Sam Appleton, Elliot Rippe, Ben Beyer

CSCI 307

Constance Royden

4/25

A Study of Lung Cancer and Life Expectancy of Smokers

For this project we chose to analyze the relation between smoking and the development of lung cancer as well as overall life expectancy. We are searching for how attributes such as gender, age, years smoked, cigarettes a day, years quit, and education affects if someone dies from lung cancer or not. In order to do this we specifically chose a dataset that contained instances for current smokers, former smokers, and non-smokers. The dataset chosen classifies the outcome “Death codes” (0=alive, 1=death from other causes, 2=lung cancer death). It consists of 187 thousand instances and 9 other attributes (6 numeric: Age, CigarettesPerDay, YearsSmoked, YearsQuit, Freq, FollowUpTime and 3 nominal: Gender, Education, Smoker ). The values meaning for these attributes is as follows:

Age: (age on January 1, 1982)

Gender: (0=Male, 1=Female)

Education: (0=no college, 1=some college)

Smoker: (1=never, 2=former, 3=current)

Cigarettes/day: (values rounded UP to the nearest 5)

Years smoked: ( number of years smoked as of January 1, 1982)

Years quit: (number of years since smoking cessation, as of January 1, 1982 (zero indicates less than one year)

Followup Time: (years from January 1, 1982 until death or last interview)

Death codes: (0=alive, 1=death from other causes, 2=lung cancer death)

Freq: (the frequency at which each combination of variables occured)

A problem that we had found with the data provided for our dataset was that it was in 7 separate arff files. We easily fixed this by putting all of the instances into one big arff file. Each one of the seven parts had the same order or attribute values for each instance so formatting was not a problem. At the beginning of each was a description of the attribute values listed above along with the breakdowns of the seven different arff files which were:

* never-smokers (0.071%)
  + current smokers: male (11.98%)
  + current smokers: female (10.91%)
  + former smokers: male, no college (20.57%)
  + former smokers: male, some college (26.65%)
  + former smokers: female, no college (12.52%)
  + former smokers: female, some college (16.63%)

For our classifying algorithms we chose OneR, Naive Bayes, and J48 decision tree. We had chosen these classification algorithms because we believed from the results from each we could derive some meaning towards a real world application such as if someone would have went to college, they would be more susceptible to smoking more cigarettes. With the amount of instances and the types of attributes in the datafile, we believed that the results from the OneR classification would give us a defining rule for the target class. We also believed that the J48 decision tree would show us definite splits on numeric attribute values such as age or years smoked.

OneR is a simple, yet fairly accurate algorithm that generates rules for each attribute based on the majority outcome for each value in the form of “if <value> then <outcome>”. It then calculates the total error rates by taking the fraction of times the rule set for an attribute was incorrect. After this, the total error rates for each attribute rule set are compared and the set with the smallest error rate is chosen as the rule set for the entire data set. This algorithm is particularly effective when the chosen rule set has a low error rate.

Naive Bayes is an algorithm that assumes all attributes to be independent and that all attributes contribute equally to the outcome. In this algorithm, we make a table with the likelihood of each outcome for each value of each attribute. It then uses this information to find the likelihood of each outcome for a new instance by multiplying each applicable value’s likelihood from the table together with the overall likelihood of that outcome. It then converts those results into probabilities to determine which outcome is more likely for the new instance.

J48 is a decision tree that creates splits at decision points of numeric or nominal attributes that the algorithm calculates to be the point of divergence between target class values. Our J48 decision tree was enormous because of the multiple splits it created from the numeric data. It was not as simple as it would have been with multiple nominal attributes where the difference is very clear cut. However, the J48 algorithm calculated that there must be multiple splits as the numeric attributes each affected the outcomes of the instance in different ways.

For each algorithm we ran 4 different training set sizes. We ran each with 100% testing and 100% training, not to learn and infer from the data so much as to set a baseline. We also used 66% / 34% train test splits and 50%/50% train test splits. Lastly we did a ten fold cross validation.

**OneR Analysis**

The OneR analysis appeared to perform well. The classifying model seemed accurate and efficient through four runs. 1) 100% testing and 100% training, 2) 66% / 34% train test splits 3) 50%/50% train test splits and 4) ten-fold cross-validation. All four runs had an accuracy close to 95% and were built quickly (between .3 and .5 seconds build and test). All four data splits produced the same Classifier Model:

=== Classifier model (full training set) ===

Followup\_Time:

< 0.5 -> 0

< 5.5 -> 1

>= 5.5 -> 0

(177920/187109 instances correct)

The only variance in the results between data splits came from the percent chosen to train versus test on. The attribute Followup\_Time (years from January 1, 1982 until death or last interview) is chosen (with the same numeric values) for each of the four splits. While follow up time produces an accuracy of 95% in classifying new instances it does not seem to have much correlation with health expectations (specifically the development of lung cancer) in participants.

=== Confusion Matrices ===

|  |  |
| --- | --- |
| **(10-fold cross-validation)**  Correctly Classified: 177920 95.089 %  Incorrectly Classified: 9189 4.911 %  a b c <-classified as  159936 36 0 | a = 0  4931 17984 0 | b = 1  965 3257 0 | c = 2  **(66-34 split)**  Correctly Classified: 60455 95.0296 %  Incorrectly Classified: 3162 4.9704 %  a b c <-classified as  54378 11 0 | a = 0  1670 6077 0 | b = 1  340 1141 0 | c = 2 | **(Use training set)**  Correctly Classified: 177920 95.089 %  Incorrectly Classified: 9189 4.911 %  a b c <-classified as  159936 36 0 | a = 0  4931 17984 0 | b = 1  965 3257 0 | c = 2  **(50-50 split)**  Correctly Classified: 88872 94.9954 %  Incorrectly Classified: 4682 5.0046 %  a b c <-classified as  79893 18 0 | a = 0  2481 8979 0 | b = 1  512 1671 0 | c = 2 |

Lung cancer is never classified in the model produced by the rule based on Followup\_Time. This is the glaring flaw in using OneR with this data set. There are only 4,222 instances with the class value of lung cancer death (2) out of the 187 thousand instances in the data set. In this way the OneR rule *almost* functions similarly to ZeroR choosing the majority outcome value as the classification. Although OneR is both efficient and accurate throughout all four data splits we conclude it fails to give any insight into the life expectancy of current, former, and non smokers, nor into the likelihood of death from lung cancer. Followup\_Time does not relate to smoking habits specifically. Because of this the classifying rule created it not useful in analyzing causal relationships with death from lung cancer. If the Followup\_Time attribute is missing, the missing value is handled by choosing a rule determined by next lowest error rate. The rule chosen when run with a ten-fold cross-validation was Years\_Smoked. Years\_Smoked is a more useful attribute in studying the relationship between smoking and death from lung cancer. This can be seen in the classifier model for the OneR ten-fold cross-validation split run without Followup\_Time.

=== Classifier model (full training set) ===

Years\_smoked:

< 2.5 -> 1

< 64.5 -> 0

< 65.5 -> 1

>= 65 -> 0

There are still two issues with the model produced after eliminating Followup\_Time (either to handle a missing value or to find a rule utilizing an attribute with closer relationship to the outcome). The OneR model based on the Years\_smoked attribute still only classifies new instances as dead or alive, not dead from lung cancer. Here it is no more helpful than Followup\_Time. It is also less accurate running with 85% accuracy (10% less than the rule generated based on Followup\_Time). Also it appears the data may be overfit between 64.5 and 65.5 as the last two rules imply contradiction. Further study would be needed to analyze this. Despite the apparent accuracy and efficiency, OneR is not the optimal classification model for our data set due to the small percentage of “Death Codes” (outcome) values being lung cancer, about 4,200 compared to 187K instances.

**Naive Bayes Analysis**

Naive Bayes succeeded in some of the areas where OnerR failed. Two characteristics of Naive Bayes, equal importance and independence, allow for more accurate classification of instances with lung cancer. The confusion matrices below show the classifications for the four runs of Naive bayes with different data splits.

=== Confusion Matrices ===

|  |  |
| --- | --- |
| **(10-fold cross-validation)**  Correctly Classified: 161198 86.1519 %  Incorrectly Classified: 25911 13.8481 %  a b c<-classified as  142973 16800 199 | a = 0  4670 18200 45 | b = 1  601 3596 25 | c = 2  **(66-34 split)**  Correctly Classified: 55256 86.8573 %  Incorrectly Classified: 8361 13.1427 %  a b c<-classified as  49267 5064 58 | a = 0  1751 5982 14 | b = 1  237 1237 7 | c = 2 | **(Use training set)**  Correctly Classified: 160703 85.8874 %  Incorrectly Classified: 26406 14.1126 %  a b c<-classified as  142376 17378 218 | a = 0  4563 18301 51 | b = 1  581 3615 26 | c = 2  **(50-50 split)**  Correctly Classified: 81735 87.3667 %  Incorrectly Classified: 11819 12.6333 %  a b c<-classified as  73026 6815 70 | a = 0  2739 8697 24 | b = 1  380 1791 12 | c = 2 |

Despite yielding seemingly worse results than the OneR algorithm, only about 86% of instances were correctly classified (almost 10% lower than the ~95% of OneR) and the model was less efficient (most models training and testing runtime > 1 second), this model is better suited to our problem due to its ability to classify new instances as lung cancer death. Naive Bayes works fairly well in studying our problem. Yet while the assumptions made by the Naive Bayes classifier are relatively beneficial, the strong assumption on the shape of our data distribution (i.e. any two features are independent given the output class), can be harmful - hence, a “naive” classifier. No further insight is gleaned on how individual attribute relate to and affect the outcome. The Naive Bayes model is almost a middle of the road classifier algorithm for our analysis, it is significantly more efficient, yet less insightful (and accurate), than J48; at the same time it, it is less efficient (an accurate) than OneR but More insightful an analyzing the data for our problem.

**J48**

=== Confusion Matrices ===

|  |  |
| --- | --- |
| **(10-fold cross-validation)**  Correctly Classified: 178825 95.5726 %  Incorrectly Classified: 8284 4.4274 %  a b c<-classified as  159612 317 43 | a = 0  3921 18911 83 | b = 1  591 3329 302 | c = 2  **(66-34 split)**  Correctly Classified: 60767 95.5201 %  Incorrectly Classified: 2850 4.4799 %  a b c<-classified as  54243 130 16 | a = 0  1311 6417 19 | b = 1  201 1173 107 | c = 2 | **(Use training set)**  Correctly Classified: 179308 95.8308 %  Incorrectly Classified: 7801 4.1692 %  a b c<-classified as  159779 164 29 | a = 0  3725 19149 41 | b = 1  556 3286 380 | c = 2  **(50-50 split)**  Correctly Classified: 89294 95.4465 %  Incorrectly Classified: 4260 4.5535 %  a b c<-classified as  79765 115 31 | a = 0  2057 9361 42 | b = 1  326 1689 168 | c = 2 |
|  |  |

Our J48 decision tree ended up with 266 leaves and 530 individual total nodes. The initial decision split at the top of the tree was follow up time. The split was if follup time was less than or equal to 5 or if it was greater than 5. These results are similar to the result that the OneR classification showed where there were splits at 0.5 and 5.5. An interesting part of the decision tree generated is that if the follow up time is greater than 5, it then moves onto a node of whether or not the the instance has never smoked, is currently a smoker, or used to smoke. This intuitively seems like one of the more important attributes but the decision tree continues to split and classify the instances based on various numeric values.

In terms of accuracy, the J48 decision tree classified around 95% average over the 4 different methods of training and testing that we used. In all of the J48 methods, it classified the target class of death by lung cancer the most times our of all three classification algorithms. This makes it seem like it is the most useful for us in this context, but it is the not efficient. The run time for OneR and Naive Bayes was less than 0.5 seconds, while the run time for J48 was a little more than 11 seconds. J48 also takes up more data in comparison to the other more efficient algorithms.

**Conclusion**

Each of the four analyses offered its own benefits and drawbacks. OneR was the most efficient algorithm and ran surprisingly accurately. However due to the lack of lung cancer outcomes in the data set for OneR to train on it was the least helpful algorithm for our purpose of studying the relationship between smoking and lung cancer.

The Naive Bayes model functioned as a compromise between OneR and J48. Ironically the assumptions of independence and equal weight account for both its success (compared to OneR) and its failure (compared to J48). Iit is efficient, yet less insightful (and accurate), than J48; at the same time it, it is less efficient (and accurate) than OneR but More insightful an analyzing the data for our problem. Less than 26 instances were correctly classified as death by lung cancer in each data split whereas over 100 were correctly classified as death by lung cancer in J48. Although J48 is the least efficient classifying model to build and generates a rather large tree with complex rules, it is this visualization and the complex structure of rules that give the most insight into the relationship between smoking (the attributes) and death by lung cancer. J48 is the best classifying algorithm to study this problem with the given data. With the results of the project, we can provide a solid analysis and evidence for how certain attributes contribute to death by lung cancer. We can also use this as evidence to deem whether quitting for a certain period of time reduces lung cancer or not. Each individual attribute can be analyzed in correlation with the target class of alive, dead, or death by lung cancer to see how much of an impact they cause on said target class.