MODIFIED ARIMA AUTO REGRESSION, KNN AND EUCLIDEAN DISTANCE: FOR PREDICTING HEART DISEASE WITH ACCURACY ENHANCMENT

Rekha Rani
Computer Engineering & Technology
Guru Nanak Dev University, Amritsar
rekhareet66@gmail.com

Kamaljit Kaur
Computer Engineering & Technology
Guru Nanak Dev University, Amritsar
kamal.aujla86@gmail.com

***Abstract*-In this paper, Data Mining is presented and in addition enormous Data in the system of Healthcare. Moreover, the Data mining for collected Data is researched. Particularly, their complexities of the different regions health care and medicinal research. Data set used for healthcare prediction can contain missing value which is analyzed initially by the use of support vector machine (SVM). Proposed literature uses ARIMA model merge with KNN and Euclidean distance to predict most probable value which can be replaced with missing value present within the Dataset At last, machine learning calculations have been utilized in order to compare the accuracy and perform prediction accordingly**.
Keywords-Data mining; SVM; ARIMA; Euclidean distance

I. INTRODUCTION

Data Mining has quickly grown with the presence of the wonder BIG Data[1]. For sure, numerous associations have begun to digitize their records, and have changed their paper-based frameworks to electronic frameworks. This change conveys a few advantages to the associations, among them time funds, a superior administration and a more tightly checking making the assignments less demanding. One of the immediate results of this change is the visit gathering of significant Data. While the Data's holders started to stress over the capacity of Data, they understood the benefits they can take from it. The Data gathered can be considered as another unformatted of structure (Raw Data) which needs to be filtered. Handling Data give a superior quality Data which contribute in request to make choice in data selection[2]. Moreover, Healthcare elements likewise choose electronic frameworks, by utilizing different strategies, among them, Electronic health Record (EHR) or Electronic Medical Records (EMR) frameworks. It implies the executing EHR frameworks, leads to an immense measure of Data gathered by doctor's facilities, centers and other health care suppliers.[3] At that point, the vast majority of these Datasets are most certainly not extremely very much organized and fitting for explanatory purposes. In expansion, health care Data are generally extremely perplexing and difficult to investigate. For instance the US Healthcare framework alone as of now achieved 150 Exabyte (1 Exabyte = 8388608 Terabit) five years prior. This pattern is because of the way that multi scale Data created from people is consistently expanding, especially with the new high-throughput sequencing stages, continuous imaging, and purpose of care gadgets, also as wearable figuring and versatile health care innovations. As needs be, Data Mining has gotten a great deal of consideration on account of its solid capacity of separating Data from Data, furthermore, winds up noticeably prevalent in Healthcare field by dint of its productive diagnostic procedure for recognizing obscure and significant Data in health care Data[1], [4].

In health care industry, Data Mining gives a few advantages for example, discovery of the extortion in medical coverage, accessibility of therapeutic answer for the patients at lower cost, discovery of reasons for ailments and recognizable proof of therapeutic treatment techniques. It likewise helps the Healthcare analysts for making productive Healthcare approaches, developing medication suggestion frameworks, creating health care profiles of people and so on[1], [2], [5], [6].Taking such a case, McKinsey gauges that enormous Data examination can empower more than 300 billion in investment funds for every year in U.S. Medicinal services, 66% of that through decreases of around 8 percent in national Healthcare consumptions. Clinical operations and R & D

(innovative work) are two of the biggest ranges for potential reserve funds with 165 billion furthermore, 108 billion in waste individually The result of Data Mining advancements are to give advantages to Healthcare association for gathering the patients having comparative sort of infections or medical problems so that Medicinal services association gives them successful medications[6]. It can likewise valuable for anticipating the length of remain of patients in healing center, for restorative conclusion and making arrangement for compelling Data framework administration. Late innovations are utilized as a part of restorative field to improve the restorative administrations in practical way. Data Mining methods are additionally used to examine the different elements that are in charge of sicknesses for instance sort of nourishment, diverse working condition, instruction level, living conditions, accessibility of unadulterated water, human services administrations, social, natural and rural variables.

In this paper, we introduce the upsides of Data Mining for health care and the reasons make Data Mining critical to be considered in health care Data examination. Data mining health care Dataset with missing values is considered to be analyzed initially through Support vector machine and accuracy is analyzed and after that ARIMA with KNN and Euclidean distance is used for rectification and analysis purpose[6]–[8][9][10]. Accuracy is observed in both the cases to prove worth of the study.

II.STUDY OF EXISTING LITERATURE

Data mining approaches is the base of this literature. Analysis of existing literature provide base for proposed literature. [11] Reviewed various models and methods used within data mining. Data mining techniques development from 2005 to 2015 is reviewed and application in regards to health care is proposed. [1] Suggests the integration of medical data with data mining strategies used to form medical information system. Patient medical condition can be analyzed along with future prediction about patient's health. Hidden possibilities can be extracted using unlimited data mining techniques to make accurate health forecast. [12] Proposed multilayer perception in order to analyze big data corresponding to health care. As literature deals with health care of patients hence high degree of accuracy is desired. To accomplish the desired goal comparison of SVM and multilayer Perception on health care data set is made. Results of SVM in terms of classification are better as compared to multilayer perception. [3] Suggests data mining techniques used for analysis of diabetics. Support Vector Machine (SVM) is used for this purpose. Genetic approach is also analyzed for diabetic's dataset in the field of data mining. Results of SVM are obtained to be better. [13] Suggests five J.48 classifiers to predict hypertension and eight other diseases. Prediction accuracy is obtained and compared against naïve bayes approach. Results in terms of J.48 are obtained to be better. [7] Suggests hybrid approach for health care to predict diseases using Big data. Pruning based KNN is used for this purpose which used density based clustering based method integrated with KNN approach. Local outlier factor of PB-KNN is better as compared to KNN. [14] Proposes SVM and neural network techniques for skin lesion detection in human body. Segmentation along with classification is performed in order to detect the diseases. [8]predict heart diseases are primary cause of death among humans in last decade. Data mining techniques are used in order to detect and predict heart diseases efficiently. [4] Proposes a mechanism through which information about patient coming for checkup at hospital is stored and algorithm is applied in order to perform predictions. Data mining algorithm considered in this approach is naïve bayes . Accuracy of prediction is obtained is significant in this case.[15] suggests intelligent heart disease prediction system. Decision tree, naïve bayes and neural network technique are used for accurate analysis and prediction of disease.

Analyzed approaches enhance performance considering datasets not including any noisy or missing values. Missing values or noisy data handling and increasing prediction accuracy is primary task of proposed approach.

III PROPOSED FORMULATION

A. Gaps in Literature

Data mining techniques used are inefficient in the handling of missing data or values. As missing values appear, accuracy reduces significantly. Data of different values may be classified distinctly. Number of classes defined decides accuracy or degree of classification. Variation hence has to be minimized to enhance accuracy. Predictions through data mining approach SVM is not done. Future predictions accuracy has to be enhanced by handling missing values. Most probable values needed to be calculated and replaced with existing values for handling missing values.

C. Problem Definitions

Data mining is not a simple assignment, as the calculations utilized can get exceptionally intricate and information is not generally accessible at one place. It should be coordinated from different heterogeneous information sources. These elements additionally make a few issues.

i. Handling of relational and complex types of data-The database may contain complex information objects, mixed media information objects, spatial information, transient information and so forth. It is unrealistic for one framework to mine all these sort of information.

ii. Efficiency and scalability of data mining algorithms −Keeping in mind the end goal to successfully separate the data from colossal measure of information in databases, information mining calculation must be productive and adaptable.

iii. Handling noisy or incomplete data − The information cleaning strategies are required to deal with the clamor and inadequate items while mining the information regularities. In the event that the information cleaning strategies are not there then the exactness of the found examples will be poor.

iv. Prediction accuracy- In this case missing value are present accuracy of prediction is lowered so it has to be improved.

D. OBEJECTIVE OF STUDY

Heart disease detection is primary concern of this literature. The problem starts to appear as missing value appear with in the dataset. The comparison with the existing classifier is made to prove worth of a study. These days, wellbeing sicknesses are expanding step by step because of way of life, genetic. Particularly, coronary illness has turned out to be more typical nowadays, i.e. life of individuals is at hazard. Every individual has diverse esteems for Blood weight, cholesterol and heartbeat rate. Be that as it may, as indicated by therapeutically demonstrated outcomes the typical estimations of Blood weight is 120/90, cholesterol is less than 200mg/dl and beat rate is 72.

The description of objective is listed as under:-

1. To analysis heterogeneous datasets.
2. Comparing existing classifier SVM with proposed modified ARIMA.
3. Handling missing value within the dataset
4. Improving accuracy of the prediction.

IV.PROPOSED METHODOLOGY

A. PROPOSED AGLORITHM

The algorithm for the proposed approach is as under:

Algorithm ARIMA (KNN+EUCLIDEAN)

* INPUT: Dataset with attribute values including artifact or noisy or missing values.

* Output: Prediction Accuracy

Step1-Perform Pre-Processing

Convert attribute values to nominal form for analysis.

Step2- Apply auto regression mechanism for calculating most probable values from each attributes

Step3- Specify value of $K=E_n$

Calculate Euclidean Distance

$$Euc_i = \sqrt{\left(X_{k+E_n+1} - X_k\right)^2 - \left(Y_{k+E_n+1} - Y_k\right)^2}$$

$$Obtain\ average\ of\ euclidean\ distance(MPV) = Euc_i/n$$

Where n is total number of rows present within dataset

Step4- Replace missing values calculated from KNN+EUCLIDEAN with most probable values (MSV).

Step5- Calculate accuracy and compare it with SMO

B.PROPOSED FLOWCHART

Figure 1:Showing Proposed Flow Chart

V. EXPERIMENTAL RESULTS AND COMPARISON

A. Dataset Description

The system being created is analyzed for accuracy. The accuracy is enhancement of the dataset is by removing the missing value from the dataset. The missing values are removed and replaced by most probable value. Most probable values are calculated by the use of auto-regression KNN and Euclidean distance

The dataset used contain 13 attributes Age, Gender, Chest pain, Resting blood Pressure, Serum Cholesterol in mg/dl, Fasting blood sugar, Resting electrocardiographic result, Maximum heart rate achieved, Exercise induced angina, Old peak, ST_ amount of exercise, Vessels, Thal.

Table 1:-Showing dataset of heart disease used in the proposed system

Age	Gender	Chest Pain Type	Resting blood Pressure	Serum Cholesterol in mg/dl	Fasting blood sugar	Maximum heart rate achieved	Exercise induced angina	Old peak	ST_amount of exercise	Vessels	Thal
21	0	1	101	121	0	0	0	1.1	1	1	1
22	1	?	102	122	1	1	1	?	2	2	6
23	0	3	103	150	0	2	0	1.3	3	3	3
24	1	4	104	144	?	0	0	1.4	1	1	4
25	0	1	105	145	1	1	1	1.5	2	?	5
---	---	---	---	---	---	---	---	---	---	---	---

This dataset contain missing values which are to be handled through the proposed literature

B. RESULTS AND PERFORMANCE ANALYSIS

Results are obtained in terms of accuracy and future Prediction . In case missing value are present the support vector machine cannot tackle this issue. In order to overcome this problem, missing value are tackled using KNN and Euclidean distance.

Table 2:-Accuracy of Classification of existing and proposed approach

NO_OF_TUPLES_ANALYSED	SMO	MODIFIED_ARIMA
300	85.5	93.3
290	82.365	92.12
250	75.235	85.212
150	65.555	70.646
100	48.999	53.457

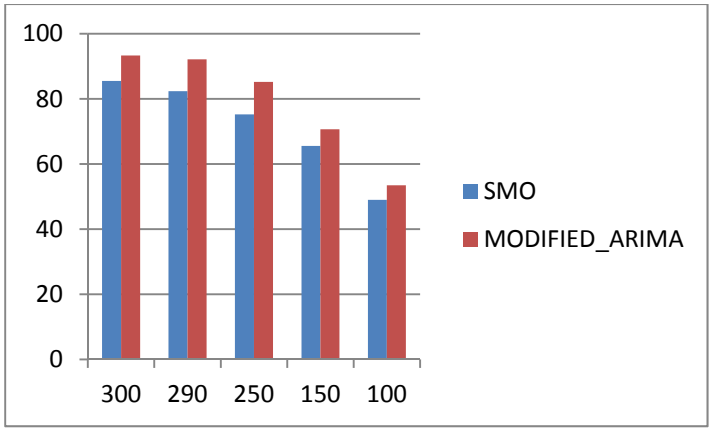

Figure 2:- Accuracy of classification of existing and proposed approach.

The time consumed of existing approach is also higher as compare to proposed approach. This is also known as latency or delay

Table 3:- Showing latency of Support vector machine classifier with modified Arima

No_of_tuples analyzed	SMO	Modified_arima
300	2013	435
290	2011	432
250	1990	399
150	1880	390
100	1473	380

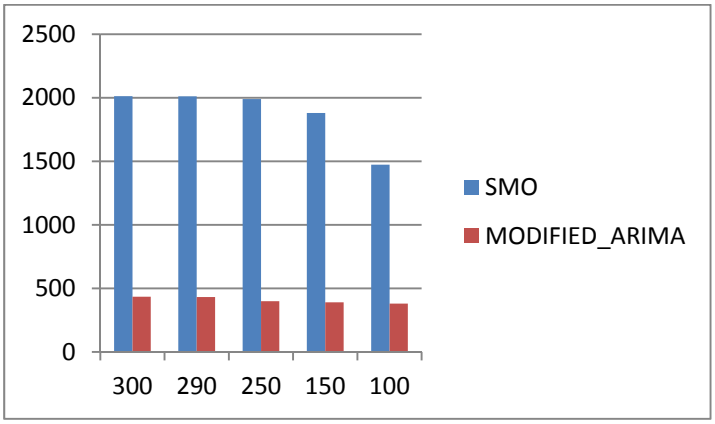

Figure 3:-Showing latency of Support vector machine classifier with modified Arima

The prediction generated through the proposed approach is going to help the person in predicted heart diseases and take preventing measures.

VI. CONCLUSION AND FUTURE SCOPE

The missing values handling is a critical part which is accomplished with the modified ARIMA model. The SVM is unsupervised classifier usually used for numerical value. The missing values are generally string in nature which cannot be classified with SVM. Hence accuracy degrades. In order to overcome this problem modified ARIMA with KNN Euclidean distance is used. The result indicates betterment in terms of accuracy, prediction, and latency.

In future KNN and Manhattan distance can be combined or hybridized to improve prediction accuracy.

REFERENCES

[1] I. Taranu, "Data mining in healthcare: decision making and precision," Database Syst. J., vol. 5, no. 4, pp. 33–40, 2015.
[2] M. E. Student, C. T. Nadu, and C. T. Nadu, "Heart disease classification and its co-morbid condition detection using WPCA genetic algorithm," pp. 287–291, 2016.
[3] "1-s2.0-S2001037016300733-main(1).pdf." .
[4] C. Anusha, S. K. Vinay, H. J. Pooja Raj, and S. Ranganatha, "Medical data mining and analysis for heart disease dataset using classification techniques," Natl. Conf. Challenges Res. Technol. Coming Decad. (CRT 2013), pp. 1.09–1.09, 2013.
[5] E. Pinheiro, W. Weber, and L. Barroso, "Failure trends in a large disk drive population," Proc. 5th USENIX Conf. File Storage Technol. (FAST 2007), no. February, pp. 17–29, 2007.
[6] A. Sharma and V. Mansotra, "Emerging applications of data mining for healthcare management - A critical review," 2014 Int. Conf. Comput. Sustain. Glob. Dev., pp. 377–382, 2014.
[7] K. Yan, X. You, X. Ji, G. Yin, and F. Yang, "A Hybrid Outlier Detection Method for Health Care Big Data," 2016 IEEE Int. Conf. Big Data Cloud Comput. (BDCloud), Soc. Comput. Netw. (SocialCom), Sustain. Comput. Commun., pp. 157–162, 2016.
[8] S. Sivagowry, M. Durairaj, and a. Persia, "An empirical study on applying data mining techniques for the analysis and prediction of heart disease," 2013 Int. Conf. Inf. Commun. Embed. Syst., pp. 265–270, 2013.
[9] W. E. Leland, W. E. Leland, D. V Wilson, and D. V Wilson, "On the Self-Similar Nature of Ethernet Traf c," Comput. Commun. Rev., vol. 2, no. August 1989, pp. 203–213, 1992.
[10] S. Jain and N. Pise, "Computer aided Melanoma skin cancer detection using Image Processing," Procedia - Procedia Comput. Sci., vol. 48, no. Iccc, pp. 735–740, 2015.
[11] N. Jothi, N. A. Rashid, and W. Husain, "Data Mining in Healthcare - A Review," Procedia Comput. Sci., vol. 72, pp. 306–313, 2015.
[12] P. Naraei, V. Street, V. Street, and V. Street,"Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data," no. December, pp. 848–852, 2016.
[13] F. Huang, S. Wang, and C. Chan, "Predicting Disease By Using Data Mining Based on Healthcare Information System," 2012 IEEE Int. Conf. Granul. Comput. Predict., pp. 12–15, 2012.
[14] M. A. Farooq, M. A. M. Azhar, and R. H. Raza, "Automatic Lesion Detection System (ALDS) for Skin Cancer Classification Using SVM and Neural Classifiers," 2016 IEEE 16th Int. Conf. Bioinforma. Bioeng., pp. 301–308, 2016.
[15] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2008 IEEE/ACS Int. Conf. Comput. Syst. Appl., pp. 108–115, 2008.