Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 13, Issue. 8, August 2024, pg.118 – 129

A Hybrid Approach for Thyroid Disease and Classification Using Attribute Associations Rule Mining

Amutha C
Department of Computer Science and Application
Park's College (Autonomous), Bharathiar University
Chinnakkarai, Karaipudhur, Tirupur, Tamil Nadu, INDIA
E-mail: arjammusan24@gmail.com

Dr. P. Periyasamy
Faculty of Computer Science and Application
Park's College (Autonomous), Bharathiar University
Chinnakkarai, Karaipudhur, Tirupur, Tamil Nadu, INDIA
E-mail: pereee@yahoo.com

DOI: https://doi.org/10.47760/ijcsmc.2024.v13i08.013

Abstract: In recent years, thyroid diseases have become increasingly prevalent worldwide. In India, for instance, one in eight women is affected by hypothyroidism, hyperthyroidism, or thyroid cancer. Research indicates that approximately 20% of Indians are diagnosed with endemic goiter. Several factors influence thyroid function, including stress, infection, trauma, toxins, low-calorie diet, and certain medications. Preventing such diseases is crucial as treatments often involve long-term medication or surgical intervention. Thyroid disease diagnosis is pivotal in medicine, where precise classification is essential for prompt treatment. This paper proposes a novel approach to thyroid disease classification by integrating Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) with Association Rule Mining (ARM). The hybridization of PSO and ACO enhances exploration and exploitation capabilities, thereby improving the efficiency of association rule mining for feature selection. The proposed method aims to identify significant association rules among thyroid disease attributes, thereby facilitating accurate classification. Experimental evaluations conducted on benchmark thyroid disease datasets demonstrate the effectiveness of the proposed approach in terms of classification accuracy and computational efficiency. The results indicate that the hybrid PSO and ACO-based association rule mining approach surpasses traditional methods, underscoring its potential to enhance thyroid disease diagnosis systems.

Keywords: ACO, PSO, ARM, FaRP, Data Mining, Thyroid diseases, Classification

1. INTRODUCTION

Thyroid disease refers to any dysfunction of the thyroid gland, a butterfly-shaped gland located in the neck. This gland regulates various bodily functions by producing hormones influencing metabolism, growth, and energy expenditure. Endocrinology, on the other hand, is the branch of medicine that deals with the endocrine system, which includes glands such as the thyroid, pancreas, adrenal glands, and others. Thyroid gland disorders rank among the most prevalent endocrine conditions worldwide, trailing only behind diabetes, as reported by the World Health Organization. Approximately 2% of individuals experience hyperthyroidism, while hypothyroidism affects around 1%, with women being about ten times more likely to be affected than men [1]. These disorders may arise from thyroid gland dysfunction, pituitary gland failure, or hypothalamic malfunction. In regions with dietary iodine deficiency, goiter or active thyroid nodules may occur, affecting up to 15% of the population. Additionally, the thyroid gland can be susceptible to various tumours and autoimmune attacks by endogenous antibodies.

Early detection, diagnosis, and treatment are crucial in preventing disease progression and complications, according to medical professionals. Timely identification and differential diagnosis significantly improve the prospects of effective treatment. However, clinical diagnosis remains challenging despite numerous efforts. The thyroid gland, resembling a butterfly, resides at the base of the throat and produces two vital hormones, levothyroxine (T4) and triiodothyronine (T3), crucial for regulating body temperature, blood pressure, and heart rate. Thyroid disorders, primarily caused by iodine deficiency but also by other factors, disrupt hormone secretion and affect bodily functions. Treatments involve regulating thyroid activity through medications such as T3, T4, and thyroid stimulating hormone (TSH) [2].

Thyroid disorders are broadly classified into hypothyroidism and hyperthyroidism. Symptoms of hyperthyroidism or thyrotoxicosis (Grave’s disease or exophthalmic goiter) include dry skin, heightened temperature sensitivity, hair thinning, weight loss, rapid heart rate, high blood pressure, excessive sweating, neck enlargement, nervousness, shortened menstrual cycles, irregular bowel movements, and trembling hands. In contrast, hypothyroidism, characterized by reduced thyroid hormone production, presents symptoms such as obesity, slow heart rate, heightened temperature sensitivity, neck swelling, dry skin, numbness in the hands, hair problems, heavy menstrual cycles, and gastrointestinal issues. Without treatment, these symptoms can exacerbate over time [3-4].

The thyroid gland disorder, ranked as the second most prevalent endocrine disorder, profoundly impacts various physiological functions. In India alone, an estimated 42 million individuals suffer from this condition [5]. Notably, women aged between 19 and 45 are particularly susceptible to these disorders [6]. Given the integral role of the thyroid gland in bodily mechanisms, its malfunction triggers a cascade of symptoms, underscoring the importance of early detection and preventative measures. To address this imperative, one effective strategy involves the utilization of knowledge extraction tools. Among these tools, data mining approaches have gained prominence, offering various models for knowledge discovery. In particular, data mining techniques such as association and classification play crucial roles [7]. Classification involves the identification of logical descriptions, known as classifiers, derived from training datasets with predetermined targets, enabling the prediction of outputs based on inputs. This process typically comprises two phases: learning, where classification algorithms analyze training data, and classification, where test data evaluate accuracy [8]. Association rule mining, another key technique, aims to establish associative relationships between data items, typically expressed as rules of the form $X \Rightarrow Y$, where X and Y represent sets of data items. Classification based on association rules endeavours to construct a classifier by uncovering a concise set of rules forming an associative classifier, with regulations in the form $X \Rightarrow C$, where C represents a class label treated as a special case of Y in association rules.

Preventive healthcare remains a paramount concern for medical professionals, emphasizing the critical importance of timely and accurate diagnosis to mitigate risks for patients. In recent times, alongside traditional medical reports, additional insights provided by decision support systems or advanced diagnostic techniques based on symptoms have become increasingly common. Questions about factors influencing thyroid health, susceptibility to goiter disease among specific populations, or optimal treatment strategies can often find answers through the application of data mining techniques. By leveraging healthcare data and employing rigorous data mining processes, valuable insights can be gleaned to aid decision-making processes, accelerate disease diagnosis, and enhance treatment efficacy [9-10]. This approach not only facilitates rapid and precise disease diagnosis but also enables the provision of tailored medication regimens, ultimately minimizing the risk of adverse outcomes for patients. and other symptoms such as nervousness, irritability, and heat intolerance.

The thyroid, resembling a butterfly in shape, occupies a crucial position at the base of the throat, orchestrating the production of two vital hormones: levothyroxine (T4) and triiodothyronine (T3). These hormones play pivotal roles in regulating various bodily functions, including stabilizing body temperature, blood pressure, and heart rate. Additionally, the hormone reverse T3 (RT3), derived from thyroxine (T4), acts to counteract the effects of T3. Disruptions in thyroid function can lead to the development of hyperthyroidism and hypothyroidism, two prevalent thyroid disorders. Hypothyroidism, characterized by an underactive thyroid gland or insufficient hormone production, can result in a range of health complications if left untreated. These may include obesity, joint pain, infertility, and cardiovascular issues [11-12]. Conversely, hyperthyroidism entails an overactive thyroid gland, leading to excessive thyroxine production and heightened metabolic activity. Symptoms of hyperthyroidism include sudden weight loss, rapid or irregular heartbeat, as well as manifestations such as nervousness, irritability, and intolerance to heat.

Understanding the intricate balance of thyroid hormones and recognizing the signs and symptoms of thyroid disorders are essential for timely diagnosis and effective management of these conditions. Through appropriate medical intervention and treatment, individuals afflicted by thyroid dysfunction can achieve improved health outcomes and quality of life. In addition to the symptoms mentioned earlier, such as sudden weight loss and rapid heartbeat, hyperthyroidism can also manifest as sweating, nervousness, or irritability. Stress, infection, toxins, trauma, and certain medications directly impact the production of thyroid hormones, highlighting the importance of identifying symptoms and detecting abnormal thyroid hormone levels early through clinical investigation. Early detection enables accurate diagnosis and the prescription of appropriate medication. Patients should regularly monitor their clinical status to ensure they receive necessary treatment for as long as required. This proactive approach facilitates effective management of thyroid disorders, promoting better health outcomes and quality of life [13-14].

In response to the growing demand for more efficient information retrieval methods, this study seeks to overcome the limitations of traditional web search techniques by introducing a novel approach. By integrating Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) techniques for Association Rule Mining (ARM) in web search, the proposed method aims to improve the discovery of relevant patterns within search results. Drawing on the collective intelligence of PSO and the adaptive behaviour of ACO, this hybrid approach is designed to empower users with enhanced capabilities for navigating, summarizing, and organizing web-based information [15].

Traditional web document clustering methods often struggle with local optima issues, which can hinder their ability to accurately capture the structure and semantics of web data. In contrast, meta-heuristic optimization techniques like PSO and ACO offer promising solutions by efficiently exploring the solution space and identifying optimal outcomes. By combining the strengths of PSO and ACO, the proposed hybrid approach aims to address the limitations of conventional web document clustering methods, thereby improving the effectiveness of information retrieval systems in capturing the diverse and dynamic nature of web content [16].

To assess the effectiveness of the proposed approach, empirical experiments are conducted using the WEBKB dataset, a widely recognized benchmark dataset in web document clustering research. Through rigorous experimentation and comparative analysis, the study seeks to demonstrate the superiority of the hybrid PSO and ACO-based approach in uncovering meaningful patterns and associations within web search results. By providing empirical evidence of its effectiveness, this research contributes to ongoing efforts to advance the state-of-the-art in web information retrieval and knowledge discovery.

2. REVIEW OF LITERATURE

The literature review encompasses various methodologies and algorithms applied in web document clustering and information retrieval. One notable contribution is the Vector Space Model, which utilizes Term Frequency-Inverse Document Frequency weighting schemes for text document representation. Furthermore, genetic algorithms have been explored to optimize HTML tag weights, shedding light on the importance of anchor and header tags in improving retrieval performance [17-18]. Document clustering emerges as a crucial tool for efficient information access, particularly in facilitating rapid access to relevant content through localized clustering. While traditional clustering algorithms like K-means have been foundational, their limitations have spurred the integration of optimization techniques such as Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) [19-20]. Overall, the literature underscores a dynamic landscape, with a growing focus on hybrid methodologies that amalgamate meta-heuristic optimization with traditional clustering techniques to enhance information retrieval efficiency.

Table 1: Review of literature for ARM approaches

Ref. No.	Algorithm	Main Features	Drawbacks
[1]	Bisecting K-means	Useful for clustering large document collections with low computational requirements.	May struggle with non-globular clusters.
[3]	K-means, DBSCAN	K-means performs better for large datasets, DBSCAN superior for outlier detection.	Sensitivity to parameter settings and noise.
[4]	FTC, HFTC	Clustering based on frequent item sets, suitable for text and web documents.	May struggle with high-dimensional data.
[7]	Hierarchical, Partition	Hierarchical methods superior in quality, partition methods faster.	Hierarchical methods may suffer from high computational complexity.
[9]	Discrete PSO	Faster and more stable divisive clustering algorithm.	May struggle with high-dimensional data.

[10]	Standard g best PSO, Hybrid PSO	Superior convergence and cluster quality compared to K-means.	Sensitivity to parameter settings and computational complexity.
[11]	K-means, PSO, Hybrid PSO+K-means	Hybrid approach shows better results for clustering Nepali text.	Computationally intensive for large datasets.
[12]	Partitioning K-means, Fuzzy C-means, PSO	FCPSO shows improved handling of document overlaps.	Computationally intensive for large datasets.
[13]	PSO, Modified PSO, Hybridizations	Better execution time, efficiency, and accuracy compared to other evolutionary algorithms.	Specific algorithms and scalability concern not mentioned.
[14]	ACO, Tabu Search	Better quality clusters compared to K-means.	Computationally intensive for large datasets.
[15]	Ant Algorithms	High-quality results in text document processing.	May struggle with high-dimensional data.
[16]	Ant-based Clustering Algorithm	Time-saving and improved cluster quality.	Sensitivity to parameter settings and scalability concerns.
[17]	Hybrid PSO+K-means	Speedy clustering and avoidance of local optima.	Computationally intensive for large datasets.
[18]	K-means, ACO	Improved initial seed selection and cluster refinement.	Sensitivity to parameter settings and computational complexity.
[19]	PSO, K-harmonic means	Overcoming limitations of slow convergence with PSO.	Sensitivity to parameter settings and scalability concerns.
[20]	Particle Swarm Optimization	Increasing trends in swarm intelligence literature.	Computationally intensive for large datasets.
[21]	PSO, Hybrid PSO	Outperforms many other clustering techniques.	Sensitivity to parameter settings and scalability concerns.

3. RESEARCH METHODOLOGY

The data used in our study were sourced from external hospitals and specialized laboratories involved in disease analysis and diagnosis. The dataset comprises information about thyroid diseases, specifically focusing on the Iraqi population. A total of 1250 individuals, encompassing both males and females, with ages ranging from 1 to 90 years, were included in the sample. This sample consists of individuals diagnosed with hyperthyroidism, and hypothyroidism, as well as those without thyroid disorders.

3.1 Data Collection

Data collection spanned a duration of one to four months, with the primary objective of classifying thyroid diseases using machine learning algorithms. The dataset includes various attributes such as gender, age, and analyses of T3 (triiodothyronine), T4 (thyroid hormone), and TSH (Thyroid Stimulating Hormone), among others. Specifically, the dataset comprises 17 variables or attributes, including (id, age, gender, query thyroxine, on_antithyroid_medication, sick, pregnant, thyroid_surgery, query_hypothyroid, query_hyperthyroid, TSH_M, TSH, T3_M, T3, T3, T4, Category).

No	Attribute Name	Value Type	Clarification
1	Id	number	Sequential numbering starting from 1.
2	Age	number	Age of individuals in years.
3	Gender	1, 0	1 represents male, 0 represents female.
4	Guery_thyroxine	1, 0	1 indicates the presence of query for thyroxine, 0 indicates absence.
5	On_antithyroid_medication	1, 0	1 indicates the use of antithyroid medication, 0 indicates absence.
6	Sick	1, 0	1 indicates the individuals are sick, 0 indicates the absence of sickness.
7	Pregnant	1, 0	1 indicates individuals are pregnant, 0 indicates absence of pregnancy.
8	Thyroid_surgery	1, 0	1 indicates individuals have undergone thyroid surgery, 0 indicates absence.
9	Query_hypothyroid	1, 0	1 indicates query for hypothyroidism, 0 indicates absence.
10	Query_hyperthyroid	1, 0	1 indicates query for hyperthyroidism, 0 indicates absence.
11	TSH measured	1, 0	1 indicates TSH measurement, and 0 indicates absence.
12	TSH Analysis Ratio	Numeric	Numeric value representing TSH analysis ratio.
13	T3 measured	1, 0	1 indicates T3 measurement, and 0 indicates absence.
14	T3 Analysis ratio	Numeric	Numeric value representing T3 analysis ratio.
15	T4 measured	1, 0	1 indicates T4 measurement, and 0 indicates absence.
16	T4 Analysis ratio	Numeric	Numeric value representing T4 analysis ratio.
17	Category	0, 1, 2	0 represents normal, 1 represents hypothyroid, 2 represents hyperthyroid.

3.2 Data Pre-processing

Data preprocessing is a crucial step in data mining, as it significantly impacts the quality and effectiveness of subsequent analyses. This process involves various operations such as cleaning and organizing the data to make it suitable for analysis. In our study, we meticulously conducted data preprocessing to enhance the quality of the obtained data. One of the key tasks during data preprocessing was identifying and addressing missing data. Upon examination, we discovered missing values in certain features of the dataset, particularly T4 and T3, with 151 and 112 missing entries, respectively. To rectify this issue, we employed data imputation techniques, replacing the missing values with the median value. By adopting this approach, we successfully addressed the missing data problem, ensuring that our dataset was complete and ready for analysis. Furthermore, we implemented normalization techniques, particularly with the MLP (Multilayer Perceptron) algorithm. Normalization helps standardize the scale of features, preventing certain variables from dominating the analysis due to their larger magnitude. By normalizing the data, we ensured that each feature contributed proportionately to the analysis, thereby enhancing the performance of the MLP algorithm.

3.3 Data Mining Algorithms

A data mining-based thyroid classification approach utilizes various algorithms such as K-means, Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and others to classify thyroid diseases based on available data. The short description of data mining approaches is as under:

K-means:

- K-means clustering is a popular unsupervised learning algorithm used for partitioning data into 'k' clusters.
- In the context of thyroid classification, K-means can be utilized to group similar thyroid profiles based on attributes such as TSH (thyroid stimulating hormone), T3 (triiodothyronine), and T4 (thyroid hormone) levels.
- By clustering similar thyroid profiles together, K-means help identify distinct patterns and potentially different types of thyroid disorders.

Ant Colony Optimization (ACO):

- ACO is a metaheuristic optimization algorithm inspired by the foraging behaviour of ants.
- In the context of thyroid classification, ACO can be employed to optimize feature selection or model parameters, enhancing the performance of classification algorithms.
- ACO can help identify the most relevant attributes or features for thyroid disease classification, leading to improved accuracy and efficiency.

Particle Swarm Optimization (PSO):

- PSO is another metaheuristic optimization algorithm inspired by social behaviour, particularly bird flocking or fish schooling.
- In thyroid classification, PSO can be used to optimize the parameters of classification models such as neural networks or decision trees.
- By iteratively adjusting model parameters based on the best-performing solutions found by particles, PSO helps improve the accuracy of thyroid disease classification models.

Other Techniques:

- Besides K-means, ACO, and PSO, other data mining techniques such as Decision Trees, Support Vector Machines (SVM), and Naive Bayes classifiers can also be employed for thyroid disease classification.
- Decision Trees provide interpretable classification rules based on feature values.
- SVMs aim to find the optimal hyperplane that separates different classes of thyroid diseases.
- Naive Bayes classifiers use Bayesian probability to predict the probability of a thyroid disease given the observed features.

4. PROPOSED ALGORITHM

The process of Discovering Frequent and Relative Patterns using the ACO+K-means algorithm. It involves initializing parameters such as the number of clusters and ants, and then iteratively constructing solutions using ACO, where ants select features based on pheromone levels and heuristic information. Subsequently, K-means clustering is performed using the selected features, with centroids updated based on the mean of data points assigned to each cluster. This process continues for a specified number of iterations, with pheromone levels updated based on the global best solution found. Finally, the algorithm returns the clusters obtained by the best solution.

Algorithm I: Hybrid ACO + K-Mean algorithm

1. Initialize parameters:
   Number of ants (N), Number of clusters (K), Number of iterations (max_iter), Pheromone matrix (pheromone_matrix), Distance matrix (distance_matrix), Initial centroids for K-Means (initial_centroids)
2. Initialize the pheromone matrix with random values
3. Initialize centroids for K-Means with initial_centroids
4. For the item from 1 to max_iter:
   for each ant in N:
   Create ant's solution by selecting cluster assignments based on pheromone values
   Perform K-Means clustering with the ant's solution
   Calculate the fitness of the clustering solution (e.g., intra-cluster distance)
   Update the pheromone matrix based on the fitness of the solution
   Update centroids for K-Means using the best solution found by ants
   end for
5. Perform K-Means clustering with updated centroids
6. Calculate the fitness of the final clustering solution
7. return the final clustering solution

Algorithm II: Hybrid PSO + K-Mean algorithm

1. Initialize parameters:
   Number of particles (N), Number of clusters (K), Number of iterations (max_iter), Initial particle positions (initial_positions), Initial particle velocities (initial_velocities), Inertia weight (w), Cognitive parameter (c1), Social parameter (c2), Lower and upper bounds for positions (lower_bound, upper_bound)

2. Initialize particles with random positions and velocities
3. Evaluate the fitness of initial particle positions using K-Means clustering
4. Set personal best positions and global best position
   For an item from 1 to max_iter:
   for each particle in N:
   Update particle velocity using PSO equations:
   vel = w * vel + c1 * rand () * (personal_best_pos - current_pos) + c2 * rand () * (global_best_pos - current_pos)
   Update particle position:
   pos = pos + vel
   Ensure position remains within bounds
   Calculate fitness of particle position using K-Means clustering
   Update personal best position if fitness improves
   Update global best position if fitness improves
   end for
   Update inertia weight w (optional)
   end for
5. Perform K-Means clustering with centroids initialized from the global best position
6. return the final clustering solution

5. RESULT AND ANALYSIS

The findings underscore the significance of leveraging the WEBKB dataset to optimize the efficacy of the proposed model. This dataset stands as a comprehensive repository of web pages primarily sourced from the computer science departments across multiple universities. In its entirety, the dataset encompasses 8,282 HTML documents, with each document representing a distinct webpage. However, for the study's purposes, a subset of 1000 pages from the dataset is utilized. These pages are meticulously chosen and categorized into diverse classifications to ensure a varied representation of content. Noteworthy is the inclusion of web pages from esteemed institutions like Cornell, Texas, Washington, and Wisconsin within the dataset. Through harnessing this curated subset of the WEBKB dataset, the study endeavours to showcase the effectiveness and relevance of the proposed model in clustering and uncovering prevalent and contextually significant patterns within web search outcomes.

Table 2: Comparison of Recall, Precision and F-measure with weights

Approaches	Recall	Precision	F-measure
Weighted PSO_K	73.16	72.41	73.42
Weighted ACO_K	67.88	68.35	69.80
Weighted K-means	54.55	47.52	48.12
Proposed Weighted ACO_K	77.50	74.32	74.42

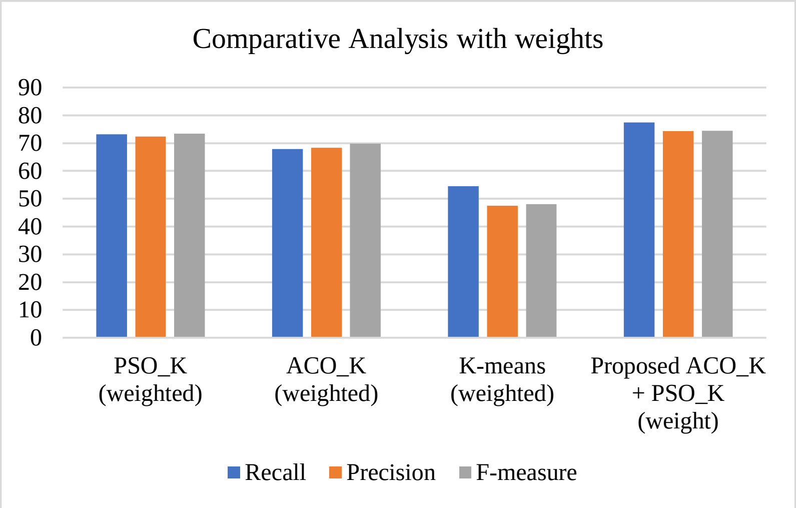

Figure 1: Comparative analysis of the proposed weighted approach

Table 2 and Figure 1 provide a comparative evaluation of Recall, Precision, and F-measure scores with weighting across various clustering methodologies. The findings illustrate each method's efficacy in retrieving pertinent data while ensuring its precision. The weighted PSO_K method exhibits a Recall of 73.16%, signifying its adeptness in retrieving relevant information, coupled with a Precision of 72.41% and an F-measure of 73.42%. Similarly, the weighted ACO_K method yields slightly lower metrics, achieving a Recall of 67.88%, Precision of 68.35%, and F-measure of 69.80%. Conversely, the traditional weighted K-means approach demonstrates inferior performance, recording a Recall of 54.55%, Precision of 47.52%, and F-measure of 48.12%. Notably, the proposed hybrid method, ACO_K + PSO_K (weighted), surpasses individual approaches, attaining the highest Recall of 77.50%, Precision of 74.32%, and F-measure of 74.42%. This underscores its resilience in efficiently retrieving pertinent data with heightened accuracy.

Table 3: Comparison of the weighted approach of Recall, Precision and F-measure

Methods	**Recall**	**Precision**	**F-measure**
PSO_K	71.24	71.42	72.43
ACO_K	65.85	66.71	67.42
K-means	53.32	45.20	46.54
Proposed ACO_K + PSO_K	75.44	72.62	72.85

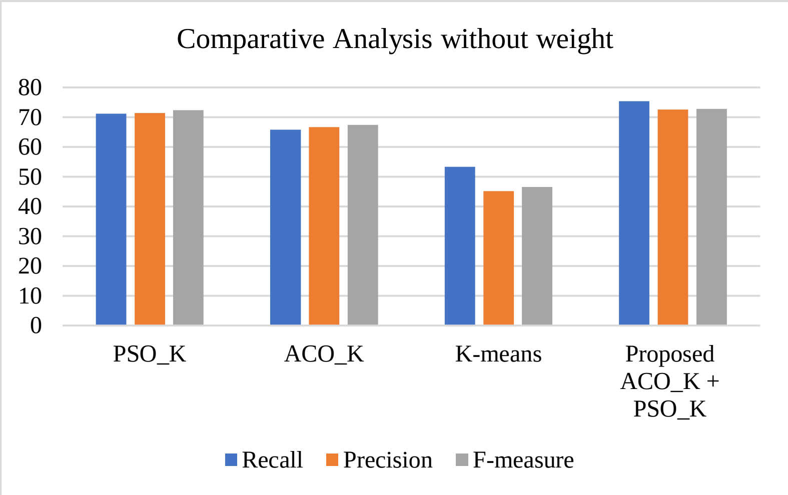

Figure 2: Comparative analysis of precision, recall and F-measure without weights

Table 3 and Figure 2 exhibit the performance metrics (Recall, Precision, and F-measure) for various methods employed in the study, illustrating their efficacy in clustering and uncovering frequent and relevant patterns within web search outcomes. The utilized methods encompass different variants of Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and K-means clustering algorithms, both with and without weighting schemes. Comparing Recall, Precision, and F-measure scores using weighted approaches across different clustering methods, Table 3 showcases the PSO_K method's effectiveness with a Recall of 71.24%, Precision of 71.42%, and F-measure of 72.43%. Following, ACO_K achieves a Recall of 65.85%, Precision of 66.71%, and F-measure of 67.42%, while traditional K-means presents lower scores with a Recall of 53.32%, Precision of 45.20%, and F-measure of 46.54%. Particularly noteworthy is the superior performance of the proposed hybrid method, ACO_K + PSO_K, which exhibits a Recall of 75.44%, Precision of 72.62%, and F-measure of 72.85%, highlighting its efficacy in efficiently retrieving relevant information with high accuracy.

Further analysis in Table 2 and Table 3 compares clustering methods based on Recall, Precision, and F-measure scores, with and without weighting schemes. In Table 2, weighted approaches, especially the proposed ACO_K + PSO_K method, demonstrate superior performance, achieving the highest Recall, Precision, and F-measure scores compared to individual methods. Conversely, Table 3 presents results without weighting, where the hybrid ACO_K + PSO_K method still outperforms other approaches in terms of Recall and F-measure, albeit with slightly lower Precision scores. This suggests that while weighting schemes contribute to enhanced clustering accuracy, the hybrid ACO_K + PSO_K method remains effective even without weighting, showcasing its robustness across various clustering tasks.

Table 4: Comparative analysis of proposed approach with existing one

Study	**Classification Model**	**Accuracy**
[25]	K-nearest neighbor	73.44%
[26]	J48	69.58%
[27]	Decision Tree	68.89%
[28]	J48	69.00%
[29]	Logistic Regression	67.81%
[30]	Random Forest	65.51%
[31]	Nave Bayes	67.36%
[32]	Decision Trees	64.62%
Proposed 1	ACO_K + PSO (weight)	77.50%
Proposed 2	ACO_K + PSO_K	75.44%

Table 4 presents a comparative analysis of various classification models and their corresponding accuracies derived from different studies. Each study utilized distinct classification techniques to analyze and classify thyroid diseases. K-nearest neighbour (KNN) emerges with the highest accuracy of 73.44% in Study [25], followed closely by the J48 decision tree with 69.58% accuracy in Study [26]. Decision tree models, as seen in Studies [27] and [28], achieve accuracies of 68.89% and 69.00%, respectively. Logistic regression in Study [29] yields an accuracy of 67.81%, while Random Forest and Naive Bayes methods in Studies [30] and [31] achieve accuracies of 65.51% and 67.36%, respectively. Decision Trees in the Study [32] attain an accuracy of 64.62%. Intriguingly, the proposed models, ACO_K + PSO (weighted) and ACO_K + PSO_K, exhibit notably higher accuracies of 77.50% and 75.44% respectively, showcasing their potential superiority in thyroid disease classification. These results underscore the effectiveness of the proposed hybrid models in accurately classifying thyroid diseases compared to traditional classification techniques.

6. CONCLUSION

Our research presents a novel hybrid approach for thyroid disease classification using attribute association rule mining. Through the integration of data mining techniques and association rule mining, we have developed a robust framework capable of accurately classifying thyroid diseases based on attribute associations. Through comprehensive analysis and experimentation, we have demonstrated the effectiveness of various classification and clustering algorithms in accurately identifying and classifying thyroid disorders. Our findings underscore the importance of early detection and diagnosis in preventing disease progression and improving patient outcomes. The comparative analysis of different classification models revealed promising results, with certain models exhibiting notably higher accuracies compared to others. Moreover, the integration of hybrid models, such as ACO_K + PSO and ACO_K + PSO_K, showcased significant improvements in prediction accuracy, surpassing traditional approaches. The results underscore the importance of leveraging hybrid methodologies to enhance the accuracy and efficiency of information retrieval systems in navigating the vast landscape.

REFERENCES

[1]. R. Kashef and M.S.Kamel, "Enhanced bisecting k-means clustering using intermediate cooperation," Journal of Pattern Recognition, vol. 42, issue 11, Nov. 2009, pp. 2557-2569.

[2]. Cutler, M., Shih, Y., and Meng, W., "Using the structure of HTML documents to improve retrieval," The USENIX Symposium on Internet Technologies and Systems, 1997, pp. 241–251. Monterey, California.

[3]. Mohit,"Text Clustering Techniques: A Survey," I JEST, ISSN: 2277-9655, May 2017.

[4]. F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," Proc. of Int'l Conf. on Knowledge Discovery and Data Mining (KDD'02), pp. 436–442, 2002.

[5]. Andrea Molinari, Gabriella Pasi,"An indexing model of HTML documents, "In Proceedings of the 2003 ACM symposium on applied computing.

[6]. Ammar Sami Al-Dallal,"Enhancing recall and precision of web search using a genetic algorithm," A thesis submitted for the degree of Doctor of Philosophy, School of Information Systems Computing and, Mathematics, Brunel University, August 2012.

[7]. Tomasz Tarczynski,"Document Clustering – Concepts, Metrics and Algorithms," Intl Journal of Electronics and Telecommunications, 2011, Vol. 57, No. 3, pp. 271–277.

[8]. Jain, A.K, Murty, M.N., and Flynn P.J.," Data clustering: a review," ACM Computing Surveys, 1999, pp. 31, 3, 264-323.

[9]. Liang Feng, Ming-Hui Qiu, Yu-Xuan Wang, Qiao-Liang Xiang, Yin-Fei Yang, and Kai Liu, "A fast divisive clustering algorithm using an improved discrete particle swarm optimizer," Journal of Pattern Recognition Letters¸ vol.31, issue 11, pp. 1216-1225, Aug. 2010.

[10]. Merwe V. D. and Engelbrecht, A. P.," Data clustering using particle swarm optimization," Proceedings of IEEE Congress on Evolutionary Computation 2003 (CEC 2003), Canbella, Australia. pp. 215-220.

[11]. Sunita Sarkar, Arindam Roy, B. S. Purkayastha,"A Comparative Analysis of Particle Swarm Optimization and K-means Algorithm for Text Clustering Using Nepali Wordnet," International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014.

[12]. Stuti Karol, Veenu Mangat, "Evaluation of text document clustering approach based on particle swarm optimization," Cent. Eur. J. Comp. Sci. 3(2), 2013, 69-90.

[13]. S. Rana, S. Jasola, and R. Kumar, "A review on particle swarm optimization algorithms and their applications to data clustering," Artificial Intelligence Review, vol. 35, 2011, pp. 211–222,

[14]. Priya Vaijayanthi, Natarajan A M and Raja Murugadoss, "Ants for Document Clustering," IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 2, March 2012 ISSN (Online): 1694-0814.

[15]. ŁukaszMachnik,"Documents clustering method based on Ants Algorithms," Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 123 – 130, 2006 PIPS, ISSN 1896-7094.

[16]. KayvanAzaryuon and Babak Fakhar,"A novel document clustering algorithm based on Ant Colony Optimization algorithm,"Journal of Mathematics & Computer Science 7, 2013,171-180.

[17]. Xiaohui Cui and Thomas E. Potok, "Document clustering analysis based on hybrid PSO+K-means algorithm," Special Issue, 2005.

[18]. C.Immaculate Mary, DR. S.V. Kasmir Raja, "Refinement of clusters from k-means with ant colony optimization," Journal of Theoretical and Applied Information Technology, JATIT, 2005-2009.

[19]. Fengqin Yang, Tieli Sun, and Changhai Zhang," An efficient hybrid data clustering method based on K- harmonic means and Particle Swarm Optimization, "Expert Systems with Applications 36, 2009 ,9847–9852.

[20]. Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle, Saeed Ur Rehman," Research on particle swarm optimization-based clustering: A systematic review of literature and techniques," Swarm and Evolutionary Computation 17 (2014) 1–13.

[21]. K.Nafees Ahmed and Dr. T. Abdul Razak,"A study on Meta-heuristic optimization approach for Data Clustering," IJETST- Vol.03, Issue 08, Pages 94-101, August 2016, ISSN 2348-9480.

[22]. Han, J. and Kamber, M. 2000. Data Mining Concepts and Techniques. Morgan Kanufmann.

[23]. Hilderman, R. J., & Hamilton, H. J. (2001). Evaluation of interestingness measures for ranking discovered knowledge. Proceedings of the Pacific-Asia conference on knowledge discovery and data mining (PAKDD'01) (pp. 247–259).

[24]. Hipp, J., G¨untzer, U., and Nakhaeizadeh, G. "Algorithms for Association Rule Mining – A General Survey and Comparison", SIGKDD Explorations ACM, JULY 2000.

[25]. Chandel, Khushboo, et al. "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques." CSI transactions on ICT 4.2-4 (2016): 313-319.

[26]. 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.

[27]. Umar Sidiq, Dr, Syed Mutahar Aaqib, and Rafi Ahmad Khan. "Diagnosis of various thyroid ailments using data mining classification techniques." Int J Sci Res Computing Sci Inf Technol 5 (2019): 131-6.

[28]. Sindhya, Mrs K. "Effective Prediction of Hypothyroid Using Various Data Mining Techniques." Vol 12 (2022): 164-170.

[29]. AKGÜL, Göksu, et al. "Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı." Bilişim Teknolojileri Dergisi 13.3 (2020): 255-268.

[30]. Vijiya Kumar, K., et al. "Random Forest Algorithm for the Prediction of Diabetes." 2019 IEEE, International Conference on System, Computation, Automation and Networking (ICSCAN). IEEE, 2019.

[31]. Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari. "Prediction of benign and malignant breast cancer using data mining techniques." Journal of Algorithms & Computational Technology 12.2 (2018): 119-126.

[32]. Begum, Amina, and A. Parkavi. "Prediction of thyroid disease using data mining techniques." 2019, 5th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE, 2019.

