How to Cite:
Parimala, S., & Vadivu, P. S. (2022). A proactive approach with feature selection for thyroid classification by adopting data mining techniques. *International Journal of Health Sciences*, *6*(S2), 15003–15012. https://doi.org/10.53730/ijhs.v6nS2.8979

A proactive approach with feature selection for thyroid classification by adopting data mining techniques

Parimala. S
Research Scholar, Department of Computer Science, Hindustan College of Arts and Science, Coimbatore, India
Corresponding author email: parimsarun@gmail.com

Dr. P. Senthil Vadivu
Head and Professor, Department of Computer Applications, Hindustan College of Arts and Science, Coimbatore, India
Email: sowjupavan@gmail.com

***Abstract*---**Data mining holds a staggering amount of promise for healthcare services due to the rapid growth of enormous amounts of electronic health records. In the past, healthcare professionals and physicians manually kept track of patient information. The data was a pain to keep up with. By modernizing and digitizing data and making it available in an easily accessible format, new strategies are introduced to reduce human labor. The main goal is to detect disease diagnosis at an earlier stage and with greater precision... Data mining is an important aspect of the classification idea in health care services. The diagnosis of diseases and the provision of necessary therapy for patients has become a major problem for healthcare providers. Thyroid disease is recognized with increased accuracy in this research using updated data mining algorithms and classification systems, which is more significant in diagnosing a chronic condition like thyroid. This study discusses the topic of dimensionality reduction and the usage of feature selection techniques, which is one of the most important tasks in feature engineering. An algorithm is proposed that provides a solution to the difficulties in the existing categorization system. According to the findings, the proposed methodology greatly improves thyroid disease classification accuracy.

***Keywords*---**data mining, thyroid, dimensionality reduction, feature selection algorithms.

Introduction

Feature selection is one of the most widely used and critical data preprocessing strategies for data mining (Asha Gowda Et al,2010) The purpose of feature selection in a classification challenge is to achieve the highest possible accuracy rate [2]. Feature selection is the process of removing superfluous or incorrect features from the original data collection. As a result, the classifier's processing time lowers, while accuracy increases, because irrelevant characteristics can include noisy data, negatively impacting classification accuracy (Doraisami Et al.,2008). Understanding can be increased, and the cost of data handling can be reduced, thanks to feature selection (Arauz Et al., 2011).

Feature selection (FS), one of the contributing commitments, has already aided data mining by allowing it to find correlated features and delete redundant or uncorrelated features from the original dataset (Zhou Et al.,2017; Bolon Et al.,2015). Feature selection, which is often used to locate correlated features and remove redundant or uncorrelated information from a feature collection, is one of the most important data analysis tools. (Bolon El al., 205). Randomized or unpredictable qualities can make it more difficult for a classification to create effective connections and duplicated or correlated features can complicate a classifier without offering any useful information (Wu Et al., 2012). As feature selection methods, filter, wrapper, and embedding approaches (Bolon Et al., 2013 & Abd Et al., 2014) have all been developed.

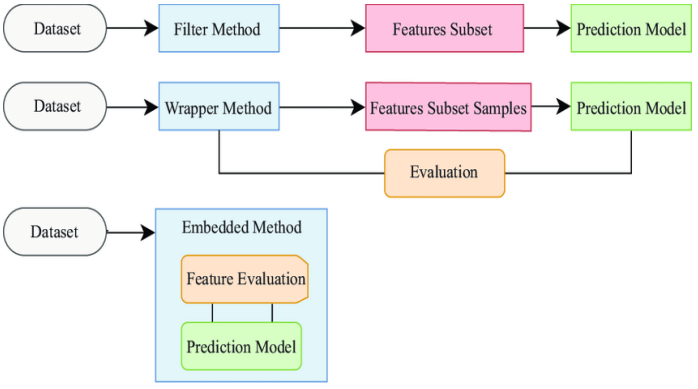

Figure 1: Different approaches to Feature Selection methods

Filter Methods

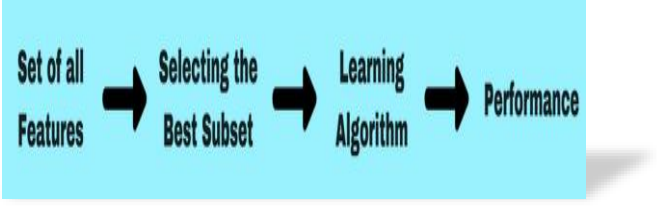

Figure 2: Filter-based Approach

Filter-based approaches are employed in the pre-processing phase. The selection of features is independent of any data mining algorithm. Instead, features are picked based on their scores in various statistical tests for their connection with the outcome variable. Different statistical tests are done depending on their scores for their association with the outcome variable. The features are scored and either maintained or eliminated from the dataset based on their score. The approaches are frequently univariate, taking into account the feature separately or about the predictor variables. We are unable to forecast which algorithm will benefit from a filter-based approach. For the feature selection technique in this study, we exclusively used a filter-based approach. Filter-based approaches are much faster than wrapper methods since they do not require training the model and rely solely on the features in the data set.

Benefits from Feature Selection

The basic goal of feature selection is to build a new dataset with no redundant or extraneous characteristics, as well as the original data pattern while keeping all of the useful information from the original dataset. Because of its ability to minimize dimension, feature selection methods are currently widely utilized in the fields of written text and DNA microarray analysis. They show their worth when the number of features is large but the number of samples is restricted. Feature selection's main purpose is to locate features that are based on the original dataset. It's worth noting that certain related qualities may be redundant because they're so closely linked to others. (Vergara Et al., 2014; Liu Et al., 2018; Bagherzadeh Et al.,2016). (1) The chosen characteristics can be used to build a brief model for describing original data, which is beneficial for improving criteria performance; (2) (3) the chosen characteristics can help the decision-maker pick meaningful information from a large number of noisy data by reflecting the core properties of the original data, and thus are beneficial for identifying notion displacements of content affirmation with stability; and (Bagherzadeh Et al., 2016; Dash Et al., 1997)

Feature selection strategies are typically characterized as filters, wrappers, or embedding based on how they combine a supervised learning technique and a classification algorithm to pick a subset of features (Guyon Et al., 2003; Liu Et al., 2018; Bagherzadeh Et al., 2016; Dash Et al., 1997), independent of whether they use supervised or unsupervised methods (Liu Et al., 2018; Chen Et al., 2018; Wang Et al., 2018). The only distinction between supervised and unsupervised learning is whether or not class labels are available. In the realm of feature selection, the available label information is used with the very first methods for evaluating the implications of features and offering ranks of these features. The latter looks for hidden structures in unlabeled data and uses intrinsic data attributes to build a feature selector (Wang Et al., 2015)

Reparability criteria, consistency criteria, and error rate criteria are useful for evaluating the performance of supervised feature selection algorithms. In contrast, the clustering validity criteria, information theoretical criteria, and feature similarity criteria are currently commonly used for unsupervised feature selection methods (Ang Et al., 2016; Li Et al., 2017; Liu Et al., 2005)

Literature Survey

Dharmarajan et al.[23], (2020) have analyzed that, the hypothyroid and hypothyroid datasets and the data mining methodology is used to determine the positive and negative values from the full dataset. When comparing male and female datasets, the experimental results show that females are more influenced than males. By comparing the Decision tree and Support vector Machine, were able to improve accuracy, precision, and recall. The use of various optimization methods or rule extraction algorithms has resulted in further improvements. Using a big data framework, the authors Vijayalakshmi et al., (2018) have investigated the Hypothyroid Prediction System. Big data may be used for predictive modeling and information retrieval more efficiently with the proposed methodology. Patients with hypothyroidism can be predicted using the Naive Bayes hypothesis. The classifier demonstrates the best result in terms of accuracy and execution time when it comes to forecasting.

Sumathi et al., (2018) have researched that thyroid hormone controls a variety of metabolic processes throughout the body. Thyroid illness affects women more than it does men. Thyroid illnesses are divided into two categories: hyperthyroidism, which generates a lot of thyroid hormone in the blood, and hypothyroidism, which produces less thyroid hormone in the blood. Hypothyroidism is a condition that causes a weaker immune system as well as chronic degenerative diseases and hormone abnormalities. People with hypothyroidism are taking the correct dosage of thyroid medication and have normal thyroid levels. For a more accurate diagnosis, subtype classification is required. Thyroid diseases and subtypes are efficiently classified using the EM clustering technique and the J48 classification algorithm. According to the authors Sidiq et al., (2019), women are 5 to 8 times more likely than males to suffer from thyroid diseases over the world. It is caused by the thyroid gland's abnormal secretion of thyroid hormones, which is one of our body's most vital organs, positioned in the front of the neck and below Adam's apple. Thyroid disease develops when the thyroid gland ceases to function normally, and it is classified as hypothyroidism or hyperthyroidism. The Decision Tree algorithm which is highly preferred gives the highest test accuracy of 98.89% when compared to other classifiers.

Yadav and Pal (2020)[have researched the hormones which have an important role in blood flow and metabolism in humans, and both excessive and low hormone levels are harmful. Thyroid hormones are produced by the thyroid gland to keep blood flowing and regulate metabolism; the thyroid gland produces three types of hormones: triiodothyronine (T3), thyroxin (T4), and thyroid-stimulating hormone (TSH) (TSH). Hyperthyroidism occurs when the thyroid gland produces more hormones, while hypothyroidism occurs when the thyroid gland produces fewer hormones. The prediction of thyroid disease is calculated using three classifiers: decision tree, random forest, and additional tree. Tyagi et al., (2018) have proposed that the thyroid gland, located in the neck, is an endocrine gland. It rises beneath the As apple in the lower region of the human neck, assisting in the secretion of thyroid hormones, which in turn affects the pace of metabolism and protein synthesis. Thyroid hormones are useful in a variety of ways to govern the body's metabolism,

Calculating how fast the heartbeats and how quickly calories are burned. The thyroid gland's production of thyroid hormones aids in the body's metabolism's dominance. The subset selection using mutual information and thyroid dataset prediction is done using the artificial neural networks (ANN) algorithm. Vinitha, (2020) has examined that thyroid disease affects a large percentage of the population nowadays. Thyroid problems can be identified early on, which will aid in recovery. The survey will aid in the early detection of thyroid issues. There are a lot of different categorization algorithms explored. The J48 algorithm, for example, delivers the highest accuracy while simultaneously being the quickest. As a result, in the future, the J48 can be used in conjunction with any other advanced methodology to provide maximum accuracy while using the least amount of time. In the healthcare industry, Begum & Parkavi, (2019) proposed that data mining is mostly utilized to make decisions, diagnose diseases, and provide better treatment to patients at a lower cost. The classification of thyroid illness is a crucial step in the disease prediction process. Dimensionality reduction could be done in the future to reduce the number of thyroid blood tests and the time it takes to diagnose disease.

Issues during Classification

During classification, the presence of irrelevant and noisy duplicate features causes two issues: I) increased computational cost and ii) Can result in overfitting. As a result, FS algorithms are utilized to address the aforementioned concerns. FS algorithms are based on the idea that not all features are necessary for data mining, and that certain features may be deleted without hurting performance. Is the challenge of choosing elements that have the greatest impact on characterizing the findings and eliminating features that have little or no impact? It's a matter of ubiquitous interwoven ordinance in data mining, which reduces the number of features, removes superfluous, chaotic, and redundant data, and boosts accuracy.

Proposed Method

To find relevant features, a single feature selection approach is typically utilized. Several feature selection methods are employed for this purpose, each with its own set of benefits and drawbacks. This research study recommends employing multiple feature selection methods to pick optimal features integrating the benefits of numerous feature selection algorithms. Four essential and often utilized algorithms are used in this study. Mutual Information, Fisher Score, Pearson Coefficient, and Chi-Square are the terms used. 'Multiple Feature Selection Algorithms (MFS)' is the name of this algorithm.

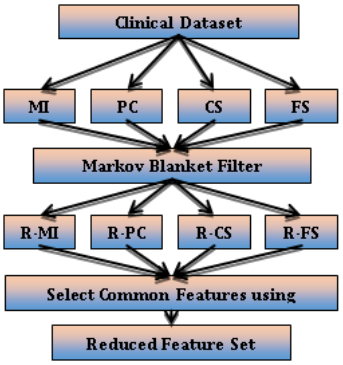

Figure 4. Proposed Multiple Feature Selection Algorithm (MFS)

Workflow of Multiple Feature Selection

The flow of the suggested filter-based technique is depicted in the diagram. Statistical tests are used in the proposed approach to finding the subset of attributes having the highest predictive potential. To identify optimal features, the M2FPS algorithm uses four feature selection filters: Mutual information, Pearson correlation, Chi-Square, and Fisher's Criterion Score. The resulting features are fed into a more computationally costly subset selection technique called Markov blanket filtering, which yields four sets of ideal reduced features: R-MI, R-PC, R-CS, and R-FS. To choose only common features that exist amongst the four subsets, these four sets are joined using the intersection technique. (M. Dash Et al., 1997).

Datasets

Dataset for this research work is taken from UCI Thyroid dataset (UTD)
Taken from: http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease
Number of Instances: 7200
Number of Attributes: 21
Number of Classes: Normal, hypo, and hyper
Performance Metrics: Accuracy
The missing percentage was varied with values ranging from 0 to 60% in steps of 10.

Results and Discussion

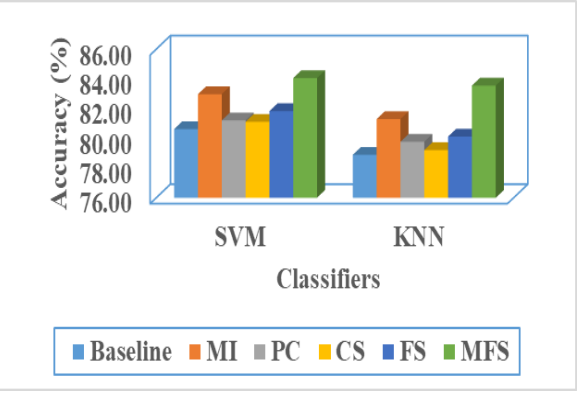

Figure 5. Analysis of Feature Selection Algorithms Accuracy (%)

Table 1: Different types of Algorithms used and their results

Algorithms	Values (%)
Baseline	80.62
Mutual Information	82.97
Person Coefficient	81.21
Chi-Square	81.11
Fisher Score	81.83
Multiple Feature Selection	84.06

We have been using SVM and KNN since they became popular in thyroid disease. The feature selection approach is thoroughly examined using SVM and KNN classifiers. Table: 1 clearly illustrates that the suggested technique, the Multiple Feature Algorithm, is the only one that provides great accuracy. The suggested technique significantly increases thyroid illness categorization accuracy

Conclusion and Future Work

The phrase "Multiple Feature Selection Algorithms (MFS)" refers to four algorithms: Mutual Information, Information Gain, Fisher Score, Pearson Coefficient, and Chi-Square. We have tackled the problem of dimensionality reduction using feature selection techniques, which is regarded as one of the most important tasks in feature engineering. The Multiple Feature Selection Algorithms algorithm is offered as a remedy to the difficulties in the current classification system. We've been using SVM and KNN since they became popular in thyroid disease. The feature selection approach is thoroughly examined using SVM and KNN classifiers. Table 1 clearly illustrates that the suggested technique, the Multiple Feature Algorithm, is the only one that provides great accuracy. The suggested technique significantly increases thyroid illness categorization accuracy. In the future, a few more difficulties and algorithms will be researched and improved to improve accuracy by implementing datasets that can aid to improve diagnosis accuracy.

References

A. Arauzo-Azofra, J. L. Aznarte, and J. M. Benítez, Empirical study of feature selection methods based on individual feature evaluation for classification problems, Expert Systems with Applications, 38 (2011) 8170-8177.

Asha Gowda Karegowda, M.A.Jayaram, and A.S. Manjunath, *—" Feature Subset Selection Problem using Wrapper Approach in Supervised Learning"*, International Journal of Computer Applications, Vol. 1, No. 7, pp. 0975–8887, 2010.

Dharmarajan, K., Balasree, K., Arunachalam, A.S. & Abirmai, K. (2020). Thyroid Disease Classification Using Decision Tree and SVM. Indian Journal of Public Health. [Online]. 11 (03). bll 229. Available from: https://www.researchgate.net/profile/K_Dharmarajan/publication/341742234_Thyroid_Disease_Classification_Using_Decision_Tree_and_SVM/links/5ed192d492851c9c5e664edb/Thyroid-Disease-Classification-Using-Decision-Tree-and-SVM.pdf.

F. Bagherzadeh-Khiabani, A. Ramezankhani, F. Azizi, F. Hadaegh, E. W. Steyerberg, and D. Khalili, "A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results," J. Clin. Epidemiol. vol. 71, pp. 76–85, Mar. 2016.

H. Liu and L. Yu, "towards integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491–502, Apr. 2005.

H. Zhou, S. Han, and Y. Liu, "A novel feature selection approach based on document frequency of segmented term frequency," IEEE Access, vol. 6, pp. 53811–53821, 2018.

Hosseinzadeh, M., Ahmed, O.H., Ghafour, M.Y., Safara, F., Hama, H. kamaran, Ali, S., VO, B. & Chiang, H.-S. (2021). A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis on the internet of medical things. The Journal of Supercomputing. [Online]. 77 (4). bll 3616–3637. Available from: http://link.springer.com/10.1007/s11227-020-03404-w.

I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, no. 6, pp. 1157–1182, Jan. 2003.

J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 13, no. 5, pp. 971–989, Sep. 2016.

J. Li and H. Liu, "Challenges of feature selection for big data analytics," IEEE Intell. Syst., vol. 32, no. 2, pp. 9–15, Mar. /Apr. 2017.

J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," Neural Comput. Appl., vol. 24, no. 1, pp. 175–186, 2014.

J. Wu and Z. Lu, "A novel hybrid genetic algorithm and simulated annealing for feature selection and kernel optimization in support vector regression," in Proc. IEEE 5th Int. Conf. Adv. Comput. Intell. (ICACI), Oct. 2012, pp. 999–1003.

M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., vol. 1, nos. 1–4, pp. 131–156, 1997.

N. Abd-Alsabour, "A review on evolutionary feature selection," in Proc. Eur Modeling Symp., 2014, pp. 20–26.

ODDS (2021). Thyroid Disease dataset. [Online]. 2021. Available from: http://odds.cs.stonybrook.edu/thyroid-disease-dataset/.

Pal, R., Anand, T. & Dubey, S.K. (2018). Evaluation and performance analysis of classification techniques for thyroid detection. International Journal of Business Information Systems. [Online]. 28 (2). bll 163. Available from: http://www.inderscience.com/link.php?id=91862.

Q. Tuo, H. Zhao, and Q. Hu, "Hierarchical feature selection with subtree based graph regularization," Knowl. - Based Syst., vol. 163, pp. 996–1008, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/artiPII/pii/S0950705118305094

R. Chen, N. Sun, X. Chen, M. Yang, and Q. Wu, "Supervised feature selection with a stratified feature weighting method," IEEE Access, vol. 6, pp. 15087–15098, 2018.

Rinartha, K., & Suryasa, W. (2017). Comparative study for better result on query suggestion of article searching with MySQL pattern matching and Jaccard similarity. In *2017 5th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-4). IEEE.

Rinartha, K., Suryasa, W., & Kartika, L. G. S. (2018). Comparative Analysis of String Similarity on Dynamic Query Suggestions. In *2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)* (pp. 399-404). IEEE.

Ron Kohavi, George H. John, ―Wrappers for feature subset Selection‖, Artificial Intelligence, Vol. 97, No. 1-2. pp. 273-324, 1997.

S. Doraisami, S. Golzari, A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music, Content-Based Retrieval, Categorization, ion, and Similarity, 2008

S. Wang and W. Zhu, "Sparse graph embedding unsupervised feature selection," IEEE Trans. Syst., Man, Cybern., Syst., vol. 48, no. 3, pp. 329–341, Mar. 2018.

S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Unsupervised feature selection via maximum projection and minimum redundancy," Knowl.-Based Syst., vol. 75, pp. 19–29, Feb. 2015.

selection and a new method based on label construction," Neurocomputing, vol. 180, pp. 3–15, Mar. 2015.

Sidiq, U., Mutahar Aaqib, S. & Khan, R.A. (2019). Diagnosis of Various Thyroid Ailments using Data Mining Classification Techniques. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. [Online]. bll 131–136. Available from: http://ijsrcseit.com/paper/CSEIT195119.pdf.

Sindhya, M.K. (2020). Effective Prediction of Hypothyroid Using Various Data Mining Techniques. International Journal of Research and Development. [Online]. 5 (2). Available from: https://eprajournals.com/jpanel/upload/143am_74.EPRA JOURNALS-4009.pdf.

Sumathi, A., Nithya, G. & Meganathan, S. (2018). Classification of thyroid disease using data mining techniques. International Journal of Pure and Applied Mathematics. [Online]. 119 (12). bll 13881–13890. Available from: http://www.acadpubl.eu/hub/2018-119-12/articles/5/1275.pdf.

Tyagi, A., Mehra, R. & Saxena, A. (2018). Interactive Thyroid Disease Prediction System Using Machine Learning Technique. In: 2018 Fifth International Conference on Parallel, Distributed, and Grid Computing

V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, “A review of feature selection methods on synthetic data,” Knowl. Inf.Syst. vol. 34, no. 3, pp. 483–519, 2013.

V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, “Recent advances and emerging challenges of feature selection in the context of big data,” Knowl.-Based Syst., vol. 86, pp. 33–45, Sep. 2015.

Vijayalakshmi, K., Dheeraj, S. & Deepthi, B.S.S. (2018). Intelligent Thyroid prediction system using Big data. Int J Comput Sci Eng. [Online]. 6. bll 321–326. Available from: https://www.researchgate.net/profile/Vijayalakshmi_Kakulapati2/publication/323166604_Intelligent_Thyroid_prediction_system_using_Big_data/links/5a913b490f7e9ba4296c221a/Intelligent-Thyroid-prediction-system-using-Big-data.pdf.

Vinitha (2020). A Survey of Thyroid Detection Using Datamining Techniques. International Journal of Research and Development. [Online]. 5 (7). Available from: https://eprajournals.com/jpanel/upload/1150pm_IJRD JULY 2020 FULL JOURNAL.pdf#page=82.

W. Zhou, C. Wu, Y. Yi, and G. Luo, “Structure preserving non-negative feature self-representation for unsupervised feature selection,” IEEE Access, vol. 5, pp. 8792–8803, 2017.

X.-Y. Liu, Y. Liang, S. Wang, Z.-Y. Yang, and H.-S. Ye, “A hybrid genetic algorithm with wrapper-embedded approaches for feature selection,” IEEE Access, vol. 6, pp. 22863–22874, 2018.

Yadav, D.C. & Pal, S. (2019). Thyroid prediction using ensemble data mining techniques. International Journal of Information Technology. [Online]. Available from: http://link.springer.com/10.1007/s41870-019-00395-7.

Yadav, D.C. & Pal, S. (2020). Prediction of thyroid disease using decision tree ensemble method. Human-Intelligent Systems Integration. [Online]. 2 (1–4). bll 89–95. Available from: http://link.springer.com/10.1007/s42454-020-00006-y.