International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES),(UGC APPROVED)
Impact Factor: 5.22 (SJIF-2017),e-ISSN:2455-2585
Research Symposium on
"Advancements in Engineering, Science, Management, and Technology"
Volume 5, Special Issue 04, April-2019.

COMPARATIVE STUDY OF VARIOUS CLASSIFIERS ON HYPOTHYROID DATA USING WEKA

Rashmee Shrestha[1], *Rahisha Pokharel*[2], *Zubair Salarzai*[3], *Pooja*[4]

[1]*M.Tech Scholar, CSE, School of Engineering and Technology, Sharda University,*
[2]*M.Tech Scholar, CSE, School of Engineering and Technology, Sharda University,*
[3]*M.Tech Scholar, CSE, School of Engineering and Technology, Sharda University,*
[4]*Associate Professor, CSE, School of Engineering and Technology, Sharda University,*

Abstract— ***In this paper, WEKA tool has been used to evaluate the performance of various classifiers on a dataset namely, hypothyroid, to come out with the optimum classifier, for a particular application. Hypothyroid is an imbalance dataset which contains 28 attributes and 3772 instances. A Classifier plays an important role in any machine learning application. There are various performance analysis measures that can be used to evaluate the efficiency of a classifier. In this paper, Naive Bayes, J48, IBK, Vote, Logistic and Random Forest classifiers along with their combination have been implemented and analysed using WEKA. Accuracy of individual classifiers along with the accuracy obtained while using the combination of these classifiers have been measured and evaluated. Use of these hybrid approach helped in combining different classifiers to get the best results.***

Keywords— WEKA, Machine Learning, Classifiers, Supervised, Hypothyroid;

I. Introduction

Classification is the method to organize the data in the efficient and effective way so that it can be used with ease. It is easy to retrieve the data when it is arranged with proper classification. The concept of classification includes Unsupervised, Semi-Supervised and Supervised learning problems. Unsupervised learning relies upon the unlabeled information while in Supervised learning every data input question is assigned a class mark. Unsupervised learning depends on the unlabeled data whereas in Supervised learning each data input object is assigned a class label. In Semi-Supervised learning problems, both labeled and unlabeled instances are available and might be of absolute significance for computation of more strong decision functions in some situations. The main objective of supervised classification is to divide the classes as wide as possible. If the variable has two values, it is known as binary classification, but if the variable has more than two values it is known as multiclass classification. [1]

In this paper, data mining and machine learning tool WEKA is used for classification of data. WEKA is the data mining tool which consists of various algorithms for preprocessing, classification, clustering etc. It is funded by the New Zealand government from 1993. It is an open source and Java based software. It is used in both academic and business field. It not only provides a toolbox for already generated algorithms but also provides the platform to build new algorithms.

Here in this paper, we have focused on the performance evaluation of the various classifiers using WEKA.

II. Classification Methods

1. Naive Bayes' Classifier

Naïve Bayes' Classifier is a simple probabilistic classifier that works based on Baye's theorem which has independent features. These are highly scalable classifiers. Here, maximum likelihood training can be done by evaluating a closed form expression, which takes linear time rather than by expensive iterative approximation as used by many other classifiers. [2]

2. J48 Classifier

This is an extension of ID3 (Iterative Dichotomiser 3) having special features for missing values, decision tree pruning, continuous attribute value ranges, derivation of rules etc. In the WEKA data mining tool, J48 is an open source java implementation of the C4.5 algorithm. [3]

3. **Instance Based Learner (IBK) Classifier**
Instance Based Learner (IBK) algorithm uses distance measure to locate close instances for the training data for each test instance and uses those selected instances to make prediction. It does not perform any generalization instead they compare new problem instances with the instances seen in the training which is basically stored in the memory. [6]

4. **Vote Classifier**
Vote classifier is an ensemble vote classifier which implements hard and soft voting. In hard voting, final class label is predicted as the final class label which has been predicted most frequently by the classification models. In soft voting, the class label is predicted by taking average of the class probabilities i.e. only recommended if the classifiers are well-calibrated. [6]

5. **Logistic Classifier**

Logistic algorithm is the basic algorithm to solve classification problems. It is a statistical learning technique in supervised machine learning methods. Binary logistic model has a dependent variable with two possible values labelled as 0 and 1. In the logistic model, the log-odds for the value labelled 1 is a linear combination one or more independent variables. These independent variables can be a binary variable or a continuous variable. [5]

6. **Random Forest**

Random forest is also known as random decision forests. These are popular ensemble method that can be used for predictive models for both classification and regression problems. Its main focus is to reduce correlation issues by choosing only subsample of the feature space. It aims to make the trees de-correlated and prune the trees by setting a stopping criteria for node splits. [4]

I. Results

Dataset Name: Hypothyroid

Data description: This is an imbalance dataset which contains 28 attributes and 3772 instances. This dataset is of four classes and its characteristics are multivariate, it has categorical, inter and real numbers attribute characteristics.

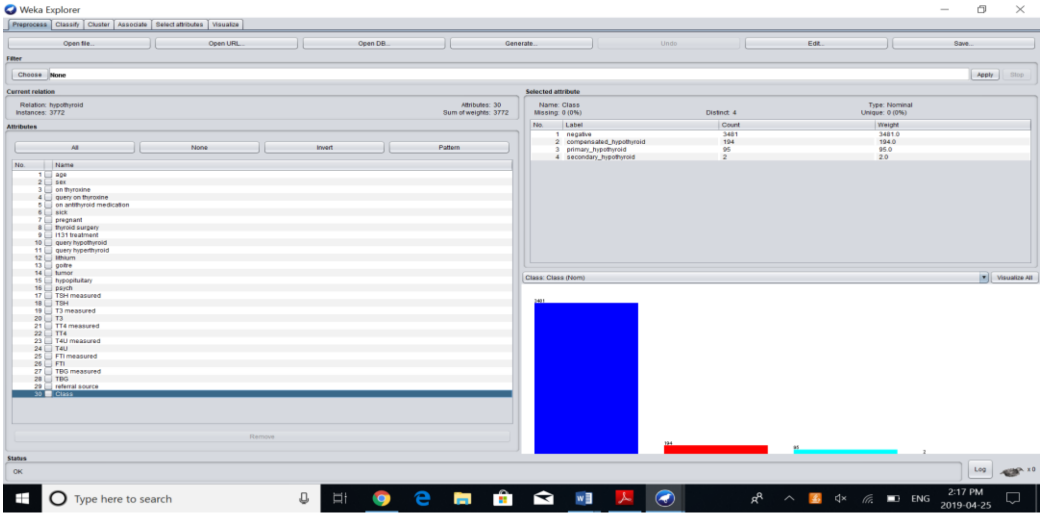
Tool: WEKA

TABLE 1
DATASET ATTRIBUTE AND VALUE TYPE

S.No.	Attribute Name	Value type
1	Age	Continuous
2	Sex	Male (M) / Female (F)
3	on_thyroxine	False (F) /True (T)
4	query_on_thyroxine	False (F) /True (T)
5	on_antithyroid_medication	False (F) /True (T)
6	Sick	False (F) /True (T)
7	Pregnant	False (F) /True (T)
8	thyroid_surgery	False (F) /True (T)
9	I131treatment	False (F) /True (T)
10	query_hypothyroid	False (F) /True (T)
11	query_hyperthyroid	False (F) /True (T)
10	Hypopituitary	False (F) /True (T)
12	Tumor	False (F) /True (T)
13	Lithium	False (F) /True (T)
14	Goitre	False (F) /True (T)
15	TSH_measured	False (F) /True (T)
16	TSH	Continuous
17	T3_measured	False (F) /True (T)
18	T3	Continuous
19	TT4_measured	False (F) /True (T)
20	TT4	Continuous
21	T4U_measured	False (F) /True (T)
22	T4U	Continuous
23	FTI_measured	False (F) /True (T)

24	FTI	Continuous
25	TBG_measured	False (F) /True (T)
26	TBG	Continuous
27	referral source	WEST, STMW, SVHC, SVI, SVHD, other.
28	Psych	False (F) /True (T)

Table 1 represents various attribute or features of the dataset. Data points are labelled into four classes which are :- primary hypothyroid, compensated hypothyroid, secondary hypothyroid and negative. Figure 1 displays all the attribute and classes. It is evident that the dataset carries imbalanced classes.



Relation: hypothyroid
Instances: 3772
Attributes: 30
Sum of weights: 3772

No.	Name
1	age
2	sex
3	on thyroxine
4	query on thyroxine
5	on antithyroid medication
6	sick
7	pregnant
8	thyroid surgery
9	I131 treatment
10	query hypothyroid
11	query hyperthyroid
12	lithium
13	goitre
14	tumor
15	hypopituitary
16	psych
17	TSH measured
18	TSH
19	T3 measured
20	T3
21	TT4 measured
22	TT4
23	T4U measured
24	T4U
25	FTI measured
26	FTI
27	TBG measured
28	TBG
29	referral source
30	Class

Name: Class
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	negative	3481	3481.0
2	compensated_hypothyroid	194	194.0
3	primary_hypothyroid	95	95.0
4	secondary_hypothyroid	2	2.0

OK

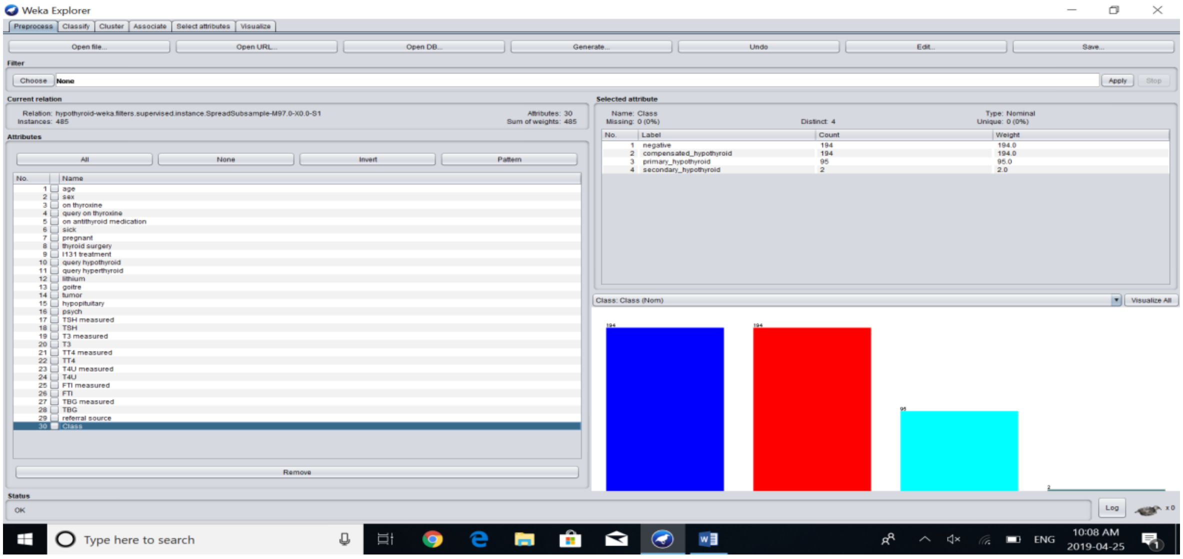
Fig1. Original Data With Imbalanced Classes

TABLE 2
PERFORMANCE OF EACH CLASSIFIERS ON TRAINING DATASET

Parameters	**Classifiers**					
	NB	**J48**	**IBK**	**Vote**	**Logistic**	**RF**
TPR	0.954	0.998	1.000	0.923	0.974	1.000
FPR	0.432	0.010	0.000	0.923	0.248	0.000
ACC (%)	95.4401	99.8144	99.213	92.285	97.401	99.982
PRE	0.954	-	0.9213	-	0.973	0.9982
ROC	0.938	0.997	0.918	0.500	0.991	0.998
TBM (sec)	0.06	0.11	0	0	2.38	0.35
TTM (sec)	0.11	0.03	1.48	0.01	0.02	0.12
RMSE	0.1353	0.0288	0.0005	0.1904	0.097	0.0249

TABLE 3
PERFORMANCE OF ENSEMBLES ON TRAINING DATASET

Evaluation metrics	**Ensemble of Classifier (Vote)**				
	Average	**Product**	**Majority Voting**	**Minimum**	**Maximum**
TPR	0.999	0.999	1.000	0.999	0.999
FPR	0.006	0.003	0.003	0.003	0.006
ACC (%)	99.947	99.894	99.993	99.894	99.867
PRE	0.999	0.999	1.000	0.999	0.999
ROC	1.000	1.00	0.998	1.000	0.999
TBM (sec)	2.63	2.46	2.47	2.67	2.62
TTM (sec)	1.57	1.5	1.64	1.54	1.5
RMSE	0.0671	0.0203	0.0115	0.0211	0.1134



Filter: None

Current relation
Relation: hypothyroid-weka.filters.supervised.instance.SpreadSubsample-M97.0-X0.0-S1
Instances: 485
Attributes: 30
Sum of weights: 485

Selected attribute
Name: Class
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	negative	194	194.0
2	compensated_hypothyroid	194	194.0
3	primary_hypothyroid	95	95.0
4	secondary_hypothyroid	2	2.0

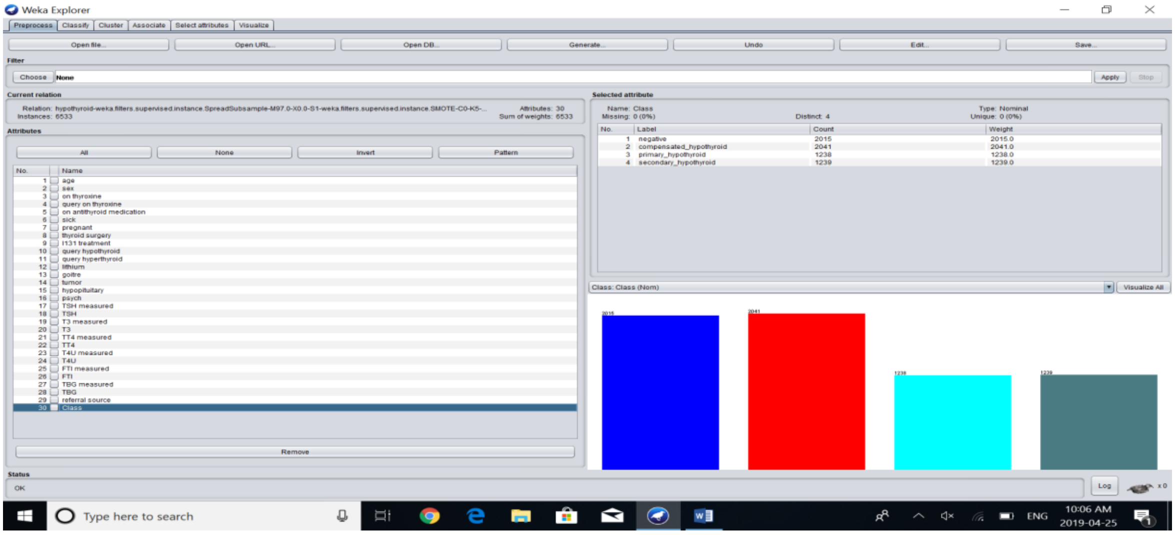
Attributes

No.	Name
1	age
2	sex
3	on thyroxine
4	query on thyroxine
5	on antithyroid medication
6	sick
7	pregnant
8	thyroid surgery
9	I131 treatment
10	query hypothyroid
11	query hyperthyroid
12	lithium
13	goitre
14	tumor
15	hypopituitary
16	psych
17	TSH measured
18	TSH
19	T3 measured
20	T3
21	TT4 measured
22	TT4
23	T4U measured
24	T4U
25	FTI measured
26	FTI
27	TBG measured
28	TBG
29	referral source
30	Class

Class: Class (Nom)
Visualize All

Status
OK

Fig 2. Under Sampling



Filter: None

Current relation
Relation: hypothyroid-weka.filters.supervised.instance.SpreadSubsample-M97.0-X0.0-S1-weka.filters.supervised.instance.SMOTE-C0-K5-...
Instances: 6533
Attributes: 30
Sum of weights: 6533

Selected attribute
Name: Class
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	negative	2015	2015.0
2	compensated_hypothyroid	2041	2041.0
3	primary_hypothyroid	1238	1238.0
4	secondary_hypothyroid	1239	1239.0

Attributes

No.	Name
1	age
2	sex
3	on thyroxine
4	query on thyroxine
5	on antithyroid medication
6	sick
7	pregnant
8	thyroid surgery
9	I131 treatment
10	query hypothyroid
11	query hyperthyroid
12	lithium
13	goitre
14	tumor
15	hypopituitary
16	psych
17	TSH measured
18	TSH
19	T3 measured
20	T3
21	TT4 measured
22	TT4
23	T4U measured
24	T4U
25	FTI measured
26	FTI
27	TBG measured
28	TBG
29	referral source
30	Class

Class: Class (Nom)
Visualize All

Status
OK

Fig 3. Over Sampling

TABLE 4

ENSEMBLE OF CLASSIFIER (VOTE)

Class Name	**Original**	**Under Sampling**	**Over Sampling**	**Final**
Negative	3481	194	2015	2015
Compensated_hypothyroid	194	194	2041	2041
Primary_hypothyroid	95	95	1238	1238
Seconday_hypothyroid	2	2	1239	1239
Total	3772	485	6533	6533

TABLE 5

PERFORMANCE EVALUATION OF ENSEMBLE OF CLASSIFIERS

Evaluation metrics	**Original**	**Under Sampled**	**Over Sampled**
TPR	0.999	0.998	0.996
FPR	0.006	0.001	0.002
ACC (%)	99.947	99.793	99.556
PRE	0.999	0.998	0.996
ROC	1.000	1.000	1.000
TBM (sec)	2.65	0.46	6.67
TTM (sec)	1.84	0.05	6.44
RMSE	0.0671	0.1161	0.1504

According to the analysis done with the above result, the original data-set having 4 classes and 3772 instances we carried out. Classification using 6 various classifiers namely; Naive Bayes, J48, IBK, Vote, Logistic and Random Forest were performed. According to the analysis we found Random Forest to have the highest accuracy of 99.982%, Combination of 6 classifiers for voting were used and the highest accuracy was given by majority voting which was 99.993%.

Next Under Sampling was done using Spread Sub Sample and Over Sampling was done using SMOTE. Under sampling and over sampling are usually performed on the data set to even up the imbalanced classes. Under sampling aims to balance class distribution by eliminating the number of majority class examples. Over sampling aims to balance class distribution by increasing the number of minority class examples. Finally a combination of the same 6 classifiers were used to determine the best one and the final accuracy given as 99.556%.

II. Conclusion

A comparative analysis of Naive Bayes, J48, IBK, Vote, Logistic and Random Forest classifiers along with a combination of these classifiers were used to see which one will give the best result using the Hypothyroid dataset. The result shows that the combination of classifiers provide better result than using these classifiers individually. However, there will always be more scope for further work to be carried out on different datasets using different classifiers in WEKA tool or other data mining tool. We can use such combination in data mining which is mostly required in areas of medical, banking, stock market and various other areas.

References

[1] "Comparative Analysis of Bayes Net Classifier, Naïve Baye's Classifier and Combination of both Classifiers using WEKA", Abhilasha Nakra, Manoj Duhan, International Journal of Information Technology and Computer Science, 2019.

[2] "Comparative Analysis of Classification Algorithms Using Weka", Sakshi Saini, Amita Dhankkar, Dr. Kamna Solanki, Vol. 08, Issue 10, October, 2018.

[3] "Comparison of Different Classification Techniques Using WEKA for Hematological Data", Md. Nurul Amin, Md. Ahsan Habib, Volume-4, Issue-3, pp-55-61, 2015.

[4] "Implementing WEKA for Medical Data Classification and Early Disease Prediction", N.Kumar and S. Khatri, IEEE Int.Conf."Computational Intell. Commun. Technol., vol.3rd, pp.1-6, 2017.

[5] "A Comparative Analysis of Meta and Tree Classification Algorithms Using Weka", T.Sathya Devi, Dr.K.Meenakshi Sundaram, Volume: 03 Issue: 11, Nov -2016

[6] https://www.wikipedia.com// accessed on April 2019.