An Integrated Architecture for Prediction of Heart Disease from the Medical Database

Gunasekar Thangarasu[1], Kayalvizhi Subramanian[2] and P. D. D. Dominic[3]
[1] Faculty of Engineering and Technology and [2] Faculty of Built Environment
Linton University College, Seremban, Malaysia.
[3] Department of Computer & Information Sciences, Universiti Teknologi PETRONAS,
32610 Seri Iskandar, Perak Darul Ridzuan, Malaysia.
[1] dr.t.gunasekar@ktg.edu.my, [2] kayalvizhi@ktg.edu.my, [3] dhanapal_d@utp.edu.my

***Abstract*—a database is a collection of data organized for storage, access and retrieval. With increasing growth in big data, especially in healthcare and biomedical communities, the techniques to analyze the medical data tends to benefit the patients by detecting the disease early. However, with the advent of incomplete data in such medical datasets, the quality tends to reduce. In addition to this, each region has its own unique disease characteristics, which further reduces the prediction quality. To overcome the difficulty in processing the medical datasets with incomplete data, the proposed method initially reconstructs the missing or incomplete data. To improve the processing capability of automated disease prediction in an uncertain environment, an integrated architecture is proposed. It controls the processing capability of medical datasets in an uncertain environment. This integrated diagnostic model generates the hesitant fuzzy based decision tree algorithm using genetic classification. The architecture is designed to process both the structure and unstructured data sets. The new and innovative prediction methods are projected in this research to predict heart disease from the medical database in a faster manner.**

Keywords—Hesitant fuzzy algorithm, Genetic algorithm, Rule optimization, Heart disease, Classification

I. INTRODUCTION

With the increasing development in the medical data, there is a substantial increase in the electronic health records. As the data is increasing in its size, the prediction of heart disease is being a major consideration in big data analytics. The classification algorithms have been developed to increase the classification accuracy in medical diagnosis [1]. With the development of machine learning algorithms, the classification process in big data analytics takes a lead in classifying the datasets as per the diagnostic application.

The Big data analytics utilizes efficient techniques for finding the insights, correlation and hidden patterns from the datasets. It leads to significant reduction of cost, improves the decision making performance and forms new items to meet the demands of customers'. Hence, with wide advantages, it addresses the problems in different applications like healthcare, plants and Bioinformatics [2]. The problems are usually addressed by the machine learning algorithms, decision making tree patterns and the formation of rules. However, most of the classification algorithms in the big data processing consider only the structured data. The processing of unstructured data is usually carried out by combining the structured and unstructured information [3, 4] that reduces the risk of heart disease prediction. The combination of both the information makes the processing deliberate; since the redundant data is present more while processing the information.

A. Problem Statement

Medical databases are widely used nowadays by many medical specialists and also new databases are rapidly emerging. Medical databases are a wide concept extending over many different medical tasks. With advent of incomplete data in such medical datasets, the quality tends to reduce. In addition to this, each region has its own unique disease characteristics, which further reduces the prediction quality.

B. Aim of the Research

The main aim of this research is to propose a method initially reconstructs the missing or incomplete data. To improve the processing capability of heart disease prediction in an uncertain environment, an integrated architecture is proposed. The early recognition and prediction can give a warning at a stage, where some medications and precautionary action can facilitate the patient to increase the period of patient's healthy life.

C. Objectives of the Research

The objectives of this research are:

- To propose a Fuzzy based decision tree system to classify the unstructured datasets from the medical database.
- To propose a Genetic algorithm used to increase the rule based decision making in a large unstructured medical database.

D. Meaning of Heart Diseases

Heart disease is an umbrella term for any type of disorder that affects the heart. Heart disease means the same as cardiac disease but not cardiovascular disease. According to WHO (World Health Organization) [5], heart disease is the leading cause of death in the UK, USA, Canada and Australia. The number of US adults diagnosed with heart disease stands at 26.6 million. 23.5% of all deaths in the USA today are caused by heart disease. Some of the most common heart diseases are Angina, Arrhythmia, Congenital heart disease, Coronary artery disease, Dilated cardiomyopathy, Myocardial infarction, Heart failure, Hypertrophic cardiomyopathy, Mitral regurgitation, Mitral valve prolapsed, Pulmonary stenosis.

II. REVIEW OF LITERATURE

It is prominent to classify the unstructured data in a separate manner using classification algorithms. This reduces the risk of predicting the disease based on the trained classifiers. Istephan, et al. [6] proposed structured data processing method for un-structured medical image data using three phases in a Single Server Architecture framework. Gardner, et al. [7] proposed an integrated framework to structure the medical text documents using a simple Bayesian classifier with a sampling and conditional random field classifier to extract the attributes. Further, k-anonymization technique is used to de-identify the data extracted, which ensures maximum data utility. Schmidt, et al. [8] proposed a novel faceted search tool to identify the correlations to structure the unstructured medical data. Kalantari, et al. [9] proposed Support Vector Machine based Artificial Immune Recognition System and utilized SVM-Genetic Algorithm, SVM-Artificial Immune System and fuzzy support vector machine to compare the performance of unstructured data classification. Here, the proposed technique attains better results than the other SVM based techniques. Further, instance-based algorithm [10], Latent Dirichlet Allocation with Topics over Time [11], linguistic rules is used to extract the temporal features in cancer and nosocomial infections applications [12], decision models with a dictionary or non-dictionary features [13], natural language processing [14] and semantic relationships [15].

III. RESEARCH METHODOLOGY

The previously used techniques fail to use rule based system to classify the medical datasets. Hence, the processing of datasets using rule based system reduces the use of pre-processing operation to eliminate the redundant data. Since, the rule based system can itself eliminate the redundant data with its rule base and structures well the unstructured data. Also, the optimization of the rule set is needed to improve the risk of classification accuracy.

In this research, propose a fuzzy based decision tree system to classify the unstructured data sets and hence to improve the novelty of the proposed work. In addition, genetic algorithm is further used to increase the rule based decision making in large unstructured medical datasets. Initially, the fundamentals of hesitant fuzzy algorithm are specified and then the introduction of Fuzzy based Decision Tree algorithms is discussed. Finally, brief discussions on discretization methods are proposed.

A. Basic Concepts of Hesitant Fuzzy Sets

Hesitant fuzzy sets have been extensively used in the literature since they effectively incorporate the different expert's ideas on a single set. Given three hesitant fuzzy elements represented by h^c, h_1, h_2, [16] defined some operations on them, which can be described as:

1. $h^c = \bigcup_{\gamma \in h}\{1-\gamma\}$.
2. $h_1 \cup h_2 = \bigcup_{\gamma_1 \in h_1, \gamma_2 \in h_2} \max\{\gamma_1, \gamma_2\}$.
3. $h_1 \cap h_2 = \bigcup_{\gamma_1 \in h_1, \gamma_2 \in h_2} \min\{\gamma_1, \gamma_2\}$.

B. Data Balancing

The imbalanced or incomplete datasets are balanced using (i) under- and (ii) over-sampling methods. The former method reduces the major class samples and hence the useful information is ignored. The latter method increases the minor class samples and hence it may undergo over-fitting problem. Thus, the major class samples can be eliminated by converting it into several classes using the process of clustering. Initially, k-means algorithm is used to obtain multiple clusters from the major class samples. The clusters are labelled or numbered to construct several pseudo classes from the new balanced dataset.

C. Hesitant Fuzzy Algorithm

The Hesitant Fuzzy Algorithm (HFA) is an improved version of fuzzy that removes the hesitation provoked due to the assignment of fuzzy sets with the membership degree of an element [17]. An HFA is defined in terms of a function that returns a set of membership values for each element in the domain.

In many medical problems, sometimes, it is difficult to define a membership grade of an element, because of a set of possible membership values [18]. This issue is very important in heart disease prediction problems, when perform heart disease predictions do not support the same membership grade for an element [19]. To deal with these cases, the hesitant fuzzy algorithm was introduced as a new generalization of fuzzy sets. The preliminaries of the HFA assumed are given below.

Definition 1 [20]**:** For a reference set (**Y**), the Hesitant fuzzy sets of D is represented as a function $hD(y)$ and if Hesitant fuzzy sets are applied over the reference set, **Y** returns a subset [0, 1].

$$D = \left\{ \left\langle y, h(y) \right\rangle \middle| y \in Y \right\} \tag{1}$$

The membership degree of element (y to A) is represented using the set of different values, $hD(y)$, where $y \in$ **Y**. For simplicity, the $h(y)$ is called as an element of hesitant fuzzy sets.

***Definition 2*:** Let **D** be the cardinality of discourse with membership value (l_i) and the member of **D** with elements (j) of discourse members (i) of D is $h^2_{D_\sigma(j)}(y_i)$, where the value

of D contains the information energy from hesitant fuzzy sets, which is defined as:

$$E_{sets}(D)=\sum_{i=1}^{n}\left(\frac{1}{l_i}\sum_{j=1}^{l_i}h^2_{D_\sigma(j)}\left(y_i\right)\right) \qquad (2)$$

D. Proposed Research Architecture

This section deals with the proposed research architecture, inclusive of hesitant fuzzy decision tree algorithm and then it deals with a genetic algorithm to optimize the rules. The architecture of the proposed system is shown in Figure.1

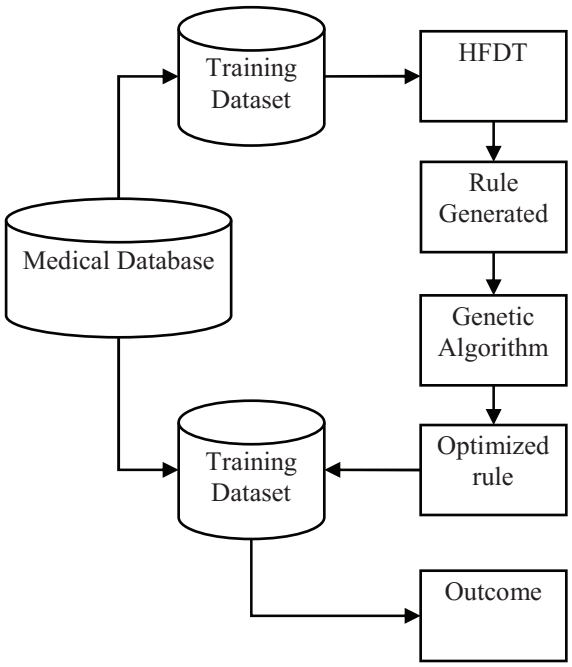

Fig. 1. Proposed Research Architecture

TABLE.1. NOTATIONS

Symbols		Descriptions
$\{A_1,..., A_k, Y\}$	-	Set of dataset attributes
A_i	-	i^{th} attribute
k	-	Total input attributes
Y	-	Target attributes
y_i	-	Value of Y
$\{A_1,..., A_k\}$	-	Set of attributes
$Y \in \{c_1,...,c_m\}$	-	Class labels or the values of target variable set
m	-	Total number of classes
$S = \{(X_1,\mu_s(X_1)),...,(X_n,\mu_s(X_n))\}$	-	Fuzzy dataset
$\mu_s(X_i)$	-	Membership degree of vector X_i with input and target attributes
$\left[x_i^{(1)},...,x_i^{(k)},y_i\right]^T$	-	$x_i^{(k)}$ is the k attribute values
n	-	Total instances
$\left\{F_i^{(1)},...,F_i^{(r_i)}\right\}$	-	$F_i^{(r_i)}$ is the fuzzy term of r_i
r_i	-	Total fuzzy terms defined by A_i
$\mu_{F_i}(j)$	-	Fuzzy membership function
$\lvert S\rvert=\sum_{i=1}^{n}\mu_s(X_i)$	-	Total instances in Fuzzy dataset (S)
$S_{y=c_i}$	-	The instance S belonging to i^{th} class
$F_i^{(j)}$	-	Fuzzy term
$S\left[F_i^{(j)}\right]$	-	Child node Fuzzy set when a parent node fuzzy set is S

E. Construction of Fuzzy Decision Tree

The instances are allowed to follow multiple branches with difference membership degree using FDT that ranges from [0, 1]. Fuzzy linguistics are used to specify the conditions of branches emerging from the nodes. The instances or leaves may fall with different membership degrees, since it falls under different child nodes at any level. In case of incomplete information or during the presence of noise, the falling behavior of leaf is quite an advantageous one. However, FDT is slower than a normal decision tree, but it provides better classification accuracy.

In general, the FDT consist of tree construction and inference from decision making. The FDT algorithm is an extension of fuzzy logic with ID3 algorithm, which employs linguistic terms that fuzzifies the attributes the medical training data. The information gain measure is used to determine the attributes of the branched node. Further, it uses the fuzzy dataset with the inclusion of membership degree, input and target attribute. The child node dataset contains the entire instances of parent nodes where the elimination of branching attributes takes place. Further, the major difference takes place in a fuzzy membership degree of entire instances.

Consider a fuzzy dataset (S) with the branching attribute (A_i), fuzzy terms, $\left\{F_i^{(1)},...,F_i^{(r_i)}\right\}$, fuzzy term of child nodes $S\left[F_i^{(j)}\right]$. Thus the membership degree of the ith instant is $X_e = \left[x_e^{(1)},...,x_e^{(k)},y_e\right]^T$ in child node fuzzy terms is given by,

$$\mu_{S\left[F_i^{(j)}\right]}\left(X_e\right)=\mu_{F_i^{(j)}}\left(x_e^{(i)}\right)\times\mu_S\left(X_e\right) \qquad (3)$$

where

$\mu_S\left(X_e\right)$ - membership degree of X_e

$\mu_{F_i^{(j)}}\left(x_e^{(i)}\right)$ - membership degree of $x_e^{(i)}$ in relevance with $F_i^{(j)}$

This algorithm considers the branching attribute based on the maximum value of Fuzzy Information Gain, where the fuzzy entropy is given by,

$$E(S)=\sum_{i=1}^{m}-\frac{\left|S_{y=c_i}\right|}{|S|}\log_2\frac{\left|S_{y=c_i}\right|}{|S|} \quad (4)$$

The attribute (A_i) from the information gain (I_G) relevant to the S is given by:

$$I_G(A_i,S)=E(S)-\sum_{j=1}^{r_i}w_j E\left(S\left[F_i^{(j)}\right]\right) \quad (5)$$

Where,

$E(S)$ – entropy of S

$E\left(S\left[F_i^{(j)}\right]\right)$ – entropy of the child node (j) or the expected value of S after the partitioning of S using A_i

w_j – instances of the child node (j), which is given by,

$$w_j=\frac{\left|S\left[F_i^{(j)}\right]\right|}{\sum_{j=1}^{r_i}\left|S\left[F_i^{(k)}\right]\right|} \quad (6)$$

F. Algorithm for construction of Hesitant Fuzzy Decision Tree

Inputs: Training data, membership function, threshold value of stopping criteria, split and stop criteria.

1. The membership function of the training data is set to one
2. Root node is generated with fuzzy set
3. For new node (N)
4. If stop criteria is reached, then
5. Consider N as a leaf
6. Fraction of N records belonging to each class is labelled.
7. Else if the stopping criteria is not reached, then,
8. Calculate the fuzzy information gain over each attribute.
9. Attribute with maximum information gain is chosen as branching attribute
10. Generate Child nodes, where fuzzy term contains the dataset with an entire dataset attributes other than S.
11. End
12. End

The construction process of HFDT is shown in the above algorithm. The construction of HFDT utilizes stopping criterion, as in line 3. This method uses normalized maximum I_G - 28 as a stopping criterion.

The HFDT performs the construction of decision tree using two discretization methods, where the hesitant fuzzy system for a given attribute A_i is given by,

$$B=\left\{\left\langle A_i,e_B(A_i)\right\rangle\middle|A_i\in A\right\} \quad (7)$$

$$e_B(A_i)=\left\{I_G\left(f_{\text{in}}(A_i,S)\right),I_G\left(f_{uf}(A_i,S)\right)\right\}$$

Finally, the HI_G in relevance with the fuzzy dataset (S) over an attribute A_i is given by,

$$HI_G(A_i,S)=0.2\left(I_G\left(f_{\text{in}}(A_i,S)\right)^2+I_G\left(f_{uf}(A_i,S)\right)^2\right) \quad (8)$$

Here,

$F(A_i,S)$ is the expected fuzzy entropy reduction caused by A_i.

$f_{in}(A_i,S)$ is the expected reduction in merging discretization fuzzy entropy caused by A_i.

$f_{uf}(A_i,S)$ is the expected reduction in uniform frequency fuzzy entropy caused by A_i.

The Eq.(8) is calculated based on two different information gain energy values, *(i)* merging discretization and *(ii)* uniform-frequency discretization method. The former one is a bottom up approach and considers the cut-points from the complete continuous values. Then the intervals are merged and the cut-points are eliminated from the final discretization process. While, the latter one considers a generated n intervals from a parameter n of equal instances and forms a supervised discretization process. The node is selected based on the fuzzy information gain approach and it uses two discretization methods to construct the fuzzy discretization method.

IV. EXPECTED RESULT

In this research, the author proposes a novel classification technique, which embeds the HFDT with genetic algorithm for rule optimization to improve the risk prediction of heart disease. The proposed algorithm will be tested over unstructured data from the heart disease datasets. To the best of our knowledge, none of the conventional work in the medical data diagnosis has deployed the rule optimization procedure to improve the classification accuracy. Hence, this work will be considered as a novel in the field of medical data analytics. Comparing with conventional prediction algorithms, the proposed method attains a better likelihood with the diagnosis of clinicians. Further, the proposed work can be improved by utilizing meta-heuristic approaches to optimize the rule set of the decision tree.

A. Theoretical Contributions

Hesitant Fuzzy based Decision tree Algorithm (HFDT) is used to classify the un-structured instances related to the given medical diagnosis. In addition, the Genetic algorithm (GA) is utilized to improve the classification rule set of the fuzzy decision tree algorithm, depending on the optimized rules from the training set using FDT with GA.

B. Community Value Contributions

The early recognition and prediction can give a warning at a stage, where some medications and precautionary action can facilitate the patient to increase the period of patient's healthy life.

V. CONCLUSION

Our current research work provides the comprehension into the design of new framework using combined of Hesitant Fuzzy based Decision tree Algorithm and Genetic algorithm for predicting heart disease. The vision of this research is to identify and develop methods and procedures for predicting heart disease that assist medical practitioners in an efficient way for people getting longer life in this world.

A. Limitations

Hesitant Fuzzy based Decision tree Algorithm and other techniques are proposed in this research capable of tolerating a certain level of noise data. However, the noise data sometimes lead to misleading the result when people health conditions are more complex.

B. Future Work

Our future work will explore the development of tool and design strategies that encourage progression. Because human health is so complex and is not solely based on individual behaviour and lifestyle, it covers Genetic and environmental factors also. With the productive implementation of this proposed work, the constant technique might even be applied in identifying different illness like diabetic disease and hypertension etc.

REFERENCES

[1] B. Qian, X. Wang, N. Cao, H. Li, & Y.G. Jiang, "A relative similarity based method for interactive patient risk prediction", Data Mining and Knowledge Discovery, 2015, Vol. 29, No. 4, pp.1070-1093.

[2] M.S. Gharajeh, "Biological Big Data Analytics", Advances in Computers, 2017, pp.321-355.

[3] J.R. Vest, S.J. Grannis, D.P Haut, P.K Halverson, & N. Menachemi, "Using structured and unstructured data to identify patients' need for services that address the social determinants of health", International Journal of Medical Informatics, 2017, Vol.107, pp.101-106.

[4] A. Fong, A.Z. Hettinger, & R.M. Ratwani, "Exploring methods for identifying related patient safety events using structured and unstructured data", Journal of biomedical informatics, 2015, Vol.58, pp.89-95.

[5] C. Nordqvist, "Heart Disease: Definition, Causes, Research", Medical NewsToday, 2016.

[6] S.Istephan, & M.R. Siadat, " Unstructured medical image query using big data–an epilepsy case study", Journal of biomedical informatics, 2016, Vol.59, pp. 218-226.

[7] J. Gardner, & I, Xiong,, " An integrated framework for de-identifying unstructured medical data", Data & Knowledge Engineering, 2009, Vol.68, No.12, pp.1441-1451

[8] D. J. Schmidt, K. Budde, D. Sonntag, H.J Profitlich, M. Ihle, & O. Staeck, " A novel tool for the identification of correlations in medical data by faceted search", Computers in Biology and Medicine, 2017, Vol.85, pp.98-105.

[9] A. Kalantari, A. Kamsin, S. Shamshirband, A. Gani, H. Alinejad-Rokny & A.T. Chronopoulos,. In Press, Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions. Neurocomputing, 2017.

[10] J. Falip, A. Aït-Younes, F. Blanchard, B. Delemer, A. Diallo, & M. Herbin, " Visual instance-based recommendation system for medical data mining", Procedia Computer Science, 2017, Vol.112, pp.1747-1754.

[11] J. Chen, W. Weia, C. Guoa, L. Tanga, & L. Suna, . In Press, Accepted Manuscript, "Textual analysis and visualization of research trends in data mining for electronic health records", Health Policy and Technology, 2017.

[12] C. Bouvry, N. Tvardik, I. Kergourlay, A. Bittar, P. Arnod-Prin, F. Segond, & M.H Metzger, "The SYNODOS Project: System for the Normalization and Organization of Textual Medical Data for Observation in Healthcare", *IRBM*, 2016, Vol. *37,* No. 2, pp.109-115.

[13] S.N. Kasthurirathne, B.E. Dixon, J. Gichoya, H. Xu, Y. Xia, B. Mamlin & S.J. Grannis, 2017, "Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data", Journal of Biomedical Informatics, 2017, Vol.69, pp.160-176.

[14] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, & T. Botsis, "Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review", Journal of Biomedical Informatics, 2017, Vol.73, pp.14-29.

[15] P.L. Hsu, H.S. Hsieh, J. H. Liang, & Y. S. Chen, "Mining various semantic relationships from unstructured user-generated web data", Web Semantics: Science, Services and Agents on the World Wide Web, 2015, Vol. 31, pp.27-38.

[16] R.M. Rodriguez, "Hesitant Fuzzy Sets: State of the Art and Future Directions", International Journal of Intelligent System, 2014, Vol.29, No. 6, pp.495-524.

[17] C. Kahraman, S.C. Onar and B. Oztaysi, " Present Worith Analysis Using Hesitant Fuzzy Sets", 9th Conference of the European Society for Fuzzy Logic and Technology [EUSFLAT), Altantis Press, 2015, pp.255-259.

[18] Z. Ding, Y. Wu, "An Improved Interval-Valued Hesitant Fuzzy multi-Criteria Group Decision-Making method and Applications", Mathematical and Computational Applications, 2016, Vol.21, pp.22-37.

[19] S. Wibowo, H. Deng, W. Xu, "Evaluation of cloud services: A fuzzy multi-criteria group decision making method", Algorithms, 2016, Vol.9, pp.84-96.

[20] V. Torra, "Hesitant fuzzy sets", International Journal of Intelligent System, 2010, Vol.25, No.6, pp. 529-539.