An Analysis of Hepatitis Through Data Mining

Thiago Clari Bryan Irlbeck

Abstract

With our collected data and calculated results, we strive to gain insight on correlating attributes that lead to serious cases of hepatitis. We also wanted to try and identify which attributes aided patients or was a detriment to them. Our methods of finding this data included two classification algorithms, one association algorithm, and one clustering algorithm. The two classification algorithms are, as follows: ZeroR and OneR. Likewise, our association algorithm used was Apriori, and our clustering algorithm was K-Means Clustering. Some of our key results were the uncovering of the strong relationship between high alkaline-phosphatase levels and low albumin levels, and another result is that the pairing of these attributes along with high bilirubin levels lead to a higher chance of death in patients. With our analysis of the dataset and given attributes, we conclude that there is a strong connection between some attributes and the wellbeing of patients that exhibit some of these symptoms as well as ones that do not.

1 Introduction

With our findings, we are looking to find common symptoms between serious hepatitis patients, as well as which attributes are present when others are as well. Along with this, we are striving to uncover leading causes of death from hepatitis in our dataset. We believe this is interesting because we will be trying to bring to light certain attributes that contribute to the well-being or death of hepatitis patients. We are hoping to find a strong underlying cause that will help identify a patient that would be at risk of dying.

Before we go in-depth with the algorithms, we want to talk about the data cleansing we performed on the data set. After some research on the attributes and exactly what hepatitis is, it was obvious we had some attributes that were skewing our data and were common amongst all the patients. After finding this out, we went ahead and trimmed the total attributes from 19 to 14. Since we had some numeric data and needed all nominal, we had to normalize some of our attributes such as bilirubin and albumin. After this normalization, we also added some key attributes to give the numeric data more weight, as they were not influencing our results as expected. These added attributes were 'highBilirubin', 'lowAlbumin', and 'highPhosphate', and will be explained in further detail later on in the paper.

2 Methods

After our data was ready to be worked on, we started our experiments with the popular classification algorithm, ZeroR. This algorithm was a great pick for our testing because it provides a useful baseline to compare other algorithms against. After this, we moved on to OneR, which is used to generate one rule for each predictor and then selects the rule with the smallest total possible error. When using this algorithm, we found that using OneR with lowAlbumin as the predictor yielded us the most accurate results. We also tested OneR with highPhosphate, and found that it also presented us with very accurate results. We later discovered that the tie between these two attributes was more important than we had previously realized, and will expand upon this later in the paper.[1]

Next in line was the famous Apriori association algorithm. We chose this algorithm to identify key associations between attributes, and later used these key attributes with OneR to confirm the importance and weight of these certain attributes. This came to be useful, as we identified that the attribute 'lowAlbumin' was a very good predictor with OneR. We conducted the algorithm in Weka, using a minimum confidence of 80% and minimum support of 65% to weed out poor associations. This was productive, as we found multiple rules with confidences all above 92% confidence, most were even higher.

After Apriori, we used K-Means Clustering to try and uncover any associations that we may have overlooked, and it was well worth it. We discovered a very strong association between previously dismissed attributes, and were able to almost perfectly cluster together the entirety of deceased patients in a separate cluster from the surviving patients. During our testing, we tried various cluster counts but decided that a k-value of 4 was the best identifier of unique and reliable clusters. We gained insight into hidden correlations that were previously dismissed, and this helped us look back to other algorithms; testing for these previously dismissed correlations helped us further confirm our findings that we were considering using. During our testing, we did further research into the attributes that were showing correlation, and these clusters continued to show convincing results each time we tested against them.[2]

3 Results

To begin our results, we started our algorithm runs with K-Means clustering to try and gain insight on some relationships that we had not yet considered. We used a percentage split of 66% with a k-value of 4, and the results were very useful. In our first cluster, we saw that all the symptoms of a serious case of hepatitis were present- namely 'highPhosphate', 'highBilirubin', and 'lowAlbumin' were all true. This was surprising, as we knew these attributes were all signs of an endangered hepatitis patient, but the key observation was that they were all present at the same time in this cluster. The reason this was so surprising is that these attributes were false in all 3 of the other clusters; any cluster that didn't have one of these attributes as true had any of the others as true either. When looking at the first cluster in the graph, we saw that most people in this cluster were the patients who died, which made these attributes telltale signs of a dying patient. This finding helped us guide our other algorithms in a different direction; instead of focusing on the class and associations with this, we focused on the relationships between these key attributes and their effects on results and ties to other attributes.

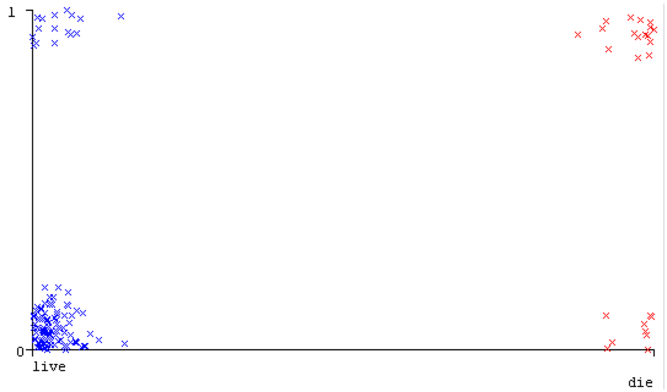
Our results carry weight over all aspects of our study, and while we did some first, some other algorithms done later helped us gain insight into other iterations of our previous algorithms. For example, our findings in K-Means with the correlation between 'highPhosphate', 'highBilirubin', and 'lowAlbumin' pushed us to retest with other previous algorithms. Using a min confidence of .80 for our apriori tests, we managed to get significant results. With our best rule sitting at a .96 confidence, we see that having low phosphate levels and high albumin levels leads to living patients; the science explaining why these levels are important will be discussed later in our analysis.

[1] All classification algorithms ran through weka used a cross-validation 5 fold test option.

[2] All clustering tests run through weka used a 66% percentage split test option.

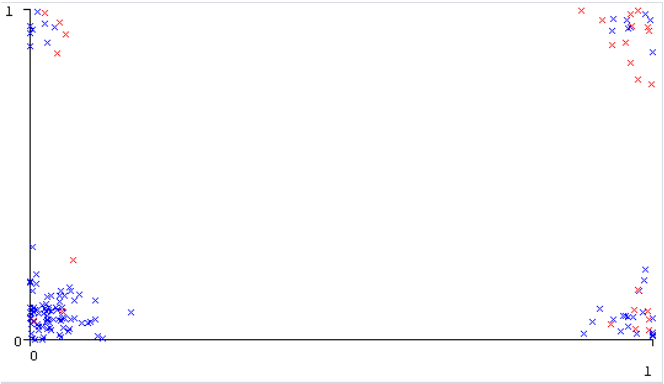
Now visualizing our data, we can see that having low albumin levels seem to have a contribution to the deaths of hepatitis patients. We can see that the blue clusters, which are the class that lives, do not normally exhibit the attribute of ‘lowAlbumin’ being true. Likewise, we can see that the red clusters, which are the class that dies, commonly exhibit the attribute of ‘lowAlbumin’.

[3]



Since our apriori rule revealed that ‘lowAblumin’ and ‘highPhosphate’ are a dangerous combo to the patients, we wanted to visualize that data and see if there are any apparent connections. Here we can see that many of our patients who died seemed to have either ‘highPhosphate’, ‘lowAlbumin’, or both.

[4]



We have promising results so far, but that’s only from our apriori and clustering tests. We started with a simple zeroR classifier so we can see how split our class is between living and dead. We get 80% correctly classified, that being 124 living, leaving 20% incorrectly classified for the 31 dead instances. After ZeroR, we started running OneR to find our best predictor, and we found that ‘lowAlbumin’ was an outstanding predictor. It managed to correctly classify 98.56% of instances and there were no incorrectly classified dead instances. This shows us that the ‘lowAlbumin’ attribute might be the key component to achieving our goal. However, further tests with OneR shows us that ‘highPhosphate’ has an amazing 89.26% of correctly classified instances, also with no incorrectly classified dead instances. Hence why we are trying to find a connection with ‘lowAlbumin’ and ‘highPhosphate’. These were great results, as they further confirmed the relationship found between them with our apriori algorithm.

Before we get started on our discussion, we wanted to dive into some of the research we did beforehand and extrapolate on how this influenced our experiments, as well as some of our data cleansing methods. First, we analyzed each attribute and discovered that some attributes were almost always going to be present in the patients, given all the data is from patients who had hepatitis. Knowing this, some data we had was skewing our results into models that revealed completely obvious relationships. One example of this was the presence of the attribute ‘varices’, which was present in 88% of all patients. This is an extremely common symptom of hepatitis, so we decided to remove this attribute, as well as ‘sex’, ‘steriod’, ‘antiviral’, ‘liver big’, ‘ascites’, ‘protime’, and ‘histology’. This ended up being a major part of our data cleansing. To make up

[3] Weka generated graph. The X-axis is the class and Y-axis is lowAlbumin.

[4] Weka generated graph. The X-axis is highPhosphate and the Y-axis is lowAlbumin.

for these lacking attributes, we did some research to give the continuous data more weight since it was not being very influential in our early experiments. For example, the attribute 'albumin' displays the level of albumin (a protein made by the liver) in a patient's bloodstream and plasma. If found in low levels, this can pose a significant threat to the patient. Based on this finding, we decided to write a program that analyzed these levels and categorized these levels in a binary attribute called 'lowAlbumin'. If patients were under a certain level in the data, the program marked this patient's 'lowAlbumin' as true, and marked false levels higher than the recommended threshold of 3.5.[5] We added this attribute and did the same with other continuous data such as 'alk-phosphate' (highPhosphate) and 'bilirubin' (highBilirubin). Just as we set a limit of differentiation with 'lowAlbumin' at 3.5, we also set a level of 130[6] for 'alkPhosphate' to differentiate between 'highPhosphate' equaling either true or false. The same was done with bilirubin levels, with a point of interest at 1.2.[7] These points of interest are not arbitrary, as we did research into the healthy levels deemed by multiple sources, which will be present in the bibliography/sources page. These attributes helped this data gain lots of weight, and opened up the door for use in other algorithms such as K-Means, which produced great results because of these additions.

Due to these changes and retesting with this new data, we discovered deeper, underlying relationships that lied in between our attributes and our class outcomes. One such relationship would be the connection between the attributes 'albumin' and 'alk-phosphate'. We noticed with our new attributes we added (lowAlbumin and highPhosphate) that when both of these are true, patients tend to die much more often than others, no matter what other attributes they may exemplify. For example, there was a patient that exhibited both of these symptoms, and while many of the other attributes that are available in our dataset were false (ex: 'spiders', 'anorexia', 'malaise'), the patient still died. We want to present the theory that these two attributes (lowAlbumin and highPhosphate), when true together, lead to a very high chance of death in hepatitis patients.

This theory is backed by all of our algorithms, and here are some examples. Our OneR, K-Means, and apriori algorithms heavily favored that low albumin levels and high alk-phosphate levels were extremely dangerous on our hepatitis patients. As stated earlier, with OneR as the algorithm and the predictor as 'lowAlbumin', the accuracy of correctly classified instances was an astonishing 98.5%, with no incorrectly classified instances of dead patients. Along with this, our theory is also backed by the OneR of 'highPhosphate' with an accuracy of correctly classified instances at 89.26%, with no incorrectly classified instances of dead patients. The apriori algorithm also supports that the other symptoms, while serious in some cases, are much more treatable than the new attributes we added. We observed that even when 'spleen pal' and 'spiders' are true or 'anorexia' and 'malaise' are true, the patient ends up living. Based on this run and the support from the other algorithms, we made the generalization that regardless of what other symptoms you are experiencing, if your albumin levels are low and your alk-phosphate levels are high, your chances of dying are much higher.

[5] Moman, Rajat N., et al. "Physiology, Albumin." *PubMed*, StatPearls Publishing, January 2021 (pg 1).

[6] Yazdi, Puya, "High Alkaline Phosphatase Symptoms & How to Reduce It." *SelfDecode,* March 2, 2021. Accessed via website at ("https://labs.selfdecode.com/blog/alkaline-phosphatase/").

[7] Gill, Karen, "What Causes High Bilirubin?" *Healthline,* November 12, 2018. Accessed via website at ("https://www.healthline.com/health/high-bilirubin").

We also observed that while ‘highPhosphate’, ‘lowAlbumin’, and ‘highBilirubin’ were telltale signs of risk for a liver patient, these attributes seemed to show up hand-in-hand with each other. In other words, if a patient has ‘highPhosphate’ as true, it would be very likely that ‘lowAlbumin’ and ‘highBilirubin’ would be true as well. Therefore, the relationship between these variables is almost undeniable. To confirm this even further, we conducted a K-Means algorithm with the intention to see the relationship between these variables within the different clusters. We observed that when ‘lowAlbumin’ was false in a cluster, ‘highPhosphate’ and ‘highBilirubin’ were false as well, and we saw this occur in three out of our four clusters. The only cluster where this was not the case further increases our confidence in this relationship. When ‘lowAlbumin’ was true, we observed that ‘highPhosphate’ and ‘highBilirubin’ were also true.

We believe that the results fairly speak for themselves. With such a distinct relationship between these key attributes, it is hard to deny that the leading cause of deaths in hepatitis patients tends to stem from these very attributes. Likewise, we also believe that the other attributes, such as ‘spiders’, ‘anorexia’, and ‘malaise’ are not as important, and this is understandable, as these symptoms are very common in the dataset and are treatable with modern health practices and therapies.

4 Conclusion

Based on the data collected and analysis of that data, there is a strong connection between the albumin/alkaline phosphatase levels and if our hepatitis patients either live or die. Likewise, we observed that these attributes, when true, tend to show up together as a group and not individually. However, there are factors to keep in mind, one being that we only had 155 instances to test on. This is a very small dataset in relation to how many individuals have hepatitis in the general population, and if given another chance, we would have picked a larger dataset with more attributes to observe. Another factor is that hepatitis has many efficient therapies and treatments, so getting enough instances of patients who died made it harder to formulate key relationships between death and different symptoms and attributes. If we were given another opportunity to replicate this study, there are a few things we would improve on. First, we would prefer to have a bigger dataset, namely with more instances and more attributes. Secondly, during our research, we discovered that while our dataset provided some significant data, there are key symptoms of hepatitis patients that were not present as attributes. Nonetheless, we believe that the theory we have presented is significant, and the results of our algorithms definitely back up our claim. We have also uncovered that while some symptoms may seem dangerous and alarming, if they are not high Phosphate levels and low Albumin levels, then the symptoms tend to be far less life-threatening than one would believe.

5 Bibliography

1. Gong, G, “Hepatitis.” (1988). *UCI Machine Learning Repository*, November 1, 1988. Accessed via website at (“https://archive-beta.ics.uci.edu/ml/datasets/hepatitis/”).
2. Moman, Rajat N., et al. “Physiology, Albumin.” *PubMed*, published by StatPearls Publishing, January 2021 (pg 1).
3. Gill, Karen, “What Causes High Bilirubin?” *Healthline,* November 12, 2018. Accessed via website at (“https://www.healthline.com/health/high-bilirubin”).
4. Yazdi, Puya, “High Alkaline Phosphatase Symptoms & How to Reduce It.” *SelfDecode,* March 2, 2021. Accessed via website at (“https://labs.selfdecode.com/blog/alkaline-phosphatase/”).
5. Lee, William M., Stravitz, Todd, Larson, Anne M., “Introduction to the Revised American Association for the Study of Liver Diseases Position Paper on Acute Liver Failure 2011.” *AASLD Position Paper,* published by AASLD, December 2, 2011 (pg 1 - 8). Accessed via web at (“https://www.aasld.org/sites/default/files/2019-06/AcuteLiverFailureUpdate201journalformat1.pdf”).
6. “Hepatitis C Information.” *Centers for Disease Control and Prevention*, published by CDC, August 7, 2020.
7. Accessed via website at (“https://www.cdc.gov/hepatitis/hcv/hcvfaq.htm#section2”).