Binary and Multi-Class Classification Using Supervised Machine Learning Algorithms and Ensemble Model

H Asela

Department of Electrical and Information Engineering, University of Ruhuna, Sri Lanka

asela.h@eie.ruh.ac.lk

Abstract— Classification is a vital aspect in data mining, where vast quantities of data are segregated into discrete classes. Models based on different statistical and machine learning approaches are used for this task. However, the classification performance depends on multiple factors like selected algorithm, domain and features of the dataset. The objective of this study is to evaluate the classification performance of widely used supervised machine learning algorithms; Decision Tree (DT), Naïve Bayes (NB) algorithm, Support Vector Classifier (SVC), K-Nearest Neighbour (KNN) algorithm and the Ensemble Model (EM) based on soft voting technique. These algorithms are tested on 6 datasets in different domains, and the datasets contain both multi-class and binary class data as well as balanced and imbalanced data. Accuracy, Precision and Recall are used as evaluation metrics to evaluate the classification performance in balanced datasets, where F1-measure is used in imbalanced dataset for the same task. The evaluation results indicate that EM outperformed single algorithms at most instances. When comparing single algorithms, KNN performed best with multi class classification, where SVC performed best in binary classification in balanced datasets. Also, KNN showed the best classification performance when it comes to imbalanced dataset. All the algorithms performed well when the data set is balanced. However, the classification performance in all models including EM is below expectation, when the data distribution is highly imbalanced.

Keywords: classification, machine learning, supervised algorithms, ensemble model, soft voting classifier

I. INTRODUCTION

Classification is the process of categorizing a given structured, unstructured or semi-structured data into classes.It is an important aspect in data mining and analysis and widely used in different domains like business, helath, education, medicine, telecommunication, security etc.

Supervised classification is one of the most frequent tasks done by Intelligent Systems. In classification, data instances are assigned to an appropriate class and there are machine learning and statistical models used as classifiers for this task. (Kotsiantis, Zaharakis and Pintelas, 2006) Supervised machine learning approaches are primarily used for this due to their ability to grasp complex patterns in datasets. Current research studies also investigate the ability of ensemble models for classification. Ensemble models combine several machine learning techniques into one predictive model thus improving the accuracy of the classification compared to individual supervised algorithms.Also classification mechanism differs with the complexity of the dataset. There are datasets with balanced or imbalanced class distributions. Also there are datasets with multiple class labels and binary labels.Balanced datasets have equal to nearly equal data points for each class where the data points in imbalanced datasets are biased towards one label. However, imbalanced data classification can be challenging since the class distribution is severely skewed and there are unequal misclassification costs. (Brownlee, 2017)

In this paper, I focus on analysing the classification performance of supervised machine learning algorithms and the ensemble models built upon them. For the evaluation I used four famously used supervised machine learning

algorithms namely Decision Tree, Naïve bayes algorithm, Suppot Vector Classifier and K-Nearest Neighbour algorithm.Ensemble model is built by combining these supervised algorithms using soft voting technique.I analyse and evalute how these single algorithms and the emsemble model perform in the classification task in mutli-class and binary class datasets as well as balanced and imbalanced data distribution datasets. Also the limitations of the study and the future direction of the research are discussed at the end of the paper.

II. RELATED WORK

Machine learning algorithms have shown their effectiveness in data classification. Supervised machine learning algorithms hold prominence in this. These algorithms use a labelled training dataset first to train the underlying algorithm and this trained algorithm is then fed on the unlabelled test dataset to categorise them into classes. (Uddin, Khan, Hossain and Moni, 2019) There are famously used supervised algorithms for classification.

1) Decision tree

Decision tree is one of the earliest and prominent supervised machine learning algorithms used for classification.It is a tree based algorithm where the data is continuously split according to a certain parameter.Each node in the tree shows a feature and each branch shows a decision or rule.Also each leaf in a decision tree shows a class label. (Patel and Prajapati, 2018)

2) Naïve bayes

Naïve bayes classification technique is based on the Bayes' theorem. This theorem considers probability of an event based on the prior knowledge of conditions related to that event. The classifier assumes that features are independent given class. Even though this assumption is considered to be weak and even if calculated probability estimates are inaccurate, Naïve bayes classifier is proven to perform well in classification tasks. (Rish, 2001) Most of current machine learning libraries provide optimizations for this algorithm therefore improving the classification performance.

3) Support vector classifier

The objective of the support vector classification algorithm is to find a hyperplane in a N-dimensional space to distinctly classify the data points into classes. Support vectors are data points that are closer to the hyperplane and these vectors influence the position and orientation of the hyperplane. These support vectors maximize the margin of the classifier. (Gandhi, 2018) This classfier convert the machine learning problem to an optimization problem and use mathematical programming to solve the problem. Support vector machines and classifiers are found to be beneficial in wide range of classification tasks like text categorization, face detection, verification, recognition, speech recognition and bioinformatics. (Tian, Shi and Liu, 2012)

4) K-nearest neighbour

K-nearest neighbour is a non-parametric classification technique. This is very simple yet very powerful algorithm based on proximity or similarity. The algorithm assumes the similarity between the new data points and and put the new data points into the class that is most similar to the available data classes. The classifier is known to be work best with numerical data. However, one needs to carefully select the features feed into the algorithms since this classifier is very sensitive to irrelevant or redundant features. However this can be avoided using proper feature selection and feature weighting. (Cunningham and Delany, 2007)

However, there are certain pros and cons in each algorithm. Ensemble models are used to yield the benefits and reduce the limitations of each single algorithm. These models combine the results from single algorthms based on multiple metrics like weight and probability. This enhances the classification performance of the model. Ensemble approaches like soft voting classifier is proved to provide superior results compared to single algorithms in different domains. (Kumari, Kumar and Mittal, 2021)

III. METHODOLOGY

E. Datasets

For the evaluation in balanced data class distribution, I used labelled multi class and binary class benchmark datasets. They are Ecoli

(Horton and Nakai, 1996), Glass identification (Evett and Spiehler, 1987), Iris (Hart and Duda, 1973), Stroke prediction (Zaki, Mohamed and Habuza, 2021) and Prima Indians diabetes (Choubey et al., 2016) datasets. These datasets are retrieved from well-known UCL machine learning repository and Kaggle.

Table 1. Balanced data distribution datasets

Dataset	Source	Data instances	Feature count	Data classes
Ecoli (D1)	UCL	336	7	8
Glass identification (D2)	UCL	214	9	7
Iris (D3)	UCL	150	5	3
Stroke prediction (D4)	Kaggle	5110	11	2
Pima Indians diabetes (D5)	Kaggle	768	8	2

For the evaluation in imbalanced data distribution, I used the Yahoo! S5 Anomaly benchmark dataset (Laptev and Amizadeh, 2015). It contains real data collected from Yahoo services and systhetically generated data separated in 4 data classes. A1 contains real data in 67 metrics where other data classes contain synthetic data in 100 metrics.

Table 2. Imalanced data distribution dataset

Data class	Number of instances	Number of features	Contamination
A1	94,866	2	0.0176
A2	142,100	8	0.0033
A3	168,000	8	0.0056
A4	168,000	8	0.0062

F. Exprimental design

Experimental design consists of two stages. First stage consist of data preprocessing and data split for training and test sets. Second stage consist of model training and testing.

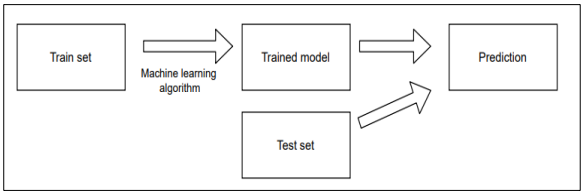

Figure 2. Data preprocessing and split stage

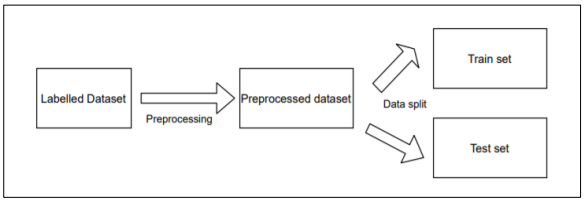

Figure 2. Model training and testing stage

At first, each machine learning algorithm was tested on each dataset. Finally, ensemble model testing was performed on each dataset. For the implementation, I used Python Scikit-learn library. (Pedregosa et al., 2011)

1) Data preprocessing and Data splitting

The data sets were preprocessed before feed into machine learning algorithm. All categorical data were converted into numeric data using label encoder. Then all the numerical data were normalized using a min-max scale. In case of missing values, those data were imputed by the mean value of the data column. Proprocessed data was split into two subsets randomly, one with 70% for the training and 30% for test set.

2) Model training and testing

Training set was used to train the classification model and the test set was used for model validation.

Models were built using famously used supervised machine learning algorithms. Hyperparameter tuning for each machine learning algorithm was done using previous literature and trial and error approach.Ensemble model was built by combining all of these supervised alrothms using soft voting technique. Soft voting is based on membership probabilities where the ensemble model sums the predicted probabilities from single algorithms for class labels and predicts the class label with the largest sum probability.

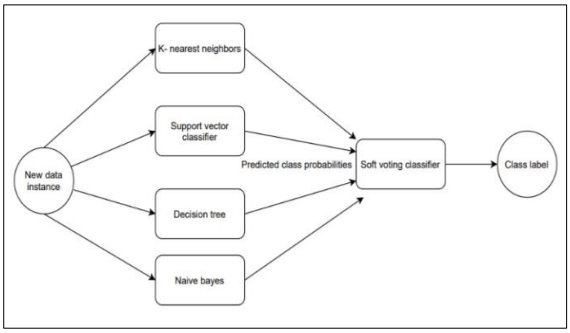

Figure 3. Ensemble soft voting classification

Hyperparameters used for each machine learning algorithms are depicted in Table 3.

Table 3. Hyperparameter values

Algorithm	Hyperparamenter	Value
DT	Minimum sample split	2
	Minimum sample leaf	1
NB	Distribution	Guassian
SVC	Kernel	RBF
	Regularization	1.0
	Gamma	Scale
	Probability estimation	True
KNN	Number of neighbours	5
EM	Voting	Soft

Other hyperparameter values are set into default values in Scikit-lLearn library.

G. Evaluation metrics

This study evaluates the classification performance of ensemble machine larning model in both balanced and imbalanced datasets.For the balnced datasets, Accuracy (A), Precision (P) and recall (R) were used as evaluation metrics. (Hossin and Sulaiman, 2015) Generalized confusion metrix was used to calculate these metrics. (Manliguez, 2016)

Table 4. Confusion matrix

		Predicted			
		Class 1	**Class 2**	**...**	**Class n**
Actual	**Class 1**	x_{11}	x_{12}	...	x_{1n}
	Class 2	x_{21}	x_{22}	...	x_{2n}
	...	...	...	...	...
	Class n	x_{n1}	x_{n2}	...	x_{nn}

The total numbers of false negative (TFN), false positive (TFP), and true negative (TTN) for each class i will be calculated using the 1,2 and 3 generalized equations respectively. The total true positive (TTP) in the system will be calculated using equation 4.

$$TFN_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} x_{ij} \quad (1)$$

$$TFP_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} x_{ji} \quad (2)$$

$$TTN_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} x_{jk} \quad (3)$$

$$TTP_{all} = \sum_{j=1}^{n} x_{jj} \quad (4)$$

Precision and recall for each class I were computed using the 5 and 6 generalized equations. For the overall precision and recall, macro average values are used. Overall accuracy is then derived using equation 7.

$$P_i = \frac{TTP_{all}}{TTP_{all} + TFP_i} \quad (5)$$

$$R_i = \frac{TTP_{all}}{TTP_{all} + TFN_i} \quad (6)$$

$$ACCURACY = \frac{TTP_{all}}{Total\ Number\ of\ Testing\ Entries} \quad (7)$$

For imbalanced dataset, accuracy will not be a suitable metric. So F1 – measure is used as evaluation metric for those datasets. F1-measure of 0 means a useless classifier where F1- measure of 1 means a perfect classifier. F1 – score is calculated based on overall Precision and recall using equation 8. (Jeni, Cohn and De La Torre, 2013)

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (8)$$

IV. RESULTS AND DISCUSSION

The results obtained for multi class classification and binary classification in balanced datasets are summarized in Table 5.

Table 5. Evaluation metrics for multi-class and binary classification in balanced datasets.

Accuracy (%)					
Dataset	**DT**	**NB**	**SVC**	**KNN**	**EM**
D1	75.24	69.3	76.23	78.21	89.1
D2	67.18	54.68	68.75	70.31	75
D3	95.55	93.33	95.55	97.77	100
D4	90.99	87.01	94.58	91.71	94.32
D5	71.86	75.75	77.92	71.86	77.92
Precision (%)					
Dataset	**DT**	**NB**	**SVC**	**KNN**	**EM**
D1	60.76	29.03	60.39	71.85	84.46
D2	66.87	55.58	64.45	70.31	70.98
D3	95.58	93.27	95.58	98.24	100
D4	55.61	57.82	47.29	52.75	66.93
D5	68.04	71.96	74.94	66.91	74.69
Recall (%)					
Dataset	**DT**	**NB**	**SVC**	**KNN**	**EM**
D1	53.45	31.76	64.94	63.54	80.22
D2	73.55	61.17	66.27	67.09	72.64
D3	93.58	93.73	95.58	97.43	100
D4	55.48	66.44	50	51.89	53.83
D5	69.11	69.75	71.7	62.48	72.43

D1, D2 and D3 are multi class data sets where D4 and D5 are binary datasets. Accuracy metric can be used to decide the high performing models since the datasts are balanced. From all the tested approaches, ensemble approach performed the best. Ensemble machine learning model provided the highest accuracy for all datasets. Also it provided the highest precision and the recall for the majority of datasets. From individual machine learning algorithms, KNN performed best with multi class classification where SVC preformed best with binary classification. NB is the worst performing algorithm since it recorded low accuracy, precision and recall for majority of the tested datasets. DT algorithm performed better compared to NB but it had less classification performance compared to SVC, KNN and ensemble approaches.

The results obtained for classification in Yahoo! S5 anomaly benchmark dataset are summarized in Table 6.

Table 6. Evaluation metrics for imbalanced data classes

Accuracy (%)					
Data class	**DT**	**NB**	**SVC**	**KNN**	**EM**
A1	99.17	99	99.35	99.14	99.2
A2	99.92	99.76	99.82	99.92	99.94
A3	99.18	99.41	99.55	99.63	99.6
A4	99.24	98.73	99.5	99.64	99.6
Precision (%)					
Data class	**DT**	**NB**	**SVC**	**KNN**	**EM**
A1	60.57	54.07	55.63	63.92	67.23
A2	91.26	57.57	63.63	90.88	93.83
A3	25.71	11.57	29.47	60	46.31
A4	18.79	15.49	15.55	55.18	42.22
Recall (%)					
Data class	**DT**	**NB**	**SVC**	**KNN**	**EM**
A1	58.75	63.62	47.77	64.84	64.99
A2	98.48	50.5	59.59	98.48	98.48
A3	20.84	3.52	21.07	34.87	26.43
A4	12.57	8.51	8.2	32.44	22.48
F1-measure (%)					
Data class	**DT**	**NB**	**SVC**	**KNN**	**EM**
A1	58.74	55.6	50.05	63.6	64.5
A2	93.5	52.42	60.75	93.54	95.47
A3	21.79	5.29	23.75	42.24	32.29
A4	14.24	8.49	10.16	38.54	27.85

When the classification classes are imbalanced in dataset, the accuracy of machine learning models are biased towards the majority class. The classification algorithm tends to predict the majority class often. Hence accuracy is not a good performance metric to evalute imbalanced datasets. It is evident from these results as I got very high accuracy values but low precision and recall values. I used F1-measure to evaluate the models. From F1-measures, it is evident that KNN and emsemble classifier had better performance in biased label classification. Classification algorithm performance decrade with the increase of data instance count and biasness of data lables. That is the reason for the poor classification performance of all algorithms in A3

and A4 classes. Adding an ensemble learning model did not help much for the classification in these data classes.

In order to improve the classification in imbalanced datasets, one can introduce oversampling or undersampling. Oversampling replicates minority class data points where understamping removes majority class data points. This can reduce the class imbalance in the dataset thus improving classification performance in machine learning algorithms.

V. CONCLUSION

In this research, I have investigated the classification performance of famously used supervised machine learning algorithms and ensemble model. Decision tree, Support vector classification, Naïve bayes classification and K-nearest neighbour classification algorithms were trained to perform classification in both balanced and imbalalnced datasets. The balanced datasets consist of both multi class and binary datasets where imbalanced dataset is a binary dataset with anomaly data. This research study also evaluated the ensemble machine learning model classification performance on these datasets. The ensemble model was developed using voting technique with aforementioned supervised learning algorithms. The experimental results show that the ensemble model performs better compared to single algorithms in classification. From individual algorithms, K-nearest neighbour algorithms performed best in multi class classification where Support vector classification algorithm performed best in binary classification for balanced datasets.Naïve bayes algorithm had the worst performance. However all algorithm models including ensemble model performed average to poor in imbalanced dataset classification.

For future work, one can investigate the effectiveness of oversampling and undersampling techniques to solve the class imbalance problems. Also it is worth investigate on optimizing hyperparameters in these machine learning algorithms to improve the classification performance.

REFERENCES

Brownlee, J. (2017) Why Is Imbalanced Classification Difficult?. [online] Machine Learning Mastery. Available at:<https://machinelearningmastery.com/imbalanced-classification-is-hard/> [Accessed 30 May 2021].

Cunningham, P. and Delany, S. (2007) k-Nearest neighbour classifiers. [online] Available at: <https://www.researchgate.net/publication/228686398_k-Nearest_neighbour_classifiers> [Accessed 30 May 2021].

Gandhi, R. (2018) Support Vector Machine — Introduction to Machine Learning Algorithms. [online] Medium. Available at: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47> [Accessed 30 May 2021].

Hossin, M. and Sulaiman, M. (2015) A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), pp.01-11.

Jeni, L., Cohn, J. and De La Torre, F. (2013) Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction

Kotsiantis, S., Zaharakis, I. and Pintelas, P. (2006) Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), pp.159-190.

Kumari, S., Kumar, D. and Mittal, M. (2021) An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering, 2, pp.40-46.

Horton, P. and Nakai, K. (1996) A Probablistic Classification System for Predicting the Cellular Localization Sites of Proteins. Intelligent Systems in Molecular Biology, pp.109–115.

Evett, W. and Spiehler, E. (1987) Rule Induction in Forensic Science. KBS in Government, pp.107–118.

Hart, P.E. and Duda, R.O. (1973) Pattern classification and scene analysis : Richard O. Duda, .Peter E. Hart, New York ; London ; Sydney: J. Wiley & Sons.

Zaki, N., Mohamed, E.A. and Habuza, T. (2021) From Tabulated Data to Knowledge Graph: A Novel Way of Improving the Performance of the Classification Models in the Healthcare Data.

Choubey, D., Paul, S., Kumar, S. and Kumar, S. (2016) Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. Communication and Computing Systems.

Laptev, N. and Amizadeh, S. (2015) Yahoo anomaly detection dataset s5. [online] Available at: https://webscope.sandbox.yahoo.com/.

Manliguez, C. (2016) Generalized Confusion Matrix for Multiple Classes. [online] Available at: <https://www.researchgate.net/publication/310799885_Generalized_Confusion_Matrix_for_Multiple_Classes> [Accessed 28 May 2021].

Patel, H. and Prajapati, P. (2018) Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering, 6(10), pp.74-78.

Pedregosa, F. et al. (2011) Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), pp.2825–2830

Rish, I. (2001) An Empirical Study of the Naïve Bayes Classifier. [online] Available at: <https://www.researchgate.net/publication/228845263_An_Empirical_Study_of_the_Naive_Bayes_Classifier> [Accessed 30 May 2021].

Tian, Y., Shi, Y. and Liu, X. (2012) Recent Advances On Support Vector Machines Research. Technological and Economic Development of Economy, 18(1), pp.5-33.

Uddin, S., Khan, A., Hossain, M. and Moni, M. (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1).

ACKNOWLEDGMENT

I would like to show my gratitude to staff of Department of Electrical and Information Enginnering, University of Ruhuna for their support provided in this research study.

AUTHOR BIOGRAPHY

Hevapathige Asela is a Lecturer (Probationary) attached to Department of Electrical and Information Enginering, University of Ruhuna. His research interests are Machine learning, Deep learning, Natural language processing, Computational mathematics and Business computing.