprocessed data was divided into training data set and test data set on 70:30 ratio. For the all the medical dataset, the initial value for minimum number of nearest neighbor was assigned to be 1 and maximum number of nearest neighbor was assigned as 99.

Table 1 shows different values of N0, N1, N2 and N3 obtained during the successive iterations of the proposed algorithm for PIDD diabetic dataset. Table 1 clearly shows that during initial 5 iterations of algorithm the maximum value N1 gradually reduces from 99 to 14 and in further interactions finally to 12. N0 , mimimum value increases to 5, 8 and finally 10. For N0 = 10 and N1 = 12, it was noticed that AICc (N2) is always less that AICc(N3) for further iterations, hence N0 takes N2 which is again the same old value of N0. This repeats infinite number of times. Further with N0 = 10 and N1 = 12, the terminating condition of proposed algorithm is reached i.e (N1-N0) <=3 . At

this point we computed the AICc for three values of k : 10,11 and 12. The k value = 11 which resulted in least AICc was taken as optimal number of nearest neighbors for KNN to categorize PIMA dataset as shown in fig 1. Root mean square error (RMSE) and the number of correctly classified test samples is shown in table 5. Table 5 shows for optimal k neighbors = 11, the RMSE is the least and the correctly classified neighbors is 100 which is only 2 less than the best classfication.

Table 2 shows different values of N0, N1, N2 and N3 obtained during the successive iterations of the proposed algorithm on Heart Statlog diabetic dataset. Table 2 clearly shows that during initial 5 iterations of algorithm the maximum value N1 gradually reduces from 99 to 14 (and finally to 11during later iterations). On further iterations, the minimum value increases from to 1 to 5 finally to 8. As explained in section 3 (b) since the (N1-N0) = 3, N2 will be always equal to N3. i.e without considering value of AIC for N1, N1 will be reduced to N3, hence to avoid this situations (with N0=8, N1 = 11) the terminating condition (N1-N0) <=3 is proposed in algorithm. At this point we computed the AICc for three values of k : 8,9,10 and 11 is computed. The k value = 10 which resulted in least AICc was taken as optimal number of nearest neighbors for KNN to categorize Heart statlog dataset as shown in fig 2. Table 5 shows for optimal k neighbors = 10, with the RMSE of –148.85 and the correctly classified neighbors is 70 out of 81 test samples, which is best compared to KNN classfication with other values of K.

The authors attempted to prove the proposed terminating condition is generalized one, by applying it to two more medical dataset namely Wisconsin-breast-cancer Dataset and Dermatology Dataset. The Table 3 & 4 shows different values of N0, N1, N2 and N3 obtained during the successive iterations of the proposed algorithm on Wisconsin-breast-cancer and Dermatology dataset respectively. For Wisconsin-breast-cancer, the final value of N0 is 10 and N1 is 13. The AIC was computed for k = 10,11,12 and 13.With k = 12 we got the least AIC (fig 3), least RMSE and the number of correctly classified test samples was found to 198 out of 205 (table 7). For this dataset the number of correctly classified test samples was almost constant for different values of K.

Finally for Dermatology dataset, the final values of N0 and N1 were found to be 4 and 7 respectively. AIC was computed for K values in between 4 and 7 (both inclusive). K = 5 resulted in the least AIC (fig 4), least RMSE of 0.0847 and number of correctly classified test samples was found to be 107 out of 110 test samples (table 8).

The optimal value of number of nearest neighbors was found to be 11,10,12 and 5 for PIMA, Heart statlog, Wisconsin-breast-cancer and Dermatology datasets respectively.

7 Conclusions

Categorization is one of the major machine-learning tools used in field of medical data mining. However, one of the most complex tasks in developing a KNN is determining the optimal number of nearest neighbors, which is usually obtained by recurring experiments for different values of K till the minimum error rate is achieved. It is not easy to know in advance the exact number of nearest neighbors to be used for KNN for categorizing any dataset. This paper presented successful application of combination of AIC and golden section algorithms for finding optimal number of nearest neighbors for KNN classifier. As a component of Data preprocessing, to begin with, he noisy data was handled by eliminating records with missing values. As the subsequent step in data processing, reduction in input dimension was effected using filter approach by cascading GA-CFS. The proposed approach to find the optimal number of K nearest neighbors for KNN classifier has been successfully applied to categorize medical data set namely, Pima Indian Diabetes Dataset, Wisconsin Breast Cancer dataset, Dermatology dataset and Heart statlog dataset.

8 REFERENCES

[1] .Akaike H. (1974). A New Look at Statistical Model Identification. IEEE Transactions on Automatic Control , AU-19, 716-723.

[2] Akaie H. (1973) Information theory as an extension of the maximum likelihood principle. Second Intl Symp Inf Theory ,267-81.

[3] Asha Gowda Karegowda and M.A.Jayaram. (2009). Cascading GA & CFS for feature subset selection in Medial data mining. IEEE International Advance Computing Conference,Patiyala, India.

[4] .D. Goldberg .1989. Genetic Algorithms in Search, Optimization, and Machine learning, Addison Wesley,

[5] J. Han And M. Kamber. (2001). Data Mining: Concepts and Techniques(. San Francisco, Morgan Kauffmann Publishers.

[6] http://en.wikipedia.org/wiki/Golden_section_search

[7] http://www1.ics.uci.edu/~mlearn/MLSummary.html

[8] Liqun Ren , Zhiye Zhao. (2002). An optimal neural network and concrete strength modeling. Advances in Engineering Software 33117-130

Table 1: N_0,N1,N2 and N_3 values obtained during optimization process for PIMA dataset

N_0	N_1	N_2	N_3
01	99	38	61
01	61	23	38
01	38	15	23
01	23	09	14
01	14	05	09
05	14	08	10
08	14	10	11
10	14	11	12
10	12		

Table 2: N_0 to N_3 obtained during optimization process for Heart Statlog Dataset

N_0	N_1	N_2	N_3
01	99	38	61
01	61	23	38
01	38	15	23
01	23	09	14
01	14	05	09
05	14	08	10
08	14	10	11
08	11		

Table 3: N_0 ,N_1,N_2 and N_3 obtained during optimization process for Wisconsin breast cancer Dataset

N0	N1	N2	N3
01	99	38	61
01	61	23	38
01	38	15	23
01	23	09	14
01	14	05	09
01	09	04	05
04	09	05	07
04	07		

Table 4: N_0 ,N_1,N_2 and N_3 obtained during optimization process for Dermatology Dataset

N_0	N_1	N_2	N_3
01	99	38	61
01	61	23	38
01	38	15	23
01	23	09	14
09	23	14	17
09	17	12	13
09	13	10	11
10	13		

Table 6: RMSE and correctly classified test samples for Heart Statlog Dataset for different K nearest neighbors.

K_value	RMSE	Correctly classified test samples (Total Test cases = 118)
61	0.3478	99
38	0.340	101
23	0.3352	100
15	0.331	102
14	0.3369	99
12	0.3362	100
11	0.3308	96
10	0.3378	99
09	0.3397	99
08	0.3451	95
05	0.3627	95

K_value	RMSE	Correctly classified test samples (Total Test cases = 81)
61	0.3807	65
38	0.362	66
23	0.364	66
15	0.3666	66
14	0.3638	66
11	0.3509	66
10	0.350	70
09	0.354	68
08	0.360	65
05	0.3768	66

Table 7: RMSE and correctly classified test samples for Wisconsin_breast_cancer Dataset for different K nearest neighbors.

K_value	RMSE	Correctly classified test samples (Total Test cases = 205)
61	0.1584	198
38	0.156	199
23	0.1526	199
17	0.1497	199
15	0.1502	199
14	0.1496	198
13	0.1482	198
12	0.1488	198
11	0.1546	198
10	0.1559	198
09	0.1545	198

Table 8: RMSE and correctly classified test samples for Dermatology Dataset for different K nearest neighbors.

K_value	RMSE	Correctly classified test samples (Total Test cases = 110)
61	0.1999	97
38	0.1565	103
23	0.1229	107
15	0.0993	107
14	0.0987	107
09	0.0901	107
07	0.0881	107
06	0.0889	107
05	0.0847	107
04	0.0864	107

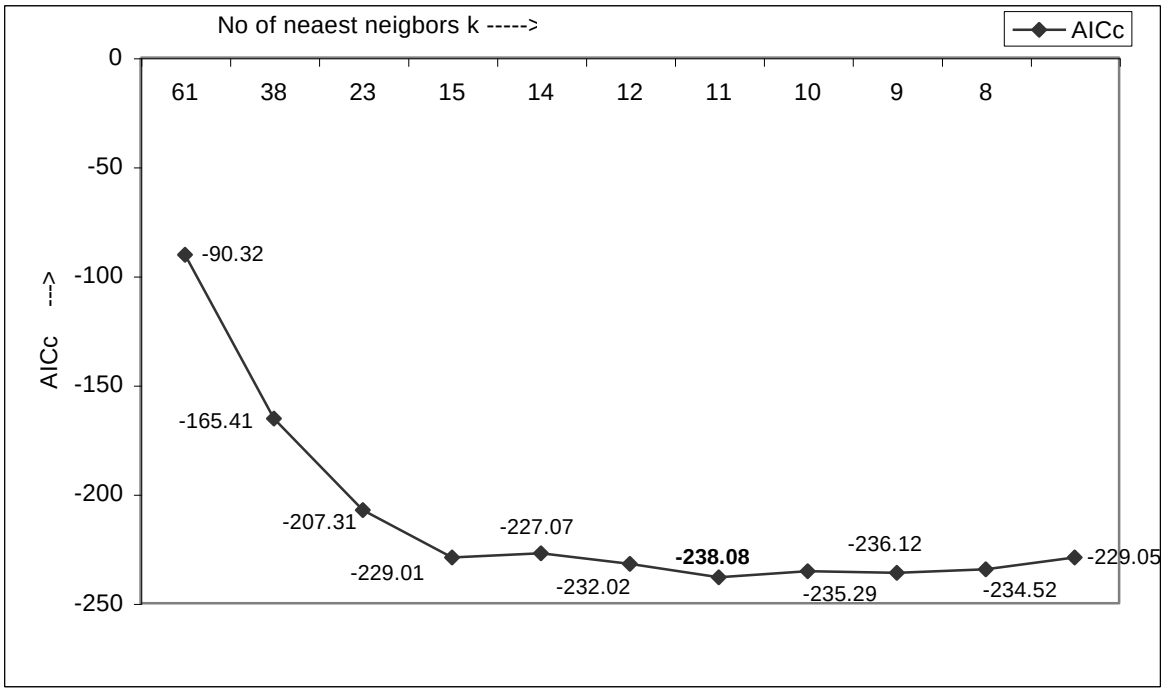

Fig 1 : Number of Nearest neighbors vs AICc for PIMA diabetic Dataset

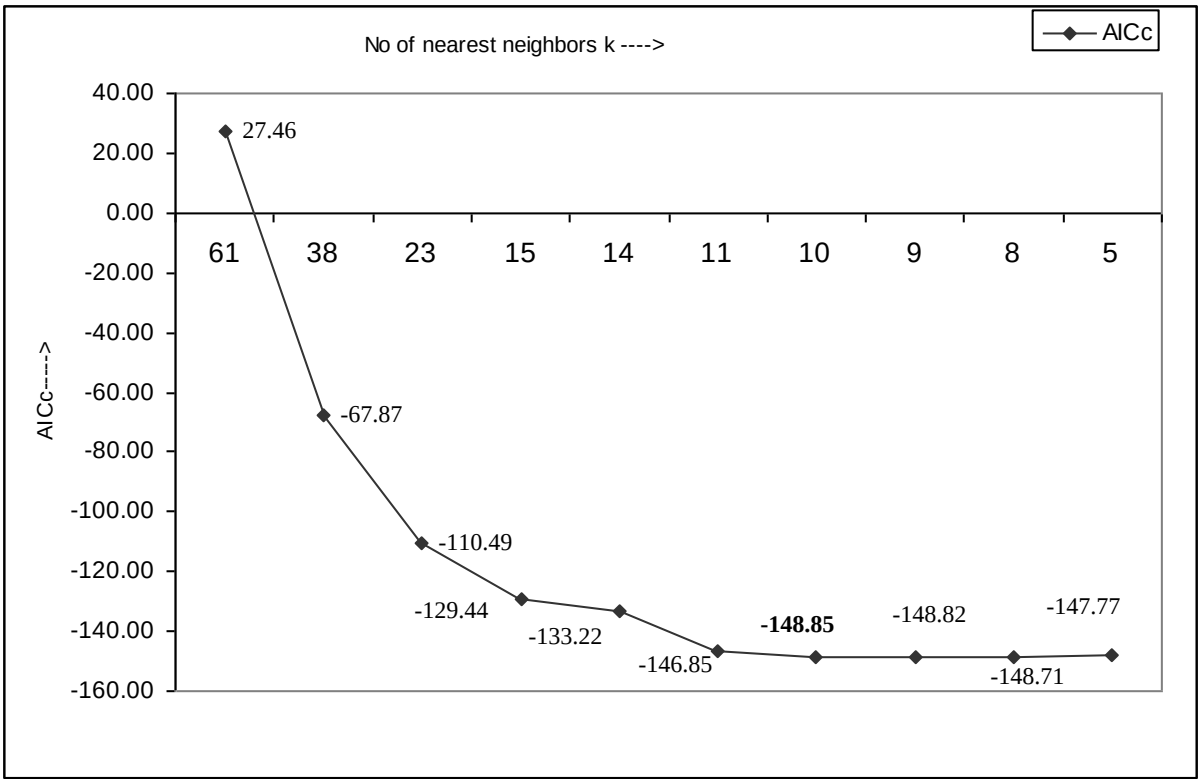

Fig 2 : Number of Nearest neighbors vs AICc for Heart Statlog Dataset

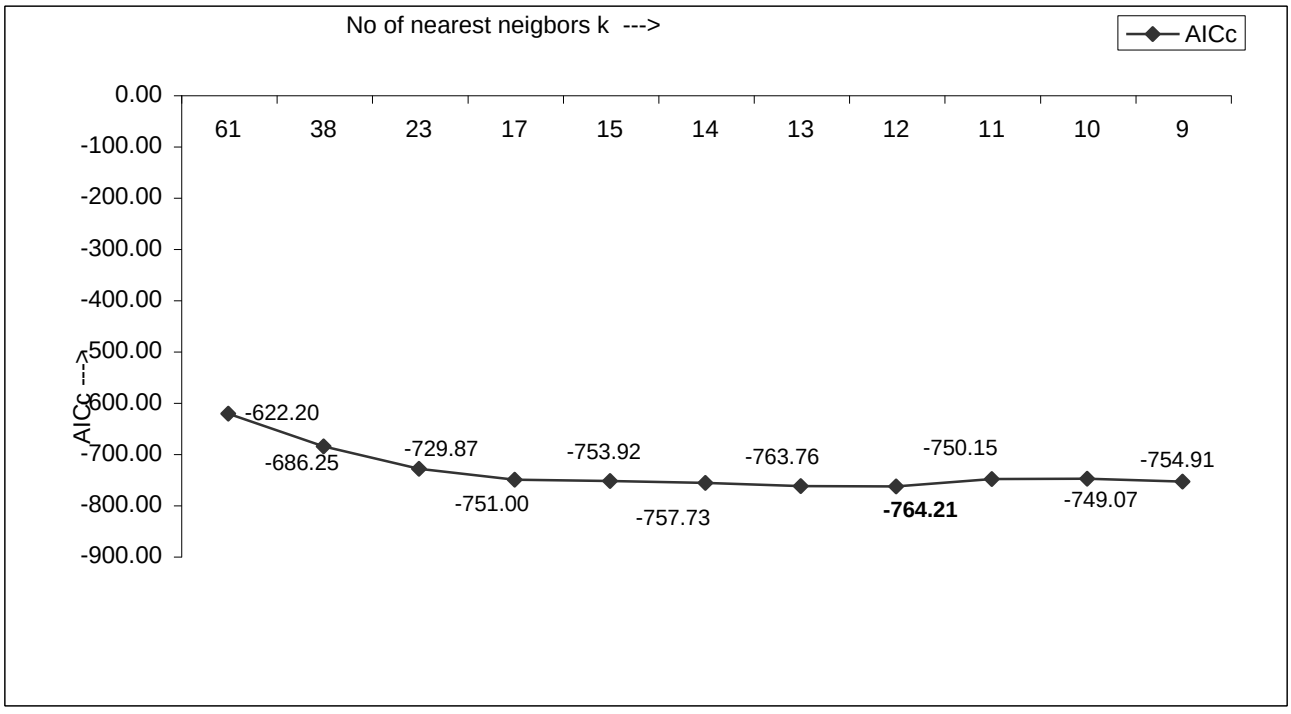

Fig 3 : Number of Nearest neighbors vs AICc for Wisconsin_breast_cancer Dataset

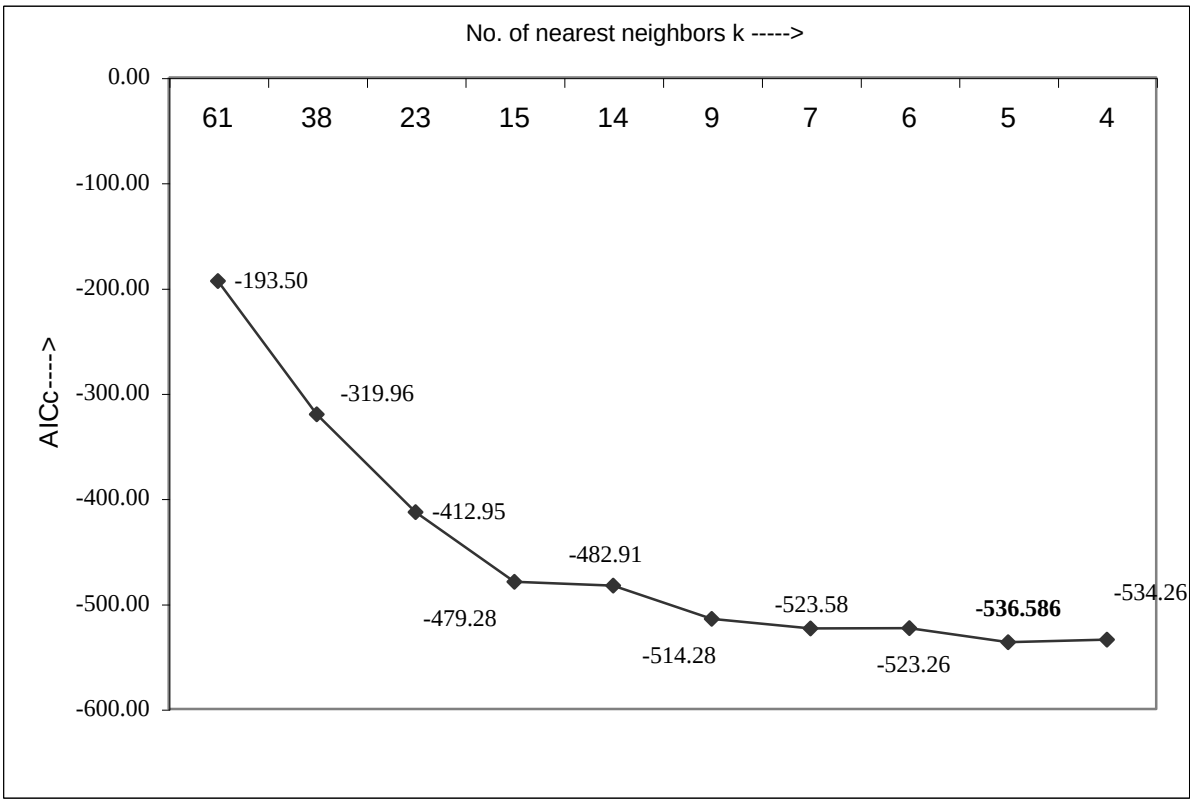

Fig 4: Number of Nearest neighbors vs AICc for Dermatology Dataset