Original Article

Hybrid Model of KNN and PCA to Pre-processor of Thyroid Dataset using Machine Learning

R.Vanitha[1]* , Dr. K. Perumal[2]

[1]Research Scholar, Department of Computer Applications, Madurai Kamaraj University, Madurai, Tamilnadu, India.

[2]Professor, Department of Computer Applications, Madurai Kamaraj University, Madurai, Tamilnadu, India.

Email: vanithachezian2004@gmail.com*

Abstract

Nowadays, Thyroid is common disease that affects most of the people due to their modern lifestyle. A Thyroid ailments are the disorders that disturbs the thyroid gland which has a butterfly shape positioned at the front of the neck. The general endocrine carcinoma that may occur in the thyroid gland is Thyroid cancer. This type of cancer might not cause any indications at chief. But as it develops, it can show major signs and symptoms, such as swelling in the neck, voice changes and difficulty in swallowing. An ample effort has been given by many researchers in improving its diagnosis and prognosis. Thyroidectomy is considered as the main treatment method for thyroid problems. This research work mainly focuses on cleansing of thyroid cancer data from the UCI machine learning repository. This cleansing process involves removing the redundant values, impute the missing values, and selecting the best features from the existing fresh medical dataset by applying and comparing the various algorithms. It helps the medical practitioner to make an accurate diagnosis of the thyroid among various people.

Keywords: Thyroid, Pre-processing, Hyperthyroidism, SVM, Random Forest.

DOI: 10.47750/pnr.2022.13.S03.170

Introduction

The most recurrent endocrine malignancy is Thyroid cancer and it shows that approximately 2.5% of all newly cancer cases are found in the United States [1]. Thyroid cancer is the most misunderstood and undiagnosed diseases [2],[3] which generally causes serious effects to women during pregnancy and the menstrual cycle. Moreover, around 42 million people in India are suffered by thyroid problems. The symptoms of thyroid includes weight gain, fatigue, faintness as well as feeling coldness, etc [4]. The thyroid gland discharges the hormones into the blood which control the entire metabolism of the body. Suppose if the hormones secreted are not in enough quantity, it slows down the overall processes of the body.

The two main types of thyroids are - Hyperthyroidism and Hypothyroidism. These are produced owing to the inappropriate production of thyroid hormones. Thyroid disease can also lead to cancer which may give death to people. Hyperthyroidism is an escalation in the functioning of the thyroid gland and vice versa. This stage is triggered due to a deficiency of iodine in the human body which hints to the thyroid and severe conditions like goitre, and cretinism.

Machine learning techniques play a vital role in analyzing and identifying diseases with reduced costs to the patients [5]. Nowadays, a precise disease diagnosis is a very difficult step in the healthcare industry because several disease types occur every year. Therefore the detection of disease types has become crucial.

Data mining plays a vigorous role in dealing with disease diagnosis and prognosis. Thyroid cancer is one of the deadly diseases which has spread throughout the world. A recent research study indicates that thyroid has spread extensively all over the world and women are more likely to be affected by thyroid disease than men.

In this paper, real thyroid cancer dataset is used for classification which contains many noisy and missing values. The machine learning algorithms played an efficient role in categorising thyroid disorders in [6] and showed high performance and efficiency in classification. Even though the application of computer techniques based on artificial intelligence [7] in healthcare industry is outdated, there has been an upliftment to consider the necessity for machine learning-based healthcare solutions. The results indicate that in the near future, machine learning will play commonplace in the healthcare industry [8]. Before, applying classification to real data, cleansing operations such as filling missing values, and selecting the best features in order to reduce the dimension of the dataset are carried out in order to generate efficient model.

The structure of the paper is designed as follows. The work relevant to thyroid cancer is presented in section 2. The various missing imputation methods and several feature selection algorithms are discussed in section 3. The results of the study are discussed in section 4 and section 5 concludes the study.

Related Works

Data pre-processing is utilized to clean the data [9] and is essential and an important phase in the healthcare industry. Data Cleaning and missing values imputation are the principal fragments of Data pre-processing. The foremost aim of data cleaning is to eliminate redundant and irrelevant items. Different techniques are available for data pre-processing. This paper provides an overview of different techniques used in thyroid data pre-processing

In [10], a survey was done for data pre-processing operations like data cleaning and data reduction by using two algorithms for pre-processing which removed the records with extension .jpg, gif, and .css from input data. The paper [11] explored various machine learning techniques and probability methods for the cleaning and rectifying process. But, these methods need the correlation details between attributes because attributes of the same records in the database may contain dirty details.

A study was carried out in [12] to perform multiclass hypothyroidism by applying selective feature selection algorithms and machine learning techniques. The classification of Hypothyroidism into four types was carried out by applying Random Forest (RF) algorithm, KNN, SVM, and Decision Tree. The results indicated that RF was superior to SVM, KNN, and DT algorithms and achieved 99.18% accuracy. The study in [13] used selection cum classification algorithms to predict hypothyroidism namely recursive feature selection (RFE), univariate feature selection (UFS), and principal component analysis (PCA) alongside SVM, DT, RF, LR, and Naive Bayes (NB). The combination of RFE with machine learning algorithms showed better performance than three feature selection methods. The accuracy attained by all five machine learning algorithms was 99.35% when combined with RFE feature selection.

The paper [14] used KNN with various distance functions to predict the thyroid disease. The optimal features were selected by using the chi-square method, and L1-based featured selection methods. The KNN applied Euclidean and Cosine distance functions. The results stated that KNN achieved promising results in thyroid detection. Mishra et al. [15] applied sequential minimal optimization (SMO), DT, RF, and K-star classifier to forecast hypothyroid disease. The data set used for this analysis contains 3772 records. The results reported that RF and DT outperformed well than the other techniques with accuracy values of 99.44% and 98.97% compared to other methods.

Materials And Methods

Dataset

The data is taken from the UCI machine learning repository and contains 9127 records and 21 attributes. The thyroid dataset contains both Boolean and continuous-valued variables which are shown in Table 1.

Table 1: Dataset Descriptions

S. No.	*Attribute Name*	*Range*
1.	Outcome	hypo, negative
2.	Age	1 to 94
3.	Sex	M, F
4.	On_thyroxine	f, t
5.	Query_ on_thyroxine	f, t
6.	On_antithyroid_medication	f, t
7.	sick	f, t
8.	Thyroid_surgery	f, t
9.	I131_treatment	f, t
10.	Query_hypothyroid	f, t
11.	Query_hyperthyroid	f, t
12.	lithium	f, t
13.	Goitre	f, t
14.	tumor	f, t
15.	hypopituitary	f, t
16.	psych	f, t
17.	TSH	0.005 to 530
18.	T31	0 to 11
19.	TT4	2 to 430
20.	T4U1	0.25 to 2.12

Missing Values Imputation

KNN Imputation

The thyroid Dataset may contain many missing values, and this will lead to problems for many data mining algorithms. It is a good method to detect and substitute missing values for each attribute in the dataset before modeling the prediction task. A popular approach to missing data imputation is to use the k-nearest neighbor (KNN) algorithm [16] which is an effective and efficient method for substituting missing values. The missing values are assigned using the mean value calculated from K- nearest neighbors found in the training data based on the Euclidian distance function. The main advantages of this method are that: a) it can be applied for both qualitative and quantitative attributes estimation; b) It is not essential to develop a predictive model for every attribute per missing data, even if it does not create visible models [17]. It uses mean value for imputing numerical value and mode for categorical data.

Mean/Median Imputation

First, it finds the mean or median of the non-missing values in the attribute and then for each column, substitutes the absent values separately from the other values. This method is suitable for numeric data only.

PCA Imputation

Principal Component Analysis (PCA) is applied frequently in machine learning as a sort of black box dimensionality reduction technique. It is grounded on the concept of principal components, the factorial analysis for mixed type data, which stabilities the effect of all the variables that are continuous and categorical in the assembly of the dimensions of inconsistency. Since the imputation process utilizes the principal axes and components, the likelihood of the missing values is evaluated based on the likeness between individuals and on the relations among variables. The eminence of the imputation process is estimated with the aid of simulation studies and real-time datasets.

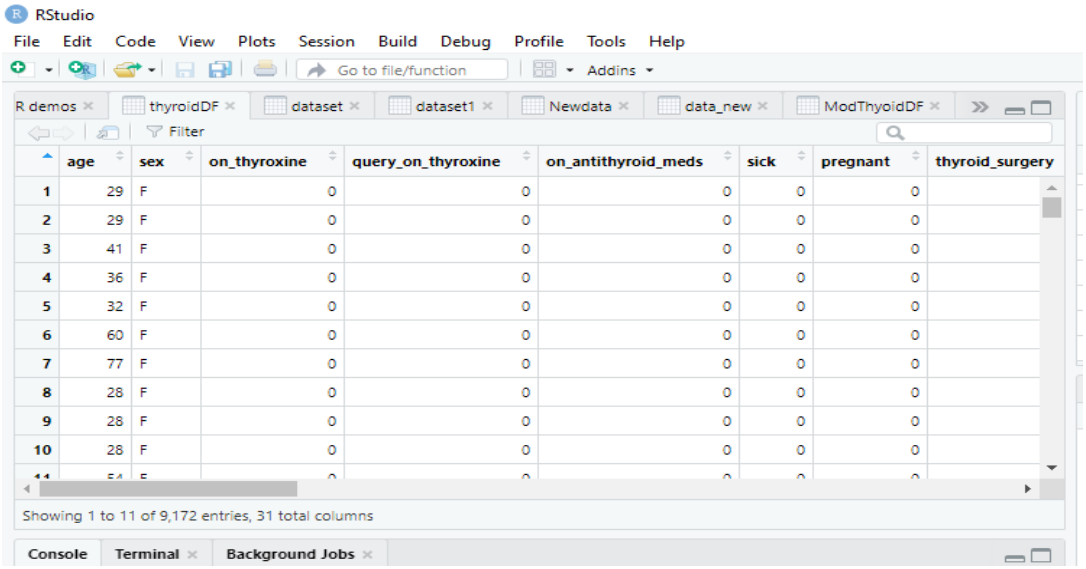
Proposed Hybrid Thyroid Imputation Technique (HTIT)

In this study, a new imputation technique called HTIT is proposed for imputing the missing values in the thyroid dataset. This method integrates the advantages of two

imputations methods such as KNN and PCA method. In the first step, the original dataset is analyzed in order to determine various missing data value patterns. In the second step, KNN imputation method is applied to the original dataset followed by PCA imputation method. This hybrid method attained the results in balanced estimates and provides extra rationality than ad hoc approaches to mislaid data in the thyroid dataset.

Performance Evaluation

The experiment was carried out using the RStudio tool. Before applying K-NN, Mean/Median and PCA method, and HTIT, the missing values in each column are found. The following Fig1. Shows the screenshot of loading the thyroid dataset in RStudio.



	age	sex	on_thyroxine	query_on_thyroxine	on_antithyroid_meds	sick	pregnant	thyroid_surgery
1	29	F	0	0	0	0	0	
2	29	F	0	0	0	0	0	
3	41	F	0	0	0	0	0	
4	36	F	0	0	0	0	0	
5	32	F	0	0	0	0	0	
6	60	F	0	0	0	0	0	
7	77	F	0	0	0	0	0	
8	28	F	0	0	0	0	0	
9	28	F	0	0	0	0	0	
10	28	F	0	0	0	0	0	

Fig.1. Loading of the dataset in RStudio

Results And Discussions

Fig.2 shows the percentage of missing values in each column of the thyroid dataset. From this figure, one can easily understand the number of missing values in the dataset before applying any imputation algorithms.

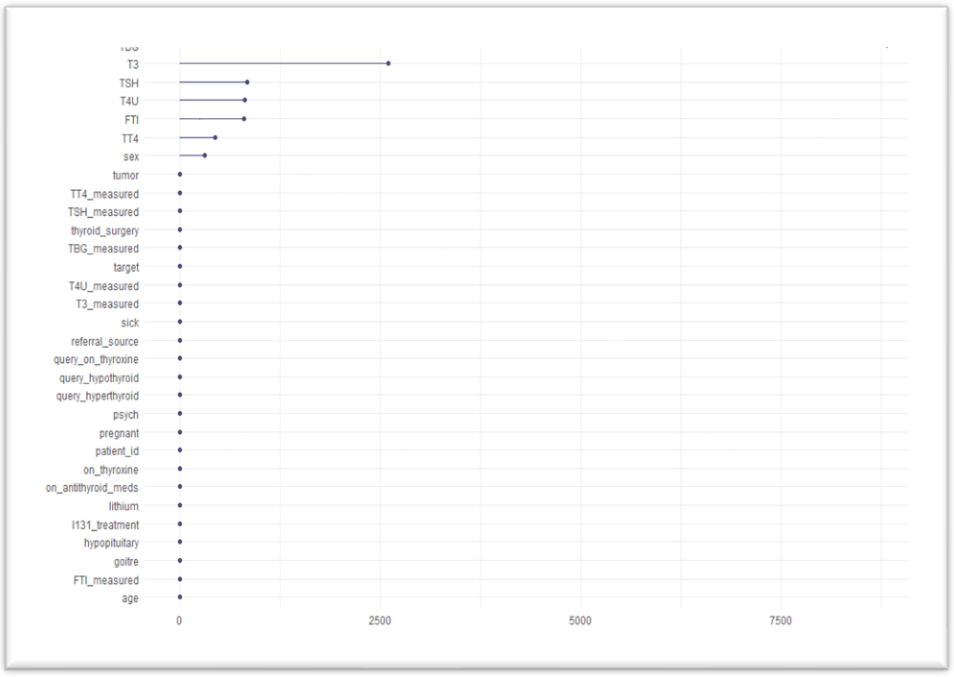


Fig.2. Percentage of Missing values

Fig.3 shows the proportion of missing values against each column in the data and original thyroid data. The missing values are shown in red colors in the graph.

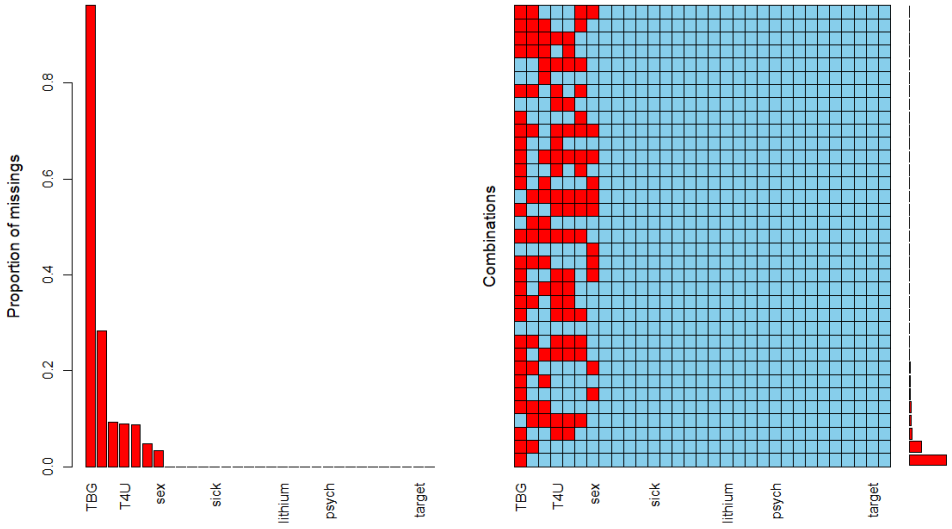


Fig.3. Proportion of missing values in Thyroid dataset

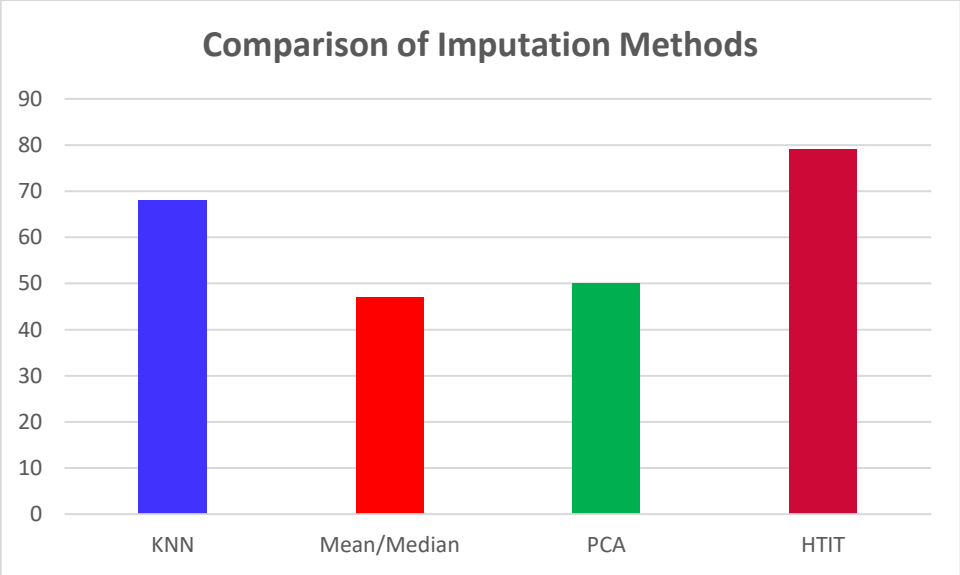


Fig.4. Efficiency of Imputing Algorithms

Fig.4 Shows the comparison of missing value imputation methods such as HTMT imputation, KNN, Mean/Median, and PCA imputation for the thyroid dataset. The Hybrid imputation is a combination of KNN imputation followed by PCA Imputation. From this figure, we found that the HTIT method is superior to Mean/Median, KNN, and PCA imputation for handling the missing values in the thyroid dataset. This method is well suited for handling both categorical and numerical data. From this graph, it is observed that the imputation rate of HTIT imputation is superior to KNN, PCA, and the Mean/Median algorithm. It is also found that the proposed method yields more sufficient results than other imputation methods. The efficiency increment by the HTIT Imputation method is 11% greater than KNN, 32% more than the Mean/Median method, and 29% more than the PCA method. Fig.5 shows the thyroid data after filling in the missing values by KNN. The graph shows the histogram pattern of the dataset.

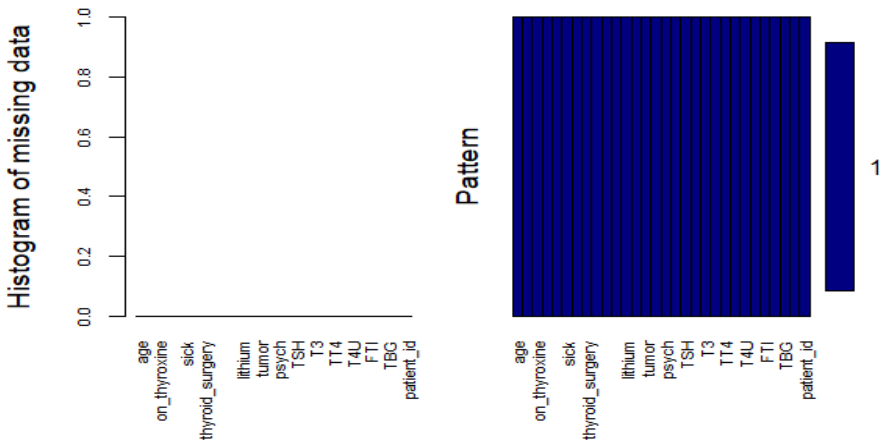


Fig.5. After Imputing Missing values

Conclusions

In this work, KNN and PCA-based pre-processing algorithms to remove the redundant and irrelevant values in the thyroid dataset and also compared with algorithms namely KNN, Mean/Median, and PCA. Among the three, Hybrid impute seems to be superior to the other three imputation algorithms. After pre-processing, efficient feature selection algorithms will be applied and various classification algorithms will be used to generate an accurate prediction model for the Thyroid dataset. In future cases, both regression and classification algorithms will be combined to produce accurate diagnosis results.

Future Enhancements

(1) In future, more imputations algorithms will be compared for this data.

(2) Dimension reduction techniques such as feature selection and extraction will be used.

(3) In order to generate an efficient prediction model, transformation techniques will be used.

References

Siegel, R. L., Miller, K. D., Fuchs, H. E. &Jemal, A. Cancer statistics, 2022. CA Cancer J. Clin. 72, 7–33 (2022).

Azar, A.T, Hassanien, A.E. and Kim, T. Expert system based on neural fuzzy rules for thyroid diseases diagnosis, Computer Science, Artificial Intelligence, arXiv:1403.0522, pp. 1-12,2012.

Keles, A. ESTDD: Expert system for thyroid diseases diagnosis, Expert Syst Appl., Vol. 34, No.1, pp.242–246,2008.

Ammulu, K., and T. Venugopal. Thyroid data prediction using data classification algorithm. Int. J. Innov. Res. Sci. Technol 4.2 (2017): 208-212.

Aswad, Salma Abdullah, and Emrullah Sonuç. "Classification of VPN Network Traffic Flow Using Time Related Features on Apache Spark." 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE, 2020.

Banu, G. Rasitha. A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease. International Journal of Computer Sciences and Engineering 4.11 (2016): 64-70.

Chandio, Jamil Ahmed, et al. "TDV: Intelligent system for thyroid disease visualization." 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube). IEEE, 2016.

Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. JAMA, 309(13):1351–1352, 2013.

Chandrama W., Devale P.R., Ravindra M., Survey on Data Preprocessing Method of Web Usage Mining, International Journal of Pure and Applied Mathematics Special Issue 792 Journal of Computer Science and Information Technologies 5(3) (2014), 3521-3524.

Navin Kumar T., Solanaski A.K., Sanjay T., An Algorithmic Approach to Data Preprocessing in Web Usage Mining, International Journal of Information Technology and Knowledge Management 2 (2010).

Yakout, Mohamed, Laure Berti-Équille, and Ahmed K. Elmagarmid. (2013) "Don't be SCAREd: use Scalable Automatic Repairing with maximal likelihood and bounded changes", in the Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, New York, USA.

Das, R.; Saraswat, S.; Chandel, D.; Karan, S.; Kirar, J.S. An AI Driven Approach for Multiclass Hypothyroidism Classification. In Proceedings of the International Conference on Advanced Network Technologies and Intelligent Computing, Varanasi, India, 17–18 December 2021; pp. 319–327.

Riajuliislam, M.; Rahim, K.Z.; Mahmud, A. Prediction of Thyroid Disease (Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques. In Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 27–28 February 2021; pp. 60–64.

Abbad Ur Rehman, H.; Lin, C.Y.; Mushtaq, Z. Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease. J. Chin. Inst. Eng. 2021, 44, 77–87.

Mishra, S.; Tadesse, Y.; Dash, A.; Jena, L.; Ranjan, P. Thyroid disorder analysis using random forest classifier. In Intelligent and Cloud Computing; Springer: Berlin/Heidelberg, Germany, 2021; pp. 385–390.

Minakshi et al., Missing Value Imputation in Multi Attribute Data Set, Intl J. of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5315-5321.

Liu Peng, Lei, A Review of Missing Data Treatment Methods.