A Hypothyroidism Prediction using Supervised Algorithm

Shalini L, Muhammad Rukunuddin Ghalib

Abstract: ***Thyroid issue are pervasive and their appearances are dictated by the dietary iodine accessibility. The most well-known reason for thyroid issue worldwide is iodine deficiency, inducing growth in goiter and hypothyroidism. In many regions, the people with thyroid issue have iodine deficiency which leads to poor have immune system. The greater parts of the populace are undiscovered or misdiagnosed. Ladies are multiple times bound to contract thyroid issues than men and almost 50% everything being equal and a fourth of all men will kick the bucket with proof of an induced thyroid. The side effects of this sickness regularly shift from individual to individual and are non-explicit, so a right finding can without much of a stretch be missed or misdiagnosed for immaterial issues. In light of the trial directed it demonstrates that Rand forest and Support Vector Machine gives result closest in anticipating the illnesses. This paper points in diagnosing the Hypothyroidism utilizing different classification. The precision of the every classifier helps in distinguishing the sicknesses. A modified Support Vector Machine (SVM) that uses Convex hull to compute the support vectors is proposed. The proposed SVM is evaluated on the UCI Thyroid dataset.***

Keywords **: Classification Algorithm, Hypothyroidism, Random forest, Support Vector Machine.**

I. INTRODUCTION

With the quick advancement of new Technologies, the issues of distinguishing the illness turned out to be increasingly more. Butterfly-shaped organ in the front of the neck influences the thyroid organ. Thyroid manages the various metabolic procedures of the body. There are unique thyroid issue that have impact on either its structure or capacity. The thyroid organ is situated beneath the Adam's apple folded over the trachea where a tissue known as the isthmus, connects the two thyroid projections on each side. The secretion of hormones uses iodine. Thyroxine, is an required hormone called as T4, which is dynamic hormone changed over to triiodothyronine (T3).

The point when thyroid hormone levels are low, the nerve center in the cerebrum delivers a hormone known as thyrotropin discharging hormone (TRH) that causes the pituitary organ to discharge thyroid invigorating hormone (TSH). TSH animates the thyroid organ to discharge more T4. Since the thyroid organ is constrained by the pituitary organ and nerve center, issue of these tissues can likewise influence thyroid level and cause thyroid problems. With the rapid development of new technologies, the problems of identifying the diseases related to thyroid have become more essential to be solved. Thyroid disorders affect the thyroid gland. It regulates the number of metabolic system of human body. Various thyroid disorders will disrupt the structure or function of thyroid gland. The vital hormone, Thyroxine (T4), produced and released from the gland is converted to triiodothyronine (T3), the most agile hormone.

The thyroid gland regulates brain. The effect of disorders of pituitary gland and hypothalamus can create problems to the functioning of thyroid. The proposed work demonstrates the prediction of Hypothyroidism. The new SVM proposed constructs a convex hull to identify the support vector and calculates the mean of the hull points. The distance method is used to find the distance from this mean.

II. RELATED WORKS

As of late, the thyroid is the most deadly disease among the people. Examination completed by contrasting four characterization models[1]: Naive Bayes, Decision Tree, Multilayer Perceptron and Radial Basis Function Network brings about noteworthy precision for all the grouping models and the best is the Decision Tree model. The data set taken from UCI machine learning repository used here to validate the classifier.

Subtype arrangement is important for better determination [2].An extra tests, for example, RT3 and basal metabolic temperature will help in effective manner to check the illness. Subsequently the paper proposed characterizes the sorts of Thyroid ailment and its subtypes in a productive manner utilizing different mining approach. In this paper EM bunching calculation and J48 order calculation were utilized to arrange the thyroid infections and its subtypes in an effective manner. The extra properties RT3 and Basel Metabolic temperature were utilized to analyze the subtypes of hypothyroidism.

This Research work centre's around the different classes of thyroid to identify the accurate diagnosis [5]. This work demonstrates prediction using the classes of thyroid [6] and Evolutionary Multivariate Bayisean classifier representation accomplishes amazing feature.

Thyroid ailment study utilizing Classification techniques [7], for example, Decision tree, Naïve Bayse which help to get rapid and precision rate when connected with huge information base. The overview recognizes how the information mining methods are useful in foreseeing the thyroid issue at prior stage.

Revised Manuscript Received on October 05, 2019

Shalini L School of Computer Science and Engineering, Vellore Institute of Technology(VIT) Deemed to be University, Vellore, Tamil Nadu, India,lshalinirao@gmail.com,

Dr. Muhammad Rukunuddin Ghalib ,School of Computer Science and Engineering, Vellore Institute of Technology(VIT) Deemed to be University, Vellore, Tamil Nadu, India,

Information mining procedure is utilized in medicinal service industry to recognize and analyse the disease to provide a better treatment in the beginning period of the illness. The algorithms utilized in this paper were CART, LDA, classification [8], clustering, decision tree and k-fold cross validation . The model given was with classification and clustering precision with less number of features.

In this paper [9] the investigations plainly demonstrate that variables, for example, various sorts of antibodies, heredity and so forth can be considered for precise finding of thyroid illness. Counterfeit Neural Networks considered for the diagonisis of the thyroid infection. In this study [10] a comparison of various decision tree algorithms carried out and analysed for their performances according to six evaluation metrics. In this paper [11] various data mining techniques were used with k-fold cross validation to diagnosis hypothyroid disease.

This study demonstrated [12], the comparative study of various algorithms where with Quick reduct and Johnson's reduct computation algorithms of Rhetorical Structure Theory (RST) produced good result. The Evolutionary Computation (EC) approach based Particle Swarm Optimization (PSO) algorithm outperformed than the Genetic Algorithm and RST based algorithms, by using minimal subset generation. It increased accuracy of the classification. The paper[13] discuss about different neural network modeling in identifying the thyroid dysfunctionality. This paper [14] provides machine learning algorithms for effective prediction of thyroid disease.

In the proposed work [15], the random forest approach predicts the hypothyroidism and the experimental result provided shows improved accuracy, precision, recall and F-measure by comparing the random forest with LDA algorithm.

III. TECHNIQUES

The dataset is classified into 7200 instances selected from UCI repository for Verification and Validation. The Analysis parameter used is Accuracy, F-measure, Kappa statistic to compare how closely the instances classified by the machine learning classifier.

In Fig 1, a comparative study on using the above said dataset shows that Kappa being the highest priority parameter while judging the classifiers, Random Forest tops the table with accuracy and F-measure. The next in the row with accuracy of 95.8% by SVM shows that these classifiers predict thyroid illness. Therefore we consider the Naive Bayes, K-Nearest Neighbour, and Neural Network as an apparent failure for the Thyroid disease prediction. But this table gives an comparative study of the algorithms with various elements. A new SVM method is tried to check for the accuracy.

Table -I Comparison of algorithms

	Accuracy	*Precision*	*Recall*	*Kappa*	*F-Measure*
KNN	90.2	52.31	50.2	32.44	70.75
NN	89.3	48.65	51.1	30.03	73.34
Naive Bayes	93.3	58.82	60.14	24.38	59.38
Decision Tree	94.1	65.24	52.21	35.6	70.81
Random Forest	96.5	64.23	50.25	42.5	86.5
SVM	95.8	57.1	53.32	40.25	86.2
Linear Regression	92.2	53.2	48.23	31.04	72.18

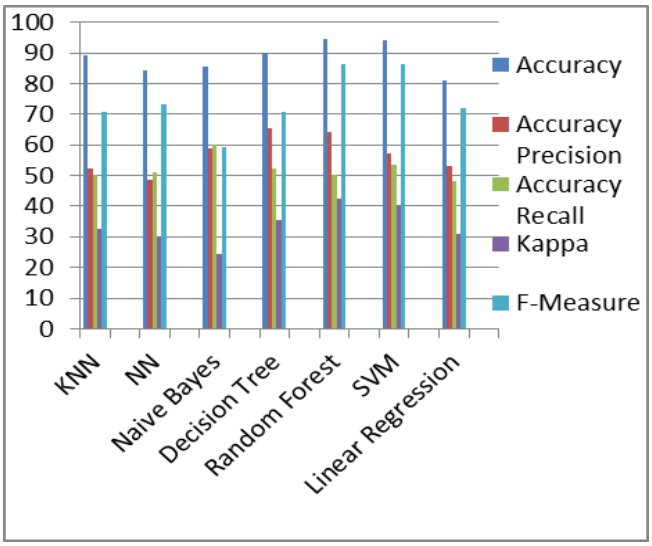


Fig 1 Comparison of Performance Measures

IV. PROPOSED WORK

Machine learning includes foreseeing and grouping information. To carry out the process of predicting utilization of different Machine learning Algorithm namely SVM or Support Vector Machine deployed. It is one of linear representation for classification and regression problems. The SVM can be used for any linear or non-linear problem to check for the evaluation of the dataset considered. In the algorithm the hyperplane isolates the data into separate classes. SVM creates a space, which is finite. The feature of each dimension represents a particular object.

SVM train the model to categorize the featured space into two categories by partitioning linearly. Categorization places the items "above" and "below" the separation plane. The location identified places the new items in respective space in non-probabilistic way.

However, the benefit of SVM not restricted to being linear classifiers but introduction of kernel will be more flexible for non- linear decision boundaries.

The support vectors are the data points that lie closer to the hyperplane. These points are difficult to locate and can change the optimum decision surface. The margin is maximized around the hyperplane of SVM. Support vectors are the decision function which is specified by the samples. Since this is a Quadratic programming problem, the normal method can be utilised to get a solution.

Input: Consider set of (input, output) training pair samples $u_1, u_2 \ldots u_n$, and the output result z. The input features taken to be u_i.

Output: The weights for each items is set as q_i, where these linear combinations predicts the value of z. To maximize the margin reduce the number of weights such that the values are nonzero. This helps in selecting the important features that decides the hyperplane. The nonzero weights correspond to the support vectors helps in separating the hyperplane.

In this method identifying which points influence optimality to be checked. The positional change of the hyperplane is based on the support vector of the dataset which divides it. Here the optimization problem is to find optimal hyperplane. In this paper Graham Scan Algorithm [28] considered to identify the support vectors by constructing the convex hull. The convex hull thus constructed will have all the boundary points in order. This algorithm is applied for both the points which are above the feature space and below the feature space to identify the boundary points of the hull . The following figure illustrates the algorithm implementation and the outcome. The algorithm of Graham Scan is given below

The Graham Scan Algorithm to identify Support Vector

The point on the feature space above and below are considered as points of array A[0..n-1].The algorithm is applied separately to identify the hull above and below so that the support vectors can be defined.

1. Select a lowest point by comparing the y coordinate of all points. The tie is resolved between two y values by selecting the smaller x coordinate value. Let the lowest point be Q_0.The first position in output hull is Q_0.
2. Sort the $n-1$ points by considering their polar angle in counter clockwise direction around points [0]. Nearest point is taken as first point, if there is same polar angle for two points.
3. Angle point further away from Q_0 is taken, if two or more points have same angle even after sorting. Let the size of the new array be k.
4. Check for k less than 3, then print "Convex Hull not possible"
5. The points[0],points[1],points[2] are pushed into the empty stack S which is defined.
6. The remaining $k-3$ points are processed one after another by following below steps for each points[i].
   4.1 The points are removed from stack while checking the orientation of points that makes a left turn or not.
   a) The next, points top in the stack
   b) The point at the top of the stack.
   c) Points[i]
   4.2 Push points[i] to stack S
7. Finally the content of the stack S gives a hull.

Below Fig 2 illustrate the definition of support vector using Graham scan algorithm. The construction of the hull provides set of boundary points.

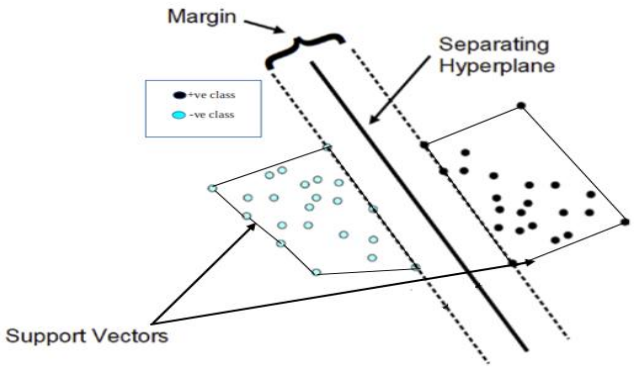


Fig 2. Support Vector Construction

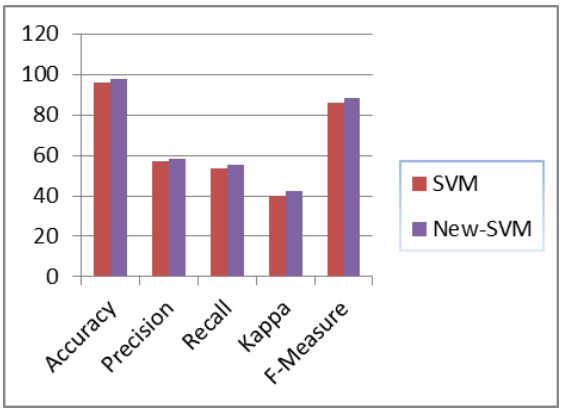
The above algorithm finds a convex hull which consists of boundary point stored in a stack. The points are then removed from the stack in counter clockwise direction. The points are then added to find the mean and which is used as support vector for new-SVM. By applying the above algorithm the new -SVM performance improved by 2%.

V. RESULT AND DISCUSSION

The Table-I given above expresses the comparison perspective of all the classifiers mentioned in the paper. Seven classifiers are analysed on four parameters namely accuracy, Kappa, F-measure and Recall.

Table- II Comparisons of SVM

	Accuracy	*Precision*	*Recall*	*Kappa*	*F-Measure*
SVM	95.8	57.1	53.32	40.25	86.2
New-SVM	97.716	58.242	55.432	42.21	88.34



From Table- II we can conclude that even though there is not much improvement in the new SVM algorithm. The novelty of this approach proved that there is chance of improvement in the new SVM by introducing various distance method. Applying various distance algorithms the evaluation may improve the accuracy of the new-SVM.

VI. CONCLUSION

Kappa being the highest priority parameter while judging the classifiers, Random Forest tops the table with accuracy and F-measure. The next in the row with accuracy of 95.8% by SVM and 97.716% with the modified SVM shows that this classifier provides prediction on thyroid disease. Therefore we consider the Naive Bayes, KNN, NN as an apparent failure for the Thyroid disease prediction. Improvement in the algorithm can be made by considering the various distance algorithms. Time can be considered to check for the time elapsed to classify the classifier for a faster prediction. In future the reduction in feature can be applied to test the data and predict the disease. Moreover the reduction in the dimension would decrease the number of test to be carried out and the time to diagnose the disease will be minimal.

REFERENCES

1. Irina IoniŃă Informatics, Liviu IoniŃă, "Prediction of Thyroid Disease Using Data Mining Techniques" BRAIN. Broad Research in Artificial Intelligence and Neuroscience Volume 7, Issue 3, August 2016, pp 119-124.
2. Sumathi A, Nithya G and Meganathan S "Classification of Thyroid disease Using Data Mining Techniques ", International Journal of Pure and Applied Mathematics Volume 119 No. 12 2018, pp.13881-13890.
3. Amit Kumar Dewangan, Akhilesh Kumar shrivas, Prem Kumar, "Classification of Thyroid Disease with Feature Selection Technique", International Journal of Engineering and Techniques - Volume 2 Issue May – June 2016 ,pp.128-131
4. **N**. Kumar and S. Khatri, "Implementing WEKA for medical data classification and early disease prediction," 2017 3rd Int. Conf. Comput. Intell. Commun. Technol., pp. 1–6..
5. Prem Kumar, Amit Kumar Dewangan, " Classification of Thyroid Disease: A Survey ", International Journal of Computer Science and Information Technologies, Vol. 7 (3), 2016, pp. 1102-1104
6. K.Geetha and Capt. S. Santhosh Baboo "An Empirical Model for Thyroid Disease Classification using Evolutionary Multivariate Bayseian Prediction Method ", Global Journal of Computer Science and Technology: E Network, Web & Security Volume 16 Issue 1 Version 1.0 Year 2016.pp
7. Dr.B.Srinivasan, K.Pavya," Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study", International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 11 Nov-2016,pp.,1191-1194.
8. Roshan Banu D, K.C.Sharmili ," A Study of Data Mining Techniques to Detect Thyroid Disease", International Journal of Innovative Research in Science, ,Engineering and Technology, Vol. 6, Special Issue 11, September 2017.pp., 549-553.
9. Ifrah Raoof, Arvind Selwal," A study on Neural Network based Thyroid disease Prediction System", International Journal of Latest Trends in Engineering and Technology Vol.(10)Issue(1), pp.337-343.
10. EBRU TURANOGLU-BEKAR1, GOZDE ULUTAGAY, SUZAN KANTARCI-SAVAS," Classification of Thyroid Disease by Using Data Mining Models: A Comparison of Decision Tree Algorithms", Oxford Journal of Intelligent Decision and Data Science 2016 No. 2 (2016) pp. 13-28.
11. Shivanee Pandey, Rohit Miri , S. R. Tandan "Diagnosis And Classification Of Hypothyroid Disease Using Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT, Vol. 2 Issue 6, June – 2013,pp. 3188-3193.
12. Surekha S, JayaSuma G," Comparison of Feature Selection Techniques for Thyroid Disease" International Conference on Intelligent Systems, Control & Manufacturing Technology (ICICMT'2015) March 16-17, 2015 Abu Dhabi (UAE),pp.20-26.
13. Shaik Razia and M. R. Narasinga Rao," Machine Learning Techniques for Thyroid Disease Diagnosis - A Review", Indian Journal of Science and Technology, Vol9(28), July 2016,pp.1-9
14. M.Shyamala, P.S.S. Akilashri," Thyroid Disease Prediction by Machine Learning Technique From Healthcare Communities" International Journal of Computer Sciences and Engineering, Vol.6, Special Issue.11, Dec 2018,pp237-242.
15. Ammulu K,Venugopal T, "Thyroid Data Prediction using Data Classification Algorithm",International Journal for Innovative Research in Science & Technology,Volume 4, Issue 2, July 2017,pp.208-212.
16. Vinod Kumar Pal, V.P.Sriram,Rashmi Mahajan, Suresh Chandara Padhy," A Literature Review on Diagnosing Thyroid disease Through Artificial Neural Network Techniques",Volume 14 Issue 5,2019,pp.1510-1517.
17. Komal Agrawal, Mradul Dhakar," Thyroid Prediction System using Auto Associative Neural Network", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 6, Issue 4, April 2017,pp . 2239-2247.
18. S. Sathya Priya, Dr. D. Anitha," Survey on Thyroid Diagnosis using Data Mining Techniques ",International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Special Issue 1, January 2017,pp. 161-164.
19. I. Kalaimani," Anaysis For the Prediction of thyroid Disease by using ICA and Optimal Kernel SVM Approach", International Journal of Emerging Technology and Innovative Engineering, Volume 5, Issue 3, March 2019,pp39-55.
20. https://en.wikipedia.org/wiki/Precision_and_recall.
21. https://en.wikipedia.org/wiki/Precision_and_recall#F-measure
22. Madhuri V. Joseph, LipsaSadath and VanajaRajan, 2013. Data Mining: A Comparative "Study on Various Techniques and Methods, International Journal of Advanced Research in Computer Science and Software Engineering, Research Paper Available online at: www.ijarcsse.com ISSN: 2277 128X, 3(2).
23. Mehmed Kantardzic, DATA MINING Concepts,Models, Methods and Algorithms, IEEE Press 445 Hoes Lane Piscataway, NJ 08854 IEEE Press Editorial Board.
24. Xindong Wu,Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, Top 10 algorithms in data mining , Knowl Inf Syst (2008).
25. **J**oão MarocoEmail author, Dina Silva, Ana Rodrigues, Manuela Guerreiro, Isabe Santana and Alexandre de Mendonça, Data mining methods in the prediction of DementiaBMC Research Notes, 2011
26. Anantha M. Prasad, Louis R. Iverson, and Andy Liaw, Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction, Techniques for Ecological Prediction, 2006.
27. Wenliang Du, Yunghsiang S. Han, Shigang Chen, Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classification, Syracuse University SURFACE, 2004.
28. www. geeksforgeeks.org

AUTHORS PROFILE

Shalini L have completed MSc(Maths) ,MCA,, MPhil(ComputerScience) and is working as Assistant Professor (Senior) in School of Computer Science and Engineering,, Vellore Institute of Technology, Vellore for a period of 12 years.

Dr. Muhammad Rukunuddin Ghalib have completed ME(ComputerScience), Phd and is working as Associate Professor in School of Computer Science and Engineering, Vellore Institute of Technology, Vellore for a period of 11 years.