International Journal of Engineering & Technology, 7 (1.1) (2018) 326-328

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A comparative study of support vector machine and logistic regression for the diagnosis of thyroid dysfunction

Deepthi Gurram [1] *, M. R. Narasinga Rao [2]

[1] *Research Scholar, Department of Computer Science, K L University, Vaddeswaram, Andhra Pradesh, India*
[2] *Department of Computer Science & Engineering , K L University, Vaddeswaram, Andhra Pradesh, India*
**Corresponding author E-mail: deepug2110@gmail.com*

Abstract

Thyroid is one of the vital diseases that influence individuals of any age group now a day. Infections of the thyroid, incorporate conditions related with extreme release of thyroid hormones (Hyper thyroidism) which is likewise called thyrotoxicosis and those related with thyroid hormone insufficiency (Hypothyroidism). Expectation of these two sorts of thyroid disease is critical for thyroid analysis. In this paper, support vector machines and logistic regression are proposed for predicting patients with thyrotoxicosis and without thyrotoxicosis. The outcomes demonstrate that, logistic regression perform well over support vector machine with 98.92% exactness.

Keywords: *Logistic Regression; Precision; Recall; Support Vector Machine; Thyrotoxicosis*.

1. Introduction

These days, by the rapid development of innovation and information in medical sciences, the software engineering experts are fit for providing expert frameworks to determine various types of diseases with high exactness. The therapeutic experts are made to utilize these frameworks because of the existence of some problems at the time of prediction process [1]. Disease diagnosis operation utilizing expert frameworks are performed in light of a set of disease manifestations. These frameworks depend on machine learning procedures which help the doctor and a physician to limit the expenses and time and can act as an expert advisor in making successful conclusions.

In the human body, the thyroid organ is an essential organ. It produces thyroid hormones to keep up our body metabolism [2]. The thyroid organ is situated in the front of the neck and underneath the Adam's apple. The thyroid produces two noteworthy hormones called T3 (triiodothyronine) and T4 (thyroxine). These T3 and T4 hormones make a trip in our blood to all parts of our body and influence practically every cell in the body, and controls our body's function. On the off chance that, the measure of thyroid hormone diminishes in our blood and our body work gets back off, this condition is called hypothyroidism [3]. The signs of hypothyroid are depression, exhaustion of body strength, tiredness, constipation, excess weight, cramps, dry skin, sexual disorders and infertility. On the off chance that, the increased measure of thyroid hormones found in our blood, our body functions will increase. This condition is called hyperthyroidism. The indications of hyperthyroid are anxiety, palpitation, exhausted body quality, tremors, loss of weight, diarrhoea, menstrual disorder and exophthalmia. Specialists can fuse various factors, including clinical assessment, blood tests, imaging tests, biopsies, and different tests to analyze thyroid illness. A typical utilized strategy is a test, called the thyroid- stimulating hormone (TSH) test, which can recognize thyroid issue even before the onset of indications.

These days, CAD frameworks are getting increasingly prominent. With the assistance of the CAD frameworks, the identified errors a doctor can make, over the span of analysis can be kept away from, and the medical information can be inspected in shorter time and more definite as well [4]. Machine learning procedures are progressively acquainted with to build the CAD frameworks inferable from its firm capacity of draw out complex relationships in the biomedical information. As of late, different strategies have been proposed to take care of this issue.

2. Related work

In 2014, Baydaa S. B. Alyas [5] planned an inventive framework which would diagnose the thyroid patients with least execution time and superior. This system should enable the healthcare experts to answer inquiries that describe the endocrine organ dysfunction and that allow them to take any clinical conclusions.

In 2013, Ahmad Taher Azar et al. [6] proposed a correlation amongst hard and fuzzy clustering algorithms for thyroid maladies informational index in order to locate the optimal number of clusters. Distinctive scalar validity measures are utilized in looking at the overall execution of the proposed clustering frameworks. To locate the optimal number of clusters, elbow paradigm is enforced. The clustering outcomes for all algorithms are then imagined by the Sammon mapping strategy to locate a low-dimensional (ordinarily 2D or 3D) portrayal of a collection of points scattered in a high dimensional pattern space.

In 2013, Ms. Wrushali Mendre Dr. Ranjana D. Raut [7] explored the potentiality of neural network to segregate the two subtypes, hypothyroid and negative form of thyroid issue based on the argument of laboratory medical information base. The best parameters are distinguished for the neural networks like Multilayer Perceptron (MLP), Radial Basis Function (RBF) and Principal Component Analysis (PCA).

In 2012, Hui-Ling Chen, et al. [8] represented a three-phase expert framework based on a hybrid support vector machines (SVM)

technique to deal with analysis of thyroid disease. The initial phase is for building varied feature subsets with various discriminative capabilities. In the second phase, the feature subsets are sustained into the designed SVM classifier for preparing an optimal predictor framework whose parameters are enhanced by particle swarm optimization (PSO). At last, the optimal SVM framework continues to carry out the thyroid disease analysis tasks utilizing the most discriminative feature subset and the optimal parameters. The proposed framework has accomplished the most astounding classification accuracy revealed so far by 10-fold cross-validation (CV) technique, with the mean accuracy of 97.49% and with the greatest precision of 98.59%.

In 2005, Kenji Hoshi et al. [9] figured out the thyroid information by statistical technique, multivariate analysis and by two sorts of neural networks. One is the self-organizing map approach, that clusters the patients and shows outwardly a characteristic of the distribution according to laboratory tests. SOM isolated the data into three clusters relating to hyperthyroid, hypothyroid and ordinary. To figure out the QSAR issue inside the thyroid information, a classification technique Bayesian Regularized Neural Network (BRNN) is enforced and discovered that its forecast precision is superior to statistical approach.

3. Materials and methods

3.1. Dataset

In this work, thyroid database [11] comprises of 2 classes and 185 samples. These classes are allocated to the values that relate to the hyper-, hypo-and original function of the thyroid organ. All samples have five features. These are: 1- T3-resin uptake test. (A percentage) 2- Total Serum thyroxin as measured by the isotopic displacement method. 3- Total serum triiodothyronine as measured by radioimmuno assay. 4- Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay[10]. 5- Maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value. Initial class has 150 samples and other class has 35 samples. A training set and a test set are prepared utilizing 185 examples.

3.2. Logistic regression

Logistic regression is one of the statistical procedures used in biomedical informatics. It is a one of the most widely recognized models for prediction. From past examinations, logistic regression is generally used as a part of therapeutic literature particularly to correlate the dichotomous results with the predictor factors, that incorporate diverse physiological information. In logistic regression, the predicted odd proportion of positive result is calculated as a sum of product. Product is calculated, by multiplying the estimations of independent factor and its coefficients [12]. The likelihood of positive result is acquired from the odd proportion through a simple transformation (Samatha, 2009). At that point, the coefficient acquired from the logistic regression is used to ascertain the predictor factors (Zhou, 2004). Logistic regression is used to forecast by fitting information to the logistic curve. It requires the model to be good enough with the output generated by it which is in accordance with the expected output. A logistic regression framework is the non-linear transformation of the linear regression framework.

Logistic regression deals with the issues by implementing the logit transformation to the dependent variable. Basically, the logistic system predicts the value of Y from X [13]. As stated before, the logit is the natural logarithm (ln) of odds of Y, and odds are proportions of probabilities (π) of Y happening to probabilities ($1 - \pi$) of Y not happening. Even though logistic regression can contain categorical results that are polytomous, in this article we focus on dichotomous outcomes only. The example shown in this article can be drawn out easily to polytomous factors with ordered (i.e., ordinal-scaled) or unordered (i.e., nominal-scaled) outcomes.

The simple logistic model is represented in the form of

$$\text{Logit}(Y) = \text{naturallog (odds)} = \ln[\pi/1-\pi] = \alpha+\beta X. \quad (1)$$

3.3. Support vector machine

Here, we quickly depict the essential thoughts behind SVM for pattern recognition, particularly for the two-class classification issue, and was as of late proposed by Vapnik and associates (Cortes and Vapnik, 1995; Vapnik, 1995, 1998) as an exceptionally viable technique for broadly useful supervised pattern recognition. The SVM is not only established model but it is based on extremely well developed machine learning theory, Statistical Learning Theory (Vapnik, 1995, 1998), and it has very wide practical applications in many domains.

At the point when utilized for classification, they isolate a given set of binary labelled training information with a hyper-plane that is maximally far off from them(known as 'the maximal margin hyper-plane')[14]. For cases in which no linear separation is conceivable, they can work in combination with the method of "kernels" that automatically understands a non-linear mapping to a feature space. The hyper-plane found by the SVM in feature space relates to a non-linear decision boundary in the input space. The documentation used to characterize formally a hyperplane:

$$f(x) = \beta_0 + \beta^T x \quad (2)$$

Where β is known as the weight vector and β_0 as the bias.

4. Results

In this section, we present the performance evaluation of the models to evaluate the prediction accuracies. Rapid Miner tool is used to evaluate these methods. SVM classification accuracy for thyroid disease dataset is compared with Logistic regression. According to Fig.2 and Fig.4, it is obvious that the result of Logistic Regression is better than SVM.

4.1. Support vector machine

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (185 examples, 4 special attributes, 5 regular attributes)

Row No.	class	confidence(...	confidence(...	prediction(c...	T3 resin	T4	T3	TSH	basal value
146	normal	0.942	0.058	normal	126	10.400	1.700	1.200	3.500
147	normal	0.982	0.018	normal	114	7.500	1.100	1.600	4.400
148	normal	0.814	0.186	normal	111	11.900	2.300	0.900	3.800
149	normal	0.898	0.102	normal	104	6.100	1.800	0.500	0.800
150	normal	0.948	0.052	normal	102	6.600	1.200	1.400	1.300
151	hyper	0.332	0.668	hyper	139	16.400	3.800	1.100	-0.200
152	hyper	0.341	0.659	hyper	111	16	2.100	0.900	-0.100
153	hyper	0.331	0.669	hyper	113	17.200	1.800	1	0
154	hyper	0.002	0.998	hyper	65	25.300	5.800	1.300	0.200
155	hyper	0.004	0.996	hyper	88	24.100	5.500	0.800	0.100
156	hyper	0.001	0.999	hyper	65	18.200	10	1.300	0.100
157	hyper	0.172	0.828	hyper	134	16.400	4.800	0.600	0.100
158	hyper	0.054	0.946	hyper	110	20.300	3.700	0.600	0.200
159	hyper	0.001	0.999	hyper	67	23.300	7.400	1.800	-0.600
160	hyper	0.563	0.437	normal	95	11.100	2.700	1.600	-0.300
161	hyper	0.113	0.887	hyper	89	14.300	4.100	0.500	0.200
162	hyper	0.004	0.996	hyper	89	23.800	5.400	0.500	0.100
163	hyper	0.239	0.761	hyper	88	12.900	2.700	0.100	0.200
164	hyper	0.235	0.765	hyper	105	17.400	1.600	0.300	0.400

Fig. 1: Prediction of Thyroid Disfunctioning Using Support Vector Machine.

4.2. Performance evaluation

Table / Plot View | Text View | Annotations

Criterion Selector: accuracy

Multiclass Classification Performance | Annotations

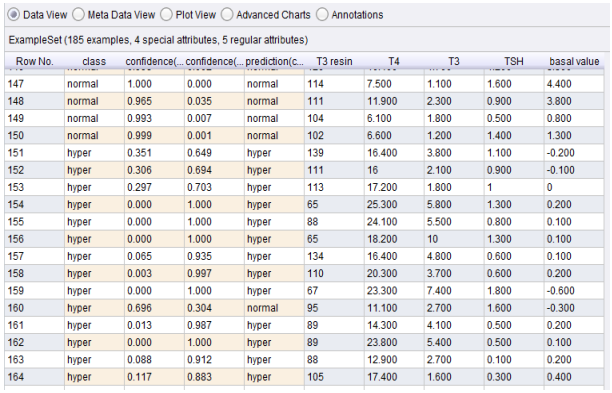
Table View | Plot View

accuracy: 98.38%

	true normal	true hyper	class precision
pred. normal	149	2	98.68%
pred. hyper	1	33	97.06%
class recall	99.33%	94.29%	

Fig. 2: Accuracy of Support Vector Machine in Classifying Thyroid Dataset.

4.3. Logistic regression



Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (185 examples, 4 special attributes, 5 regular attributes)

Row No.	class	confidence(...	confidence(...	prediction(c...	T3 resin	T4	T3	TSH	basal value
147	normal	1.000	0.000	normal	114	7.500	1.100	1.600	4.400
148	normal	0.965	0.035	normal	111	11.900	2.300	0.900	3.800
149	normal	0.993	0.007	normal	104	6.100	1.800	0.500	0.800
150	normal	0.999	0.001	normal	102	6.600	1.200	1.400	1.300
151	hyper	0.351	0.649	hyper	139	16.400	3.800	1.100	-0.200
152	hyper	0.306	0.694	hyper	111	16	2.100	0.900	-0.100
153	hyper	0.297	0.703	hyper	113	17.200	1.800	1	0
154	hyper	0.000	1.000	hyper	65	25.300	5.800	1.300	0.200
155	hyper	0.000	1.000	hyper	88	24.100	5.500	0.800	0.100
156	hyper	0.000	1.000	hyper	65	18.200	10	1.300	0.100
157	hyper	0.065	0.935	hyper	134	16.400	4.800	0.600	0.100
158	hyper	0.003	0.997	hyper	110	20.300	3.700	0.600	0.200
159	hyper	0.000	1.000	hyper	67	23.300	7.400	1.800	-0.600
160	hyper	0.696	0.304	normal	95	11.100	2.700	1.600	-0.300
161	hyper	0.013	0.987	hyper	89	14.300	4.100	0.500	0.200
162	hyper	0.000	1.000	hyper	89	23.800	5.400	0.500	0.100
163	hyper	0.088	0.912	hyper	88	12.900	2.700	0.100	0.200
164	hyper	0.117	0.883	hyper	105	17.400	1.600	0.300	0.400

Fig. 3: Prediction of Thyroid Disfunctioning Using Logistic Regression.

4.4. Performance Evaluation

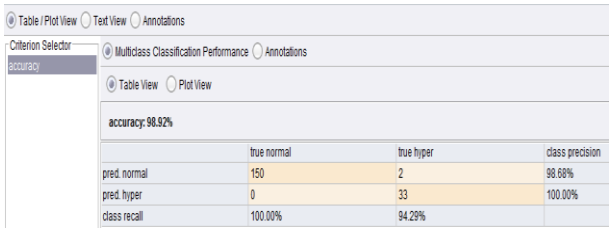


Table / Plot View | Text View | Annotations

Criterion Selector: accuracy

Multiclass Classification Performance | Annotations

Table View | Plot View

accuracy: 98.92%

	true normal	true hyper	class precision
pred. normal	150	2	98.68%
pred. hyper	0	33	100.00%
class recall	100.00%	94.29%	

Fig. 4: Accuracy of Logistic Regression in Classifying the Data.

Given the results, shows the accuracy of the model: precision, recall. According to the Fig.2 & Fig.4, the highest precision, recall accuracy belongs to Logistic Regression with 98.68 and 100 respectively.

According to Table 1, the comparison of data classification accuracy for diagnosis of thyrotoxicosis is shown in Fig.5.

Table 1: Data Classification Accuracy of Thyrotoxicosis by Evaluating Experimental Data

Model	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)
Support Vector Machine	98.3	1.7
Logistic Regression	98.9	1.1

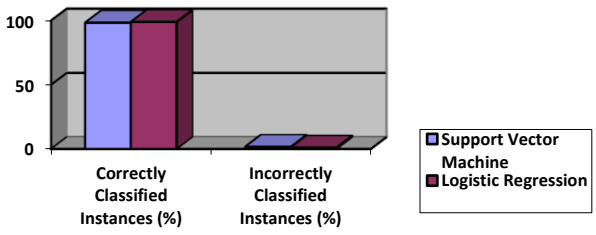


Fig. 5: The Comparison of Data Classification Accuracy.

5. Conclusion and discussion

In this study, Logistic Regression method and support vector machines have been employed for thyroid diseases diagnosis is proposed. The performance evaluation of these models had been proposed with respect to the precision and recall in Fig 2 & Fig 4. As shown from these results, the Logistic regression obtains promising results in classifying the possible thyrotoxicosis patients. The efficiency of logistic regression on thyroid disease diagnosis is 98.92% and the efficiency of the Support Vector machine is found to be 98.3%. This high rate of accuracy can be utilized to support the Doctor's decision to avoid Biopsy. The results show that Logistic regression technique can assist in the diagnosis of thyroid diseases than support vector machines.

References

[1] Farhad Soleimanian Gharehchopogh, et al. "Using Artificial Neural Network in Diagnosis of Thyroid Disease: A Case Study" .International Journal on Computational Sciences & Applications 2013; 3 (4).

[2] K.Saravana Kumar, Dr. R. Manicka Chezian, "Support Vector Machine and K- Nearest Neighbor Based Analysis for the Prediction of Hypothyroid", International Journal of Pharma and Bio Sciences 2014; 5(4), pp 447 – 453.

[3] M. R. NazariKousarrizi, F.Seiti, and M. Teshnehlab, "An Experimental Comparative Study on Thyroid Disease Diagnosis Based on Feature Subset Selection and classification", International Journal of Electrical & Computer Sciences IJECS, Vol: 12 No. 01, February (2012), pp 13-20.

[4] Li-Na Li, et al. "A Computer Aided Diagnosis System for Thyroid Disease Using Extreme Learning Machine", Journal of Medical Systems 2012; 36(5), pp 3327-3337. https://doi.org/10.1007/s10916-012-9825-3.

[5] Baydaa S. B. Alyas, "Design an Intelligent System for Thyroid Diseases Diagnosis", International Journal of Enhanced Research in Science Technology & Engineering 2014; 3 (4),pp 217-229.

[6] Ahmad Taher Azar, et al. "Fuzzy and hard clustering analysis for thyroid disease", Computer Methods and Programs in Biomedicine 2013; 111(1), pp 1-16. https://doi.org/10.1016/j.cmpb.2013.01.002.

[7] Ms.Wrushali Mendre, Dr.Ranjana D.Raut, "Neural Network based Decision Support System for the Diagnosis of Thyroid Diseases", International Journal of Computer Science and Applications 2013; 6(2), pp 102-106.

[8] Hui-Ling Chen, et al. "A Three-Stage Expert System Based on Support Vector Machines for Thyroid Disease Diagnosis" Journal of Medical Systems 2012; 36, pp 1953-1963. https://doi.org/10.1007/s10916-011-9655-8.

[9] Kenji Hoshi, et al. "An Analysis of Thyroid Function Diagnosis Using Bayesian-Type and SOM-Type Neural Networks", Chemical & Pharmaceutical Bulletin; 200653(12), pp 1570-1574. https://doi.org/10.1248/cpb.53.1570.

[10] L. Ozyılmaz, T. Yıldırım, "Diagnosis of thyroid disease using artificial neural network methods", Proceedings of ICONIP'02 9th International Conference on Neural Information Processing. Orchid Country Club. Singapore 2002, pp 2033– 2036. https://doi.org/10.1109/ICONIP.2002.1199031.

[11] www.ics.uci.edu/pub/ml-repos/machine-learningdatabases/, 2001.

[12] H. Yusuff, et al. "Breast Cancer Analysis Using Logistic Regression", IJRRAS 10 (1) January 2012,pp 14-22.

[13] CHAO-YING JOANNE PENG,et al. "An Introduction to Logistic Regression Analysis and Reporting", The Journal of Educational Research, September/October 2002 [Vol. 96(No. 1)], pp 3-14.

[14] Terrence S. Furey, et al. "Support vector machine classification and validation of cancer tissue samples using microarray expression data", Bioinformatics, 2000, Vol.16, no.10, pp 906-914. https://doi.org/10.1093/bioinformatics/16.10.906.