RESEARCH ARTICLE OPEN ACCESS

Enhanced Classification Model for Cervical Cancer Dataset based on Cost Sensitive Classifier

Hayder K. Fatlawi
Information Technology Research and Development Center , University of Kufa, Najaf, Iraq

--*************************--------------------------------

Abstract:

Cervical cancer threatens the lives of many women in our world today. In 2014, the number of women infected with this disease in the United States was 12,578, of which 4,115 died, with a death rate of nearly 32%. Cancer data, including cervical cancer datasets, represent a significant challenge data mining techniques because absence of different costs for error cases. The proposed model present a cost sensitive classifiers that has three main stages; the first stage is prepressing the original data to prepare it for classification model which is build based on decision tree classifier with cost selectivity and finallyevaluation the proposed model based on many metrics in addition to apply a cross validation.The proposed model provides more accurate result in both binary class and multi class classification. It has a TP rate (0.429) comparing with (0.160) for typical decision tree in binary class task.

Keywords **—Decision Tree, Cost Sensitive Classifier, Cervical Cancer, Imbalance Class.**

--*************************--------------------------------

I. INTRODUCTION

Cervical cancer threatens the lives of many women in our world today. In 2014, the number of women infected with this disease in the United States was 12,578, of which 4,115 died, with a death rate of nearly 32% [1]. Cancer data, including cervical cancer datasets, represent a significant challenge to the techniques of data mining. The challenge is that these techniques utilize measures of the accuracy for the models extracted from the data, not taking into account the difference between the accuracy of the patient's classification as infected and the accuracy of the classification as uninfected. In this work we suggest a classification model to deal with this problem using data mining techniques.

Data mining is gaining useful knowledge from data using machine learning techniques and statistical methods [2]. It has three main phases; the data is prepared to mining process then applying machine learning techniques for knowledge extraction and finally the results will processed in understandable form to be helpful for decision making [Jiaw06].

Medical databases are often big, complex and unstructured. The size of these data are huge because they are associated with the lives of the public peoples, their medical history, hospital and health center information, health insurance, medical staff, etc. The complexity of this data is related with the number of attributes and the correlation of each attribute with the target. The structure of medical dataset is often designed for archiving and information retrieval, not for mining purposes. Therefore, the researcher faces a number of difficulties in dealing with it directly without pre-processing [3].

On the other hand, cancer data differ from the rest of the medical data according to the importance or weight of each value of the target attribute values. For example, the importance of the value (infected) much more than the value of (non-infected) so the cost of error in the first value is much higher than the second value. This difference requires that the data mining technique is capable of dealing with the values of different weights which lead to develop cost sensitive classifier.

II. METHODOLOGY

The proposed model has three main stages; the first stage is prepressing the original data to prepare it for classification. The second stage is building a classification model based on decision tree classifier with cost selectivity. The final stage is evaluate the proposed model based on many metrics in addition to apply a cross validation.

A. Data Pre-processing

Real-world databases mostly tend to contain low quality data which could not be used directly in mining process without pre-processing [4]. Data pre-processing techniques are classified into two groups; the first group is concerned with cleaning the data from noisy, missing, duplicate, and inconsistent data samples. The second group focuses on reconstructing the data by binarization, construction the attributes and aggregate data rows [4, 5]. This section explains the concepts of the required pre-processing steps to improve efficiency the mining processand reduce the complexity of the required resources (i.e. storage and time).

1. Attributes Construction

The original set of attributes may not be useful directly for mining; so many methods are applied to produce new set of attributes. Attribute construction is one of these methods; it includes constructing new attributes from original attributes. It aims to improve the accuracy and understanding in high-dimensional data. Attribute construction can provide knowledge discovery by discovering hidden information about the relationships among data attributes [4].

2.Data Normalization

The target of data normalization is to reduce the range of the values to be mostly in [-1,1] or [0,1] range. The purpose of this step is to prevent the dominant of the attribute that has large value. For example, age attribute usually has value between 1 and 120 while number of pregnancies attribute may has values between 0and 12.

3. Best attributes selection

Even the computation efforts (i.e. execution time and memory size) is a considerable issue in the classification task but the accuracy of the results is very important especially with cancer classification tasks. From computation calculations viewpoint, A large number of attribute requires more computational resources in classification, from accuracy viewpoint, theoretically, the increment of the number of attributes lead more discriminating power, but in practice, presence some irrelevant attribute may degrade the performance of classification model [6].

Correlation can be defined as a measurement of association amongdata. If values of prediction's target Y increase when values of attribute xincreased, the correlation would be called a positive correlation. The termnegative correlation is used when values of Y decreased after increasingvalues of x . It is used usually for attribute selection in pre-processing. Thereare many type of correlation and using of it depends on the nature of data [7, 8, 9].

B. Cost sensitive classifier

Decision tree is a simple and powerful form of data analysis which allows us to predict, explain, describe, or classify a target. [10]. A decision tree represents a flowchart-like tree structure, where every non terminal node denotes a condition on an attribute, to split data records which have different characteristics. Each branch represents the result of that condition, and each leaf node (i.e. terminal node) holds a class label. The first node in a tree is the root node.

i. Binary Class Decision Tree

In binary decision tree; each internal node branches to exactly only two other nodes [Pang05, Jiaw06].For new data record X, which has unknown target class label y, the values of attributes of X could be tested against the decision tree. Tracing starts from root node to leaf node, which assign the value for class for X. The building of decision tree classifiers does not require any domain knowledge, and therefore is suitable for exploratory knowledge discovery. Representation of acquired knowledge by using binary tree form is self-evident and generally easy to understand by humans [11].

ii. Multi Class Decision Tree

If the target of classification task has k possible values where k > 2, techniques of classification should be extended to allow for multiclass classification task. One-versus-all (OVA) is simple approach which treat with k classes by training k binary classifiers, one for each class [11].

The targetattribute may have multi possible values for infection (surely infected, likely infected, and not infected). Three classifiers could be used to solve this task, first classifier is trained for value surely infected as a class andothers values likely infected, and not infected) would treated together as one class. The other two classifiers could be trained in the same manner.

iii. Cost sensitive classifier

Mostly, the ratio of infected patients to the not infected patients in medical dataset is not equal especially in cancer dataset. Imbalance class problem lead the classification model to focus on the major class (i.e. not infected) which has less value on decision making. The solution for this problem is to modify the mechanism of model building and the measures that used to evaluate the performance of the classification.

For binary class task, the minor class is named as a positive class while the major class is named as a negative class. The confusion matrix is a representation tool that summarizes the numbers correctly classified data rows in addition to the incorrectly ones. The true positive TPis the number of data row that have a positive class and classified as a positive class, if they are classified as a negative class, they are named the false negative FN. The true negative TN is the number of data rows that have a negative class and classified as negative class, if they are classified as a positive class, they are named the false negative FP. In a cost matrix, each one from four measures (TP,FN,TN an FP) has a specific weight according to the risk of wrong classification. Four metrics is used for evaluating the proposed model (TP rate, F-measure, Recall and ROC curve).

TP rate = TP/(TP + FN) ……. …… (1)

F-measure = (2×TP) / (2×TP + FP + FN) … (2)

Recall = TP / (TP + FN) ………… (3)

In ROC curve, true positive rate is plotted along they axis while false positive rate is shown on the x axis.

Dividing available dataset between testing and training process may lead to unreliable evaluation for the model [2, 4]. The problem is happened because the selected part of data could be not representative for all data. The solution is to use cross-validation; it repeats the whole process (i.e. training and testing) many times with different samples of data records.In ten cross-validation, classification model would repeat ten times, In each iteration, 10% of data selected for testing and the remainder would be used for training. Error of all iterations is averaged to get an overall error rate [6, 12].

III. EXPERIMENT AND RESULTS

A. Database Description

Cervical cancer dataset 2017 [13] consist of 858 data rows, each one has 36 attributes. Four medical test attributes constitute the target of this database. The description of each attribute is shown in Table (1).

TABLE I
CERVICAL CANCER DATASETDESCRIPTION

Attribute Name	Attribute Type	Attribute Name	Attribute Type
Age	Integer	STDs:pelvic inflammatory disease	Boolean
Number of sexual partners	Integer	STDs:genital herpes	Boolean
Age of First sexual intercourse	Integer	STDs:molluscum contagiosum	Boolean
No. of pregnancies	Integer	STDs:AIDS	Boolean
Smokes	Boolean	STDs:HIV	Boolean
No. of smoking years	Real	STDs:Hepatitis B	Boolean
Smokes (packs/year)	Real	STDs:HPV	Boolean
Hormonal Contraceptives	Boolean	STDs: Number of diagnosis	Integer
Years of Hormonal Contraceptives	Real	STDs: Time since first diagnosis	Integer
IUD	Boolean	STDs: Time since last diagnosis	Integer
Years of IUD	Real	Dx:Cancer	Boolean
STDs	Boolean	Dx:CIN	Boolean
No. of STDs	Integer	Dx:HPV	Boolean
STDs:condylomatosis	Boolean	Dx	Boolean
STDs:cervical condylomatosis	Boolean	Hinselmann (target)	Boolean
STDs:vaginal condylomatosis	Boolean	Schiller (target)	Boolean
STDs: vulvoperineal condylomatosis	Boolean	Cytology (target)	Boolean
STDs:syphilis	Boolean	Biopsy (target)	Boolean

B. Pre-processing Stage

At the beginning, the absence of four target attributes complicates the classification task.

Attribute construction presents a solution to the above problem by generate a new attribute from the information of other attributes. A combination of four targets (Hinselmann, Schiller, Cytology and Biopsy) is made to produce one target with five values (0, 1,2,3,4 and 5). The values of new target represent the number of medical tests which indicate infection of cancer.

The second step in pre-processing of Cervical Cancer dataset is data normalization. The values of all attributes except the target attribute are converted to the range (0-1). For example, before normalization, the values of Age attributes were (13-84) with an average (26.8), after normalization the minimum value (13) became (0) and maximum value (84) became (1) with an average (0.195).

The final step in pre-processing is to reduce the number of attributes by best attributes selection. Correlation based selection is performed between each attribute and the target.The high value of correlation represents the best preferred attribute. According to this concept, two of Cervical Cancer dataset attributes are ignored because they have a zero correlation with the target in both binary class and multi class classification model as shown in Table (2) and Table (3).

TABLE 2
RANKING OF CERVICAL CANCER ATTRIBUTES ACCORDING TO CORRELATION WITH BINARY CLASS CLASSIFICATION

Attribute Name	**Correlation with target**	**Attribute Name**	**Correlation with target**
Dx	0.42274	STDs:condylomatosis	0.08282
Dx:HPV	0.36479	Smokes	0.06715
Dx:Cancer	0.36479	Number of pregnancies	0.06152
Dx:CIN	0.25658	Smokes (packs/year)	0.05328
STDs:HPV	0.11888	First sexual intercourse	0.03195
STDs: Number of diagnosis	0.11819	STDs:vaginal condylomatosis	0.02972
IUD	0.11136	STDs: Time since last diagnosis	0.01944
STDs	0.10742	STDs:syphilis	0.01652
IUD (years)	0.10407	Number of sexual partners	0.0159
STDs (number)	0.10281	STDs:pelvic inflammatory disease	0.01483
STDs:HIV	0.1013	STDs:molluscum contagiosum	0.01483
Smokes (years)	0.09029	STDs:Hepatitis B	0.01483
STDs:vulvo-perineal condylomatosis	0.08605	STDs: Time since first diagnosis	0.01066
Age	0.0859	Hormonal Contraceptives	0.00938
STDs:genital herpes	0.08401	STDs:cervical condylomatosis	0
Hormonal Contraceptives	0.08332	STDs:AIDS	0

TABLE 3
RANKING OF CERVICAL CANCER ATTRIBUTES ACCORDING TO CORRELATION WITH MULTI CLASS CLASSIFICATION

Attribute Name	**Correlation with target**	**Attribute Name**	**Correlation with target**
Dx:Cancer	0.16247	Number of pregnancies	0.04725
Dx:HPV	0.16247	Age	0.03691
Dx	0.1475	First sexual intercourse	0.02911
STDs: Number of diagnosis	0.1253	STDs:vaginal condylomatosis	0.02572
STDs:HIV	0.11132	STDs:HPV	0.01816
STDs (number)	0.10982	STDs:pelvic inflammatory disease	0.01283
STDs	0.10559	STDs:molluscum contagiosum	0.01283
STDs:vulvo-perineal condylomatosis	0.10184	STDs:Hepatitis B	0.01283
STDs:condylomatosis	0.09876	STDs: Time since first diagnosis	0.011
STDs:genital herpes	0.08878	Number of sexual partners	0.00979
Hormonal Contraceptives (years)	0.08354	STDs:syphilis	0.00932
IUD (years)	0.07292	Smokes (packs/year)	0.00848
IUD	0.07247	STDs: Time since last diagnosis	0.00375
Smokes (years)	0.06732	Hormonal Contraceptives	0.00175
Dx:CIN	0.06641	STDs:AIDS	0
Smokes	0.06292	STDs:cervical condylomatosis	0

C. Cost Sensitive Classification

The classification model building is performed using binary decision tree in which each node produce two child nodes during tree growth. According to the number of classes, binary class classification has two classesincludes two values for the target attributes; No for not infected patients (i.e. The value of constructed target is zero) and Yes if The value of constructed target is (1 or 2 or 3 or 4). In multi class classification, there are five classes (0, 1, 2, 3, and 4).

The error in detecting of infected patient as not infected patient is very dangerous and may lead to death as a result of staying without necessary medical procedures. Wherefore, the cost of those cases must be higher, in this step the factor of (10:1) is used for binary class task as shown in cost matrix in Table (4).

TABLE 4

THE COST MATRIX WITH BINARY CLASS CLASSIFICATION

Actual Class	Classified Class	
	Positive Class	Negative Class
Positive Class	0.0	1.0
Negative Class	10.0	0

According to four evaluation measures (TP rate, F-Measure, ROC Area, Recall), the cost sensitive classifier produce more accurate result from typical decision tree for infected patients. Table (5) shows a comparison between two models and focusing on the error rate of positive class.

TABLE 5

COMPARISON BETWEEN TYPICAL DECISION TREE AND COST SENSITIVE DECISION TREE FOR POSITIVE CLASS (INFECTED WITH CANCER) IN BINARY CLASS

Model	TP rate	F-Measure	ROC Area	Recall
Decision Tree	0.160	0.229	0.533	0.160
Cost Sensitive Decision Tree	**0.429**	**0.306**	**0.609**	**0.429**

For multi class task, the cost matrix consists of (5× 5) values. The maximum cost value (8) is used for class 4 that classified as class 1, i.e. the patient which has 4 positive medical test as infected should be given a higher cost when he is classified as not infected. Table (6) shows the cost matrix for multi class task.

TABLE 6

THE COST MATRIX WITH MULTI CLASS CLASSIFICATION

Actual Class	Classified class				
	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	0.0	1.0	1.0	1.0	1.0
Class 1	2.0	0.0	1.0	1.0	1.0
Class 2	4.0	2.0	0.0	1.0	1.0
Class 3	6.0	4.0	2.0	0.0	1.0
Class 4	8.0	6.0	4.0	2.0	0.0

The evaluation measures show some different result for multi class task, for class 1 and class 3 , all four measures detect an improvement with cost sensitive model comparing with typical decision tree. For class 2 and class 4 ROC area measure point to this improvement. Table (7) shows a comparison between two models for four classes of positive medical tests.

TABLE 7

COMPRESSION BETWEEN TYPICAL DECISION TREE AND COST SENSITIVE DECISION TREE FOR POSITIVE CLASS (INFECTED WITH CANCER) IN MULTI CLASS

Model	Class	TP rate	F-Measure	ROC Area	Recall
Decision Tree	**1**	0.000	0.000	0.497	0.000
	2	0.000	0.000	0.420	0.000
	3	0.030	0.038	0.355	0.030
	4	0.000	0.000	0.364	0.000
Cost Sensitive Decision Tree	**1**	**0.122**	**0.100**	**0.557**	**0.122**
	2	0.000	0.000	**0.442**	0.000
	3	**0.061**	**0.073**	**0.527**	**0.061**
	4	0.000	0.000	**0.465**	0.000

IV. CONCLUSIONS

Improving a classification model without considering the real cost for each error case may lead to unreliable results. The proposed model depends on a decision tree classifier with a cost matrix that different cost values. It contain a higher cost for error in cases that have a positive medical tests as infected patients but classified as not infected patients. The proposed model provides more accurate result in both binary class and multi class classification. It has a TP rate (0.429) comparing with (0.160) for typical decision tree in binary class task.

REFERENCES

[1] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2014 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2017. Available at: http://www.cdc.gov/uscs.

[2] Pang-Ning Tan, Michael Steinbach , and Vipin Kumar, "Introduction to Data Mining", ISBN-13: 978-0321321367, Addison- Wesley, 2005.

[3] Jared Dean, "Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners", ISBN-13: 978-1118618042, Wiley Publishing, 2014.

[4] Jiawei Han and Micheline Kamber, "Data Mining: Conceptsand Techniques", 2nd Edition, ISBN-10: 1-55860-901-6, Elsevier, 2006.

[5] Krzysztof Cios, Witold Pedrycz, Roman Swiniarski, and LukaszKurgan, "Data Mining A Knowledge Discovery Approach", Springer, ISBN-13: 978-0-387-33333-5, 2007.

[6] Ian H. Witten, Eibe Frank, and Mark A. Hall, "Data Mining:Practical Machine Learning Tools and Technique", 3nd Edition, ISBN-13: 978-0123748560, Morgan Kaufmann, 2011.

[7] Thomas Dietz and Linda Kalof,"Introduction to Social Statistics: The Logic of Statistical Reasoning" , ISBN-13: 978- 1405169028, Wiley-Blackwell, 2009.

[8] Anthony Graziano and Michael Raulin , "Research Methods: AProcess of Inquiry" , ISBN-13: 978-0205907694 , 8th Edition,Pearson publishing, 2012.

[9] Thomas Cleff, " Exploratory Data Analysis in Business and Economics: An Introduction Using SPSS, Stata, and Excel", PrintISBN-13: 978-3319015163, Springer; 2014.

[10] Barry deVille, “Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner”, ISBN-13: 978-1590475676, SAS Publishing, 2006.

[11] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, 3th Edition, ISBN 978-0-12-381479-1, Elsevier, 2013.

[12] Max Kuhn and Kjell Johnson, “Applied Predictive Modeling”, ISBN 978-1-4614-6849-3, Springer, 2013.

[13] Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.' Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017.