**An Experimental Study on Medical Survey and Machine Learning Algorithms**

# Md. Arifur Rahman, A. G. M. Zaman

Department of Computer Science American International University-Bangladesh

Dhaka, Bangladesh

{acreage, agmzaman}@gmail.com

# Md. Monsur Ali

Department of Computer Science American International University-Bangladesh

Dhaka, Bangladesh [monsuralirana@gmail.com](mailto:monsuralirana@gmail.com)

***Abstract*— Many medical research are based on statistical surveys. These surveys reveal risk factors and other crucial assumptions regarding a disease. However, machine learning and predictive algorithms may agree or disagree with these surveys. In this paper, we tried to find what predictive algorithms predict about the risk factors of a disease. We checked J48 and Naïve Bayes classifiers’ stand on the risk factors of cardiovascular disease. In our approach, we found that these classifiers are not always consistent with the medical surveys on risk factors.**

***Keywords— medical surveys; machine learning; classifiers; risk factors; cardiovascular diseases; predictive algorithms***

1. INTRODUCTION

Certain traits of a person are considered as risk factors for certain diseases. In case of cardiovascular diseases, many international journals, organizations have pointed some factors as the determinants of Cardiovascular Disease (CVD). Ages of the patient, gender, cholesterol in serum, fasting blood sugar, systolic and diastolic blood pressure are few of them. Many researchers claim different significance of this information about a person. These claims are based on data derived from various surveys. Machine learning and predictive algorithms can also analyze these data.

If we check medical surveys, the risks of CVDs vary between male and female, ages below and above 65, and so on. If the age attribute is split into two parts based on medical surveys, and analyzed with machine including other attributes, then algorithms may or may not support the research. With this approach, we might get to know, given the conditions, with which medical survey, the algorithms strongly agree, with which survey, the algorithms narrowly agree or disagree, and with which claim the algorithms strongly disagree.

1. BACKGROUND

Surveys are the most frequently used method used in medical education research [1]. This is done on the basis of questions and respondents. The number of respondents can be very low. For example, the medical community is the most frequently surveyed population in the US and therefore response rates are typically low [1]. Moreover, communication failures are commonplace in questionnaires [2]. Therefore, these surveys are not very dependable if the survey depends on questionnaires.

Cardiovascular diseases have risk factors such as age and gender. According to World Heart Foundation, the risk factors involved in cardiovascular diseases can be of both modifiable and Non-modifiable types [3].

Modifiable risk factors include Hypertension (high blood pressure), Tobacco use, Raised blood glucose(diabetes), Physical inactivity, Unhealthy diets, Cholesterol/lipids, Overweight and obesity [3].

Non-modifiable or the risk factors that cannot be changed include Age, Gender, Family history [3].

These risk factors were mostly found by using statistics. From this, we get to know that people above age 65 have a greater percentage of death if they have CVDs [4]. Men after 45 and women after 55 years of age or the menopause, have the same risk [5]. Taking these facts into account, we can take age 65 as a risk factor for the narrowing of the coronary arteries more than 50%.

People with diabetes or high blood sugar [3], cholesterols above 225 mg/dl [6], with systolic blood pressure above 140 or hypertension are more likely to have CVDs. According to WHF, after menopause woman have the same risk of CVDs as men.

1. *Machine Learning and Statistical Surveys*

Machine learning is mostly dependent on mathematical basis. Many machine learning algorithms incorporate statistics directly. Some statistics techniques have algorithms such as, CART which are very similar to machine learning. Both work on the same principle, i.e., analyzing previous data.

A statistical survey works on a population concentrating on opinions or factual data. Data analysis can be exploratory or confirmatory. However, in a typical statistical research we usually concentrate on one attribute only. Machine learning process can be confirmatory or exploratory. This includes analyzing the datasets with algorithms. A learning process takes place with a training dataset. After the learning been done, test datasets are used for reaching classifications and predictions. The algorithms used in machine might be derived from statistics, but the focus of a typical statistical surveys and results from different machine learning algorithms can vary.

**978-1-5386-4304-4/18/$31.00 ©2018 IEEE.**

1. *Significance of Machines' Outcomes about Risk Factors*

The respondents of a survey can be selected by age, gender, whether a person has diabetes and so on. While concentrating on a particular attribute, the final result might lose its touch from other factors. So it becomes hard to tell if a person's age is playing a bigger part in CVD or the fact that patient has high cholesterol. Statistical models and machine learning algorithms can include multiple attributes and values. So, significant information about the risk factors can be found by the machines’ outcome.

1. APPROACH FOR CHECKING CONSISTENCY

To check the consistency between statistics and predictive algorithms' assumptions, we needed both the medical statistics and a dataset for using with algorithms. For medical statistics, we got our references from different journals, studies and websites that were available.

The prediction, classification or accuracy of a classifier algorithm highly depends on the arrangement of data, values of attributes in the dataset which it uses as a training dataset. Based on the statistics, we split our dataset into two parts for each risk factor. For example, dataset was divided into two parts based on the cholesterol values. One dataset contained data of people who had lipid 225 mg/dl or above and another dataset contained data of rest of the people. This was done as people having cholesterol/lipid 225 mg/dl are reported to be at more risk of having CVDs.

After dividing the dataset for each experiment, we trained two algorithms with the full dataset by Weka tool. Then for each part of the split dataset, the classifiers were tested. Observing the confusion matrix of each dataset, we checked the results from various perspectives.

1. *Data Collection*

We collected the dataset from UCI Machine Learning Repository [9]. It includes a total of 14 selected attributes, Table I shows all attribute details.

TABLE I: ATTIBUTE LIST

|  |  |  |  |
| --- | --- | --- | --- |
| No | **Attribute Names** | **Attributes Details** | **Value** |
| 1. | Age | Age in years | Number (1, 2, 3……) |
| 2. | Sex | Gender | 1 = male;  0 = female |
| 3. | Cp | Chest pain | -- Value 1: typical angina  -- Value 2: atypical angina  -- Value 3: non-anginal pain  -- Value 4: asymptomatic |
| 4. | Trestbps | resting blood pressure (in mm Hg on admission to the  hospital) | Number (1, 2, 3……) |
| 5. | Chol | serum cholesterol in mg/dl | Number (1, 2, 3……) |

|  |  |  |  |
| --- | --- | --- | --- |
| 6. | Fbs | fasting blood sugar > 120 mg/dl | 1 = true;  0 = false |
| 7. | Restecg | resting electrocardiographic results | -- Value 0: normal  -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)  -- Value 2: showing probable or definite left  ventricular hypertrophy by Estes' criteria |
| 8. | thalach | maximum heart rate achieved | Number (1, 2, 3……) |
| 9. | exange | exercise induced angina | 1 = yes;  0 = no |
| 10. | oldpeak | ST depression induced by exercise  relative to rest | Number (1, 2, 3……) |
| 11. | slope | the slope of the peak exercise ST segment | -- Value 1: upsloping  -- Value 2: flat  -- Value 3: down sloping |
| 12. | ca | number of major vessels (0-3) colored  by fluoroscopy | Number (1, 2, 3……) |
| 13. | thal |  | 3 = normal;  6 = fixed defect;  7 = reversible defect |
| 14. | num | diagnosis of heart disease (angiographic disease status) | -- Value 0: < 50% diameter narrowing  -- Value 1: > 50% diameter narrowing |

We selected 219 data that are valid. The class to be reached is the 14th attribute, which can give either the value 0 or 1. The value 0 means less than 50% diameter of the heart vessel of that person is narrowing and otherwise the value is 1. The value 1 was interpreted as *at risk* and 0 was interpreted as *no risk*.

1. *Algorithms Used*

We chose two classifiers to work with the data. One is C4.5 and another one is Naïve Bayes. We used J48 (which is a java implementation of C4.5) in Weka tool. The classifiers are different in their approaches. Thus, we can get results from two different perspectives.

C4.5 is an expansion of the ID3 algorithm. J48 is the Java implementation of the C4.5 classifier in Weka data mining tool. The C4.5 is competent of handling continuous attributes, which are necessary in case of medical data, such as, blood pressure, temperature, etc. The worth of C4.5 was widely proven in medicine [11]. The efficiency of C4.5 was revealed in breast cancer and prostate cancer classification [11] to generate a decision tree and rules that may be helpful in medical diagnosis process [2].

Naive Bayes classifier uses probability values for classifications. It is different in various ways from C4.5. It is simple, popular and has a decent accuracy level. Naïve Bayes

assumes of mutual independence of attributes. The algorithm works on the assumption that variables provided to the classifier are independent [7]. It is possible to present interactions among variables using Bayesian networks [7]. The Naïve Bayes method’s performance was tested against a colorectal cancer in [8] in which the authors enhanced the effectiveness of this method [2].

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Risk Factor** | **J48 Algorithm** | | **Naïve Bayes** | | ***Medical Referen ce*** | ***Results*** |
| ***Risk***  ***%(TAB LE*** II**)** | ***Non- Risk%( TABLE***  *III***)** | ***Risk***  ***%(TAB LE*** II**)** | ***Non- Risk%( TABLE***  *III***)** |
| Ages | 80.08% | 96.15% | 76.17% | 76.92% | Age 65 or above | -J48  Inconsistent  -Naïve Byes  Inconsistent |
| Hyper tension | 86.15% | 90.26% | 69.23% | 79.22% | Blood pressur e 140  *mm Hg*  or above | -J48  Inconsistent  -Naïve Byes Inconsistent |
| Diebete s | 88.89% | 89.06% | 77.78% | 76.04% | Blood sugar 140mg/ dl or  above | -J48  Inconsistent  -Naïve Byes  consistent |
| Gender | 87.05% | 92.05% | 67.63% | 91.25% | Male | -J48  Inconsistent  -Naïve  Byes Inconsistent |
| Cholest erol/lip ids | 87.23% | 84.62% | 71.63% | 85.90% | 220mg/ dl or above | -J48  consistent  -Naïve Byes consistent |

We used J48 and Naïve Bayes programs available in the Weka tool. We used the whole dataset as training dataset and for the ‘Supplied test set’ option we used the datasets which were split from the whole dataset based on risk factors.

1. RESULT ANALYSIS

From our experiments, we created two result tables. Table II shows the results for datasets which contain risk factor values of an attribute. And Table III shows the results for datasets which does not contain risk factor values of an attribute. If a classifier classified more instances to be at risk for the dataset which contained risk factor values of an attribute than it does for the other dataset that doesn’t contain any risk factor value of that attribute, then we considered that it was consistent with medical surveys about the risk factor.

TABLE II: DATASETS CONTAINING RISK FACTOR VALUES OF AN ATTRIBUTE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Risk Factor** | **J48 Algorithm** | | **Naive Bayes** | |
| ***Yes, with percentage*** | ***No, with percentage*** | ***Yes, with percentage*** | ***No, with percentage*** |
| Age | 170(80.08%) | 23(11.92%) | 147(76.17%) | 46(23.83%) |
| Hyper- tension | 56(86.15%) | 15(13.85%) | 45(69.23%) | 20(30.77%) |
| Diebetes | 24(88.89%) | 3(11.11%) | 21(77.78%) | 6(22.22%) |
| Gender | 121(87.05%) | 18(12.95%) | 94(67.63%) | 45(32.37%) |
| Chlester ol/lipids | 123(87.23%) | 18(12.77%) | 101(71.63%) | 40(28.37%) |

TABLE III: DATASETS NOT CONTAINING RISK FACTOR VALUES OF AN ATTRIBUTE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Risk Factor** | **J48 Algorithm** | | **Naive Bayes** | |
| ***Yes, with***  ***percentage*** | ***No, with***  ***percentage*** | ***Yes, with***  ***percentage*** | ***No, with***  ***percentage*** |
| Age | 25(96.15%) | 1(3.85%) | 250(76.92%) | 6(23.08%) |
| Hyper tension | 139(90.26%) | 15(9.74%) | 122(79.22%) | 32(20.78%) |
| Diebetes | 171(89.06%) | 21(10.94%) | 146(76.04%) | 46(23.96%) |
| Gender | 74(92.05%) | 6(7.50%) | 73(91.25%) | 7(8.75%) |
| Chlester ol/lipids | 66(84.62%) | 12(15.38%) | 67(85.90%) | 11(14.10%) |

1. *Summary of experiment*

From Table IV we can observe that the results are mostly inconsistent with the medical researches. In some cases, the result for Naïve Bayes and the result for J48 are different.

TABLE IV: CONSISTENCY COMPARISON OF ALGORITHMS

From the result consistency table, we can conclude the following:

* + As both the classifiers gave similar result for cholesterol, we can conclude that the classifiers agreed on the medical survey of having cholesterol 225mg/dl or above increases the risk of narrowed artery or simply heart problems.
  + Risk factors cannot be determined by considering only one attribute of the people being surveyed. Other traits of patients should be considered.
  + In our experiments with Weka tool, most of the times J48 gave a more accurate classification. However, despite of c not giving accurate classifications, it still gave a statistical result which was more consistent with medical research than J48 was. If we consider the results of naïve Bayes for age risk factor in the two datasets, they are almost same. So, for age, diabetes and cholesterol, naïve Bayes gives a result that is more consistent with medical research.

1. CONCLUSION AND FUTURE WORK

Datasets play a huge role in classification by any classifier. Therefore, it's tough to get a clear picture from any classification by an algorithm whether it is consistent with medical research or not, regarding risk factors of a disease.

The legacy of the work done in this paper can be applied two separate datasets for training and testing. Although we used two different datasets here, the testing dataset is a subset of the training dataset. Therefore, if there were two datasets which

are no way part of each other and still have same attributes, the results will be more acceptable.

***References***

1. Toni Peters, PhD & Julie Irish, PhD ,“ Introduction to Survey Research Methods”, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73. [Online]. Available:https://hms.harvard.edu/sites/default/files/assets/Sites/Academ y/files/CTL%20Resource%20Survey%20Research.doc. [Accessed: 10- Oct- 2016]
2. Kamila Aftarczuk, “Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems”, Thesis no: MSE- 2007-21, September 2007, School of Engineering, Blekinge Institute of Technology, Sweden.
3. R. Nicole, J. Name Stand. Abbrev “Cardiovascular disease risk factors,”[Online].Available:http://www.world-heart- federation.org/cardiovascular-health/cardiovascular-disease-risk-factors. [Accessed: 10- Oct- 2016]
4. American Heart Association."Understand Your Risk of Heart Attack".[Online].Available[:http://www.heart.org/HEARTORG/Conditio ns/HeartAttack/UnderstandYourRiskofHeartAttack/Understand-Your- Risk-of-Heart-Attack\_UCM\_002040\_Article.jsp#.](http://www.heart.org/HEARTORG/Conditions/HeartAttack/UnderstandYourRiskofHeartAttack/Understand-Your-Risk-of-Heart-Attack_UCM_002040_Article.jsp) [Accessed: 10- Oct- 2016]
5. Tourangeau R, Rips LJ, Rasinski K. “The Psychology of Survey Response.” New York, NY: Cambridge University Press; 2000
6. Castelli WP. “The new pathophysiology of coronary artery disease.” Am J Cardiol.1998;82:60T-65T
7. Prakash Mahindrakar, Dr.M. Hanumanthappa, “Data Mining In Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges” 6, Nov-Dec 2013.
8. Sarkar M., XinZhi Q., Peng-Kheong L., TzeYun L., “Performance of existing prognostic factors for colorectal cancer prognosis - a critical view”, Engineering in Medicine and Biology Society, 2001 vol. 4, 3909- 3912
9. Andras Janosi, M.D.,William Steinbrunn, M.D, Matthias Pfisterer, M.D, Robert Detrano, M.D., Ph.D., “UCI MachineLearningDataset”, 1988-07- 01,[Online].Available:[http://archive.ics.uci.edu/ml/datasets/Heart+Disea](http://archive.ics.uci.edu/ml/datasets/Heart%2BDisea) se. [Accessed: 10- Oct- 2016]
10. Mitchell T. M., “Machine Learning” (Redmond, McGraw-Hill, 1997)
11. Tahir M.A., Bouridane A., “Novel Round-Robin Tabu Search Algorithm for Prostate” 4th October 2006
12. Fayyad, U, “Data Mining and Knowledge Discovery in Databases: Implications for scientific databases,” Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington,

USA, 2-11, 1997

1. National Cholesterol Education Program (NCEP): Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III final report, 2002. Circulation 106:3143.