experiment

Keith Howell

Contents

Introduction

Once the data has been prepared and the purpose of the data mining is. The next step is deciding what algorithm to choose. There are many algorithms which can be used for this purpose, however to find the most suitable one for the data in this paper a number of experiments must be done in order to find

one which best fits. To start this, choose a few algorithms. Then create a model based on the algorithm. After each model has been made analysis of each will be done to decide which is best suited for the data.

Dataset

The dataset being used has eight attributes with yield being used at the target. The seven other attributes are Year, Country Area Harvested, Production, Seed, Rainfall and Temp. All the attributes other then Country are numerical. Country just displays the country which the data was recorded from it doesn't give the model any information. It is only in the dataset to ensure that one dataset can quickly be disgusted from another.
Year is the year in which the data was recorded ranging from 1961 to 2013. Area harvested is the total area in hectares which was harvested in a given year. Production is measured in tonnes and is the amount of the crops grown that were processed after farming. Seed is in tonnes and is the amount of seed used for growing each year. Rainfall and Temp are both weather attributes they were not part of the same dataset as they both added after the collection of the crop data. Rainfall is the total amount of rain that for each year. The Temp is the total temperature for the year given. The target attribute is the yield it is recorded in hectogram/hectare. The yield is the return rate of crops in a season.

Before putting the data through a algorithm it will first be discretized. This is a process that changes the continues predictors into bins or bands. Rather than getting the exact value of the yield, binning the data allows to the algorithms to find the category that the it best suits. This experiment will discretizes the data into 5 bins and will use equal width binning. Equal width binning will try to split up the dataset into groups that are intervals of the same size apart. This method can be effect by values with outliers however for this experiment it will be suitable as there aren't any extreme values.

After this process is done visualising the data in a histogram will the output variable overlaid of each value gives a better understanding which variables as the most effect on the overall yields and which doesn't.

Visualization

Ireland

Just observing the yield graph shows what each colour represents. In this case pink shows the highest yield and blue shows the lowest. Looking at the year shows that the later years all had the highest category of yields. Comparing the two weather variables Temp and Rainfall, shows that rainfall may not affect the yields like may be expected as each bin has a spread out through the histogram.

While the same can't be said for Temp in the last two bins the most amount of high yield outputs.

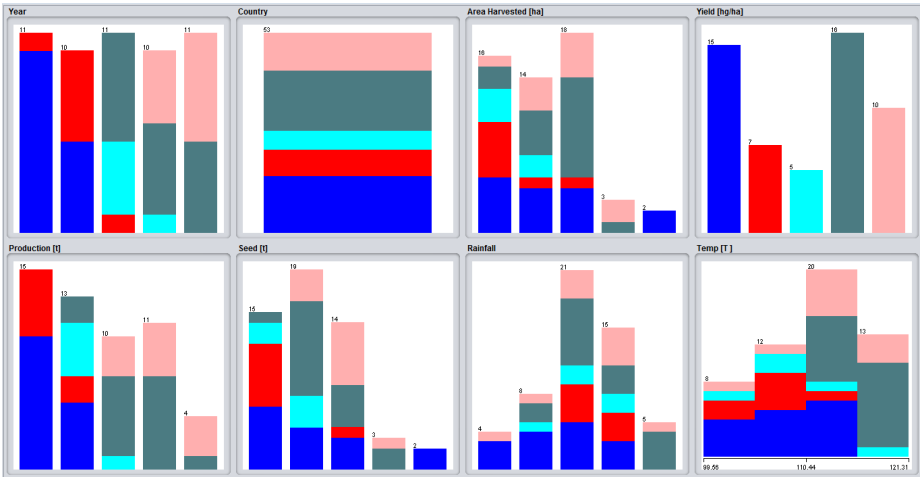

Figure 1: Ireland

India

Seeing the data give a good picture as to what has effected the yields over time. The year plays a big part in the higher yields, as the later year having the higher yields. The area harvested is interesting as it shows the highest area didn't have the highest yield. While the production shows the opposite. Of the two rainfall attributes rainfall didn't appear to have a big impact with the high yields spread out over the range, the temp attribute appears to have more of an impact as the higher total amount of sunshine the better the yields.

Algorithms

ZeroR

ZeroR is a algorithm that has no machine learning in it. It chooses the most popular value of the target variable and returns that value. If the dataset has 10 predictor variables the ZeroR return the most popular value. With a dataset of 10 it would be expected to correctly classify approximately 10% of instances as the most popular variable should always have at least 1/10th of the correct answers. The Zero R method is very useful for getting baseline for the dataset to compare other models against. As it will show it show what the worse model

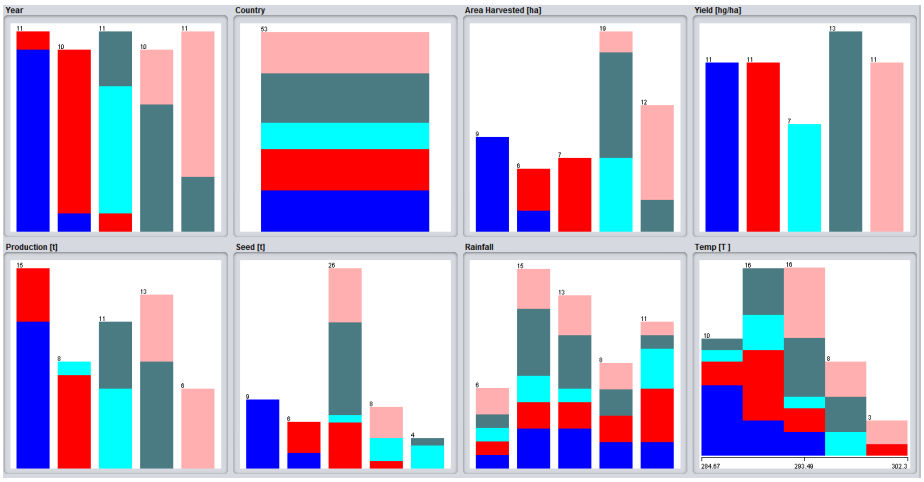

Figure 2: India

will predict, if there is a model that gets worse results then the baseline its no use as it doesn't show anything.

J48

Heuristic Is a very popular machine learning model to use known as C4.5 outside of weka. When the model is finished it produces a tree which can display the findings very clearly. This also makes it easy not from a data since background to understand the results. It works by creating a top down tree and splitting data into leaves. The goal when splitting the data into leafs to get a pure leaf. This means it only has one value from the target variable so can't be split, this shows which values are important to target variable. J48 uses information gain which shows which of the predictor variables has the target. This how the model which values to split on.

K-nearest neighbour(kNN)

K-nearest neighbour(kNN) is a widely known model and one of the most simply machine learning algorithms. It is an instance-based or lazy learner. This means it doesn't require any training as it does its leaning from the data in the currently being used. It works by analysing values around the target value. k is the number of values the model will consideration. If k = 1 then the model will only take one value and use that value as the output. If k = 3 the model will look at the 3 closes values and use the value that has the majority. For problems with only two classes k should always be an odd value to avoid ties.

In other cases, k should not be a multiple of the number of classes again this is to avoid ties.

Naive Bayes

Naive Bayes is a probabilistic classifier. It treats each value as equal and gives the probability of the outcome being in one class based on that attribute. The model will then make its prediction based on the best probability of the total attributes. Each attribute is statistically independent of each other, but the assumption made from the one attribute is rarely correct however, together it performs well in really world tests.

Results for wheat Ireland

ZeroR results Ireland Wheat

Correctly Classified Instances 11 20.7547 %

Incorrectly Classified Instances 42 79.2453 %

Confusion Matrix

a	b	c	d	e	<– classified as
5	0	0	10	0	a = '(-inf-44891.4]'
3	0	0	4	0	b = '(44891.4-58478.8]'
2	0	0	3	0	c = '(58478.8-72066.2]'
10	0	0	6	0	d = '(72066.2-85653.6]'
5	0	0	5	0	e = '(85653.6-inf)'

The results of the baseline test show that the zero R model got 20% instances correctly classified. This is a little over 1/5th which is the minimum expected for this model with only five classes. The results show that the model got 11 out of 53 predictions correct. The Confusion Matrix gives a clearer picture as to what the model chooses for each prediction. With a and d having the most values they were the only ones predicted. The Matrix shows that ZeroR got 5 correct for a and 6 for d. This means that out of the times it classified the output as a, 25 in total it got 5 correct. For d it selected 28 times and got it correct 6 times.

J48 Results Ireland Wheat

Correctly Classified Instances 29 54.717 %

Incorrectly Classified Instances 24 45.283 %

Confusion Matrix

a	b	c	d	e	<– classified as
11	4	0	0	0	a = '(-inf-44891.4]'
5	1	0	1	0	b = '(44891.4-58478.8]'
0	0	0	4	1	c = '(58478.8-72066.2]'
0	0	2	12	2	d = '(72066.2-85653.6]'
0	0	0	5	5	e = '(85653.6-inf)'

The results from the J48 model give a much better percentage of correctly classified instances then the baseline at 54 almost 55%. This is 2 and 1/2 times better with 29 correctly predicted. The root node of the model is the year this is what the model found to be the biggest factor in yields. This confirms what was observed when the data was visualised in the histogram with the yields overlaid. The next leaves show the Rainfall and Production is where the data can be split. Again, when the data was visualised it was clear that there was a split of the lowest to highest yields, with the highest yields being towards the left and lowest to the right. The rainfall however is surprising as from the graph it was not clear that it would provide a lot of information as the data was spread out within each group.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	C
0.733	0.132	0.688	0.733	0.710	0.590	0.880	0.741	'
0.143	0.087	0.200	0.143	0.167	0.065	0.727	0.371	'
0.000	0.042	0.000	0.000	0.000	-0.064	0.788	0.242	'
0.750	0.270	0.545	0.750	0.632	0.447	0.750	0.515	'
0.500	0.070	0.625	0.500	0.556	0.470	0.792	0.537	'

Weighted Avg. 0.547 0.147 0.504 0.547 0.518 0.393 0.795 0.539

Looking at the ROC area shows that the .795 is a good score however from the confusion matrix the model never picks a value for c. However, looking closer at the results to see what the model gave shows that it didn't do well at predicting values for b of c only give one to b and none for c. The model got a good ROC area of 0.795, it would require some work before being deployed in a live environment but could be reproduced.

Results kNN (k=1)

Correctly Classified Instances 24 45.283 %

Incorrectly Classified Instances 29 54.717 %

=== Confusion Matrix ===

a	b	c	d	e	<– classified as
9	5	1	0	0	a = '(-inf-44891.4]'
3	3	0	1	0	b = '(44891.4-58478.8]'
0	0	1	3	1	c = '(58478.8-72066.2]'
0	0	3	7	6	d = '(72066.2-85653.6]'
0	0	2	4	4	e = '(85653.6-inf)'

The correctly classified instances are lower than the J48 model however it is gives much better results than the baseline test at 45% correct. As kNN is an instance-based model it doesn't provide any information as to why values are classified how they were. The confusion matrix shows the values chosen were spread out more than J48. This is expected as it will look to the closes value as assign it that value.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	C
0.600	0.079	0.750	0.600	0.667	0.561	0.828	0.629	
0.429	0.109	0.375	0.429	0.400	0.303	0.720	0.357	
0.200	0.125	0.143	0.200	0.167	0.065	0.685	0.152	
0.438	0.216	0.467	0.438	0.452	0.226	0.607	0.429	
0.400	0.163	0.364	0.400	0.381	0.229	0.624	0.288	

Weighted Avg. 0.453 0.144 0.485 0.453 0.465 0.316 0.695 0.423

kNN doesn't have a high percentage of correctly classified instances, however it gives more answers for b and c then J48. Knowing how kNN works this is not a surprise. Looking at the TP rate shows that the model gets worse as the years go on, starting off with a high 0.600 and ended with 0.400. The ROC Area also goes down as the years become more recent with it stating at 0.828 ending at 0.624 with the overall 0.695. This suggest the model is better than guessing but needs some tuning before it could be deployed.

Results Naive Bayes

Correctly Classified Instances 33 62.2642 %

Incorrectly Classified Instances 20 37.7358 %

Confusion Matrix

a	b	c	d	e	<– classified as
12	3	0	0	0	a = '(-inf-44891.4]'
1	5	1	0	0	b = '(44891.4-58478.8]'
1	1	1	1	1	c = '(58478.8-72066.2]'
0	0	2	10	4	d = '(72066.2-85653.6]'
1	0	0	4	5	e = '(85653.6-inf)'

The results from the Naive Bayes model are the best with 33 values correctly classified at 62%. This is a good improvement over the baseline. The confusion matrix shows the breakdown of the values with the model correctly classifying 12 out of 15 for a True Positive rate of 80% this was the highest out of each category, with the next best being d with 10 out of 16. The weighted avg ROC Area for this model is 0.840 which gives a good indication that the model is doing a good job classifying the data.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	C
0.800	0.079	0.800	0.800	0.800	0.721	0.918	0.758	
0.714	0.087	0.556	0.714	0.625	0.566	0.820	0.402	
0.200	0.063	0.250	0.200	0.222	0.152	0.817	0.266	
0.625	0.135	0.667	0.625	0.645	0.499	0.814	0.601	
0.500	0.116	0.500	0.500	0.500	0.384	0.788	0.416	

Weighted Avg. 0.623 0.102 0.619 0.623 0.619 0.516 0.840 0.553

Results for wheat in India

ZeroR Baseline

Correctly Classified Instances 13 24.5283 %

Incorrectly Classified Instances 40 75.4717 %

Confusion Matrix

a	b	c	d	e	<– classified as
0	0	0	11	0	a = '(-inf-12194.2]'
0	0	0	11	0	b = '(12194.2-17089.4]'
0	0	0	7	0	c = '(17089.4-21984.6]'
0	0	0	13	0	d = '(21984.6-26879.8]'

a	b	c	d	e	<– classified as
0	0	0	11	0	e = '(26879.8-inf)'

For India the baseline had a better percentage then Ireland at 24% getting 13 out of 53 correctly classified. The class with the most instances was c with 13 so this was chosen as the result passed each time. The ROC Area is very low at 0.413 which is worse that just guessing what the outcome would be, however it is not expected to any better.

J48 Wheat India

Correctly Classified Instances 36 67.9245 %

Incorrectly Classified Instances 17 32.0755 %

Confusion Matrix

a	b	c	d	e	<– classified as
9	2	0	0	0	a = '(-inf-12194.2]'
1	9	1	0	0	b = '(12194.2-17089.4]'
0	1	5	1	0	c = '(17089.4-21984.6]'
0	0	4	7	2	d = '(21984.6-26879.8]'
0	0	0	5	6	e = '(26879.8-inf)'

67% correctly classified instances is an improvement over the model for Ireland. The TP rate overall was 0.679 which got worse at for the most recent years the same as the model for Ireland. Unlike like then model of Irish data the root node was not year, but Area harvested. The sub nodes were Temp and Year, this is completely different to Ireland. Also the classified instances had a different spread to the Irish model, however this is down to the classification then the model. The model had a good ROC Area score of 0.886 so could be reproduced in a live environment.

kNN (k=1) Wheat India

Correctly Classified Instances 41 77.3585 %

Incorrectly Classified Instances 12 22.6415 %

Confusion Matrix

a	b	c	d	e	<– classified as
9	2	0	0	0	a = '(-inf-12194.2]'
2	8	1	0	0	b = '(12194.2-17089.4]'
0	0	5	2	0	c = '(17089.4-21984.6]'
0	0	3	9	1	d = '(21984.6-26879.8]'
0	0	0	1	10	e = '(26879.8-inf)'

kNN gets a better result in India then Ireland at 77% correctly classified instances. The TP rate is 0.774 this may have been slightly inflated as a and e both had high scores of 0.818 and 0.909 while the b,c and d were 0.727, 0.714 and 0.692. One point of interest was the ROC area is 0.859 which is good was lower the J48 while having a higher correct percentage then J48.

Naive Bayes Wheat India

Correctly Classified Instances 42 79.2453 %

Incorrectly Classified Instances 11 20.7547 %

Confusion Matrix

a	b	c	d	e	<– classified as
10	1	0	0	0	a = '(-inf-12194.2]'
0	9	1	0	1	b = '(12194.2-17089.4]'
0	0	7	0	0	c = '(17089.4-21984.6]'
0	0	2	8	3	d = '(21984.6-26879.8]'
0	0	0	3	8	e = '(26879.8-inf)'

The naive bayes model had a very strong performance when tested with the data from India getting 42 of 53 Instances correct at 79%. The class that had the worst TP rate was also the class with the most instances, d. This could be down to there being more instances as c the class with the lowest had a perfect TP rate.

Conclusion

Testing the same models on the similar data gives interesting results. All the models including the baseline performed better when using the data from India then from Ireland. In both cases Naive Bayes model preformed the best, proving a better correctly classified percentage and better ROC area. This shows

that not only was the model able to give the correct answer most often but the results could be repeated for again for another dataset of a similar type. J48 and kNN showed interesting results as they performed differently in each experiment. When tested with the Irish data, J48 had a better percentage of correctly classified instances than kNN. When tested with the Indian data kNN was better. However the ROC Area for is worse kNN then J48. This suggests that although kNN has a better classification score that it may not be able to reproduce the same results with a different dataset. Performing the same test with a dataset with attributes that are closer to climate to Ireland. There is a large different between the temperature which don't effect the models as much as you would expect in Ireland. Also there is a difference in how Ireland and India developed as countries which may have also effected the tests. We see this as area harvested is the top node in the J48 tree for India rather then year in Ireland. This suggests that the quality of seed is not as good as that in Ireland. When the model was tested with the Ireland dataset it shows that year is the most important, suggesting that the farming methods changed which caused an increase in yields.