IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 12 (December. 2018), ||V (I) || PP 42-49

A Effective Disease Prediction Model Using Enhanced Principle Component Analysis And Mdrp Algorithm

Deepthi krishnan. K[1], Senthil Kumar. B[2]
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[1]
Associate Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[2]
Corresponding Author: Deepthi Krishnan. K

Abstract: The rapid growth of data in biomedical and healthcare community's data analysis plays an important role. In recent scenario, health issues are huge, due to this nature predicting and classifying into different conditions are very tedious. The field of data mining has involved in those domains to predict and to classify the abnormality along with its risk level. The previous studies have used several features to Disease Prediction, which has been collected from patients. By applying different data mining algorithms, the patient data can be used for Disease Prediction. Those techniques achieve high accuracy but this technique cannot apply for big data analytics. The main drawbacks of the previous studies are that need accurate and more number of features. The study proposed a Data mining model has been developed using Enhanced Principle Component Analysis to improve the Disease prediction accuracy and to investigate the risk level of the disease. The system implements a new EPCA and MDRP algorithm for effective prediction of disease. This also creates a new advanced latent factor model to reconstruct the missing data for fast and accurate disease classification. The system developed with the intension of high accuracy and less training overhead. From the experimental results, prediction accuracy of our proposed algorithm reaches 98.8% with a convergence speed which is faster than the existing system.

Keywords: Big data analytics, Machine Learning, Healthcare, Disease Detection, Medical Data Analytics.

Date of Submission: 19-11-2018 Date of acceptance: 04-12-2018

I. INTRODUCTION

The main purpose of the big data is not a new concept it is persistently changing. Big data is not anything it is a collection of large data. There are three major points of big data that is velocity, volume and variety. Healthcare is a better template for the three of big data. The healthcare data is outspread among to involving several parts of medical systems, healthcare parts, and government hospitals with the advantage of a big data and a greater awareness is paid to the Disease Prediction. The number of investigations has been managed to selecting the attribute of a disease prediction from a great volume of a data. The greatest amount of the existing work is based on a structured data. The unstructured solitary data can use a convolution neural network. Convolution neural network are make a nerve cell, each nerve cell collect some information entered system and execute operations and the full network indicates a single differentiable result functions. The precise of a disease prediction can be decreased because there is an additional difference in a various related disease because of the weather conditions prevailing in an area in general and living habits of climate the peoples in their particular regions. To reduce this difficulty to combines both the structured and unstructured data. To correctly predict the disease control the problem of a losing and insufficient data. The Big data technology can be used latent characteristic model. In the prior effort only structured data can be used but for the perfect results. In this system can use the unstructured data. In this technique can choose characteristic spontaneously using CNN algorithm. In this proposed system can base on CNN-MDRP algorithm to apply the data types. And the healthcare community can use machine learning algorithm for more perfect results. Disease Predictions for structured data use to conventional machine learning algorithm i.e., Naïve Bayesian.

A big-data is an instance of revolving system it is under way in health care. It begins to a very great extent increased to providing the information. Extending the last decade, pharmaceutical companies have been collecting years of investigations and implementation of data into medical databases, while providers have convert their patient data. Meanwhile, the associated government and additional public stakeholders have been open their expansive stores of health-care knowledge, counting data from clinical testing and information on patients protected downward the public insurance programs. In similar, modern technical advances have made it easier to receive and inspect information from lot of sources. An important benefit in health care, since data for

one patient may come from different payers, laboratories, physician offices and hospitals. This system perhaps more than any other elements that is driving the requesting for big-data applications. To disappoint overutilization, lot of payers has moved from fee-for-service for repayment, which advantage physicians for treatment capacity, to risk-sharing processes those prerequisite outcomes. Under the upcoming methods, when treatments distribute the strong results, distributor suffering may be less than before. Payers are also introducing similar concurrence with pharmaceutical companies and based on the reimbursement on a drug's capacity to increase patient health. In this new organization, health-care stakeholders have more benefits to compile and exchange information.

1. Problem Definition

In this paper (**Ayon Dey, Jyoti Singh, 2016**) has define the topic for "Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis ". In this paper author explains the heart disease is the main cause for the death. The 23.6 millions of people may die for the heart disease. In the health care industry capture the heart disease information which unfortunately the hidden data are not mined to take a decision. The study of this paper PCA has been done to find a minimum number of required attributes to increase the accuracy for supervised machine learning techniques. The main theme of this paper is to analyze the supervised machine learning technique to predict the heart disease.

In This paper (**Shirsath, 2018**) concept of big data is provides the beneficial merits like, accurate medical data analysis before the disease prediction of perfect data can be securely stored and used. And the accuracy of a medical data analysis can reduced the incomplete medical data; the genetic disease can break for using the medical data prediction. In the purpose of disease prediction can collect the hospital data in the particular region. The misplaced data can be used inactive factor model to aim the incomplete data. In medical data prediction have three types such as: a) Hospital data, b) Structured data, c) Unstructured Data. a) Hospital data: The hospital data is a highest volume of datasets of a patient can be given by a hospital and the data can be saved in the data centre to protect the patient privacy and security of saved data, can be create a security access mechanism. b) Structured data: The structured data is nothing but the scientific experiments data, patient's basic information like patient's age, life habits, gender, weight , height etc. c) Unstructured Data :Unstructured Data is a information of patients medical history, patients weakness, and doctors interrogation and diagnosis**.**

In this paper (**Shakuntala Jatav, 2018**) have gives the contents based on" An Algorithm for Predictive Data Mining Approach in Medical Diagnosis". It contains large and composite data is required extremely interesting pattern of diseases & different types of machine learning techniques are used to makes perfect decisions. For the medical research the advanced data mining techniques are used to discover knowledge in database. This paper has to be used to predict the disease for Kidney, Diabetes, and Liver disease using lot of input functionalities. The data mining categorization techniques, such as Support Vector Machine (SVM) and Random Forest (RF) are examined on Diabetes, Kidney and Liver disease datasets. The presentation of these techniques is interconnected, based on precision, accuracy, recall, f measure as well as time.

In this paper (**P. Suresh, 2018**) has given the concept for "Study and Analysis of Prediction Model for Heart Disease: An Optimization Approach using Genetic Algorithm". In the medical field the heart disease diagnosis is the biggest task. The grouping large clinical and pathological data heart disease diagnosis is very difficult. In heart disease prediction the complication, increased amount and clinical professionals about the good and perfect. Machine learning is used to provide the efficient support for predicting heart disease with perfect case of training and testing. The important of this work is to study diverse prediction models for the heart disease and choosing important heart disease feature using genetic algorithm.

In this paper (**Sreekanth Rallapalli, Suryakanthi T, 2016**) has provide the "Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm". In this concept explains Electronic Health Care records can be stored in an enterprise database or cloud databases. In this data can be efficiently processed. The predictive analysis helps the doctors, physicians to identify the patient's admission details or other information. The main challenging task is identifying the strong indicators for an accurate disease prediction.

II. PROPOSED SYSTEM

The term discusses details and brief explanation about the proposed methodology and the steps involved in that proposed system. The optimized EPCA and MDRP (Multimodal Disease Risk Prediction) algorithm has been expanded with the new optimal classification algorithms, which can handle large category dataset more rapidly, accurately and effectively, and keeps like good scalability at the same time. This term discuss about the algorithms and methodologies.

Contributions of the proposed System:

The followings are the contributions of the proposed system.

- The system implements a new EPCA and MDRP algorithm for effective prediction of disease.
- This also creates a new advanced latent factor model to reconstruct the missing data for fast and accurate disease classification. The system developed with the intension of high accuracy and less training overhead.
- So the system initially collects and make score for every label, this partially makes an ensemble approach to improve the detection speed.
- EPCA for feature selection and dimensionality reduction

A Proposed model is generated or selected to predict the best possibility of an outcome.

III. METHODOLOGIES

Latent factor model

Data preprocessing is a data mining real world data is often incomplete, inconsistent, or lacking in certain behaviors or trends, and is likely to contain many errors in that data. Data preprocessing is one of the proven methods of resolving such issues, problems and so on. Data preprocessing prepares raw data for further processing. Proposed used latent factor model to reconstruct the missing data from the medical records collected from a hospital in different Medical unit.

PSEUDO CODE

INPUT: String S = S [1.n]
Step1: Z2 = n − 1 // Value if no mismatch found
Step2: for i ← 2 to n
Do if S (i) 6= S (1)
Then Have found a mismatch — a character not matching S (1)
Step 3: Z2 ← i − 2
Step 4:Z2 > 0
Then l2 ← 2 and r2 ← Z2 + 1
Step 5: else l2 ← r2 ← 0
Step 6: General case: 3 ≤ k ≤ n
Step 7: for k ← 3 to n
Do if rk−1 < k Have not matched S (k) yet
Then Zk = n − k + 1 Value if no mismatch found
Step 8: break

EPCA (Enhanced Principle Component Analysis)

Perhaps the most widely used algorithm for manifold learning is EPCA. The proposed system utilizes a model for disease classification and prediction. It is a combination of Principal component analysis and the non linear trick. EPCA begins by computing the covariance matrix of the $m \times n$ matrix $\mathbf{X}$WPCA steps:

The Step By Step Approach For EPCA

1. Taking the entire dataset ignoring the initial class labels
2. Find initial and starting component
3. Compute the d-dimensional mean vector values continuously Compute the covariance matrix of the original or standardized d-dimensional dataset X (here: d=3); alternatively, compute the correlation matrix values effectively.
4. Compute the eigenvectors and eigen values of the covariance matrix (or correlation matrix).
5. Sort the values in descending /ascending order.
6. Choose the k vectors that correspond to the largest values where the number of dimensions of the new feature subspace (k≤d).
7. Construct the projection matrix W from the k selected eigenvectors.
8. Transform the original dataset X to obtain the k dimensional feature subspace Y (Y=WT·X).

The important process of disease Classification and prediction is the analysis of patterns and grouping those into different subset.

Predictive model

A predictive model analysis in data mining is a procedure by which a model is generated or selected to predict the best likelihood of an outcome. In certain scenarios, the model is selected on the basis of detection theory to guess the probability of an outcome given with a group of input data. For example, given patients data, ranging from disease attributes values and classes which decide how likely a data classified. A predictive analysis in data mining, such as classification, starts from a given classification of the data items. From that it

derives a situation based on the properties of the data objects that permit to predict the association to a specific class. For example, the prediction could be based on a partitioning of the attribute values along with each measurement. Predictive data mining comprises of combining the predicted classifications from different models, or from similar type of models for the purpose of different learning data. At the same time, predictive data mining is also used to tackle the intrinsic volatility of outcome when applying composite models to compare small data sets. Suppose, if the task of data mining is to construct a model for classification of predictive types of data, and the data set that is involved in mining is

relatively small and then the data and predictive data mining is the most common type of data mining procedure.

The proposed system performs the prediction model based on the medical dataset. The proposed system successfully analyses the prediction based on the given training dataset. The system also predicts the score for the chance based on the prediction. The proposed system implements a semi supervised classifier which does not depends on the training dataset completely. The system performs the statistical properties to estimate the score of every attribute. The system finally provides the prediction accuracy over the given dataset.

Data category	**Item**	**Description**
	Patient Demographics	The Details such as Patient's gender, age, height, weight, etc.
Structured data	**Living habits**	The patient smokes, has a genetic history, etc. So, Whether these details are exposed.
	Diseases	Patient's disease, such as cerebral infarction, etc
Unstructured text data	**Patient's readme ill health**	Patient's readme sickness and medical history

Table 4.1: Item Taxonomy In Hospital Data

MDRP (**Multimodal Disease Risk Prediction**)

Proposed this model effectively use the text data to predict whereas the patient is at high risk or not. The output value is C, which indicates whether the patient is amongst the high-risk population.

C0 indicates the patient is at high-risk.

C1 indicates the patient is at low-risk.

The Structured data (S-data): Uses the patient's structured data then to predict whether the patient is at high-risk. Text data (T-data): Uses the patient's unstructured text data then to predict whereas the patient is at high-risk

Step 1: deftrain_nn_SGD (nn_structure, X, y, iter_num=3000,):

Step 2: W, b= setup_and_init_weights (nn_structure)

Cnt = 0

m = len(y)

Step 3:avg_cost_func = []

Iterations'.format (iter_num))

Step 4: while cnt<iter_num:

If cnt%50 == 0:

For i in range (len(y)):

Delta = {}

Step 5: for l in range (len (nn_structure), 0, -1):

If l == len (nn_structure):

Delta[l] =calculate_out_layer_delta(y[i,:], h[l], z[l])

avg_ += np.linalg.norm ((y[i,:]-h[l]))

Step 6: else:
If l > 1:
Delta[l] = calculate_hidden_delta (delta[l+1], W[l], z[l])
tri_b[l] = delta [l+1]
# complete the average calculation
Avg = 1.0/m * avg
avg_ _func.append (avg_value)
C += 1
Return
The output value is C, which indicates whether the patient is high-risk or Not.

IV. RESULT AND ANALYSIS

Experimental Results

This section describes the implementation process. Implementation is the realization of an application, or execution of plan, idea, model, design of a research. This section explains the software, datasets and modules which are used to develop the research. Then experimental term is performed on an Intel I3 Processor with a RAM capacity 4GB. The algorithms are implemented in Dot net and are run under Windows platform.

Data SET:

The data used in this study contains real time hospital data, and the data store in database. The dataset is general composed of structured and unstructured text data. The structured data is which includes the laboratory data and the patient's basic information such as the patient's age, gender and life habits, and then etc. Whereas, the unstructured text data includes the patient's narration of his/her illness, the doctor's interrogation records and diagnosis, etc. Patient database is collected from Disease Dataset (DD) available on the UCI Repository. The attributes considered are age: age, sex, height, weight (resting blood pressure), chol (cholesterol in mg/dl), FBS (fasting blood sugar > 120 mg/dl), smokes, blood and disease info. There are a total of 500 patient records in the database.

Metrics	**Dataset**	**Existing**	**Proposed System**
Detection Accuracy (%)	**DS1(50)**	95	99
	DS2(100)	93	98.8
	DS3(120)	93	98.5
	DS4(150)	90	98

Table 5.1 Performance Evaluation

Data Preprocessing

This phase includes extraction of data from disease Dataset in a uniform format. The step involves transforming the data, which involves removal of missing fields, normalization of data, and removal of anomalies, which refers not important data. Out of the 500 available records, 25 tuples have missing attributes. These have been excluded from the data set. For proposed system, data points were automatically centered at their mean and scaled to have unit standard deviation. No changes need be made to the data sets for EPCA.

Results And Analysis

The experiments are designed so that the different parts of the work could be evaluated. These include the evaluation of the features of the above dataset, the feature selection and also the feature creation methods. To this aim, first the features which were selected by the feature selection method named as EPCA and their importance are discussed. Second, all the four possible combinations of the feature selection and creation methods are theoretically analyzed over the dataset. Finally the performance of this proposed work Scheme was compared with the existing algorithms based on the following parameters.

- **Accuracy** – Determines the correctness
- **Precision** –Repeated process same result
- **Time taken** – Determines the processing time involved.

TP, TN, FN and FP these terms are described by Sensitivity, specificity and accuracy.

A. Accuracy:

The accuracy is measured by following formula this measured in terms of percentage.

Accuracy = (TN + TP)/ (TN+TP+FN+FP) (Number of correct evaluations)/Number of all evaluations)

B. Precision:

A class is a number of true positives (i.e. the number of instances correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class) that is called Precision.

The equation is:

Precision = TP / (TP + FP)

C. **Time taken**

This Determines the processing time involved for completed entire data set prediction process.

This experiment has been done through the hospital Dataset. The dataset is preprocessed by latent factor model and features are selected effectively and finally the prediction process is made by MDRP.

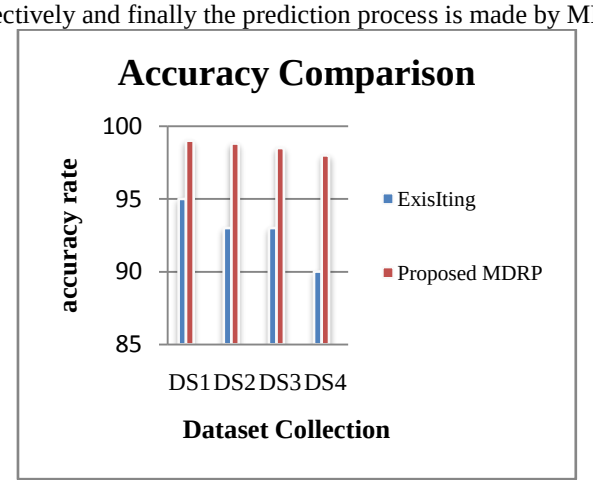

Figure 5.1 Accuracy Comparison

Performance comparison of proposed system with existing approaches based On Disease prediction Result accuracy

`From the results shown in the graphs, it can be observed that the proposed MDRP based approaches provides better accuracy and increased true positive rate when it is analyzed with different number of datasets. The system finally performs the analysis to show the accuracy of the proposed system.

Precision Comparison Chart:

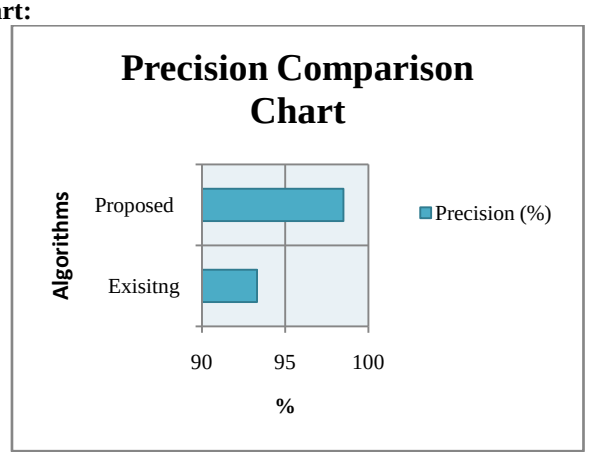

Figure 5.2 Precision comparison Graph between existing and proposed

Execution Time comparison chart:

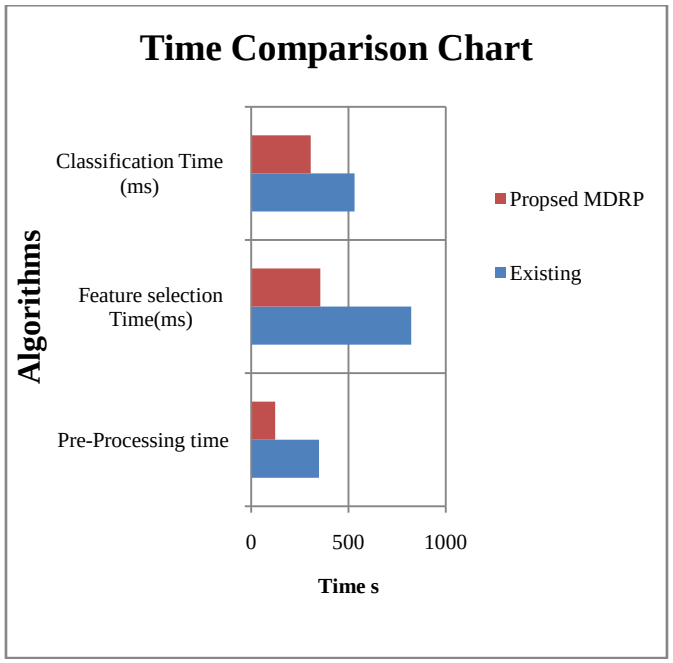

Figure 5.3 Time comparison between existing and proposed

This graph observed that the performance is very promising compared to the existing methods that have been explored in the previous term. The next term deals with the presentation of the conclusion and enhancements.

V. CONCLUSION

The study and research proposed a new classification and prediction scheme for hospital medical disease data. The system studied the main two problems in the literature survey, which are prediction accuracy and classification delay. The study overcomes the above two problem by applying the effective enhanced MDRP. The system effectively identifies and prediction the disease, the sub type which is referred as the percentage of class such as normal and disease. The experimental result shows that integrated extended proposed algorithm shows better quality assessment compared to traditional research techniques. From the experimental results, prediction accuracy of our proposed algorithm reaches 98.8% with a convergence speed which is faster than the existing system. As further work, improvements can easily be done since the coding is mainly structured or modular in nature. In the system can changing the existing modules or adding new modules can append improvements. Further enhancements can be made to the application by expanding the existing modules future research may use the model to identify the existing area of research in the field of data mining in other dataset and use of other classification algorithms. As further work, use this model as a functional base to develop an appropriate data mining system for classification performance.

REFERENCES:

[1]. Ayon Dey, Jyoti Singh, Neeta Singh. "Analysis Of Supervised Machine Learning Algorithms For Heart Disease Prediction With The Reduced Number Of Attributes Using Principal Component Analysis". International Journal Of Computer Applications (0975 – 8887) Volume 140 – No.2, April 2016.

[2]. Nagamani T, Gokul Rajhen V, Deventheran B. "A Survey Lying On Disease Prediction From Healthcare Communities Over Big Data". Ciit International Journal Data Mining And Knowledge Engineering, 2018.

[3]. Shakuntala Jatav And Vivek Sharma. "An Algorithm For Predictive Data Mining Approach In The Medical Diagnosis". International Journal Of Computer Science & Information Technology (Ijcsit) Vol 10, No 1, February 2018.

[4]. Suresh P And 2m Ananda Raj D. "Study Of Prediction And Analysis Of Prediction Model For Heart Disease: An Optimization Approach Using The Effective Genetic Algorithm". International Journal Of Pure And Applied Mathematics Volume 119 No. 16, 5323-5336, 2018.

[5]. Sreekanth Rallapalli, Suryakanthi T. "Predicting The Risk Of Diabetes In Big Data Electronic Health Records By Using The Scalable Random Forest Classification Algorithm". International Conference On Advances In Computing And Communication Engineering (Icacce). Doi: 10.1109/Icacce.2016.8073762, 2016.

[6]. Shraddha Subhash Shirsath, Prof. Shubhangi Patil. "Disease Prediction Using The Machine Learning Over Big Data". International Journal Of Innovative Research In Science, Engineering And Technology. Vol. 7, Issue 6, June 2018.
[7]. Vinitha S, Sweetlin S, Vinusha H And Sajini S. "Disease Prediction Using Machine Learning Over Big Data". Computer Science & Engineering: An International Journal (Cseij), Vol.8, No.1, February 2018.

Deepthi Krishnan. K "A Effective Disease Prediction Model Using Enhanced Principle Component Analysis And Mdrp Algorithm "IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 12, 2018, pp. 42-49