Hypothyroid Analysis with Emcluster using Frequent Pattern Mining Techniques

[1]Muni Sankar Revilla, [2]Naveen Kumar, [3]Anima Pradhan

[1]Associate Professor in Dept. of Computer Science, GIET Gangapatna, Bhubaneswar
[2]Associate Professor in Dept. of Computer Science, GIET Gangapatna, Bhubaneswar
[3]Associate Professor in Dept. of Computer Science, GIET Gangapatna, Bhubaneswar

Abstract: Data mining is a process that involves identifying sequential patterns within large datasets. In the data analysis process, cluster classification analysis and machine intelligence techniques are commonly employed. Data mining plays a crucial role in enhancing the analysis of medical bioinformatics data. Classification algorithms are used to predict outcomes, while association analysis helps in identifying rules associated with items that frequently co-occur. The Weka software is a powerful tool that encompasses data pre-processing tools, classification and regression algorithms, clustering algorithms, and association rule mining algorithms. It also includes attribute and subset evaluation methods for feature selection. Weka supports multiple platforms and is written in Java, making it widely accessible and versatile. It allows users to define filters for transforming data through processes such as discretization, normalization, resampling, and attribute selection. In the context of bioinformatics, gene expression analysis can be performed using Weka to predict the accuracy of frequent pattern mining algorithms in diagnosing conditions such as hypothyroidism. By leveraging the capabilities of Weka and data mining techniques, researchers can gain valuable insights and make informed decisions based on complex biological data.

Keywords: Bioinformatics, Co-occurrences, Gene expression, Sequential patterns.

Introduction

Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases such as association rules, correlations, sequences, classifiers, clusters and many more of which the mining of association rules is one of the most popular problems. Frequent Pattern Mining (FPM) is used for finding relationships among the items in a large database obtained from the cloud environment. Frequent itemset mining is a study in data analysis techniques for bioinformatics. Satish kumar david Amr, T.M.Saeb(2013) ,suggested that data mining "a major way of creating knowledge", is a useful way in the field of medicine, genetics, bioinformatics. Guerra L, McGarry M, Robles V, Bielza C, Larrañaga P, Yuste R (2011) in their paper they explained the classification of techniques as unsupervised and supervised learning techniques.

In weka software we have filters as "Weka.filters.supervised" that is the Classes below weka.filters.supervised in the class hierarchy are for supervised filtering, i.e. taking advantage of the class information. A class must be assigned. Then the "Weka.filters.supervised" in this class hierarchy are for unsupervised filtering, e.g. the non-stratified version of Resample. A class should not be assigned here. The present study focus on the clustering and classification techniques of the frequent patterns in the gene database. Data clustering is the task of discovering groups of objects ina data set that exhibit high similarity. The goal of classification is predicting the target class accuracyfor the case in data. This paper analyzed the algorithms such as naive Bayes, Decision Tree, and REPTree as shown in Table 1.

Data Mining Software - WEKA

WEKA software is a useful tool for data mining tasks suggested by Satish kumar david Amr, T.M.Saeb (2013). It is used in environment for Knowledge Analysis. Weka means data mining /machine learning tool developed by Computer Science department, University of Waikato, NewZealand. It is the set of machine learning algorithms applied directly to a dataset for data mining tasks as, data pre-processing, classification, regression, clustering, association rules, andvisualization. Weka is open source software issued under the General Public License (GNU). Satishkumar david Amr, T.M.Saeb(2013) in their paper they have listed the WEKA applications as probe selection of gene expression arrays ,automated protein data annotation, automatic cancer diagnosis, plant genotype discrimination, classifying gene expression profiles and computational model forframe- shifting sites and

extracting rules from them. The main features are data pre-processing tools,learning algorithms and evaluation methods, Graphical user interfaces (including data visualization), Environment for comparing learning algorithms.

A. Weka Interface

Weka GUI has four main Interface Explorer, Experimenter, Knowledge Flow and Command Line Interface. The Explorer (exploratory data analysis) performs the operations of data pre- processing, classification, and clustering, association rule mining, attribute selection, data visualization technique. The Experimenter (experimental environment) helps to compare the performance of different learning schemes. The Knowledge Flow model is used for setting up and running machine learning experiments. The Command line Interface is the Simple CLI helps to communicate the name of the training file to the learning algorithms.

Table No.1: Comparison Of Frequent Pattern Mining Techniques

Algorithms	Inferences	Working Principles	Phases	Metrics	Drawbacks	Appln.
NaiveBayes	pattern recognition communities [2]	• supervised learning • Use initialset of objects with known class memberships. • It will construct core. • Large score with class 1 object. Smaller score with class of objects [2]	Training phase Testing phase classifier decision	Flexible modification can be easily incorporated	Modifications detract from its simplicity [2]	• Text classification Spam Filtering
Decision Tree	Straight Forward rules based on if-then- else condition to classify data items	• Recursive partition of dataset using depth-first search structure made up of root, internal nodes, leaf nodes [3]	Tree building Tree Pruning [3]	• work with continuous attributes avoid over fitting used to generate frequency tables no domain knowledge is required to construct. handle high dimensional data	• wok with missing values the output attribute should be categorized limited to one output attribute [3]	Bioinformatics, Gene Expression, AI [4] [5]
REPTree	Regression Tee Selects the best tee from the generated trees	It builds the tree based oninformation gain technique [6], apply reduce error pruning [6]	• Build Regression tree apply splitting criterion	• Fast decision and regression tree[6]	• Sort numeric attributes [6]	Feature selection, Prediction [6]

Weka Explorer

The WEKA Preprocess functionality is used to choose and modify the data being acted on. The Classification technique is applied to train and test learning schemes that classify or perform regression. The Clustering technique helps to learn clusters for the data. The Association rule means to select the most relevant attributes in the data. The Visualize tab is applied for viewing aninteractive 2D plot of the data. Weka uses flat text files data format to describe the data. It can be imported from a file support various formats such as ARFF (Attribute-Relation File Format (ARFF)), CSV (comma-separated values), and C4.5, binary. It can also be read from a URL or from a SQL database (using JDBC) as shown in Figure 1.

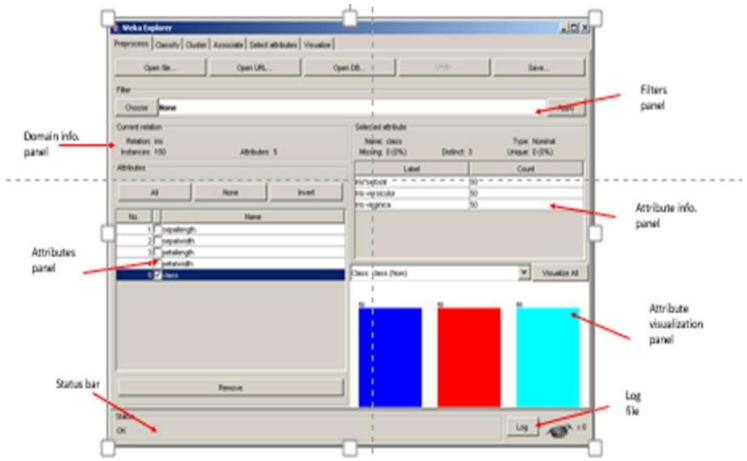

Fig. 1: WEKA Explorer

The significance of Weka are Relation, as given in the file to be loaded. Filters will modify the name of a relation. An instance is the number data points / records / rows in the data. Attributesis the number of attributes (features/ columns) in the data. Name is the name of the attribute, thesame as that given in the attribute list. Type is the type of attribute, most commonly Nominal or Numeric. Missing is the number (and percentage) of instances in the data for which this attribute is missing (unspecified). Distinct refers to different values that the data contains for this attribute. Unique is the number (and percentage) of instances in the data having a value for this attribute that no other

instances have. The attribute types are data type can be any of the four types current they are Nominal is one of a predefined list of values − e.g. red, green, blue. Numeric is a real or integer number. String is whatever enclosed in double quotes and Date If the attribute is nominal, the list consists of each possible value for the attribute along with the number of instances that have that value. If numeric attribute is in the list it has four statistics for data distribution the values are, minimum, maximum, mean and standard deviation.

Bioinformatics Dataset And Weka

The dataset "hypothroid.arff" dataset generate gene data for the thyroid disease sample. From the analysis the evaluation is to be made as to whether the patient suffering from secondary hypothyroid, primary hypothyroid, compensated hypothyroid or negative result. This dataset contained 3772 instances and 30 attributes. This dataset was analyzed using WEKA data mining software utilizing naive Bayes, Decision Tree, REPTree techniques as shown in Figure 2. Preprocessing techniques such as Attribute Selection, Class Order, and Replace Missing Values has been applied to improve the efficiency of the algorithms.

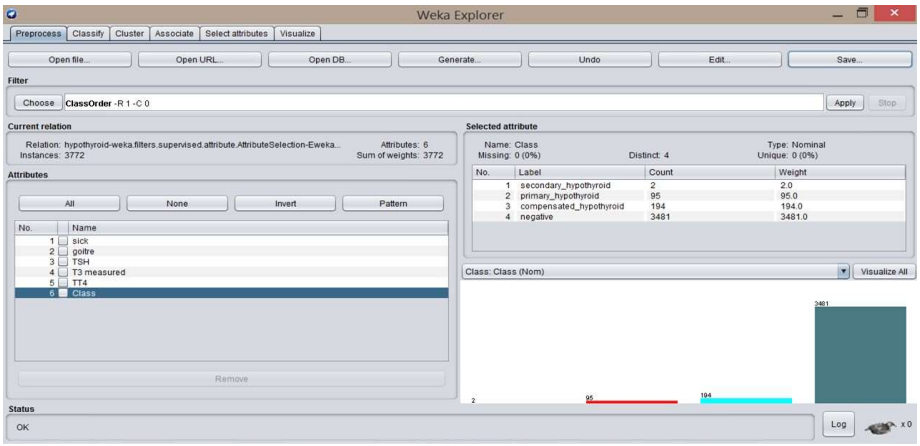

Fig. 2: Weka with Preprocessing techniques

Experimental Results

In WEKA a comparative analysis of classification algorithms such as naive Bayes, Decision Tree, REPTree techniques is done in this paper. The dataset "hypothroid.arff" is used to test the accuracy and time taken by these algorithms. This dataset has 3772 instances and 30 attributes. As a preprocessing work for this dataset we have applied three techniques (1) Attribute Selection (2) Class Order (3) ReplaceMissingValues as shown in Figure 3.

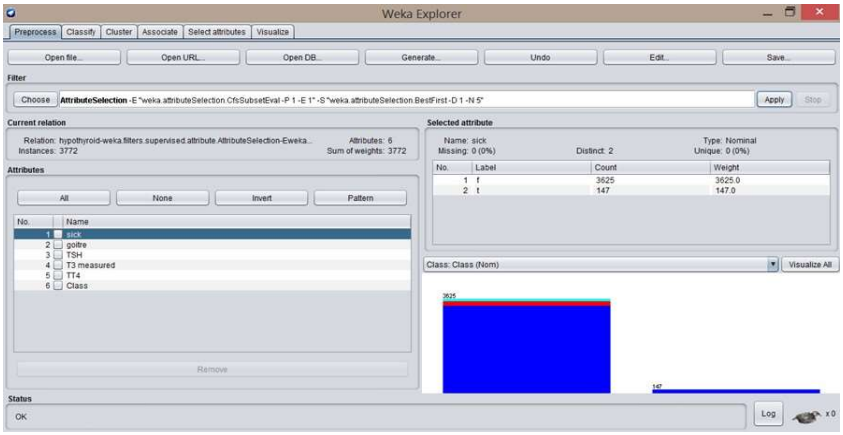

Fig. 3: Attribute Selection

A. **Attribute Selection:** As shown in Figure 4. It is a supervised filter to select attributes. It allows various search and evaluation methods to be combined. It belongs to filters. supervised. Attribute Selection. After applying this filter the dataset has 3772 instances and 6 attributes which make our analysis more efficient.

B. **Class Order:** This filter will change the order of the class according to the class frequency.

C. **Replace Missing Values:** Replaces all missing values for nominal and numeric attributes with mode and mean from the training data. The attributes TSH and TT4 has 369 and 231 missing data and after applying this filter it has been replaced with training data. This filter belongs to filters. Unsupervised. Replace Missing Values. For analysis and evaluation dataset contains incorrect classified instances, correct classified instances, Kappa statistics mean, Mean absolute Error, root mean squared error is considered. The dataset is analyzed with 10 folds cross validation. It computes parameters with given instances by applying these algorithms. Table2 shows the highest accuracy is 78.9%.

D.

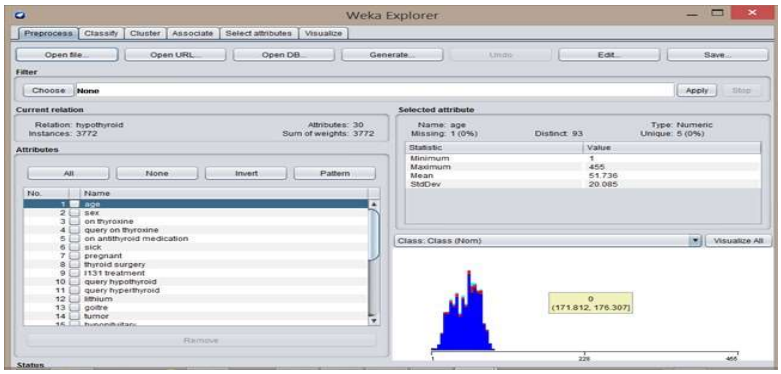

Fig. 4: Hypothyroid dataset loaded in WEKA Explorer

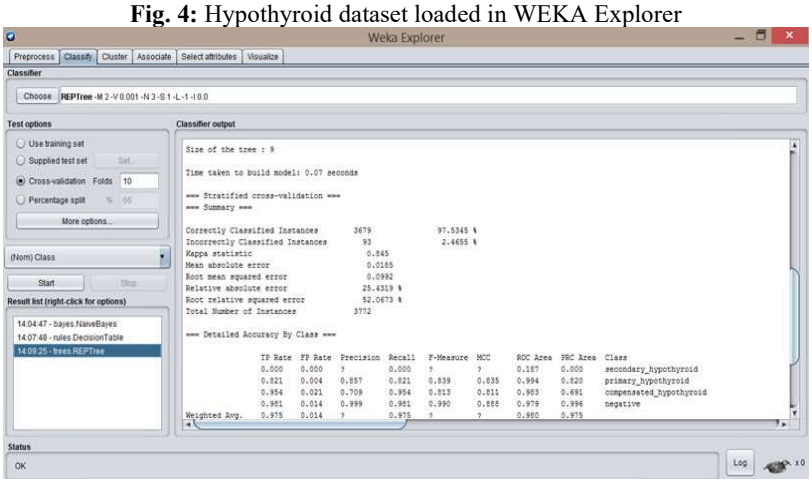

Fig. 5: Result generated

Table 2: Analysis Result of Algorithms

Classifier	Algorithm Implemented	correct classified instances (%)	Incorrect classified	Time Taken (seconds)	Kappa statistics	Mean absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative squared Error (%)
Bayes	NaiveBayes	94.591	5.408	0.08	0.523	0.033	0.150	46.292	78.986
Rules	Decision Tree	97.561	2.439	0.06	0.845	0.022	0.098	31.181	51.7193
Trees	REPTre	97.534	2.465	0.07	0.845	0.018	0.099	25.431	52.0673

Total Instances: 3772

Data Analysis In Gene Expression

EMCluster is applied to the dataset and the results are shown in table 3 accordingly based on the attribute values. Time taken to build the model: 250.51 seconds

Table 3: Performance Analysis Result

	Compensated hypothyroid (0.25)	Negative (0.34)	Primary hypothyroid (0.05)	Secondary hypothyroid (0.15)
Sick	47.4209	59.9928	1.2552	16.0803
Goitre	13.5993	7.2522	1.0181	5.2072
TSH	4.9546	1.7015	61.645	0.0982
T3	522.9159	1175.7286	158.7736	470.8151
TT4	99.1488	103.6905	63.2239	136.1847

Table 4: Category wise Analysis Result

S. No	Category	Count
1	Secondary hypothyroid	2
2	Primary hypothyroid	95
3	Compensated hypothyroid	194
4	Negative	3481

Conclusion

In this paper, various algorithms were compared and evaluated based on their accuracy, error rate, and the time required to build the model. Key metrics such as the relationship between incorrect and correct classified instances, Kappa statistics mean, Mean Absolute Error, and Root Mean Squared Error were considered during the analysis.The experiments revealed that the Decision Tree algorithm exhibited the lowest error rate at 2.439%, with an accuracy level of 97.561% and a model building time of 0.06 seconds, outperforming the other two algorithms under comparison. Based on these findings, it can be concluded that the Decision Tree algorithm offers superior classification accuracy compared to the alternative algorithms assessed in the study.

References

[1]. Irina Ionina, Liviu Ionina, Informatics, Computer Science, Mathematics and Physics, Petroleum-Gas University of Ploiesti, Ploiesti, Romania, "Prediction of Thyroid Disease Using Data Mining Techniques" Published in: Research Gate, August 2016.

[2]. Xindong Wu, Vipin kumar, "Top 10 algorithms in Data Mining", Published in: IEEE Conference on Data Mining (ICDM) Dec 2007, Spinger-Verlag London 2007.

[3]. Satish Kumar David, Amr T.M. Saeb, Department of Information Technology, Diabetes Strategic Research Center, King Saud University, "Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics", published in: Computer Engineering and Intelligent Systems ,ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online),Vol.4,2013

[4]. Farhad Soleimanian Gharehchopogh, Peyman Mohammadi, Department of Computer Engineering, Science and Research Branch, Islamic Azad University, West Azerbaijan, Iran, "Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study", Published in: International Journal of Computer Applications (0975 – 8887) Volume 52 – No. 6, August 2012.

[5]. Sushilkumar Kalmegh, Associate Professor, Department of Computer Science, Sant Gadge Baba Amravati University Amravati, Maharashtra,"Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and Random Tree for Classification of Indian News ",Published in: IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015.

[6]. Satish Kumar David, Amr T.M. Saeb, Khalid Al Rubeaan, Department of Information Technology, Diabetes Strategic Research Center, King Saud University, P.O, "Comparative

[7]. Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics", Published in: Computer Engineering and Intelligent Systems www.iiste.org ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) Vol.4, No.13, 2013.

[8]. Tobler JB, Molla MN, Nuwaysir EF, Green RD, Shavlik JW. "Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. Bioinformatics", Published in:

Bioinformatics, Volume 18, Issue suppl_1, July 2002, Pages S164–S171, (2002).

[9]. Guerra L, McGarry M, Robles V, Bielza C, Larrañaga P, Yuste R, "Comparison between supervised and unsupervised classifications of neuronal cell types:A case study", Published in : Developmental neurobiology , 71 (1): 71-82, (2011).

[10]. Sharon Christa, K. Lakshmi Madhuri, Post Graduate Programme, Dept. of ISE, Dayananda Sagar College of Engineering, Bangalore, India, "A Comparative Analysis of Data Mining Tools in Agent Based Systems", Published in: ICSCI, (2012).

[11]. Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004), "Data mining in bioinformatics using Weka, Published in: Bioinformatics (oxford, England), 20(15), 2479 – 2481. doi: 10.1093/bioinformatics/bth261

[12]. Web References

[13]. Comparison-Summary-http://voyagememoirs.com/pharmine/2008/05/18/summary.

[14]. 10 algorithms - https://www.slideshare.net/gnap/10-algorithms-in-data-mining

[15]. Decision tree - https://www.saedsayad.com/decision_tree.htm

[16]. Difference bet. Seq. and frequent mining-https://www.quora.com/Is-sequential-pattern-mining-the-same-as-frequent-pattern-mining

[17]. FPM with Cspade - https://www.youtube.com/watch?reload=9&v=uxuWCwSLjMc

[18]. Seq. pattern mining - https://www.youtube.com/watch?v=GhEteXWNIXc

[19]. Sequence databases in bioinformatics weka - https://www.youtube.com/watch?v=hCUVfqHiCeE

[20]. Dataset-https://www.kaggle.com/kevinarvai/clinvar-conflicting#clinvar_conflicting.csv