Predicting Trachoma Using Machine Learning Techniques

Akbar khan[1], Abdul Samad[2], Faizullah Khan[3], Surat Khan[3]

[1]Department of Computer Engineering, [2]Department of Computer Science, [3]Department of Telecom Engineering, Balochistan University of Information Technology, Engineering and Management Sciences, Quetta, Pakistan

Abstract

Machine learning is the area of artificial intelligence which uses statistical methods for data classifications. It is now usually applied in different areas like business, government, education and health. In health sector, it is almost used for the prediction, risk factor identification and many more. Among these applications it is used for different eye diseases. Trachoma is a common eye disease that causes blindness. This paper aims to use different techniques of machine learning algorithms and to find out the most effected factor causing trachoma.

Keywords: Machine learning; WEKA; Decision Tree; Random Forest; J48; Support Vector Machine

Corresponding author's email: *akbarkhannasar80@gmail.com*

INTRODUCTION

Machine learning has widely used in health. Machine learning is a sub field of computer science which has taken from determined learning hypothesis in medical field (Morris, 2004). Trachoma is responsible for causing blindness, according to world health organization it causes more than 3% of visual deficiency. Trachoma is usually found in youngsters and adults and normally found among nomads and unprivileged areas where the climatic condition is not good (M. Alemayehu et al., 2015). There are some several factors that lead the emergence of trachoma namely poor financial status, lack of cleanness, and social irresponsibility (Mowafy et al., 2014).

Machine learning calculates simple information and changes it into useful information. Machine learning can be divided into two unique classifications specifically supervised learning and unsupervised learning Vallejo-Alonso et al., 2011). Trachoma causes visual impairment while chlamydia tachometric is the world driving desirable reason for visual impairment. This disease causes coating of the upper eyelid which twists the top edge and causes the lashes to touch the surface of eye (Emerson et al., 2006). Trachoma is observed very high in rural areas and too low in urban areas where the sanitation is too reliable (Hsieh et al., 2000).

MATERIALS AND METHODS

Data source and data collection

The primary and essential data was obtained from Helper Eye Hospital Quetta and authorized authority in the domain in hard form and then I converted it into a digital form to make an excel sheet as CSV file which can be readable by WEKA a machine learning tool. The collected data 1000 patients were examined by and expert in the area and out of 1000 instances there were 12 positive class and 988 were negative class. The complete process was a questionnaire based in which there were multiple question in the original questionnaire Performa and I have selected the most important question which is widely used while doing such type of research

Data structure

The data structure is like a table in which each and every variable and its description and type has shown. The classification of target class contained two binary classes namely:

- Patients has trachoma
- Patients has no trachoma means normal

This generates a problem known as binary class.

Table 1: The elaboration of all variables and their description

S.No	Variables	Description	Type
1	Page	1-5, 6-10, 11-15, 16-20, 21-40, 41-60 and 60+	Nominal
2	District	KSF, Kech, Loralai, Quetta	Nominal
3	DFW	Daily face washing	Nominal
4	Sex	Male and Female	Nominal
5	UoL	Use of Latrine	Nominal
6	Trch	Trachoma	Nominal
7	WFW	Water for face washing	Nominal
8	SWCD	Solid waste collected disposed	Nominal
9	SWCS	Solid waste collection system	Nominal
10	TDW	Type of drinking water	Nominal
11	SDW	Source of drinking water	Nominal
12	T&D	Time and distance	Nominal
13	HC	House connected	Nominal
14	LoH	Latrine of house hold	Nominal
15	ToL	Type of latrine	Nominal

Data Pre-Processing

The data pre-processing steps include the removal of unnecessary questions from the given questionnaire Performa and elaboration of all those attributes which are most important while doing this type of research. Every attribute holds a categorical value and represented individually in WEKA environment.

SMOTE

Synthetic Minority Oversampling Technique is powerful too which is widely used in machine learning. In SMOTE basically there are two most important techniques are used named as oversampling and under sampling. The concept of oversampling is for minority class where the number of positive class is less than that of negative class. Here the main focus is to balance the class imbalance while using these both techniques (He and Garcia, 2009).

RESULTS AND DISCUSSION

In experiments, first of all, I have selected 12 positive classes and 988 negative classes and applied all four algorithms to find out the required results as shown in the given tables and figures.

Table 2: The result of sensitivity, specificity and AUC for SMOTE

Sensitivity	Specificity	AUC	Algorithms
0	1	0.5	SVM
0	0.997	0.394	LR
0	1	0.433	J48
0	1	0.568	RF

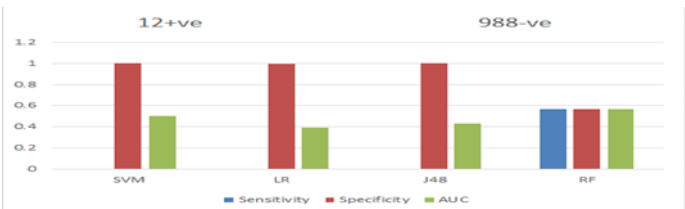

Figure 1: The graphical view results for 12+ve 988-ve.

Similarly, I applied oversampling techniques to increase the number of minority class up to 24 and repeat the same process to achieve results

Table 3: The values of sensitivity, specificity and AUC

Sensitivity	Specificity	AUC	Algorithms
0	1	0.5	SVM
0	0.997	0.699	LR
0	1	0.449	J48
0	1	0.745	RF

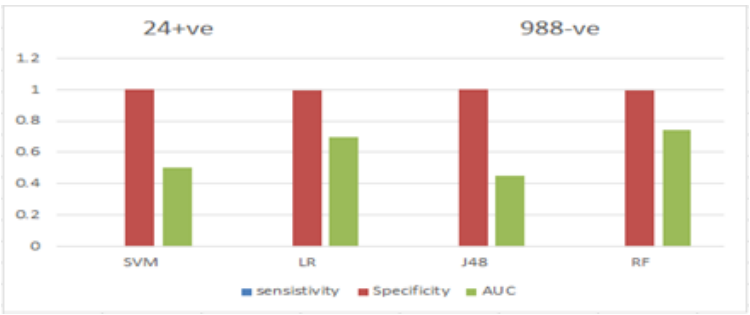

Figure 2: The graphical view result for 24+ 988-ve.

Table 4: The values of sensitivity, specificity and AUC

Sensitivity	Specificity	AUC	Algorithms
0	1	0.5	SVM
0	0.997	0.811	LR
0	1	0.482	J48
0	0.998	0.841	RF

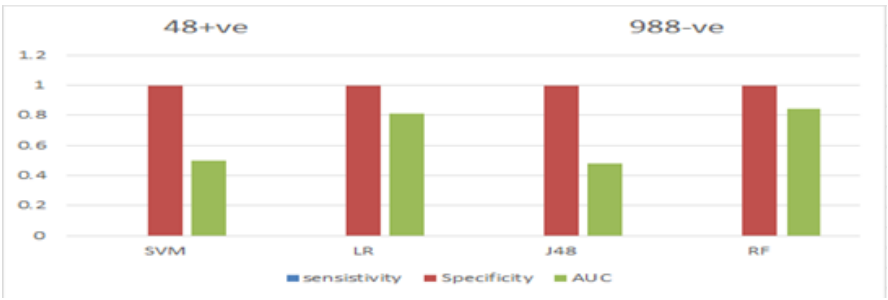

Figure 3: The graphical view result for 24+ 988-ve.

Table 3: The values of sensitivity, specificity and AUC

Sensitivity	Specificity	AUC	Algorithms
0	1	0.5	SVM
0	0.968	0.867	LR
0	0.979	0.699	J48
0	0.969	0.896	RF

Table 6: The values of sensitivity, specificity, AUC and Algorithms

Confusion matrix	Sensitivity	Specificity	AUC	Algorithms
0 1 1 0	0	1	0.5	SVM
	0	0.97	0.394	LR
	0	1	0.433	J48
	0	1	0.568	RF
0 2 1 0	0	1	0.5	SVM
	0	0.994	0.397	LR
	0	1	0.433	J48
	0	0.998	0.562	RF
0 5 1 0	0	1	0.5	SVM
	0.985	0.083	0.393	LR
	0	0.995	0.401	J48
	0	0.994	0.556	RF
0 10 1 0	0	1	0.5	SVM
	0.083	0.982	0.407	LR
	0	0.986	0.454	J48
	0	0.987	0.548	RF
0 15 1 0	0	0.997	0.499	SVM
	0.166	0.972	0.373	LR
	0	0.960	0.383	J48
	0	0.960	0.545	RF
0 20 1 0	0	0.980	0.49	SVM
	0.166	0.962	0.408	LR
	0	0.951	0.43	J48
	0	0.949	0.543	RF

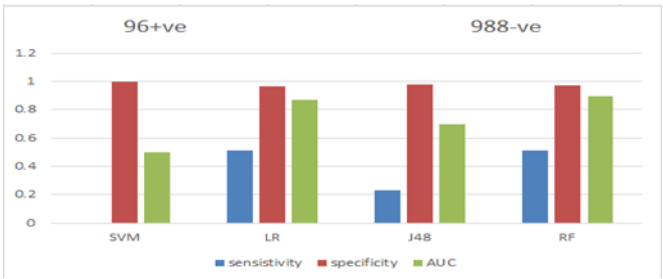

Figure 4: The graphical view of SVM, LR, J48 and RF

Cost sensitive classifier

Cost sensitive classifier shows penalty on incorrect classification. And the main function of cost sensitive classifier is hypothesis that reduces the cost of data (H. He and E. Garcia .2009). It is very useful for sampling method and attempts to balance the distribution. By applying cost sensitive classifier on imbalanced data putting different penalties the result was weak as compared to the results of SMOTE so in this regard the whole results of cost sensitive classifier have shown in the given table.

CONCLUSION

This paper examines the factor identification and ordered of variables which has been cleared from table 3.1 and 3.2 to evaluate machine learning algorithms for imbalance data AUC is a good measure with random forest and SMOTE performs better than cost sensitive classifier for 96+ve instances and 988 –ve instances. It was also found that the variable UoL is the main causing factor of Trachoma and identified with huge amount of confidence.

REFERENCES

- Alemayehu M, Koye DN, Tariku A, Yimam K. (2015). Prevalence of active trachoma and its associated factors among rural and urban children in Dera Woreda, Northwest Ethiopia: a comparative cross-sectional study. Biomed research international 2015 (2015):1-8.
- Emerson PM, Burton M, Solomon AW, Bailey R, Mabey D. (2006). The SAFE strategy for trachoma control: using operational research for policy, planning and implementation. Bulletin of the World Health Organization 84(8):613-619.
- He H, Garcia EA. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21(9):1263-1284.
- Hsieh YH, Bobo LD, Quinn TC, West S. K. (2000). Risk factors for trachoma: 6-year follow-up of children aged 1 and 2 years. American Journal of Epidemiology 152(3):204-211.
- Morris J. (2004). Beyond clinical documentation: using the EMR as a quality tool. Health management technology 25(11):20-25.
- Mowafy MA., Saad NE, El-Mofty HM, ElAnany MG, Mohamed MS. (2014). The prevalence of Chlamydia trachomatis among patients with acute conjunctivitis in Kasr Alainy ophthalmology clinic. The Pan African medical journal 17(151):1-6.
- Vallejo-Alonso B, Rodrigues-Castellanos A, Arregui-Ayastuy G. (2011). Identifying, measuring, and valuing knowledge-based intangible assets. New perspectives. New York, IGI Global.