**A Statistical Approach on Software Reuse in Medical Database for Cardiac Patients**

M. Bhanu Sridhar1, Y. Srinivas2, M. H. M. Krishna Prasad3

1Department of Computer Science and Engineering, GVP College of Engineering for Women, Visakhapatnam; E-mail: sridharbhanu@gmail.com

2 Department of Information Technology, GITAM University, Visakhapatnam

E-mail: ysrinivasit@rediffmail.com

3Department of Computer Science & Engineering, JNTUK, Kakinada

E-mail:  krishnaprasad.mhm@gmail.com

**Abstract** – Software reuse is a subfield of software engineering that is used to adopt the existing software for similar purposes. In this paper, medical database related to cardiology is considered. The Pearson Type-I Distribution is utilized for clustering the data. Further, coupling methodology is used to bring out the similarity among symptoms related to a new patient data by comparing it with the existing data. By this, the concerned treatment to be followed for the new patient is deduced. In order to ratify a disease, the patient’s symptoms are identified and are correlated with a benchmark factor i.e. T-wave, obtained from ECG data. Specificity and sensitivity is carried out next and the performance of the system developed is computed by using precision and recall. This methodology is useful in the field of medical sciences, in particular to patients of remote areas, to save the precious life at the right moment.

**Keywords** –Reuse, Coupling, Pearson Distribution, Cardiology, Bayesian theorem

1. **INTRODUCTION**

Software Engineering is a prominent field of Computer Science that can be used to organize, evaluate, and implement the software with a practical methodology. This process saves time, money and energy and also improves the quality of the new software. The success of the reuse process depends on many criteria, such as, the reusability of the existing software [1], availability of efficient algorithms to obtain the maximum reusability and so on. Software reuse process started on programs and expanded to other fields like analysis, design, testing etc. Researchers have long been attracted to software reuse for the richness of the field in topics and its worthiness in terms of saving energy and time. However, software reuse has not received the deserved importance in the medical domain. In this paper, the concept of software reuse is utilized on medical data, particularly with cardiology patients’ data. The paper is presented in two parts: part-1of the paper deals with identifying the cardiac, pro-cardiac and non-cardiac patients by consuming the Pearson Type-1 family of distribution. The clusters of cardiac and pro-cardiac are considered for further analysis. These clusters are further analyzed for ratification of a patient so that effective treatment can be suggested. The most significant parameter is the T-wave of ECG; a patient is identified as cardiac if he/she possesses the benchmark value. The correlation is calculated to prove the discrimination.

In the second part, to have a more precise discovery of the patient’s status, specificity and sensitivity are calculated. Additional parameters like cholesterol, lipid value, history and T-wave value are considered for classification at this juncture. Basing on these calculations the patient’s data is classified into two groups (cardiac, pro-cardiac). The concept is further advanced by using Bayesian theorem to test the impact of history while discriminating the patient to be a cardiac or non-cardiac. These calculations will be vital in order to categorize the patient to be a cardiac or non-cardiac, since these diseases have been proven to be hereditary.

The article is wrapped-up with a conclusion that opens the doors for a new idea and its application, particularly in rural areas, so as to save the precious life of an ailing patient.

1. **DATASET**

The dataset considered here is that of patients who have symptoms of cardiac problems. The cardiac problems are briefly discussed below.

**2.1 Cardiac Problems**

The heart is a myogenic (cell-related) muscular organ with a circulatory system (including all vertebrates), that is responsible for pumping blood throughout the blood vessels by repeated, rhythmic contradictions [3]. Among the problems related to heart, the major problem is cardiac arrest, which is the cessation of normal blood circulation due to failure of the heart to contract effectively. It should be effectively realised that cardiac arrest is different from a heart attack where blood supply is interrupted to a part of the heart which may**/**may not lead to the patient’s death.

The patients who approach a doctor can be classified into three categories taking into consideration results of different tests conducted with the existing symptoms. The properties taken into consideration are Atherosclerosis (due to Cholesterol), Myocardial Infarction (heart attack), different medical signs like blood cell count and skin rashness, various symptoms like head ache and body pain, and other facts like Diabetes, Triglyceride, Migraine and so on [2]. The three categories are:

1. Normal: A patient can be declared ‘normal’ when no signs or symptoms of a cardiovascular/coronary disease are found within the results of various tests conducted. The general factors considered are the blood pressure (BP), sugar level in blood, results of Electrocardiography (ECG), Cholesterol level, Triglyceride, and other sensations.
2. Pro-Cardiac: Pro-cardiac category keeps the account of those patients who are suspected to have some signs and/or symptoms of heart-problems. These can be observed from the BP tests slightly exceeding the normal levels, sugar levels in blood also rising, ECG suspecting (though not deducing) problems in future and some signs and symptoms like light chest pain, high cholesterol, severe headaches often turning up etc. do surface.
3. Cardiac: A cardiac is surely in the range of trouble – prone to abnormal BP conditions, having severe pain the chest region, burning sensations, sweating, pain along the left arm and finally having already had a light heart attack. A cardiac must be immediately taken into consideration for regular treatment with constant observation of all concerned positions in and around the heart and those that affect the heart.

The predominant features considered for the cardiology database are: blood pressure (BP), heartbeat (HB), pulse rate (PR), ECG (normal/abnormal), pain in the left shoulder region, sweating, nausea/vomiting, overweight, chest pain and breathlessness. Eleven core symptoms are considered while preparing the patient data dissimilarity matrix.

**2.2 Database**

After the above discussion, it should also be noted that effective medical data of the concerned patient should be readily available with the Doctors and should be regularly updated. This data forms the pillar of the patient’s classification level, severity level and the chance of saving his/her life [4]. An attempt is made here to correctly judge the patient’s disease level to suggest the suitable drugs. Apparently, this is a reuse methodology which is very handy in many situations.

Depending on the latest reports of the patient, he can be recognised to be prone by a known disease and further classified into what level of that disease he is in. A dissimilarity matrix is constructed with the readings from the data and identifying the most leading symptoms. The various readings considered are categorized into the above mentioned three groups for a patient who had hit trouble with cardiology. A database is formulated from the realistic data obtained from medical patients from the data referred in [5].

For the testing purpose a database of twenty patients has been used, with the symptoms of cardiac problems, taking into consideration eleven core features/symptoms. If a symptom does exist it has been represented by using a unique value. Following this procedure for all the inputs, a dissimilarity matrix [1] is obtained and this matrix is to be categorized. To classify a patient, the dissimilarity matrix is formulated and classified by using the clustering technique in the context of coupling.

**Table 1: Symptoms (→) of the patients**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Patient ID(↓) | BP | Heart beat (HB) | Pulse Rate (PR) | ECG | Left Shoulder pain | Sweating | Vomiting | Over Weight | Chest Pain | Breathlessness | Obesity |
| P1 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 |
| P2 | 0 | 0 | 0 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 |
| P3 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 9 | 0 | 0 |
| P4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P5 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 |
| P6 | 0 | 0 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| P7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 |
| P8 | 0 | 0 | 0 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 |
| P9 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 10 | 0 |
| P10 | 0 | 0 | 3 | 0 | 5 | 0 | 7 | 0 | 9 | 0 | 11 |
| P11 | 0 | 2 | 0 | 0 | 0 | 6 | 7 | 8 | 9 | 0 | 11 |
| P12 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 9 | 10 | 0 |
| P13 | 0 | 0 | 0 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 |
| P14 | 0 | 2 | 0 | 0 | 5 | 6 | 7 | 8 | 0 | 0 | 0 |
| P15 | 0 | 2 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 0 | 11 |
| P16 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 |
| P17 | 0 | 2 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 0 | 11 |
| P18 | 0 | 0 | 0 | 4 | 0 | 0 | 7 | 8 | 9 | 10 | 0 |
| P19 | 0 | 2 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 0 | 11 |
| P20 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 |

1. **METHODOLOGIES**

**3.1 The Pearson Family of Probability Distributions**

Karl Pearson obtained a family of distributions known as the Pearson distributions, which would represent satisfactorily almost all practical situations. The probability density function of Pearson type-1 distribution is proposed by Pearson [6]. Pearson's examples comprise survival data, which are usually asymmetric.

The Probability Density function of theType-1, Pearson family [7] is given by Eqn. (1)

 (1)

The variable *X* denotes the grey level intensity value of the echo-cardio-graphic (ECG) speckle of a cardiac patient and *f*(*x*) represents the probability density function (pdf). The *b*, *a0*, *a1*, and *a2* are parameters of the distribution. The calculations of these parameters and the clustering of the patients’ data using this model is presented by the authors in [2, 3].

These models are used to assess the symptoms of the patients and utilizing this model for developing a methodology that can be reused**.** By the usage of Pearson Type-1 distribution equations, we calculate the probability density function (pdf) for each patient. The output generated by using Pearson family Type-1 equations is presented by authors of [2, 3] is utilized and is shown below.



**Figure 1: Showing the Results of Pearson Type-1 Equations**

The reuse metrics proposed by Chidamber & Kemerer [8] are utilized in coupling the data and thus classifying the patients into normal, pro-cardiac, and cardiac clusters. A metric of this suite, the metric-4, is adopted which is concerned about ‘coupling between objects’ (CBO). Here it is to be once again noted that excessive coupling is negative to the basic component design and prevents reuse. The CBO for a class is defined to be the number of other classes to which it is coupled. In this context a patient exhibiting different symptoms is considered to be a class and the symptoms are the objects. The dataset considered above is used in this context to apply the CBO metric to view the results.

**Table 2: Coupling count of each class (patient)**

|  |  |  |
| --- | --- | --- |
| Patient ID | Associated with | Coupling Count |
| 1 | 5,16,20 | 3 |
| 2 | 8,13 | 2 |
| 5 | 1,16,20 | 3 |
| 8 | 2,13 | 2 |
| 13 | 2,8 | 2 |
| 15 | 17,19 | 2 |
| 16 | 1,5,20 | 3 |
| 17 | 15,19 | 2 |
| 19 | 15,17 | 2 |
| 20 | 1,5,16 | 3 |
| 3,4,6,7,9,10,11,12,14,18 | NONE | 10 |

By this developed methodology, one can notice that the number of patients with no common symptoms is 10 and can classify them as ‘normal’. Next the ‘pro-cardiac’ group is obtained as patients in the group with CBO metric value of 2. Finally the all-dangerous ‘cardiac’ group is obtained from the remaining CBO metric value of 3. At this juncture, armed with full data calculations and conclusions, if a new patient’s data is forwarded, one can easily place him in a cluster, and treatment can be started as it was done for the previous patients of that cluster [2].

From table-2, the cardiac and pro-cardiac patients are considered and the data concerned is presented below in table-3.

**Table-3: Data of cardiac & pro-cardiac patients**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Patient ID | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | A | B | d(A-B) | d2 |
| P1 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| P2 | 0 | 0 | 0 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| P5 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| P8 | 0 | 0 | 0 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| P13 | 0 | 0 | 0 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| P15 | 0 | 2 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 0 | 11 | 1 | 0 | 0 | 1 |
| P16 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| P17 | 0 | 2 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 0 | 11 | 1 | 1 | 0 | 0 |
| P19 | 0 | 2 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 0 | 11 | 1 | 1 | 0 | 0 |
| P20 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |

Patients 2, 8, 13, 15, 17, 19 have CBO of 2 and are classified as pro-cardiac. Patients 1, 5, 16, 20 have CBO of 3 and are identified as the cluster cardiac as explained before.

Let A🡪 Values identified on patients

B 🡪 T-wave (Benchmark value from ECG)

Correlation = 1- 6∑$\frac{d2}{n(n2-1)}$ = 1- 6 \* (1)/(4 \*15) = 0.9

Since the patient P15 has a high correlation value compared to others, he/she doesn’t belong to this group and hence it can be considered that P15 is a pro-cardiac – not a cardiac.

* 1. **Application of Latent Class Analysis:**

The latest technological updates in medical diagnosis helped to estimate the sensitivity and specificity of the concerned disease. The probability of a person suffering with a disease having positive on the indicator is called sensitivity and specificity is the probability that a person without a disease will have negative indicator. Basing on the symptoms, the specificity and sensitivity are calculated to estimate the extent of the disease on the patient.

Latent class analysis will help to identify the underlying categorical variables through which a relationship can be established to diagnose a disease. Note that these variables are mostly dichotomous. The latent class analysis is a statistical analysis based on two parameters – conditional probability and unconditional probability.

Unconditional probabilities are those probabilities in which a latent variable will be present in each class. Conditional probabilities are associated with the test results based on the membership of the latent class.

After identifying the probability, Bayes theorem is used to examine the results. Basing on the test results the data can be categorized into three groups. The examination is mostly carried out in the presence of indicators (ECG/T-wave), blood cholesterol test (HDL, LDL).

Write about blood test (LPH, CPK), ECG-T-wave (T), unconditional probability, conditional probability, history (H) etc. {C – Cholesterol, L – Lipid Value} Blood test values

**Table-4: Error calculation**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | C | L | H | T | 1 | 2 | d | E |
| P1 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| P2 | 0 | 0 | 0 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0.99 | 0.01 | 1 | 0.01 |
| P5 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| P8 | 0 | 0 | 0 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.89 | 0.11 | 1 | 0.11 |
| P13 | 0 | 0 | 0 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| P15 | 0 | 2 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 0 | 11 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| P16 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0.42 | 0.58 | 2 | 0.42 |
| P17 | 0 | 2 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 0 | 11 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| P19 | 0 | 2 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 0 | 11 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| P20 | 0 | 2 | 0 | 4 | 0 | 6 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.04 | 0.96 | 2 | 0.04 |

Note:

1 => Having; 2 => not having; d => diagnosis; E => error.

In the dataset considered above, assume that there are two classes:

Class-1 (Cardiac) => High probability for being +ve

Class-2 (Pro-cardiac) => Low probability for being +ve

* If T=1 then high sensitivity exists => belongs to class-1
* If T=0 and H=1 => belongs to class-2 (mostly)
* If H=0 and blood=high & T=0 => belongs to class-2
* If H=1 and blood=1 and T & cholesterol =0 then it belongs to class-2
* If H=1& all others are 0 then class-2
* If blood=1 & others = 0 then class-2
* If T=1 & all others = 0 then class-1

Errors:

For P2--- T=0 => belongs to class-2; but we have 0.01. $∴$ It is an error.

For P8--- T=0 => belongs to class-2; but we have 0.11. $∴$ It is an error.

For P16--- T=1 => belongs to class-1; but we have 0.42. $∴$ It is an error.

For P20---All are 0 except T=> belongs to class-1; but we have 0.04. $∴$ It is an error.

**Bayes Theorem:**

H and T values/results can be obtained more easily within a short period of time when compared with blood test results. At this juncture, we can use Bayes theorem to obtain/set the final result without the blood test values in our hand. The usage of Bayes theorem is very handy in the medical field since time is very precious in the context of saving the life of the concerned patient.

Cases to be considered:

1. T=1, H=0 => Patient has a disease (Class-1)
2. T=0, H=0 => No disease (Class-2)
3. T=0, H=1 => No disease (Class-2)
4. T=1, H=1 => Patient has a disease (Class-1)

P (Y|X) = $\frac{P(X|Y) P(Y)}{P(X)}$

 **Table-5: Values from Bayesian theorem application**

|  |  |  |  |
| --- | --- | --- | --- |
| T-wave | History | Class-1 | Class-2 |
| 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 0 |

1. **Performance Evaluation:**

In order to evaluate the methodology, the formula specified by [9, 10] has been used to identify the precision and recall values.

|  |  |
| --- | --- |
| Category of Object **xi**  1  0 |  Categories of Object **xj** |
|  1 0 |
|  a (true +ve) |  b (false +ve) |
|  c (false –ve) |  d(true –ve) |

 **Table 6 – Frequency Table**

In this notation,

a -> no. of attributes with both i and j present

b -> no. of attributes with only i present

c -> no. of attributes with only j present

d -> no. of attributes with both i and j absent

Precision = $\frac{a}{a+b}$ ; Recall = $\frac{a}{a+c}$

Note that ‘Recall’ in this context is also referred to as the True Positive Rate or *Sensitivity*, and ‘precision’ is also referred to as Positive predictive value (PPV); other related measures used in classification include True Negative Rate and Accuracy. True Negative Rate is also called *Specificity*.

These values for table-5 are calculated and presented below:

|  |  |  |
| --- | --- | --- |
| T-wave | History | Category |
| 0 | 0 | d |
| 0 | 1 | c |
| 1 | 0 | b |
| 1 | 1 | a |

Precision = 1/(1+1) = 0.5 ; Recall = 1/(1+1) = 0.5

1. **Conclusion**:

In this work, the concept of software reuse, for the field of medical domain, is presented as a novel methodology. This paper helps to identify the disease-prone and normal people, thereby helping the patients in remote places by suggesting and assisting the cardiology patients, in this case. The paper is presented using two models, both highlighting towards the identification pro-cardiac and cardiac patients, leaving off the normal people. With a suggestion of medicine that evidently has more base, confidence of the concerned paramedics or the user will be high when time comes for using the suggested medicine. Evidently, this work is aimed to be used in remote areas with lesser availability of doctors and a developed app can surely be put in place for saving the precious lives of the ailing patients at critical junctures.

**REFERENCES**

1. Robert Godin et al, Applying Concept Formation Methods to Software Reuse, IJSEKE, 5(1), 119-142, 1995.
2. Bhanu Sridhar et al, ‘Software Reuse in Medical Database for Cardiac Patients using Pearson Family Equations’, International Journal of Computer Applications (0975-8887), Volume 58-No, 14, November 2012.
3. Bhanu Sridhar et al, Software Reuse in Cardiology Related Medical Database Using K-Means Clustering Technique, Journal of Software Engineering and Applications, 2012, Vol.5, No. 9, 682-686.
4. Shantakumar Patil and Y. S. Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research, Vol. 31, no. 4, 2009.
5. Disease datasets from Machine Learning Repository, <http://archive.ics.uci.edu/ml/datasets>
6. Quoted from: <http://en.wikipedia.org/wiki/Pearson_distribution>
7. Abhinav Gupta and Karmeshu, Statistical Characterization of Speckle in Clinical Echocardiographic Images with Pearson Family of Distributions, Defence Science Journal, Vol. 61, No. 5, September 2011, pp. 473-478.
8. Shyam Chidamber and Chris Kemerer, A Metrics Suite for Object Oriented Design, IEEE Transactions on Software Engineering, Vol. 20, No. 6, June 1994.
9. Spearman’s Rank Correlation Coefficient, Rank Correlation, <http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient>, 20/4/2013.
10. Precision and Recall, <https://en.wikipedia.org/wiki/Precision_and_recall>, Wikipedia.