Enhancing Accuracy of Disease Prediction of KNN and Euclidean Distance using Hybrid Approach

Amandeep Kaur ,Varinder Kaur Attri
College Name:GNDU,RC Jalandhar (Department computer science)
Postal Address: vill-kohala,potarsikka,the-baba bakala,disttAmritsar,Punjab
Pin code- 143116
Mobile no: 8728807357
Email: sangharamandeep55@gmail.com

Abstract:- Health monitoring is critical issue associated with now day lifestyle. Lack of time is causing serious issues corresponding to health. Proposed literature focus on this key aspect and provide mechanism to generate accurate predictions corresponding to parameters fetched from dataset. Hybrid approach of K Nearest neighbour and Euclidean distance is used for enhancement in health prediction. For demonstration dataset derived from UCI is utilized. Simulation results suggest considerable improvement over KNN and Euclidean distance mechanism during prediction.

Keywords:- *Health Monitoring, Prediction, dataset, K nearest neighbour, Euclidean distance, UCI*

*****

INTRODUCTION

Our health care system is literally losing patients killing more and saving less. Enormous reasons for this gigantic problem exist. Most prominent reason is lack of time causing disease to spread without noticing. Hence when the disease is detected it is beyond the scope of cure. This paper describes models which are utilized to enhance health care environment as introduction, parameters utilized in each in the next section, highlight pros and cons in next section and then comprehensive comparison of techniques in last section.

Techniques has been created and utilized to minimize the problem in hand. One of the techniques is personal health care monitoring and emergency response system.

Personal health care system earlier relies on emergency phone call system. The service agent built a call service center at one side. This system generally deals with old age persons by providing them with telephone with special keys. In emergency user just need to press that emergency key and call to doctors and related persons is made. This system is efficient enough to handle problems of fertile old aged persons. The enhancement to Personnel health care system is also made by including speech recognition. According to emergency response team Personal Health care system requires four essential components.

1. The Membership Functions including Ids of valid participant utilizing the applications provided through personal health care system.
2. Hot line allowing users to connect with medical associates.
3. Contact list should be present where call can be placed in case of emergency
4. Database containing information about doctors and medical personals. (1)

Resource requirements associated with users are enhancing day by day. It is not possible to provide such massive resource requirements though standalone physical machine. Technology is enhancing day by day. One product of technology is cloud computing. Cloud computing can be utilized to provide resources to the user which are earlier beyond the reach of user with stand alone physical machine.

The proposed literature tackles following objectives

1. Analyze techniques of data mining used to predict diseases.
2. Improving performance of prediction by hybridizing KNN and Euclidean distance.
3. Minimizing Error rate in prediction.
4. Increasing accuracy of prediction.

Next Section describes the utilization of various Data mining mechanisms in disease prediction.

LITERATURE SURVEY

This section describes existing mechanisms like KNN, Euclidean distance and ARIMA model. The proposed approach is described after words.

Data Mining Approach for Liver Injury Detection

(2)Age-contrasts in the recurrence and signs of medication incited liver damage are not completely portrayed.

Information mining investigations were performed to evaluate the effect of age on liver occasion revealing recurrence with various phenotypes and specialists. Techniques: 236 medications related with hepatotoxicity were assessed utilizing the Empirical Bayes Geometric Mean (EBGM) of the relative revealing proportion with 90% certainty interim (EB05 and EB95) computed for the age gatherings: 0–17, 18–64, and P 65 years (or elderly), for generally, genuine (intense liver disappointment), hepatocellular, and cholestatic liver damage, utilizing the WHO Safety Report Database. Comes about: Overall, instances of age 0–17, 18–64, and 65 years or more established involved 6%, 62%, and 32% of liver occasion reports. Intense liver disappointment and hepatocellular harm were all the more every now and again announced among youngsters contrasted with grown-ups and the elderly while reports with cholestatic damage were more regular among the elderly ($p < 0.00001$). A possibility to bring about mitochondrial brokenness was more pervasive among the medications with expanded pediatric announcing recurrence while high lipophilicity and biliary discharge were more typical among the medications related with higher detailing recurrence in the elderly. Conclusion: Age-particular phenotypes and potential medication properties related with age-particular hepatotoxicity were recognized in revealed liver occasions; additionally examinations are justified.

K Nearest Neighbourhood

(3)(4) uses a KNN technique for detecting heart disease and performing prediction accurately by simplifying parameters. The nearest neighbourhood algorithm is used to identify elements having similar attributes values. These attribute values are grouped together using grouping functions. Grouping function generates certain value which is compared against the threshold value to determine problems. Problems are reflected in the form of deviation. The process is described by considering two points ‘A’ and ‘B’. Let distance(A,B) is the distance between points A and B then

a. distance(A,B)=0 and distance(A,B) >=0 iff A=B
b. distance(A,B)=distance(B,A)
c. distance(A,C)<=distance(A,C)+distance(C,B)

Property 3 is also known as transitive dependency. Distance if close to zero then prediction is accurate otherwise error is recorded. Error calculating metric is applied to determine accuracy of the approach. Accuracy is given as

Accuracy=1-Error_rate

whereError_rate is given as

$$\text{Error\_rate}=\frac{|X-X_a|}{X_a}$$

KNN is used in many distinct environments such as classification, interpolation, problem solving, teaching and learning etc. Major limitation of KNN is that its performance depends upon value of k. Accuracy is low and further work is required to be done to improve accuracy.

EUCLIDEAN DISTANCE

(5)Euclidean distance is one of the simplest mechanisms for classification and prediction. Distance is the prime criteria used to evaluate the deviation in this case. Distance can be defined in several ways. Let $[x_1, x_2, ----, x_n]$ is the distance of points in terms of x coordinate and $[y_1, y_2, ----, y_n]$ is the distance in terms of y coordinate. The Euclidean distance is defined as

$$Euclidean_{distance} = \sum (x_i - y_i)^2$$

Where i define range of values from 1 to n. All the components of vectors are taken equally and no correlation is evaluated in this case. The result of Euclidean distance equation can be normalized. This is accomplished as

$$M_i = (x_i)^2$$

Where averaging is taken over all the vectors in the dataset. The scaled distance is obtained using the following equation

$$D^2 = \sum \frac{(x_i - y_i)^2}{M_i}$$

The scaled distance is adjusted value so that obtained result lie between the specified range. The metric is used to evaluate errors. (6–8)Mean root square error is one such mechanism for observing accuracy. Accuracy and error rate is inversely proportional to each other.

$$RMS = \sqrt{(x - x_a)^2}$$

This equation is used to evaluate Root Mean square error. Lower the value of RMS more accurate a prediction. Advantage of this approach is, convergence rate is better but disadvantage is that it can work over limited values. Non negative values are allowed and hence result always lies between 0 and 1.

2.3 ARIMA

(9)(10,11)Auto regressive moving average model is used for accurate forecasting in case of disease detection. Changes in time series using mathematical model is used in ARIMA. This model is based on adjustment of observed values. The goal is to obtain the differences of observed value and value obtained from the model close to zero. This model can predict accurately difference between the stationery and non stationery series. Working of this model is described through the following diagram

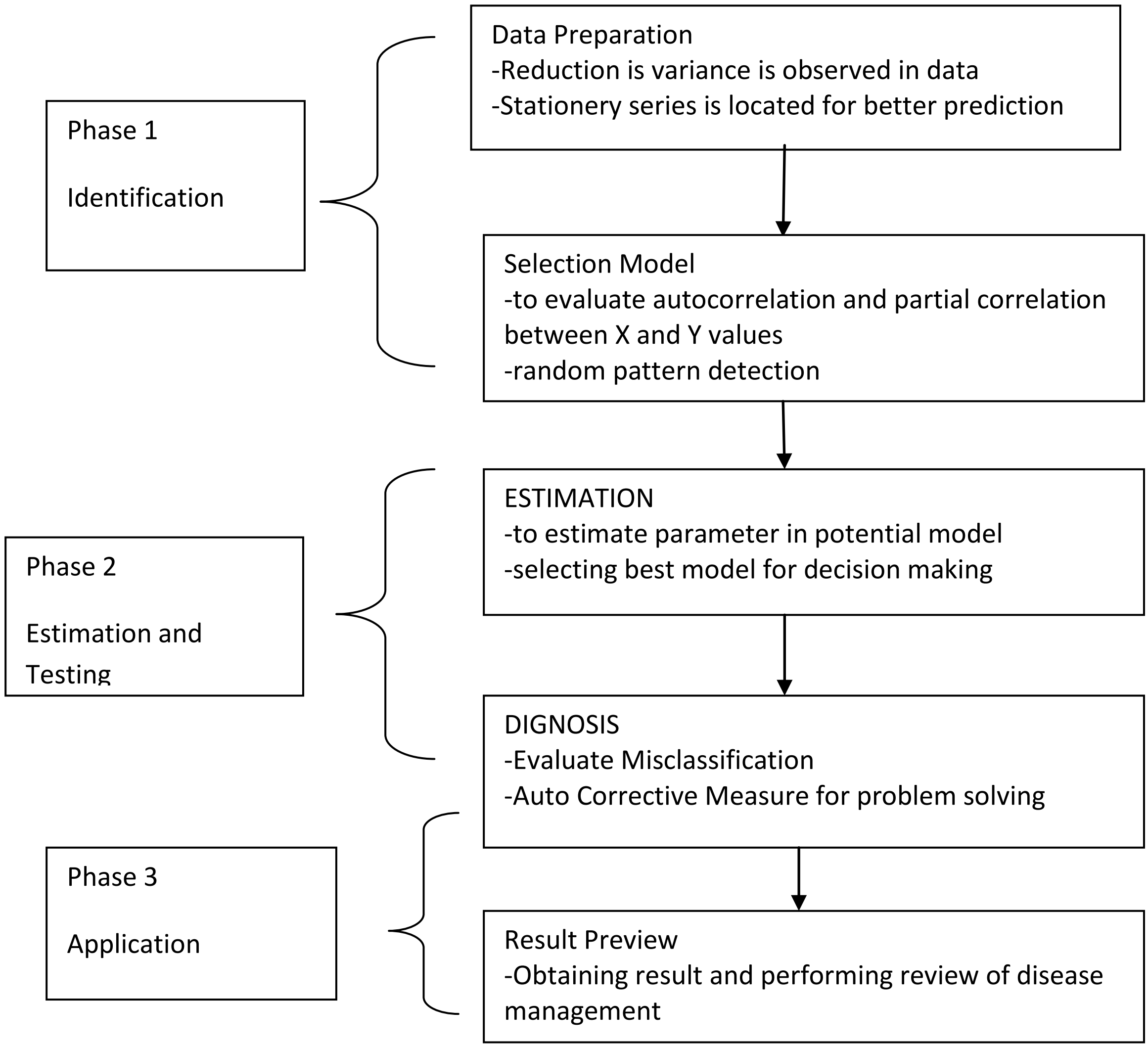
Phase 1
Identification
Phase 2
Estimation and Testing
Phase 3
Application
Data Preparation
-Reduction is variance is observed in data
-Stationery series is located for better prediction
Selection Model
-to evaluate autocorrelation and partial correlation between X and Y values
-random pattern detection
ESTIMATION
-to estimate parameter in potential model
-selecting best model for decision making
DIGNOSIS
-Evaluate Misclassification
-Auto Corrective Measure for problem solving
Result Preview
-Obtaining result and performing review of disease management

Figure 1: Working of ARIMA model for Health prediction

ARIMA model has multiple phases associated with it. This model can be merged with Euclidean distance and KNN for better performance in health prediction.

By studying the existing literature scope of further improvement originates. Next section describes proposed work of hybridization(KNN+Euclidean distance).

3. PROPOSED SYSTEM: The proposed literature enhances the accuracy of prediction by the use of hybrid approach of KNN and Euclidean distance mechanism. Though Euclidean distance, values from dataset are grouped into nearest similar index values and through KNN clusters are formed. The prediction accuracy is greatly enhanced by the use of suggested mechanism. The algorithm for proposed approach is listed as under

Algorithm: KNN+Euclidean

a. Fetch Values from Dataset
   Set X=Dataset$_x$ Time values are fetched into X variables
   Set Y=Dataset$_y$ Parametric Values corresponding to liver are fetched into Y variable
b. Calculate Euclidean distance between X and Y values

$$Euclidean_{distance} = \sum (x_i - y_i)^2$$

c. Group Values having minimum Distance
   Group$_i$=Min($Euclidean_{distance}$)
d. Perform Clustering and evaluate nearest neighbour using KNN. To evaluate Clustering Value of K is chosen as 'n'.

e. Evaluate Accuracy

Accuracy=1-Error_rate

f. Evaluate Error_Rate

Error_rate= $\frac{|X-X_a|}{X_a}$

Next Section describes performance comparison of Hybrid approach of KNN and Euclidean distance.

4. PERFORMANCE ANALYSIS

The performance of existing and proposed systems is compared against each other. The performance of proposed approach improves by 12% in terms of accuracy. Results are listed as follows

Technique	Accuracy	Error_Rate
KNN	87.5	12.5
Euclidean_Distance	75.8	24.2
KNN+Euclidean	98.2	1.8

Table 1: performance Analysis of various classifier

The plots of the performance comparison is as follows

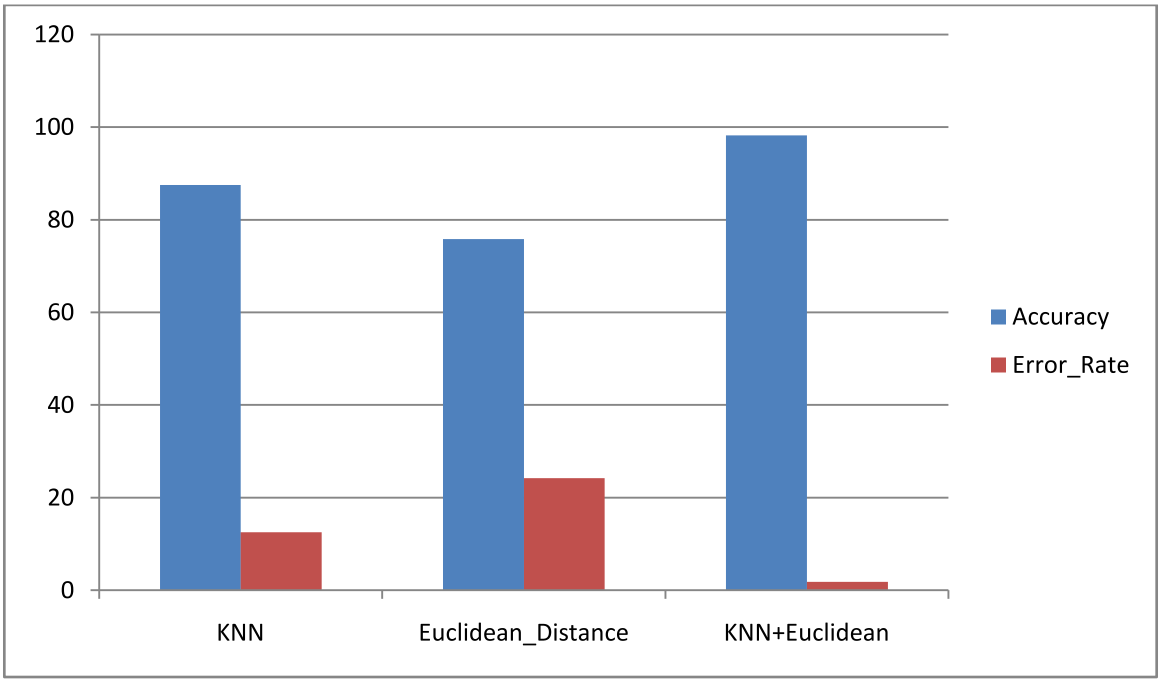

Figure 2: Performance comparison of various classifiers.

5. CONCLUSION

The performance analysis indicates that KNN+Euclidean distance approach produces better accuracy in terms of classification. The approach also produces better recognition rate. Value of K depends upon groups which are formed by Euclidean distance. Although this approach is producing optimal result but it is computationally expensive. In future, Support Vector machine and KNN can be hybridized for improving recognition rate further.

6. REFERENCES

1. Lin Y, Lu X, Fang F. Personal Health Care Monitoring and Emergency Response Mechanisms. 2013;
2. Hunt CM, Yuen NA, Stirnadel-Farrant HA, Suzuki A. Age-related differences in reporting of drug-associated liver injury: Data-mining of WHO safety report database. Regul Toxicol Pharmacol [Internet]. Elsevier Inc.; 2014;70(2):519–26. Available from: http://dx.doi.org/10.1016/j.yrtph.2014.09.007
3. Jabbar MA, Deekshatulu BL, Chandra P. Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. Procedia Technol [Internet]. Elsevier B.V.; 2013;10:85–94. Available from: http://dx.doi.org/10.1016/j.protcy.2013.12.340\nhttp://www.sciencedirect.com/science/article/pii/S2212017313004945
4. Enriko IKA, Suryanegara M, Gunawan D. Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient ’ s Health Parameters. 1843;8(12).
5. Veytsman B, Wang L, Cui T, Bruskin S, Baranova A. Distance-based classifiers as potential diagnostic and prediction tools for human diseases. BMC

Genomics [Internet]. BioMed Central Ltd; 2014;15 Suppl 1(Suppl 12):S10. Available from: http://www.biomedcentral.com/1471-2164/15/S12/S10

6. El-Hattab MM. Applying post classification change detection technique to monitor an Egyptian coastal zone (Abu Qir Bay). Egypt J Remote Sens Sp Sci [Internet]. Authority for Remote Sensing and Space Sciences; 2016;19(1):23–36. Available from: http://dx.doi.org/10.1016/j.ejrs.2016.02.002
7. Chen C, Won M, Stoleru R, Member GGX. Energy-Efficient Fault-Tolerant Data Storage & Processing in Mobile Cloud. 2014;3(1):1–14.
8. Bui D, Hussain S, Huh E, Lee S. Adaptive Replication Management in HDFS based on Supervised Learning. 2016;4347(c):1–14.
9. Jose D V, Sadashivappa G. a N Ovel E Nergy E Fficient R Outing a Lgorithm for W Ireless S Ensor N Etworks. Int J Wirel Mob Networks. 2014;6(6):15–25.
10. Pan Y, Zhang M, Chen Z, Zhou M, Zhang Z. An ARIMA based model for forecasting the patient number of epidemic disease. 2016 13th Int Conf Serv Syst Serv Manag ICSSSM 2016. 2016;31–4.
11. Permanasari, A.E. Hidayah, I.Bustoni IA. SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence. Inf Technol Electr Eng (ICITEE),2013 Int Conf . 2013;(2):2–6.