Health Care Cost Prediction using the Data Mining Approach

Radheshyam Acholiya[1]
[1]*Student, LNCT, Indore*

Amit Vajpayee[2]
Assistant Professor, LNCT, Indore

Abstract: In last ten years the human is growing much rapidly and their impact on the life styles is also observed. Rapidly changing life style of humans inviting a number of different diseases (physical and health relevant issues), in this presented work the effect of life style in the human life is investigated. During the investigation it is found for automated data analysis and prediction the data mining techniques are very useful. Thus a data mining model for health care cost estimation is proposed for design and development. Thus in order to make effective investigation a data mining based health care cost prediction model is proposed for work. This model utilizes three different kinds of set of data i.e. health attributes, user life style attributes and the cost of care. In addition of that using the classification and regression tree based data analysis is used to predict the risk level of the end user health. The CART algorithm is used to first prepare a tree using the applied attributes and then the rules from data is extracted. These rules are used to classify the end user's profile attributes in terms of risk. If the risk of issues available in classification then the hospitals data set is used for providing the suggestions of the hospital and their approximate cost of care. The system prepared in form of a web application, where the user can create their profile on the basis of considered attributes and can estimate the possible cost for their health care. The implementation of the work is demonstrated using the JSP technology. Finally using different experiments the performance of the system is also computed. The results demonstrate the proposed health care cost prediction is acceptable for the real world use.

Keywords: data mining, health care, cost prediction, CART, non-linear regression

I. INTRODUCTION

Analyzing and predicting the knowledge from the big data plays a critical role in the real world environment which needs to be concerned more to provide an efficient treatment for the health care patients. Handling and treating the patients online may generate large volume of data dynamically. These large volume of data need to be handled more carefully to predict accurate result. The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year [1]. The ability to use these data to extract useful information for quality healthcare is crucial.

Data Mining is one of the most vital and motivating area of research. Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. In health industry, Data Mining provides several benefits such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals etc. [2].

However, in this thesis we look at various data-mining tools, as all data is considered as simple data, to perform automatic vote based classification on different four datasets and also provide accuracy and other performance parameter in terms of percentage with regard to the number of classification rule of dataset that were classified correctly.

II. PROPOSED WORK

Health is wealth, everybody cares about self and his/her family health. On the other hand the health care cost is increases rapidly. In this presented work a health care cost prediction model using the CART algorithm presented which is able to predict the approximated cost of health care and the budget. This chapter offers the design and their functional aspects of prediction and data modeling.

A. System Overview

Prediction is a technique of computing values on the basis of historical data analysis. Here the term history is used for denoting the previous experience or available data which is previously collected from real world examples. Basically the computational algorithms are implemented for analyzing the data and recovering the patterns or logics for identifying the similar kinds of data in newly appeared data. That process is called pattern recognition or classification of data. This is a part of data mining stream and that is the work of supervised learning in data mining. Supervised learning techniques are applied in such conditions where some predefined data patterns are available to learn with. In this presented work the data mining and their supervised learning concept is studied and an application of data mining is health care cost prediction is proposed for simulation and modeling.

Data mining and their applications are growing in various domains i.e. business intelligence, production, marketing and others. But there is very less work available for health care industry at the user or client point of view. Most of the time the health care industries are usages the mining techniques for finding the client records, their visit patterns and other works. Therefore to provide the advantage of data mining techniques capability for managing the health of self and their loved once a data model using the data mining technique is presented in this work. The proposed technique involves the two different kinds of data analysis based on daily life style and user's health profile. By using CART decision tree rules from both the dataset is extracted, in form of "if then else". These rules are utilized with the nonlinear regression and query data for finding the decisions of the health care cost. This section provides the basic overview of the proposed work. Next section provides the detailed understanding about the data modeling.

B. Methodology

The proposed working model for health care cost prediction is demonstrated in figure 2.1. Additionally the functional aspects of the proposed model are also described in detail. Each component of the given system is working as individual unit and generates the outcome for next.

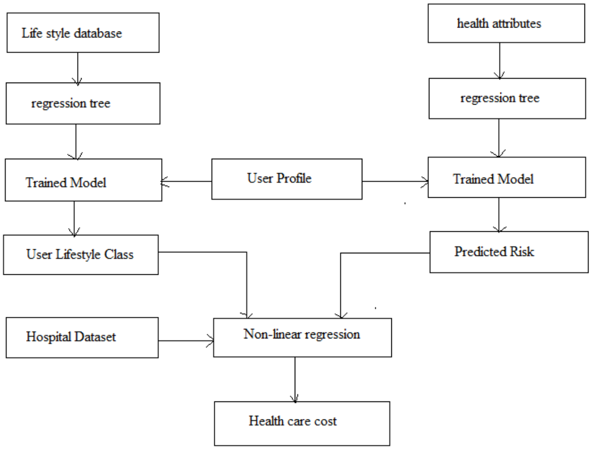

Figure 2.1 proposed system architecture

Life style dataset: the daily routine of users also impact on the health of the person. Therefore a survey on life style and their fitness is performed. Using this survey 900 instance of daily life style data is collected from working/ non-working and males/females. In addition of their current health status is also noticed during the survey. The collected data is used as the life style dataset that is first part of information collection.

Health attributes: that is the second and essential part of data collection. According to the suggestions in this phase the using the UCI repository the standard datasets are collected for different deceases. For experimentation and design aspects the heart and diabetic dataset is considered.

Regression tree: basically the CART decision tree is a binary tree which works on the basis of splitting of data in two parts. The basic idea of tree growing is to choose a split among all the possible splits at each node by which resulting child nodes becomes "purest". Each split depends on the value of only one predictor variable. All possible splits consist of possible splits of each predictor. If X is a nominal categorical variable of I categories, there are $2^{I-1} - 1$ possible splits. If X is an ordinal categorical or continuous variable with K different values, there are K-1 splits on X. A tree is starting from a root node by repeatedly using the following steps.

1. Find each predictor's best split.

For each continuous and ordinal predictor, sort its values from the smallest to the largest. For the sorted predictor, go through each value from top to examine each candidate split point (call it v, if x ≤ v, the case goes to the left child node, otherwise, goes to the right.) to determine the best. The best split point is the one that maximize the splitting criterion the most when the node is split according to it. The definition of splitting criterion is Gini index. The Gini impurity measure at a node t is defined as:

$$Gini(t) = \sum_{i,j} C(i|j)P(i|t)P(j|t)$$

The Gini splitting criterion is the decrease of impurity defined as:

$$\Delta Gini(s,t) = Gini(t) - P_l\, Gini(t_l) - P_r\, Gini(t_r)$$

Where P_l and P_r are probabilities of sending a case to the left child node t_l and/or right t_r. That is estimated using:

$$P_l = \frac{P(t_l)}{P(t)}$$

And

$$P_r = \frac{P(t_r)}{P(t)}$$

For each nominal predictor, examine each possible subset of categories (call it A, if x ∈ A, the case goes to the left child node, otherwise, goes to the right.) to find the best split.

2. Find the node's best split.

Among the best splits found in step 1, choose the one that maximizes the splitting criterion.

3. Split the node

Using its best split found in step 2 if the stopping rules are not satisfied.

Trained model: after processing of data using the CART (classification and regression tree) algorithm the data is demonstrated using the tree structure. An example of CART generated tree is demonstrated using figure 2.2. In the given decision tree the data set is modeled in form of tree structure. In addition of that these tree structure can also converted into the "if then else" rules. The representation of the above given tree in form of "if then else rule" is given as:

```
1. If weight = = heavy then
    a. High mileage
2. Else
    a. If horse power < = 86 then
        i. High mileage
    b. Else
        i. Low mileage
    c. End if
3. End if
```

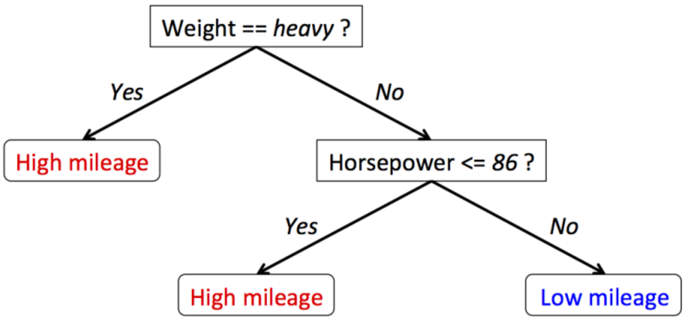

Figure 2.2 decision tree example

User profile: using both the kinds of input dataset the CART algorithm prepare the decision tree. Additionally the rules from the data model are also extracted. Now that is the final input for the proposed system as the individual user profile which is required to evaluated for finding the health care cost. Thus in this phase the user profile attributes are supplied for performing classification.

User life style class: the user life style based decision tree is used for classifying the user's personal life style attributes and a class label is generated.

Predicted risk: in the similar manner the health attribute of user is classified according to the health attribute based tree.

Non-linear regression: Nonlinear regression extends linear regression for use with a much larger and more general class of functions. Almost any function that can be written in closed form can be incorporated in a nonlinear regression model. Unlike linear regression, there are very few limitations on the way parameters can be used in the functional part of a nonlinear regression model. The way in which the unknown parameters in the function are estimated, however, is conceptually the same as it is in linear regression. The basic form of non-linear regression model is given as:

$$y = f(\vec{x};\vec{\beta}) + \varepsilon$$

Where,

1. The functional part of the model is not linear with respect to the unknown parameters, β_0, β_1, …, and
2. The method of least squares is used to estimate the values of the unknown parameters.

Hospital dataset: that is additionally prepared dataset which contains the list of different hospitals and their treatment cost. If the significance of a person found as positive then the health care cost is suggested according to the preference of user.

Health care cost: that is the final outcome of the system which contains the cost suggested from different hospitals and by different doctors.

C. Proposed Algorithm

This section provides the introduction about proposed algorithm, where each of the components are described as the process involve for optimizing the data according to the need of application. The table 2.1 contains the proposed algorithm which defines the proposed system.

```
Input: health attribute dataset H, life style Dataset L, user profile attributes U, Doctors dataset D
Output: health care cost C
Process:
    $RH = readDataset(H)$
    $HT_{model} = CART.makeTree(RH)$
    $RL = readDataset(L)$
    $LT_{model} = CART.makeTree(RL)$
    $L_{rule} = ExtractRules(LT_{model})$
    $H_{rules} = ExtractRules(HT_{model})$
    $UL_{class} = L_{rule}.InvokeRules(U)$
    $UH_{class} = H_{rule}.InvokeRules(U)$
    $Prisk = Regression.predict(UH_{class}, UL_{class})$
    $C = Prisk.SuggestCost(Prisk, D)$
    Return C
```

Figure 2.1 proposed algorithm.

III. RESULTS ANALYSIS

This chapter provides the details about the results analysis of the proposed machine learning based health care cost prediction system. Therefore different parameters that are evaluated during experimentation are reported in this section.

A. Accuracy

Accuracy measurement provides the correctness of the data model for classification or prediction. The accuracy is basically a ratio among the correctly distinguished pattern by the classifier among the total patterns provided for classification. That can be estimated using the formula given below.

$$accuracy = \frac{correctly\ classified\ pattern}{total\ pattern\ to\ classify} X100$$

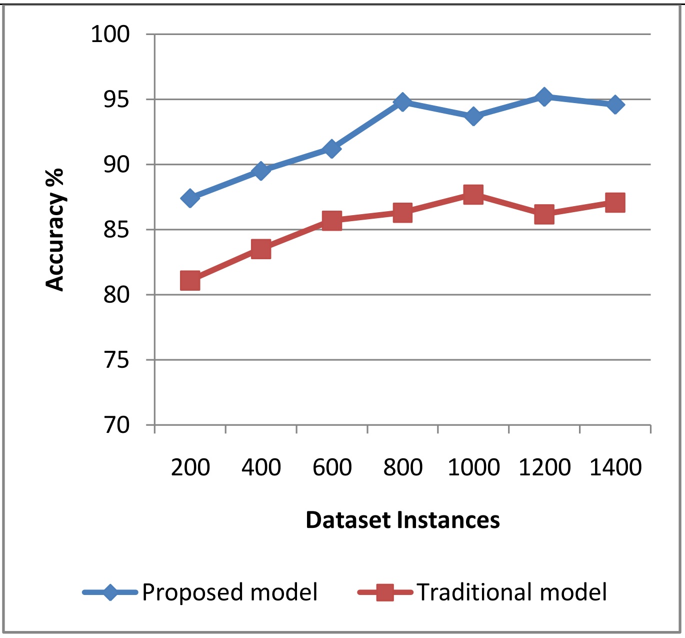

Figure 3.1 accuracy

Dataset Instances	Proposed model	Traditional model
200	87.4	81.1
400	89.5	83.5
600	91.2	85.7
800	94.8	86.3
1000	93.7	87.7
1200	95.2	86.2
1400	94.6	87.1

Table 3.1 accuracy

The performance of the proposed and traditional health care cost prediction data model is represented in table 3.1 and the figure 3.1. The given performance of the system is estimated in percentage values. For graphical representation of the performance the X axis of graph contains the instances of data for analysis and the Y axis shows the percentage accuracy value of the proposed and traditional model. According to the results with the different instances of data the accuracy of the proposed model is clearly higher enough as compared to the traditional linear regression based data model.Thus the proposed system is acceptable for real world use of application.

B. Error Rate

The error rate is amount of data which is misclassified or misrecognized during the classification operations. The error rate is measurement of incorrectness of a classification data model. The error rate of algorithm is computed using the following formula:

$$Error\ Rate = 100 - accuracy$$

Or

$$Error\ Rate = \frac{incorrectly\ classified\ pattern}{total\ pattern\ to\ classify} X100$$

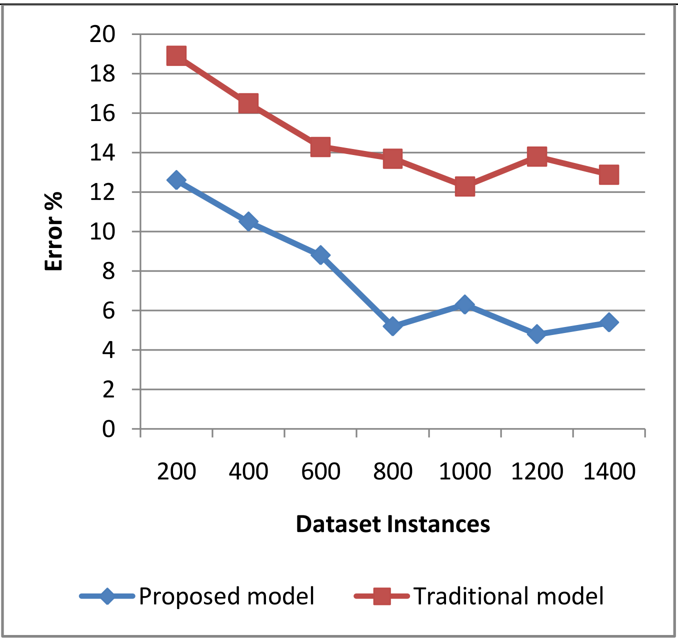

Figure 3.2 error rate

Dataset Instances	Proposed model	Traditional model
200	12.6	18.9
400	10.5	16.5
600	8.8	14.3
800	5.2	13.7
1000	6.3	12.3
1200	4.8	13.8
1400	5.4	12.9

Table 3.2 error rate

The percentage error rate of the proposed and traditional health care cost prediction algorithms are described in figure 3.2 and table 3.2. The Y axis of the graph demonstrates the computed error rate of the algorithm and the X axis shows the dataset instances for classification. According to results computed the proposed non-linear regression based technique demonstrates higher accurate outcomes and low error rate as compared to the traditional linear regression based data model. Thus the proposed technique is efficient and accurate as compared to the previously available technique.

C. Memory Usage

In any computational system when the computations performed the data is stored on the main memory and utilized by the process. The amount of space required to store data in main memory is termed as memory usage or space complexity of algorithm.

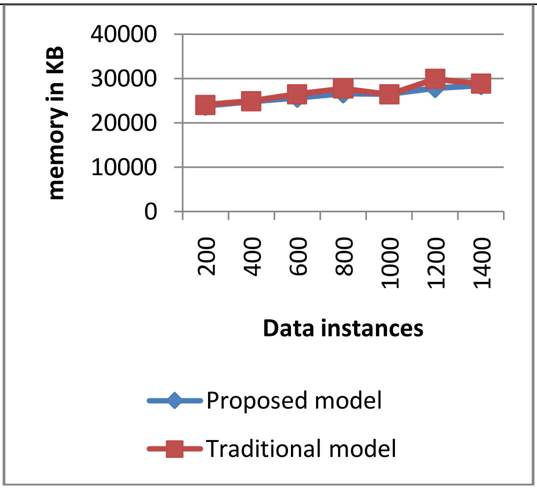

Figure 3.3 memory usage

Dataset Instances	Proposed model	Traditional model
200	23845	24083
400	24837	24894
600	25648	26451
800	26615	27746
1000	26471	26448
1200	27817	29911
1400	28472	28817

Table 3.3 memory usage

The memory requirements of the proposed and traditional technique for health care cost prediction are demonstrated using figure 3.3 and table 3.3. The blue line of this diagram shows the performance of proposed approach and the red line demonstrates the performance of traditional technique of cost prediction. To visualize the performance the X axis consist of the data instances provided for classification and prediction additionally the respective obtained memory requirements is given in Y axis. The measurement of memory is performed in terms of KB (kilobytes). According to the obtained performance both the algorithms demonstrate the similar amount of memory consumption but sometimes the traditional approach requires additional amount of space. Thus proposed technique is cost effective than tradition technique.

D. Time Consumption

The process requires sometime for analyzing the data according to the data models and input size of data. This time requirements of algorithm are known as the time complexity or the time requirements of the algorithm. The time is measured using the following formula:

$$time\ required = end\ time - start\ time$$

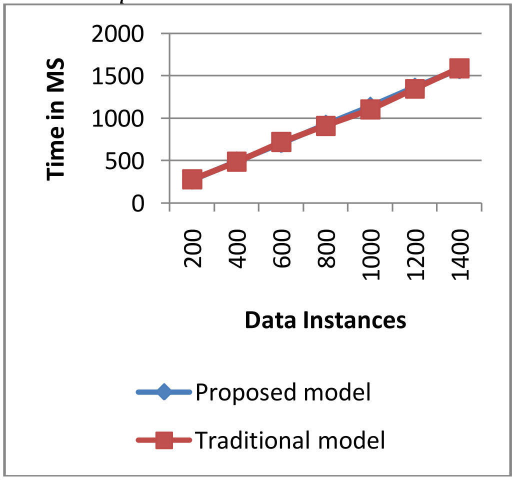

Figure 3.4 time consumption

Dataset Instances	Proposed model	Traditional model
200	273	278
400	491	485
600	707	719
800	924	911
1000	1137	1103
1200	1362	1348
1400	1578	1591

Table 3.4 time consumption

The time requirements of the algorithms namely proposed and traditional algorithm is demonstrated using the table 3.4 and figure 3.4. The given time is measured here in milliseconds (MS). The blue line of the graph shows the time consumption of the proposed algorithm for health care cost estimations and red line shows the time requirements of traditional data model. According to the results both the techniques requires similar amount of time and the requirements of time is increases in each steps.

IV. CONCLUSION AND FUTURE WORK

This chapter addresses the conclusion of the presented work in this research work. Therefore to justify the objectives of the proposed work the summary of the work is presented in this chapter. In addition of that the possible future extensions are also included in this chapter.

A. Conclusion

Data mining techniques are enabling us for analyzing the data and obtaining data patterns according to the application usage. These patterns are helpful for recognizing the similar kinds of other or new patterns from raw data or classifying the patterns in some defined groups. Additionally sometimes these techniques are also used for the predicting the approximate values for the given problem.

In this presented work the data mining technique is used for preparing the working model for health care cost prediction. The concept is based on the health care cost is affected according to the user’s life style. Therefore initially a UCI repository based dataset for health attribute is considered. From these datasets the diabetic dataset and heart dataset is considered. On the other hand for finding the effect of life style on the health some user life style attributes are also considered and a database for 900 users is created. These two data is initially used for training of the decision tree classifier. Additionally using obtained classes of the classifier is used to find likely attributes from the datasets. This task is performed using the nonlinear regression technique additionally that produces the risk of the disease. If risk is found the third input is used as the hospital dataset which contains the list of hospitals and the cost of health care. That is suggested as the final outcome of the system.

The implementation of the proposed technique is performed as the web application. Therefore the JSP (JAVA server Pages) are involved as development technology. After the implementation of data mining based model the performance is also measured and compared with the traditional model. That comparison is summarized using table 6.1.

S. No.	Parameters	Proposed	Traditional
1	Time consumption	273-1578 MS	278-1591 MS
2	Memory usages	23845-28742 KB	24083-28817 KB
3	Error rate	4.8-12.6 %	12.3-18.9 %
4	Accuracy	87.4-95.2 %	81.1-87.7%

Table 6.1 performance summary

According to the obtained experimental results the proposed technique is suitable for utilized in real world applications for predicting the approximate cost of health care for the specific domains.

B. Future Work

The main aim of improving the traditional design of the health care cost prediction is accomplished successfully. In order to improve the technique the proposed work can be extended in the following manner.

1. Currently the proposed system involve only two kinds of issues relevant to health in near future required to involve various different kinds of disease also

2. There are some improvement on the existing methodology can also be required to improve the classification accuracy
3. Involve the big data and cloud platform for involving large volume of data and user's profile.

REFERENCES

[1] Huang, H. et al. "Business rule extraction from legacy code", Proceedings of 20 th International Conference on Computer Software and Applications, IEEE COMPSAC'96, 1996, pp.162-167

[2] Han, J. Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006

[3] Marquardt, Ames, et al. "Healthscope: An interactive distributed data mining framework for scalable prediction of healthcare costs", 2014 IEEE International Conference on Data Mining Workshop (ICDMW), IEEE, 2014.

[4] Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms, Cambridge University press, 2014.

[5] Nilsson, Nils J. "Introduction to machine learning, An early draft of a proposed textbook." (1996).

[6] "Machine Learning: What it is and why it matters", available online at: http://www.sas.com/en_us/insights/analytics/machine-learning.html

[7] Alex Smola and S. V. N. Vishwanathan, "Introduction to Machine Learning", Yahoo! Labs, Ph.D thesis, Cambridge University Press.

[8] Kotsiantis S.B. Supervised machine learning: a review of classification techniques. Informatica 31:249–268, 2007.

[9] Duda, R. O., Hart, P. E., and Stork, D. G. Pattern Classification, Wiley-Interscience, 2nd edition, (2001).

[10] Ghahramani, Z. Unsupervised learning. In O. Bousquet, G. Raetsch, & U. von Luxburg (Eds.), Advanced lectures on machine learning. Berlin: Springer-Verlag (2004).