Your name: Keiryn Ward Hart , 300428418 Course code (e.g. COMP309, COMP440): COMP309

**COMP 309 | Machine Learning Tools and Techniques**

**2.1 Core: Investigate Basic Use of the Different Tribes of AI**

**Symbolists:**

**J48:**

A picture containing text

Description automatically generated

A picture containing text, receipt

Description automatically generated

**Representation:**

J48 is an algorithm designed to generate a decision tree, the program creates a classifier which takes the form of a tree where a leaf indicates a class and a decision node identifies a test using deduction which will be carried out on a single attribute with branches for each possible outcome. The program can generate both pruned and unpruned trees.

This technique represents the symbolists because it is an algorithm the hinges on propositional logic e.g. true or false, and the tree continues working this logic until a classification is made. Another key factors that show this is a symbolist technique is the fact that it is a decision tree.

**Evaluation Method:**

The evaluation method for Symbolists is Accuracy, J48 splits the dataset and evaluates the accuracy of the initially produced decisions tree and through the pruning phase aims to lower the error rate and create a more accurate decision tree. It also does pairwise comparisons on the importance of the performance measures to increase accuracy of the decision tree.

**Optimization:**

The Optimization method is inverse deduction or induction, this process essentially uses a supervised process in which knowledge that has already been gathered from the model about classes in the database is used to aid the process in order to create a more accurate classification.

**Important Aspects:**

The results from the J48 decision tree present a pruned decision tree where uniformity of cell size, Bare Nuclei and Clump thickness appear to be the only aspects of the data that the algorithm deemed to be necessary in classification. The algorithm correctly classified around 94% of the data and the kappa statistic is 0.88 indicating that the model is working well and is not random.

The precision of the model by class appears to be pretty good with a weighted average of 0.944, additionally the ROC Area which describes the weighted accuracy of the model is 0.955 indicating that the model is right 95% of the time.

Interestingly when looking at the confusion matrix it appears that roughly the same amount of instances were misclassified for both classes (malignant and benign), This could be due to the imbalanced classes as 65% of the data is benign.

**Connectionists:**

**Multilayer Perceptron:**

A picture containing text, receipt

Description automatically generated

**Representation:**

A multilayer perceptron is a feed forward artificial neural network composed of multiple perceptron’s, they are composed of an input layer to receive the signal, an output layer which makes a prediction based on the input while in between a number of hidden layers. it uses backpropagation while training the data to adjust the weights and biases of the model in order to reduce error.

This technique represents the connectionist tribe because it is a neural network which is modelled off human brains it learning by adjusting the connections between its neurons just like the brain does.

**Evaluation method:**

It evaluates its results by continuing the backpropagation until an error minimum is met or until the error reaches a specified value. While this back and forth action is taking place a gradient begins to form, this is sometimes called the landscape of error.

**Optimization:**

the optimization method is using gradient descent which is an optimization algorithm that randomly initializes the parameters then iterates through the data points adjusting the parameters accordingly, while doing this it calculates and updates the gradient and adjusts the parameters, when the error reaches a certain value it stops.

**Important Aspects:**

The multilayer perceptron classifier has correctly classified just under 96% of instances correctly and Has a Kappa statistic of 0.91 which is slightly better than that of J48 which indicates that the model is less random than J48 but not by much and is also working well.

The weighted average of precision sits at 0.96 with the precision for benign being slightly more accurate than malignant. The ROC Area is 0.988 which is very good and hints that the model is correct roughly 98% of the time which is interesting given it has correctly classified 95% of instances.

The confusion matrix for this classifier is slightly better than J48 and appears to have a more representative spread of correct to incorrect classifications and doesn’t seem to be affected by the imbalance of class sizes.

**Bayesians:**

**Naïve Bayes:**

A picture containing text, receipt

Description automatically generated

**Representation:**

Naïve Bayes is a classification technique that assumes independence among its predictors treating every feature independently and equally. it uses Bayes theorem to calculate probabilities and the algorithm will choose the larger probability considering that to be the best. The reason that it falls into the Bayesian tribe is because it uses

**Evaluation Method:**

Naïve Bayes evaluates through posterior probability using the Naïve Bayesian Equation to calculate the posterior probability value for each class, the performance is measured by this value and the highest value is the final prediction.

**Optimization:**

Its optimization driver is probabilistic inference involving variations of Bayes’ theorem. The inference is drawn based on the probability that a certain variable takes on a specific value where the higher the probability, the more likely that variable holds that value.

**Important Aspects:**

the naïve Bayes classifier has correctly classified 97.3% of instances, which is the highest percentage of correct classifications out of any of the classifiers so far, it also has the highest Kappa statistic so far which again indicates that the model is not randomly selecting and I working well.

A similar trend seems to follow for the Precision of the model with the classification for benign instances being slightly more accurate than that of malignant instances but overall the weighted average is very good at 0.97.

The ROC area in the classifier is again better than the 2 previous classifiers sitting at a weighted average of 0.99, this shows that almost every classification should be correct.

The confusion matrix is similar to that of the Multilayer perceptron in the sense that the weight of correct classifications to incorrect classifications for both classes seems to be more representative of the class imbalance.

**Analogizers:**

**IBk:**

A close up of text on a white background

Description automatically generated

**Representation:**

IBk is a form of instance based learning that is derived from the nearest neighbour pattern classifier. it assumes that similar instances have similar classifications based on their most similar neighbour’s classification. It gradually improves with training more accurately classifying the data points and the parameter k describes the number of similar items used to determine what that data point is.

This technique represents the Analogizers tribe as they are concerned with recognising similarities between different situations and determining which things are related based on those similarities. Analogizers use Nearest-neighbour algorithms which therefore means this technique fits.

**Evaluation Method:**

IBk uses k as its evaluation method, k describes the number of data points used to classify a new point, when k is equal to 1 the data is noisy and overfitting but when k increases the predictions become more robust and accurate. If k is too large though detail can be lost, when k increases variance decreases and bias increase.

**Optimization:**

The optimization method for this is constrained optimization which essentially is where you optimize a function in favour of variables while constraints are placed on the others, this is important with the similarities between data points as some similarities may be meaningless to the classification and some may be much more important which is why the constrained optimization comes into play.

**Important Aspects:**

IBk is only using 1 nearest neighbour for the classifier at the moment, it has correctly classified 95.7% of instances in the dataset which is not the best out of the four classifiers but it is also not the worst, its kappa statistic is 0.9 which again is good and shows that the model is working well and is not randomly classifying instances.

In terms of precision just like the other models the weighted average for IBk is good sitting at 0.96. again this classifier also has high precision sitting at around 0.96 as a weighted average across both classes.

ROC area is 0.98 which is again very high and shows that most of the instances will be classified correctly. The confusion matrix is interesting for this classifier as it seems to have incorrectly classified more instances for malignant instances than it has benign instances which is most likely a result of the class imbalance again as you would expect more misclassifications for the larger class.

**2.2 Completion 1: Consider a Pipeline for Dataset Processing**

The business aspect of this dataset is to identify a technique that can accurately predict if a tumour is malignant using data that was gathered using fine needle aspiration. Breast cancer is a serious problem for women around the world and although technology is getting better which is a cause for a smaller death toll, more can be done to improve people’s chances of survival. The acquisition of this data hopes to achieve an accurate way of predicting if a tumour is benign (not cancerous) or malignant (cancerous).

The Data arrived periodically over a number of years as Dr. Wolberg reported his clinical cases from January 1989 through to November 1991. The dataset contains 699 instances and no missing values or outliers seem to be apparent in the data. The original dataset contains 11 attributes being;

* Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromantin, Normal Nuclei, Mitosis and Class.

Each attribute except for Sample code number and Class has a value of 1 – 10 (where 1 indicates a normal state and 10 indicates a highly abnormal state) and Class has a value of either 2 (benign) or 4 (malignant).

Data preparation - state how the pipeline could assist in the preparation of the data prior to the technique being applied.

the project can find assistance from the pipeline as It can be useful for other datasets that are similar in the future, filters can be added in the pipeline prior to the technique that can do a variety of different preparatory adjustments to the dataset in order to get it ready for the specific techniques, for example filters such as Numeric to nominal or resample and many more. These allow the data to be sorted, filtered and prepared properly for whatever technique may be used and it also means that if a similar dataset is brought in with the same purpose, the pipeline already has a preparation process ready for the new data.

The pipeline can also automatically create folds for cross validation or training and test splits prior to the classifier which can be tuned to certain needs and preferences. Something I have used in my pipeline which is also very useful is principle component analysis which can be helpful for certain techniques.

All but one of the five tribes of AI are suitable for the pipeline, the one tribe it does not suit is the evolutionary tribe as the weka knowledgeflow does not support this, for all other tribes of AI the pipeline is suitable and can be used.

The pipeline supports multiple methods to evaluate a solution, such evaluation methods include things like the Classifier performance evaluator which can be used to assess the performance of the different classifiers you have chosen to use, this then outputs information to a text viewer which can be used to check accuracy, classification percentages, ROC area, confusion matrices and much more. Evaluation in the pipeline also comes in the form of class assigner which is another evaluation method that allows you to evaluate which attribute should be used as the class for the classification.

The model produced can be relatively easily deployed as the filters and pre-processing involved in the pipeline are relatively straightforward and as long as a dataset is not full of noise and/ or missing values then the deployment of this model is quite easy. If additional effort was required it would be in the form of cleaning the data or slight adjustments to some parameters in the classifiers to adjust for different datasets.

**2.3 Completion 2: Use the pipeline to re-evaluate the selected techniques** in Part 2.1 used to classify the datasetA close up of a map

Description automatically generated

A picture containing text, newspaper

Description automatically generated**Symbolist:**

**J48:**

A close up of a map

Description automatically generatedA picture containing text, receipt

Description automatically generated

J48 has produced another pruned decision tree (which was quite hard to clearly present) and output that shows it successfully classified 97.84% of instances in the dataset which leaves 15 instances incorrectly classified. The kappa statistic of 0.95 indicates that the model is not randomly classifying instances and shows an early indication that the model is running well.

Looking at the precision of the model, both classes show a precision value of 0.98 which is very good and reinforces the point that the model is classifying almost all of the instances correctly.

ROC area has a high value of 0.96, reinforcing that the model is classifying almost all of the instances correctly alternatively the weighted accuracy for the model is 96%. The confusion matrix shows that the model is misclassifying more instances of malignant tumours than benign tumours which is again interesting given attempts to balance out the classes.

In comparison to the initial classification before the pipeline we can see that it is classifying around 3% better than the original model and the kappa statistic is far higher than the 0.88 of the initial J48 classifier. the ROC area value is not much different between the two with the pipeline value being 0.01 higher that the initial. It is interesting to note though that the pipeline seems to have helped with the problem of misclassification of malignant tumours and it seems to be much more representative of the class imbalance than before.

**Connectionists:**

**Multilayer Perceptron:**

A close up of text on a white surface

Description automatically generated

The pipelined Multilayer perceptron has produced a classification output that shows the correct classification of 97.6% of instances in the dataset only misclassifying 17 instances. It has a kappa statistic of 0.95 which shows that the model is not randomly classifying.

The precision of the model has a weighted average of 0.98 showing that the model is again showing that the pipeline is classifying almost all of the instances correctly, ROC area has a value of 0.99 which is incredibly close to 1 and shows again that this model is making very accurate classifications. Similarly to the J48 model it is interesting to note that this model is again misclassifying more instances out of the malignant class than the benign class.

In comparison to the initial multilayer perceptron model we can see that this pipeline model is almost 2% more accurate at classifying instances correctly. It has a larger kappa statistic than the original 0.91 and its values for precision and ROC area are again higher than the initial classifier. it is interesting to note though that the initial classifier’s misclassifications were more representative of the class imbalance than the pipeline classifier’s, I believe this could possibly be due to my resampling method which may have worked for some classifiers better than others. Overall there was still an improvement in classification from the original because of the pipeline.

**Bayesians:**

**Naïve Bayes:**

A picture containing receipt, text

Description automatically generated

Naïve Bayes algorithm through the pipeline saw a very similar accuracy in correct classification with a percentage of 97.4% and a kappa statistic of 0.94 again showing the model is not randomly classifying. The precision of the model is high and is more representative of the class imbalance with a weighted average of 0.98, ROC area has a similar value of 0.99 to the previous classifiers which shows very good diagnostic power. The confusion matrix for this classifier is much more representative of the class imbalance but does indicate that the resampling methods to try and fix this imbalance may not be working as good as it could do.

In comparison to the original Naïve Bayes classifier we can see that the pipeline really has not made any huge changes to the accuracy of classification, the pipeline classifier is only marginally better than the original where it has correctly classified 1 instance more than the other, it has a slightly higher kappa statistic and precision value but apart from this it is largely unchanged, interestingly given all the other models so far have seen some improvement.

**Analogizers:**

**IBk:**

**A picture containing receipt, text

Description automatically generated**

The final classifier included in the pipeline was IBk and this classifier was extremely accurate with a correct classification percentage of 98.9% where it only misclassified 8 instances, making it by far the most successful of the 4 classifiers, IBk’s kappa statistic was 9.97 which just like the other classifiers shows that the model is not randomly classifying. Just like the other models the precision was very high with a weighted average of 0.989 showing again the accuracy of the classifications, the ROC area was the highest out of any of the 4 classifiers in the pipeline at 0.99 indicating excellent diagnostic power. Similarly to the other classifiers except for Naïve Bayes the confusion matrix misclassifications are not really representative of the class imbalance, this is interesting and may be something that I would need to look into when adjusting the parameters of the pipeline to hopefully lower misclassifications and represent the data more.

In comparison to the original IBk we can see one of the larger differences of the classifiers with just over a 3% increase in correctly classified instances. The kappa statistic for the pipeline IBk is a lot larger than the original of 0.9, and again obviously just like the previous classifiers we can see that this models precision and ROC area values are larger than that of the original, when playing around with the IBk parameters and adjusting the value for k it was interesting to note that similarly to the original a k value of 1 produced the most accurate classifications.

Overall prior to the pipeline, the classifier that I would have originally gone with would have been the Bayesian tribe’s Naïve Bayes because it was making the most accurate classifications, but, after the implementation of the pipeline I would not go with the Bayesians anymore and instead would have to choose the Analogizers’ IBk as it provided the most accurate classifications post pipeline whereas the Naïve Bayes really did not change at all.

**2.4:**

Couldn’t get Heuristic lab to work on my laptop?

# References

Akosa, J. S. (2016). *Application of Data Mining Techniques in Improving Breast Cancer Diagnosis*. SAS Customer Support Site | SAS Support. <https://support.sas.com/resources/papers/proceedings16/9420-2016.pdf>

*The five tribes of machine learning*. (2019, December 12). Fringeling Notes. <https://notes.fringeling.com/OnTheFiveTribesOfMachineLearning/>

Kaur, G., & Chhabra, A. (2014). *Improved J48 Classification Algorithm for the Prediction of Diabetes*. CiteSeerX. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.678.9273&rep=rep1&type=pdf>

Naive Bayesians. (n.d.). *Naive Bayesian*. Data Mining Map. <https://www.saedsayad.com/naive_bayesian.htm>