IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

An Experimental Study of Applying Machine Learning in Prediction of Thyroid Disease

Hetal Patel[1*]

[1]Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, CHARUSAT, Changa, India

**Corresponding Author: hetalpatel.mca@charusat.ac.in Tel.: 78743-45024*

Available online at: www.ijcseonline.org

Accepted: 15/Jan/2019, Published: 31/Jan/2019

***Abstract*—**New advancements have made it workable for an extensive variety of individuals – including humanities and sociology scholastics, advertisers, legislative associations, instructive foundations – to deliver, share, collaborate and arrange data. Monstrous informational collections that were once dark and particular are being amassed and made effectively open. The Huge volumes of heterogeneous therapeutic information these days expanding and easily obtainable from various healthcare organizations. Nowadays, the Thyroid disease is one of the common diseases found in human. The Thyroid hormones created by the thyroid organ to help the control of the body's digestion. Because of the variations from the norm of thyroid capacity, there might be a lower production of thyroid hormone, which is known as hypothyroidism, or higher production of thyroid hormone, which is known as hyperthyroidism. In this paper, an examination of thyroid disease is carried out by performing experiment of various Machine Learning algorithms techniques such as Naïve Bayes, Support Vector Machine, Multiclass Classifier, Logistic and K Nearest Neighbour. The informational index utilized for this investigation on hypothyroid is taken from UCI information store. The experiment is also completed with WEKA and RConsole. The comparison of various parameters are done and as a result the execution and investigation of different grouping calculation is determined. In the result, it is found that Multiclass Classifier gives preferable exactness over other embraced calculations.

***Keywords*—**Machine Learning, Health Care, Thyroid Disease, Prediction

I. INTRODUCTION

Day by the, the growth of data is increasing in the entire domain like education, business, civil services, scientific research, healthcare and so on. The most major challenge for Machine Learning is to explore the huge volumes of data and find out the meaningful information or knowledge and do the prediction for future actions.

Machine learning (ML) is a one of the branch of artificial intelligence. It employs a statistical, probabilistic and optimization methods that supports to "learn" from past patterns and to recognize difficult pattern from expansive, uproarious or complex datasets. The machine learning is employed in every domain like business, health, education, agriculture etc. Here in this work, the Thyroid data set is used for doing the experiments using various machine-learning algorithms. As the medical dataset are very complex, the ML technique is most suitable for it [1]. Besides, optimization techniques, such as genetic algorithm (GA) and particle swarm optimization (PSO), the machine learning techniques is used for further improvement in the prediction accuracy for large data sets.

ML strategies gives an arrangement of apparatuses that can naturally identify pattern from data, which would then be able to be used for prediction. Based on this, it is also possible to derive the model. Advancements in ML calculations and computational abilities have now made it conceivable to scale building investigation, basic leadership and structure quickly.

Healthcare sector is no exception amongst the numerous applications of Machine Learning, contributing for making the world a better place to live. Healthcare can directly affect the quality of life and hence Machine Learning applications in healthcare are of keen interest to researchers. During last few years, due to automation in the healthcare industry, we are facing explosion of data, which can be channelized for analysis to reach for important conclusions useful to the doctors and patients. It is widely used to predict epidemics, cure disease, improve quality of life and avoid preventable deaths. Through understanding about a patient in lesser time and at the early age can help prevent further damage in a patient's condition, as corrective measures can be initiated early and will be more effective.

The paper is organized as follows: Section II describes about how machine learning in health care is employed with its applications; Section III is about related work; Section IV presents the tools and methodology; Section V describes the experimental results and discussion; Finally,
Section VI includes the conclusions.

II. MACHINE LEARNING IN HEALTH CARE

To start with, there are Machine Generated Data generated from different devices used in the healthcare systems, like, remote sensors, wearable devices, smart meters etc. Then we have Biometric Data, obtained from individuals' physical scanning. A huge amount of Human Generated Data is available in the healthcare including patient's case history, laboratory records, hospital admission records etc. Transactional Data come from the financial details related to patient like billing and insurance. In addition, Epidemiological Data are available through surveys and disease registries. The data generated regarding the patient is complex as it is collected through various diagnosis method like clinical examination, report before the treatment, report after the treatment, hospital resource management records, medicine history, x-ray report, MRI report, patient history etc. which has become difficult to organize correctly. There might be a chance of taken wrong decision of treatment if the data is not organized properly. This expansion in information volume requires manners by which information can be separated and handled effectively. The precise finding of perilous sicknesses such as liver disease, various type of cancer, heart disease, etc is a very crucial task in medical science. The researcher and technoilogy both can be incorporated together to accomplish best results for precise diagnosis of diseases. This sort of trouble could be settled with the assistance of ML technique [2].

The largest endocrine glands, which has two connected lobes, is the thyroid gland. When the thyroid gland fails to function properly, it can result into two undesirable medical conditions - hypothyroidism and hyperthyroidism. Hypothyroidism means an underactive thyroid gland - incapable of secreting the sufficient hormone while hyperthyroidism means the overactive thyroid gland - producing more hormones than necessary. If both these conditions can be detected early, we can prevent the further damage to human body and help fast recovery. ML can be used to analyse the patients 'data regarding the numerous symptoms and predict whether the patient is likely to have the dysfunction of thyroid or not.

III. RELATED WORK

Exponential growth in the size of data being generated, collected and stored every minute along with the development of technology in recent year. By using the Machine Learning has given a way to Big Data Analytics, as it can contribute towards better decision making in crucial situations and can have significant impact on the aftermath.

Machine Learning has obvious applications in the scientific fields such as astronomy, atmospheric science, medicine, genomics, biologic, biogeochemistry, healthcare and other complex and interdisciplinary scientific researches where bulk of data are to be dealt with. Internet-based applications also face large volume of data which are generated through search engines, social network analysis, online communities, recommender systems, reputation systems, and prediction markets [3].

The most two decades, a wide range of different Machine Learning algorithms have been applied for the disease forecast and prediction [4].
The capabilities and limitations of artificial neural networks in medical diagnosis have been reviewed by the author Amato, Filippo, et al by using selected examples [5]. The comparison of the performance of the SVM and k-nearest neighbour of machine learning for the classification of respiratory pathologies on the RALE lung sound dataset had done [6]. Various ML algorithms have been applied to do classification of dataset. The authors Tapak, Lily, et al. have done the comparison of statistical method and ML algorithms in terms of performance for diagnosis of the diabetes[7]. In the prediction of survival and diagnosis of patients having cancer diseases, the ML have applied.
The classification model have been adopted by the author Delen D to predict the survival of prostate cancer in patients [8] and to predict survival of lung cancer in patients [9]. Apart from this, the nature of cancer was also identified by applying Machine learning algorithms to the patient's data [10]. The author Chen and et al had adopted the ML tehcniques to identify the pattern of breast cancer [11].

Tools and Methodology

Here in this paper, from the UCI machine learning repository, I have used the data of thyroid. Classification is a supervised learning method to extract models describing important data classes or to predict future trends. We have adopted Naïve Bayes, Support Vector Machine, Multiclass Classifier, Logistic and K Nearest Neighbour. classification algorithms for the analysis of the thyroid dataset. The total number of attributes in the data set are 30 and number of records are 3622 after doing the pre-processing. In the pre-process phase, the missing values are eliminated and attributes are transferred from numeric to nominal. The attribute description is given in below fig. 2.

IV. EXPERIMENT, RESULT AND DISCUSSION

The experiment is performed using Weka with RConsole tools. In figure 1 the accuracy generated using the experiment is presented. It can be seen that, the different algorithms of machine learning are applied to the dataset. Among all the algorithms, 99.5% accuracy is achieved by Multiclass Classifier algorithm on the thyroid dataset. The Kappa Statistics In Fig. 3, the Kappa Statistic of various algorithm is noted. Again, Multiclass Classifier algorithm is

performing well as compared to the rest of the algorithms. The Table 1 shows the comparative analysis based on various parameters like TP Rate, FP Rate, Precision, Recall, F-Measure and ROC area along with its class labels.

```
@relation hypothyroid-weka.filters.AllFilter-weka.filters.MultiFilter-
Fweka.filters.AllFilter
@attribute age numeric
@attribute sex {F,M}
@attribute 'on thyroxine' {f,t}
@attribute 'query on thyroxine' {f,t}
@attribute 'on antithyroid medication' {f,t}
@attribute sick {f,t}
@attribute pregnant {f,t}
@attribute 'thyroid surgery' {f,t}
@attribute 'I131 treatment' {f,t}
@attribute 'query hypothyroid' {f,t}
@attribute 'query hyperthyroid' {f,t}
@attribute lithium {f,t}
@attribute goitre {f,t}
@attribute tumor {f,t}
@attribute hypopituitary {f,t}
@attribute psych {f,t}
@attribute 'TSH measured' {t,f}
@attribute TSH numeric
@attribute 'T3 measured' {t,f}
@attribute T3 numeric
@attribute 'TT4 measured' {t,f}
@attribute TT4 numeric
@attribute 'T4U measured' {t,f}
@attribute T4U numeric
@attribute 'FTI measured' {t,f}
@attribute FTI numeric
@attribute 'TBG measured' {f}
@attribute TBG numeric
@attribute 'referral source' {SVHC,other,SVI,STMW,SVHD}
@attribute Class
{negative,compensated_hypothyroid,primary_hypothyroid,secondary_hypothyroid}
```

Figure 1. Description of Attribute with values

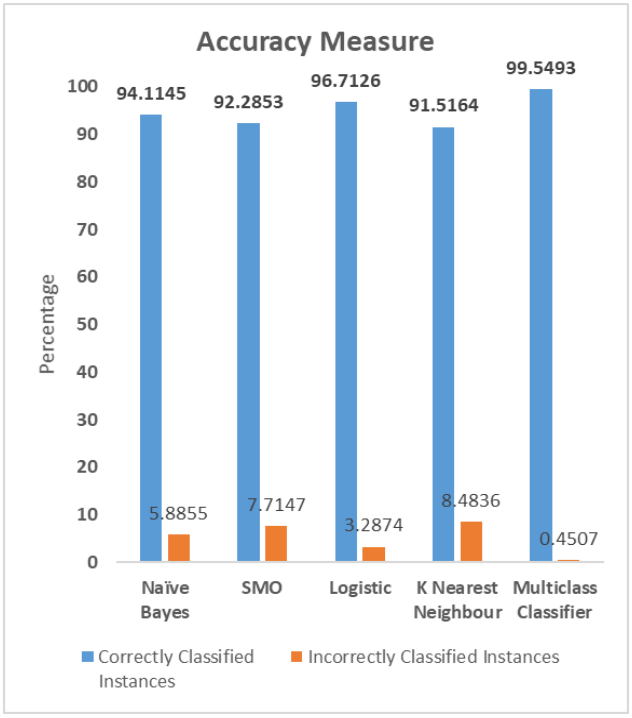

Figure 2. Accuracy Measures of different classifcation algorithms

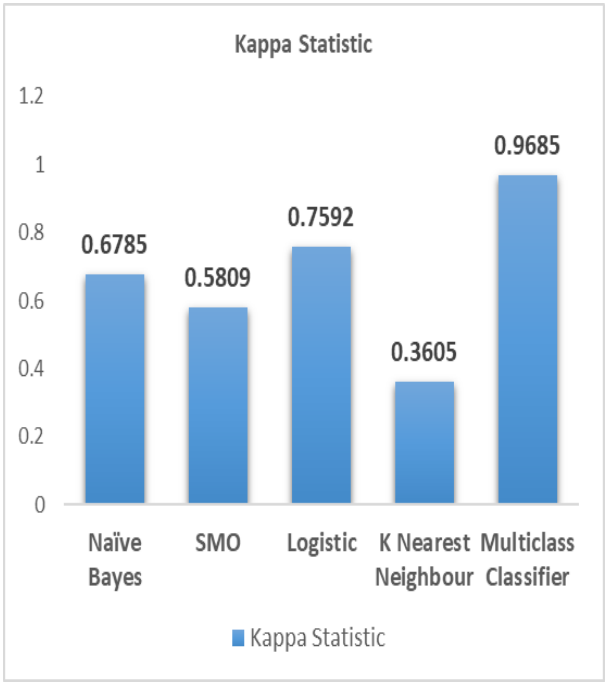

Figure 3. Kappa Statistics of different classifcation algorithms

Table 1. TP Rate, FP Rate, Precision, Recall, F-Measure and ROC Area of Algorithms

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Naïve Bayes	0.980	0.443	0.964	0.980	0.972	0.960	Negative
	0.356	0.015	0.566	0.356	0.437	0.953	compensated_hypothyroid
	0.716	0.010	0.642	0.716	0.677	0.985	primary_hypothyroid
	0.000	0.001	0.000	0.000	0.000	0.669	secondary_hypothyroid
	0.941	0.410	0.934	0.941	0.936	0.962	**Weighted Avg.**
SMO	1.000	1.000	0.923	1.000	0.960	0.500	Negative
	0.000	0.000	0.000	0.000	0.000	0.500	compensated_hypothyroid
	0.000	0.000	0.000	0.000	0.000	0.500	primary_hypothyroid
	0.000	0.000	0.000	0.000	0.000	0.500	secondary_hypothyroid
	0.923	0.923	0.923	0.923	0.960	0.500	**Weighted Avg.**
Multiclass Classifier	0.999	0.034	0.997	0.999	0.998	0.998	Negative
	0.938	0.001	0.989	0.938	0.963	1.000	compensated_hypothyroid
	0.989	0.001	0.949	0.989	0.969	1.000	primary_hypothyroid
	0.000	0.000	0.000	0.000	0.000	0.639	secondary_hypothyroid
	0.995	0.032	0.000	0.995	0.000	0.998	**Weighted Avg.**
Logistic	.99 1	0.237	0.980	0.991	0.985	0.939	Negative
	0.825	0.009	0.838	0.825	0.831	0.958	compensated_hypothyroid
	0.421	0.006	0.656	0.421	0.513	0.858	primary_hypothyroid
	0.000	0.001	0.000	0.000	0.000	0.371	secondary_hypothyroid
	0.967	0.219	0.964	0.967	0.965	0.938	**Weighted Avg.**
K Nearest Neighbour	0.964	0.619	0.949	0.964	0.956	0.682	Negative
	0.191	0.035	0.227	0.191	0.207	0.587	compensated_hypothyroid
	0.642	0.004	0.813	0.652	0.718	0.829	primary_hypothyroid
	0.000	0.000	0.000	0.000	0.000	0.902	secondary_hypothyroid
	0.915	0.573	0.715	0.915	0.843	0.681	**Weighted Avg.**

V. CONCLUSION

In this paper, the experiment has been carried out using the various ML classification algorithms like Naïve Bayes, Support Vector Machine, Multiclass Classifier, Logistic and K Nearest Neighbour. The experiment shows that each algorithm has its property and based on that the prediction of Thyroid disease is done. Based on this result, in future by using the historical data, it will be ease to take accurate decision regarding the Thyroid disease for the patient. As the conclusion, it is summarized that the Multilayer Classifier can be used to predict the hypothyroid disease, as its performance and accuracy are better than the rest of the algorithms.

REFERENCES

[1] I. Mandal and N. Sairam, "Accurate prediction of coronary artery disease using reliable diagnosis system," *J. Med. Syst.*, vol. 36, no. 5, pp. 3353–3373, 2012.

[2] J. Sun and C. K. Reddy, "Big data analytics for healthcare," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, 2013.

[3] C. L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny).*, vol. 275, pp. 314–347, 2014.

[4] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informatics*. 2006.

[5] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *Journal of Applied Biomedicine*. 2013.

[6] R. Palaniappan, K. Sundaraj, and S. Sundaraj, "A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals," *BMC Bioinformatics*, 2014.

[7] L. Tapak, H. Mahjub, O. Hamidi, and J. Poorolajal, "Real-data comparison of data mining methods in prediction of diabetes in Iran," *Healthc. Inform. Res.*, 2013.

[8] D. Delen, "Analysis of cancer data: A data mining approach," *Expert Syst.*, 2009.

[9] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "A lung cancer outcome calculator using ensemble data mining on SEER data," in *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics - BIOKDD '11*, 2011.

[10] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Classification of healthcare data using genetic fuzzy logic system and wavelets," *Expert Syst. Appl.*, 2015.

[11] T. C. Chen and T. C. Hsu, "A GAs based approach for mining breast cancer pattern," *Expert Syst. Appl.*, 2006.

Authors Profile

Dr. Hetal Patel pursed Bachelor of Computer Science and Master of Computer Application from Sardar Patel University, Gujarat in 2004 and 2006 respectively. She completed her Ph.D. in the field of Data Mining. She is currently working as Assistant Professor in Faculty of Computer Science and Applications, Charotar University of Science and Technology, Changa. She is having 11 years of experience. She has published 08 research papers in various national/international journals. Her area of interest is Data Mining and Analytics.

