Raj Pokhrel
CMPSC 497
Project 2
Report

Task 1 Normalize and Un-normalize models

Model 1: Normalized

Precision	Recall	F-Measure	Accuracy Metric
72.03%	59.07%	61.46%	86.57%
72.96%	60.22%	62.89%	86.80%
73.29%	60.82%	63.60%	86.90%
73.80%	61.52%	64.31%	87.04%
58.45%	50.95%	48.81%	85.43%

Model 2: Un-Normalized

Precision	Recall	F-Measure	Accuracy Metric
88.8%	50.21%	47.48%	88.88%
83.0%	50.0%	47.0%	88.832%
77.96%	52.12%	51.37%	89.09%
76.93%	52.93%	52.861%	89.159%
79.1%	50.0%	47.10%	88.84%

Observations

- The table about shows the two different model with normalized and Un-normalized datasets. From the results, we can say that the normalized dataset increase accuracy for some datasets for not for all. However, un-normalized datasets was stable with the prediction score. Therefore, It's better to use normalized data because it shows how the model accuracy change overtime and does not give us bias results since the data is evenly distributed.

Task 2: Resampling Datasets

Model 1: Up-sample

Precision	Recall	F-Measure	Accuracy Metric
91.3%	89.0%	90.0%	91.0%
91.13%	89.88	90.1%	91.01%
91.130%	89.18%	90.2%	90.05%
91.35%	88.90%	89.2%	90.106%
90.12%	89.20%	90.08%	91.11%

Model 2: Not Sample

Precision	Recall Metric	F-Measure	Accuracy Metric
84.0%	50.21%	47.48%	88.88%
61.0%	50.0%	47.15%	87.40%
77.96%	52.14%	51.36%	89.09%
76.93%	52.93%	52.86%	89.15%
77.20%	50.0%	47.04%	88.84%

Observation

- From the Up-sample and not sample data, we can say that the accuracy rate for the resample data is consistent with precision, recall, f-measure and accuracy. In addition, unbalanced datasets does not provide consistent accuracy in different metrics.

Question

Note that data imbalance exists in this dataset. Please explain why we want to avoid imbalance issue in training classifiers? Briefly summarize at least 3 methods deal with data imbalance issue

- The imbalance datasets will not give us consistent accuracy when dividing the data into test and training sets. Therefore, it's very important to balance the datasets before training your model.
- The three methods of dealing with imbalance datasets is by resampling, down/up sampling, cross validation, and change the imbalanced classes. Resampling the datasets will give you more precision when predicting the results. Cross validation will resample the datasets and evaluate the model by splitting the data into training and testing to provide better prediction rate.

Task 3: Feature Selection

- Features selection is really important in data mining when it comes to making a decent model because it will automatically select the attributes that are most relevant to your model and use it to predict the results with the results that is precise.

Mode: K=1

Precision	Recall Metric	F-Measure	Accuracy Metric
39.48%	50.00%	44.12%	78.86%
39.0%	50.10%	44.09%	78.52%
39.43%	49.20%	44.91%	79.72%
38.15%	49.09%	43.90%	77.46%
39.43%	50.00%	43.09%	75.66%

Mode: K=3

Precision	Recall	F-Measure	Accuracy Metric
77.0%	73.0%	68.0%	81.26%
83.0%	81.0%	81.0%	81.97%
79.90%	82.04%	80.0%	80.56%
84.0%	83.0%	82.4%	83.05%
79.72%	81.92%	79.72%	80.0%

Mode: K=5

Precision	Recall Metric	F-Measure	Accuracy Metric
68.33%	56.46%	57.29%	85.88%
68.33%	55.43%	56.46%	85.91%
67.71%	55.90%	57.14%	85.79%
73.28	60.63%	63.3%	86.89
60.23%	52.00%	50.98%	85.19%

Observation

- The accuracy of the model increases as we go from K1, K2 to K3 because we have more features to train on our model to get more precious results. Overall, with K5 features we have the most accurate prediction.

Task 4

Neuron Network

Precision	Recall	F-Measure	Accuracy Metric
91.0%	85.64%	87.56%	89.0%
83.88%	84.22%	84.03%	85.0%
92.22%	86.63%	88.63%	90.0%
93.27%	87.27%	89.15%	90.45%
86.19%	83.09%	84.23%	85.97%

Decision Tree

Precision	Recall	F-Measure	Accuracy Metric
69.0%	65.17%	56.66%	57.31%
70.87%	68.79%	61.89%	62.07%
70.944%	68.95%	62.14%	62.30%
70.00%	65.75%	57.05%	57.73%
68.94%	62.23%	51.13%	52.94%

Observation

- Overall, the neuron network perform really well than the decision tree. The accuracy score for the neuron network were much higher than the decision tree with the balance and unbalance datasets.