Assessment of critical behavior for MI, PCI and CABG events

C.Premalatha, X.Rexeena, S.Saranya

Abstract— Coronary Heart Disease (CHD) is one of the main causes of death in and around countries. Several studies with different technologies have been made in diagnosis and treatment of CHD, which includes association rules, logistic regression, fuzzy modeling, and neural network. The existing techniques are confined to small datasets that are specific to one particular disease and this knowledge mined is not indispensible for classification of risk factors for the CHD events. The implemented methodology uses C4.5 decision tree algorithm for identification of CHD related risk factors for the events that includes Myocardial Infarction, Percutaneous Coronary Intervention, and Coronary Artery bypass graft surgery based on five different splitting criteria that includes Information Gain, Gini Index, Likelihood Ratio Chi-Squared Statistics, Gain Ratio, and Distance Measure. Using performance measures, correctly classified values have been found for each splitting criteria's and accuracy is calculated. The criterion which has highest accuracy is distance measure and it is used for classification of risk factors and CHD diagnosis. The implemented methodology, C4.5 decision tree algorithm gives high classification accuracy compared to the aforementioned existing techniques

Index Terms— Coronary Heart Disease (CHD), C4.5 decision tree algorithm, classified values, high classification accuracy, performance measures, risk factors, Splitting criteria

1 INTRODUCTION

The objective of the implemented system was to develop a data mining system based on decision trees for the assessment of CHD related risk factors[8], targeting in the reduction of CHD events. Data-mining analysis was carried out using the C4.5 decision tree algorithm with five different splitting criteria for extracting rules based on the riskfactors (age, sex, FH, SMBEF, SMAFT, TC, TG, HDLM, HDLW, GLU, HXHTN, HXDM, SBP, DBP, and LDL). Data mining facilitates data exploration using data analysis methods with sophisticated algorithms in order to discover unknown patterns. Such algorithms include decision trees that have been used extensively in medicine.

Decision-tree-based algorithms give reliable and effective results that provide high-classification accuracy with a simple representation of gathered knowledge, and are especially appropriate to support decision-making processes in medicine. The C4.5 algorithm [5], which uses the divide-and-conquer approach to decision tree induction, was employed. The algorithm uses a selected criterion to build the tree. It works top–down, seeking at each stage an attribute to split on that which best separates the classes, and then recursively processing the sub problems that result from the split.

- *C.Premalatha is currently pursuing master degree program in Computer Science and engineering in Ranganathan Engineering College, India, MOB-9150451180. E-mail: premalathajck@mail.com*
- *X.Rexeena, S.Saranya are currently pursuing master degree program in Computer science and engineering in Ranganathan Engineering College, India, MOB-09048901879. E-mail: rexeena@mail.com*

In the implemented system, the following splitting criteria [8], were used: Information Gain, Gini Index, Likelihood Ratio Chi-Squared Statistics, Gain Ratio, and Distance Measure. Based on these splitting criteria, five different decision trees are constructed. Using performance measures, training and testing datasets are compared and accuracy is calculated. The criterion which has highest accuracy is used as best splitting criteria for decision tree construction such that risk factors are classified for CHD diagnosis.

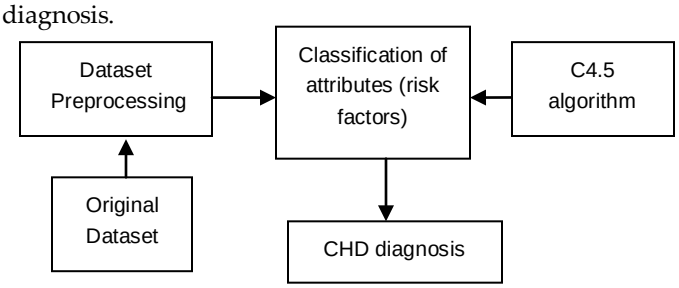

Fig .1. Block Diagram of the Coronary Heart Disease diagnosis system

2 DATA SET PREPROCESSING [4]

The data preprocessing is the first processing module in the project. Analyzing data that has not been carefully screened for such problems can produce misleading results. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult.

TABLE I. ORIGINAL DATASET

Age	Sex	FH	SMBEF	HXHTN	HXDM	SMAFT	SBP	DBP	TC	HDLW	HDLM	LDL	TG	GLU	CL
65	2	1	1	2	1	2	80	90	200	50	30	80	67	112	1
31	1	1	1	2	1	1	100	80	45	60	50	100	56	110	2
45	1	2	2	2	1	2	149	60	80	70	40	120	100	90	3
45	1	2	2	2	1	2	149	60	80	70	40	120	100	90	3
80	2	2	1	1	1	1	150	?	190	80	60	23	150	150	4

TABLE II. PREPROCESSED DATASET

Age	Sex	FH	SMBEF	HXHTN	HXDM	SMAFT	SBP	DBP	TC	HDLW	HDLM	LDL	TG	GLU	CL
65	2	1	1	2	1	2	80	90	200	50	30	80	67	112	1
31	1	1	1	2	1	1	100	80	45	60	50	100	56	110	2
45	1	2	2	2	1	2	149	60	80	70	40	120	100	90	3
80	2	2	1	1	1	1	150	**70**	190	80	60	23	150	150	4

TABLE III. CODED DATASET

Age	Sex	FH	SMBEF	HXHTN	HXDM	SMAFT	SBP	DBP	TC	HDLW	HDLM	LDL	TG	GLU	CL
3	N	Y	Y	N	Y	N	L	H	H	N	L	N	N	N	1
1	Y	Y	Y	N	Y	Y	N	N	N	N	N	H	N	H	2
1	Y	N	N	N	Y	N	H	N	N	H	N	H	N	N	3
4	N	N	Y	Y	Y	Y	H	N	H	H	H	N	H	H	4

Thus, the representation and quality of data is first and foremost before any process. Steps involved in dataset preprocessing are as follows,

- Missing values are filled using K-Nearest Neighbor algorithm [9]
- Duplications were removed
- Data were coded

The Steps involved in filling up the missing values are:

1) Determine parameter K = number of nearest neighbors
2) Calculate the distance between the query-instance and all the training samples
3) Sort the distance and determine nearest neighbors based on the K-th minimum distance
4) Gather the values of 'y' of the nearest neighbors
5) Use average of nearest neighbors as the prediction value of the query instance

If both the row has same value that is, the values duplicated, then any one of the row is removed from the dataset. None of the row is removed if at least one value differs in any column of the tuple.

3 Classification of Risk Factors [8]

The C4.5 algorithm, which uses the divide-and-conquer approach to decision tree induction, was employed. The algorithm uses a selected criterion to build the tree. It works top–down, seeking at each stage an attribute to split on that which best separates the classes, and then recursively processing the sub problems that result from the split.

Input:

1) Training dataset *D*, which is a set of training observations and their associated class value.
2) Attribute list *A*, the set of candidate attributes.
3) Selected splitting criteria method.

Output: A decision tree.

C4.5 decision tree construction module having the following 5 splitting criteria are to be investigated for training the dataset.

1) Information Gain (IG)

Information gain is based on Claude Shannon's work on information theory. InfoGain of an attribute *A* is used to select the best splitting criterion attribute. The highest InfoGain is selected to build the decision tree

$$\text{InfoGain}(A) = \text{Info}(D) - \text{Info}_A(D) \quad (1)$$

Where,

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (2)$$

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Info}(D_j) \quad (3)$$

2) Gini Index (GI)

The Gini index is an impurity-based criterion that measures the divergence between the probability distributions of the target attributes values

$$\text{GiniIndex}(D) = \text{Gini}(D) - \sum_{j=1}^{v} p_j \times \text{Gini}(D_j)$$

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2. \quad (4)$$

3) Likelihood Ratio Chi-Squared Statistics

The likelihood ratio chi-squared statistic is useful for measuring the statistical significance of the information gain criterion

$$G^2(A, D) = 2 \times \ln(2) \times |D| \times \text{InfoGain}(A). \quad (5)$$

4) Gain Ratio (GR)

Gain ratio biases the decision tree against considering attributes with a large number of distinct values. So it solves the drawback of information gain

$$\text{GainRatio}(A) = \frac{\text{InfoGain}(A)}{\text{SplitInfo}_A(D)}$$

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right). \quad (6)$$

5) Distance Measure (DM)

Distance measure, like GR, normalizes the impurity criterion (GI). It suggests normalizing it in a different way

$$\text{DM}(A) = \frac{\text{Gini}(D)}{-\sum_{j=1}^{v}\sum_{i=1}^{m} p_{ij} \times \log_2(p_{ij})}. \quad (7)$$

4 Performance Evaluation

In order to evaluate the performance of above techniques, the following factors are to be investigated.

1) Correct classifications (%CC): is the percentage of the correctly classified records; equals to (TP + TN)/N.
2) True positive rate (%TP): corresponds to the number of positive examples correctly predicted by the classification model.
3) False positive rate (%FP): corresponds to the number of negative examples wrongly predicted as positive by the classification model.
4) True negative rate (%TN): corresponds to the number of

negative examples correctly predicted by the classification model.

5) False negative rate (%FN): corresponds to the number of positive examples wrongly predicted as negative by the classification model.

6) Sensitivity: is defined as the fraction of positive examples predicted correctly by the model, equals to TP/(TP +FN).

7) Specificity: is defined as the fraction of negative examples predicted correctly by the model, equals to TN/(TN+FP).

8) Support: is the number of cases for which the rule applies (or predicts correctly; that is, if we have the rule $X \rightarrow Z$, Support is the probability that a transaction contains {X, Z} [26]
Support = P (XZ) =no of cases that satisfyX andZ/|D|

9) Confidence: is the number of cases for which the rule applies (or predicts correctly), expressed as a percentage of all instances to which it applies (that is, if we have the rule
$X \rightarrow Z$, Confidence is the conditional probability that a transaction having X also contains Z)

$$\text{CONFIDENCE} = P(Z|X) = P(XZ)/P(X)$$

.5 RESULT ANALYSIS

C4.5 algorithm used five different splitting criteria for constructing five different decision trees. The training and testing datasets were compared after decision tree construction for finding out correctly classified values.

Using Performance measures, the dataset's attribute value has been correctly classified and accuracy is calculated.

The criterion which has obtained highest accuracy is Distance measure and it is used for classification of risk factors that is, decision tree construction and CHD diagnosis

6 ADVANTAGE OF IMPLEMENTED SYSTEM

1) The highest percentages of correct classifications are achieved using this method.

2) The initial no of attribute values are also reduced using preprocessing technique.

3) Both discrete and continues values can be evaluated

7 CONCLUSION

The implemented methodology uses decision tree for assessment of CHD related risk factors and reduction of CHD events that includes Myocardial Infarction, Percutaneous Coronary Intervention, and Coronary Artery bypass graft surgery. C4.5 Decision tree technique identifies most important risk factors for the events using five different splitting criteria which provide high-classification accuracy. Based on different splitting criteria, different decision trees are constructed and those trained datasets and new testing datasets are compared, which gives the dataset values that have been correctly classified and accuracy is calculated. The criterion which has highest accuracy is used for further classification of risk factors that is decision tree construction and CHD diagnosis.

8 FUTURE WORK

Future work involves in decision tree construction for more events instead of finding for limited number of events with large dataset values. For duplication removal, here only simple technique of elimination of low values is applied but it can be extended to some other techniques or algorithmic approach.

REFERENCES

[1] C. A. Pena-Reyes,(2004) 'Evolutionary fuzzy modeling human diagnostic decisions,' Ann. NY Acad. Sci., vol. 1020, pp. 190–211.

[2] C. L. Tsien, H. S. F. Fraser, W. J. Long, and R. L. Kennedy, (1998) 'Using classification trees and logistic regression methods to diagnose myocardial infarction,' in Proc. 9th World Congr. Med. Inf., vol. 52, pp. 493–497.

[3] C. Ordonez, E. Omiecinski, L. de Braal, C. A. Santana, N. Ezquerra,J. A. Taboada, D. Cooke, E. Krawczvnska, and E. V. Garcia,(2001) 'Mining constrained association rules to predict heart disease,' in Proc. IEEE Int.Conf. Data Mining (ICDM 2001), pp. 431–440.

[4] J. Han and M. Kamber, (2001) 'Data Mining, Concepts and Techniques', 2nd ed. San Francisco, CA: Morgan Kaufmann.

[5] J. R. Quinlan,(1987) 'Simplifying decision trees', Int. J. Man-Mach. Stud.,vol. 27, pp. 221–234.

[6] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, (1984),' Classification and Regression Trees', Belmont, CA: Wadsworth Int. Group.

[7] M. Karaolis, J. A. Moutiris, L. Papaconstantinou, and C. S. Pattichis, (2009),'Association rule analysis for the assessment of the risk of coronary heart events', in Proc. 31st Annu. Int. IEEE Eng. Med. Biol. Soc. Conf., Minneapolis, MN, Sep. 2–6, 2009, pp. 6238–6241.

[8] M. Karaolis, J. A.Moutiris, and C. S. Pattichis, (2008) 'Assessment of the risk of coronary heart event based on data mining', in Proc. 8th IEEE Int. Conf.Bioinformatics Bioeng, pp. 1–5.

[9] Phayung Meesad and Kairung Hengpraprohm, (2008) 'Combination of KNN-Based Feature Selection and KNN-Based Missing-Value Imputation of Microarray Data,'the 3rd Intetnational Conference on Innovative Computing Information and Control (ICICIC'08) IEEE computer society.

[10] R. B. Rao, S. Krishan, and R. S. Niculescu,(2006) 'Data mining for improved cardiac care', ACM SIGKDD Explorations Newslett., vol. 8, no. 1, pp. 3–10.

[11] R. Lopez de Mantras, (1991) 'A distance-based attribute selection measure for decision tree induction', Mach. Learn., vol. 6, pp. 81–92, 1991.

[12] S. A. Pavlopoulos, A. Ch. Stasis, and E. N. Loukis, (2004) 'A decision treebased method for the differential diagnosis of aortic

stenosis from mitral regurgitation using heart sounds', Biomed. Eng. OnLine, vol. 3, p. 21.

[13] Subrata Paramanik,Utpala Nanda Chowdhury,(2010) 'A Comparative Study of Bagging,Boosting and C4.5: The Recent Improvements indecision Tree Learning Algorithm',Asian Journal of Information Technology.

[14] S. M. Grundy, R. Pasternak, P. Greenland, S. Smith, and V. Fuster, (1999) 'Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations', Amer. Heart Assoc., vol. 100, pp. 1481–1492, 1999.

[15] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman,(2002) 'Decision trees: An overview and their use in medicine', J. Med. Syst., vol. 26, no. 5, pp. 445– 463.