Predicting Heart-disease from Medical Data by Applying Naïve Bayes and Apriori Algorithm

Aieman Quadir Siddique, Md. Saddam Hossain

Abstract-Data mining could simply be explained at the non-trivial extraction of the potentially useful, inherent, and previously unknown information from data. The foremost feature of data mining technique is that it provides a user oriented approach to hidden and novel patterns in the data. Data mining techniques have played a major role in the designing and development of the support system exclusively because of its abilities for discovering hidden patterns and relationships in the medical data. The purpose of this research study is to compare different data mining algorithms that could be used for the prediction of heart disease.

Index Terms-Data mining, Heart-disease, Naïve Bayes, Apriori, Prediction, Medical data mining, Classification

1 Introduction

Technological advancements and health care awareness have led towards the development of huge number of health care facilities and hospitals. However, the providing a high quality of health care services at a low cost is becoming a challenging issues inside the developing countries of the world [12]. Although, many countries have taken some firm steps towards the ensuring that healthcare services are provided to everyone [9]. Medical data mining possesses a great ability for the exploration of the hidden patterns in the existing datasets of the medical domain. Such patterns could be utilized for the purpose of clinical diagnosis [8].

Data mining requires a collection of data in an organized form, the data collected was integrated in the formation of a hospital information system [2]. The technology of data mining provides the users with a user oriented approach towards hidden and novel patterns in

- *Aieman Quadir Siddique is currently a first year student of Masters degree program in Software Engineering at The University of Adelaide, Australia.*
- *Md. Saddam Hossain has completed bachelor degree program in electrical and electronics engineering from American International University-Bangladesh, Bangladesh.*

data. Efficient and effective automated heart disease system capable of predicting heart disease could prove to be immensely beneficial for the healthcare sector [11]. This research study attempts to present a detailed analysis of different data mining algorithms for the prediction of heart disease using the data consisting of patient's history. These automation systems will play a major role in reducing the overall number of tests that a patient has to take concerning heart disease [15].

2 Methodology

This research paper exhibits a critical analysis of two well-known data mining algorithm that could prove to be beneficial for the medical practitioners and analysts for accurately predicting the heart disease diagnosis [6]. The methodology used for this research study includes the examination of the various journals, reviews and publications in the field of the software engineering, cardiovascular diseases, and data mining in past few years.

3 Discussion

3.1 Applying Naïve Bayes Algorithm over medical data

The following can be the input attributes:

- Age in Year
- Sex (value 1: Male; value 0: Female)
- Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain;
- value 4: asymptomatic)
- Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
- Restecg – resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2:
- showing probable or definite left ventricular hypertrophy)
- Exang - exercise induced angina (value 1: yes; value 0: no)
- Slope – the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: downsloping)
- CA – number of major vessels colored by floursopy (value 0-3)
- Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)

- Trest Blood Pressure (mm Hg on admission to the hospital)
- Serum Cholestrol (mg/dl)
- Thalach – maximum heart rate achieved
- Oldpeak – ST depression induced by exercise
- Smoking – (value 1: past; value 2: current; value 3: never)
- Obesity – (value 1: yes; value 0: no)

Explanation of the Algorithm:

Calculate diagnosis="yes", diagnosis="no" probabilities Pyes,

Pno from training input.

For Each Test Input Record

For Each Attribute

Calculate Category of Attribute Based On Categorical

Division

Calculate Probabilities Of Diagnosis="Yes",

Diagnosis="No" Corresponds To That Category P(Attr,Yes),

P(Attr,No) From Training Input .

For Each Attribute

Calculate The Resultyes= Resultyes*

P(Attr,Yes),Resultno= Resultno*P(Attr,No);

Calculate Resultyes= Resultyes *Pyes

Resultno= Resultno*Pno;

If(Resultyes > Resultno) Then Diagnosis="Yes";

Else Then Diagnosis ="No";

In the algorithm given above the collected sample is represented with help of n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.

Consider there a m number of classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive probability assigns an unknown sample X to the class Ci if and only if:

P(Ci/X)>P(Cj/X) for all 1< = j< = m and j!= i

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis.

P(Ci/X)= (P(X/Ci)P(Ci))/P(X)

As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e.

P(C1) = P(C2) = …..= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = si/s , where Si is the number of training samples of class Ci, and s is the total number of training samples.

Formulae:

- Pyes=total number of yes/total number of records;
- Pno=total number of no/total number of records;
- P(attr,yes)=total number of yes in corresponding category/total number of yes;
- P(attr,no)=total number of yes in corresponding category/total number of yes;

3.2 Applying Apriori Algorithm over medical data

The following can be the input attributes:

- Sex (value 1: Male; value 0 : Female)
- Chest Pain Type (value 1: typical type 1 angina, value2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
- Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
- Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value2: showing probable or
- definite left ventricular hypertrophy)
- Exang – exercise induced angina (value 1: yes; value 0: no)
- Slope – the slope of the peak exercise ST segment (value1: unsloping; value 2: flat; value 3: downsloping)
- CA – number of major vessels colored by fluoroscopy (value 0 – 3)
- 8.Thal (value 3: normal; value 6: fixed defect; value7:reversible defect)
- Trest Blood Pressure (mm Hg on admission to the hospital)
- Serum Cholesterol (mg/dl)
- Thalach- maximum heart rate achieved
- Oldpeak – ST depression induced by exercise relative to rest
- Age in Year.

Explanation of the Algorithm:

This data mining algorithm could be used for finding the frequent item sets from a transactional dataset, and then generate association rules. However, under several circumstances finding item sets is not trivial due to the combinational explosion. One the frequent item sets are obtained, they automatically generate an association rule that is either equal or greater than the minimum number of users confidence. .Apriori is a seminal algorithm for finding frequent itemsets using candidate generation [3]. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". In this algorithm the system assumes that the items existing within a transaction are stored in lexicographic order. The algorithm then lets the set of frequent item set to be of size k be Fk and their candidates be of size Ck. Then in the next step the algorithm searches for a frequent number of the itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets.

1. Generate Ck+1, candidates of frequent itemsets of size k +1, from the frequent itemsets of size k.
2. Scan the database and calculate the support of each candidate of frequent itemsets.
3. Add those itemsets that satisfies the minimum support requirement to Fk+1.

1. Function apriori generates Ck+1 from Fk in the following two step process:

1. Join step: Generate RK+1, the initial candidates of frequent itemsets of size k + 1 by
2. taking the union of the two frequent itemsets of size k, Pk and Qk that have the first k−1
3. elements in common.
4. RK+1 = Pk ∪ Qk ={item1 , item2, . . . , itemk−1, itemk , itemk' }
5. Pk = {item1 , item2, . . . , itemk−1, itemk }
6. Qk = {item1 , item2, . . . , itemk−1, itemk' }
7. where, item1 < item2 < · · · < itemk < itemk'_ .

2. Prune step: Check if all the itemsets of size k in Rk+1 are frequent and generate Ck+1 by removing those that do not pass this
8. requirement from Rk+1. This is because any subset of size k of Ck+1 that is not frequent cannot be a subset of a frequent itemset of size
9. k + 1. Function subset finds all the candidates of the frequent item sets included in transaction t. Apriori, then, calculates frequency only for those candidates generated this way by scanning the database. It is evident that Apriori scans the database at most kmax+1 times when the maximum size of frequent itemsets is set at kmax.

4 CONCLUSION

In the end, it could rightfully be said that the major purpose of the data mining algorithm is the find the best available algorithm for describing the given data from multiple aspects. The development of efficient algorithm is immensely important for intending it to an automatic classification tool [4]. As with the successful implementation of these algorithms will enable the automatic designs to work efficiently, and at the same time it would save a lot of time for both the patients and the doctors as they would not have to wait in long lines to gte to the experts for diagnosis. The proposed work can be further enhanced and expanded for the automation of Heart disease prediction. In the future studies that researcher can use real data from Health care organizations and agencies and they use the available techniques for achieving optimum accuracy.

REFERENCES

[1] Anbarasi. M, Anupriya. E, N.Ch.S.N.Iyengar, ―Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm‖; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[2] Feng Tao, Fionn Murtagh, Mohsen Farid. Weighted Association Rule Mining using Weighted Support and Significance Framework, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining 2003, 2003: pp. 661-666.

[3] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, "Knowledge Management, Data Mining, and Text Mining In Medical Informatics", Chapter 1, eds. Medical Informatics: Knowledge Management And Data Mining In Biomedicine, New York, Springer, 2005: pp. 3-34.

[4] Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 2006: pp. 194-200.

[5] Larose, Daniel T. *Discovering knowledge in data: an introduction to data mining*. Wiley. com, 2005.

[6] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 2004: pp. 690–695.

[7] Ordonez, Carlos, et al. "Mining constrained association rules to predict heart disease." *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, 2001.

[8] Palaniappan, Rafiah Awang, —Intelligent Heart Disease Prediction System Using Data Mining Techniques‖; 978-1-4244-1968-5/08/$25.00©2008 IEEE: pp.34-56.

[9] Rupa G. Mehta, Dipti P. Rana, Mukesh A. Zaveri, —A Novel Fuzzy Based Classification for Data Mining using Fuzzy Discretization‖; World Congress on Computer Science and Information Engineering, 2009: pp. 90-109

[10] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.

[11] Sellappan, P., Chua, S.L.: "Model-based Healthcare Decision Support System", Proc. Of Int. Conf. on Information Technology in Asia CITA'05, Kuching, Sarawak, Malaysia, 2005: pp. 45-50.

[12] Shantakumar B.Patil, Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.

[13] Srinivas, K., B. Kavihta Rani, and A. Govrdhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." *International Journal on Computer Science and Engineering (IJCSE)* 2.02 (2010): pp. 250-255.

[14] Wilson, Andrew M., Lehana Thabane, and Anne Holbrook. "Application of data mining techniques in pharmacovigilance." *British journal of clinical pharmacology* 57.2 (2004): pp. 127-134.

[15] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management.6(4),50, , 2002: pp. 45-67.