Rosalie Sherry

OMSBA5067

May 9th, 2021

Milestone 2

**Review 1**

*Supervised Machine Learning: A Review of Classification Techniques* by SB Kotsiantis

I choose this article to look at classification techniques as with my wine dataset I want to be able to classify the different types of wine. This looked like a good start in my understanding. I was pleasantly surprised to find that a good portion of the techniques for classification discussed were ones we had learned about in this class. To begin, this article discusses the general issue with supervised learning. Here, I would like to point out the author does not mention having “bad data” as an issue. This to me is pretty important as someone who knows that my dataset will already be complicated to classify because there is a type of wine that is a “red blend” which means it uses a large variety of different grapes but just because a wine uses more than one grape type, does not automatically mean it is a “red blend”. “Red Blends” can even be heavy on a certain grape which means I would need a massive dataset in order to properly classify “red blends” versus say a cab. The article does not touch on not having a big enough dataset. The article then goes into logic-based algorithms, followed by statistical learning algorithms. This made me realize how useful a decision tree can be in manually checking my classification algorithm later. I have gotten the most out of the combining classifier section and comparing learning techniques section as they started to really cement how I was going to go about my project. It concludes by discussing the pros of using multiple methods and discussing how to best tackle “databases”.

**Review 2**

*Prediction of Diabetes using Classification Algorithms* by Deepti Sisodia and Dilip Singh Sisodia

I choose this article because it also uses the UCI database, though a different dataset, to classify data using three methods we’re familiar with, Naïve Bayes, Decision Tree, and SVM. This paper is specifically looking at the emergence of diabetes in pregnant women, but they don’t classify it as gestational diabetes, which is known to be more likely in certain races and ethnicities than others. I think they could improve their algorithm should they also take that into account. The paper then goes through the three method, describes how they perform it measure which one is “the best” by calculating the precision, recall, F-Measure, accuracy, and ROC. Both the F-Measure and ROC were new terms to me. Looking at the F-Measure I think that it would be a good way to measure my own classification algorithms for my wine dataset since it is the weighted average of precision and recall. It concludes that Naïve Bayes is the best individual algorithm to use in classifying. On another note, I also want to critique the graphs they use to display their information, coming from a data science/analytics program their Figure 4 graph is hard to view. I would think a side by side versus a 3D graph would be a better way to display something but putting two graphs on a 3D plane make it hard to compare each data point.

**Review 3**

*Comparative analysis of image classification algorithms based on traditional machine learning and deep learning* by Pin Wang, En Fan, and Peng Wang

This article I chose because of this week’s topic on CNN and when I read further found it did use CNN as a classification method (as well as SVM). This paper looks at classification of images which is interesting to me because I have a background in image classification for machine learning algorithms (not the coding though, manually verifying, *much* ***more fun***). I appreciated this article because I think it helped me in understanding neural networks better. A thing I appreciated in this article is how it does not assume someone reading it has a vast knowledge of understanding of the field. Every term was explained I think in a way that someone with little to no knowledge would be able to follow it. Again, I am going to critique the graphs they use to display their data because it has a color scale off to the side which I assumed was initially to show one of the scores but looking further it looks like it serves no purpose. The paper goes on to compare not only accuracy but also time spent on the algorithms which is a subject brought up in the initial paper I read (weighing times of algorithms) and I find it interesting possibly diving further into the subject of time and accuracy of the different classification algorithms. It goes on to say that traditional machine learning algorithms work great for small datasets, but you need more advanced ones for large scale datasets. Again, this brings me back to the first paper I reviewed which discussed at the end the error and possibility of testing large databases.

**Review 4**

*Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set* by C. Lakshmi Devasena, T. Sumathi, V.V. Gormathi, and M. Hermalatha

This one uses another UCI dataset which again was part of the appeal. This one used rule-based classifier which I did not see myself much using in my own before, but this paper basically assured me that this would be a poor choice for my own classification. Maybe in part due to not having been familiar with many of the rule-based classifiers used before this point. They use 9 different classifiers, a few of them drawing from Naïve Bayes. This paper uses Mean Absolute error, Root Mean-Squared Error, and Classification Accuracy to measure effectiveness which are errors we use frequently in class. I think my overall critique of this paper is it is rather short. It has a lot to it and goes into detail explaining what each part means but where it could be expansive in the results category and conclusion it is fairly short. I think they could go further into what everything means and why they believe certain things worked over others especially since they lay the groundwork of what all 9 classifiers are. I have no idea why they might believe the NNGE classifier works the best and ultimately don’t feel better about rule based classifiers, just more aware that they exist and are much more complex than I initially believed.

**Review 5**

*Analysis of white wine using machine learning algorithms* by Manisha Koranga, Richa Pandey, Mayurika Joshi, Manish Kumar

I chose this paper because it aligns closely with my own thoughts of my DTC. Instead of using Italian Red Wine, this one uses white wine from Portugal. I thought I could learn a lot from what they use in order to better do my own classification later. My mine issue in this paper is it uses a score of 1 – 10 to classify wine as “good” or “bad” which is incredibly subjective. I find an issue in this not being an objective dataset to begin with with this number making it hard to continue using the same algorithm because other scorers may provide different scores to the same wine or score differently new wines that the original scorers would not have. With this I find the dataset inherently flawed. This paper uses 5different ML classifiers of which we are familiar with in this class is Naïve Bayes, Support Vectors, and MLP (which I think is a variation of Neural Networks from its description). Again this paper like paper 2 uses the F1-Measure in addition to the ROC to measure the successfulness of the different classifiers which I think I will use in my paper as well seeing its uses across these paper. These classifiers almost all were rather successful besides the SVM which makes me question using it in my own classifier problem. This one uses side by side graphs which makes me a lot happier compared to the other papers who used weird 3D graphs to display their data. They concluded that ultimately there was no “best” algorithm for classification and regression but the best for regression was J48 and MLP and the best in comparison to everyone else for classification is J48 and Random Forest (two I was not familiar with before this paper).