Alyssa Dalton

Executive Summary

This report evaluates the use of fuzzy logic compared to hard classification methods. The manipulated cluster data set was used for this comparison. The manipulated cluster set is an artificial data set with four attributes, all which have very similar standard deviations. (Aleshunas, 2011) The advantage of using an artificial data set is that it can be manipulated to clearly test the performance of the fuzzy logic algorithm. For the sake of simplicity, Class 1 was removed because the maximum and minimum values of its attributes overlapped the other classes, sometimes almost entirely. Also, only the attribute with the highest correlation to class was used, again to simplify classification. Three extreme outliers were removed, to simplify classification.

The fuzzy logic algorithm first separates the data into a training and test set. The training set is composed of 80% of the data and the test set contains the remaining 20%. Once the data is randomly separated, the training set is used to develop a rule set for classification. The attribute with the highest correlation to class is used, and the data is sorted by this attributed. When numeric ranges give a clear indication of class, a rule is created and that data is removed. The remaining data is then sorted by the attribute with the highest correlation to class. Again, where numeric ranges give a clear indication of class, a rule is created and the data is removed. This is repeated until numeric ranges do not give a clear indication of class. When this was the case, the numbers of instances per value is calculated and rules giving the instance partial classification to class are developed. When the data is composed of continuous values, the overlap region is divided into discrete ranges. For each range, the number of instances is counted for each class. Once this is done, the number of instances per class is divided by the total number of instances to find the percent membership an instance would have; and this result is expressed as a rule.

When the manipulated cluster set was run through the fuzzy logic algorithm, the following rule set was the result:

1. If C < .-.44 the instance belongs to Class 4
2. If -.44≤ C ≤ .25 the instance belongs to Class 3
3. If C > 9.41 the instance belongs to Class 2
4. If .26 ≤ C ≤ 2.54 instance belongs 10% to Class 2 and 90% to Class 3
5. If 2.54 < C < 4.82 instance belongs 50% to Class 2 and 50% to Class 3
6. If 4.82 ≤ C ≤ 7.10 instance belongs 78 % to Class 2 and 22% to Class 3
7. If 7.10 < C < 9.41 instance belongs 96% to Class 2 and 4% to Class 3

When the test data is run through the rule set, an instance that is classified into a partial classification will be considered part of the class with the highest partial classification value. This rule set gave a 5.4% error for the test data set.

This result was then compared to the C4.5 decision tree induction algorithm, which is based on entropy. The same training data was run through this algorithm and the following rule set was the result:

```
Rule 1:
Attribute C <= -17.8466
-> class4 [98.4%]
Rule 7:
Attribute C > 6.28904
-> class2 [95.7%]
Rule 6:
Attribute C > -17.8466
Attribute C <= 6.28904
-> class3 [88.3%]
Default class: class3
```

This rule set also gave a 5.4% error rate.

Although both rule sets gave the same error rate, the fuzzy logic instances that were considered “errors” cannot be considered strictly errors. The instances that gave error had partial membership to their actual class according to the fuzzy logic rules. Therefore, both methods give accurate classification rules; however, the fuzzy logic rules provide a higher level of informational detail about instances that fall into the numeric overlap region.

Problem Description

This analysis evaluates the use of fuzzy logic compared to hard classification methods. This is done by determining the results of the fuzzy logic algorithm when applied to the manipulated cluster data set compared to the results determined by the C4.5 decision tree induction algorithm.

Analysis Technique

The data mining technique used in this analysis falls under the category of classification. “Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”. Popular classification techniques include decision trees and neural networks.” (Chapple, 2011)

Traditionally classification methods predict that a particular instance can strictly be categorized into one class. Using the example above, a day can only be predicted to be sunny or cloudy- not both. This is known as hard classification. Fuzzy logic, on the other hand, is a machine learning technique that allows an instance to belong partially to more than one class. For example, a day can be classified as partly sunny and partly cloudy.

The fact that fuzzy logic allows partial membership helps address logical issues that hard classification does not. For instance, a bank may want to come up with a rule set to determine whether or not someone is a good candidate for a loan (low risk). One of the criteria might suggest that someone who is under the age of 25 poses much more risk than someone who is currently 25 or above. A hard classifier would make a rule that says that a person is not a good candidate for a loan if they are under the age of 25. However, if the person's 25th birthday is the next day, they do not suddenly go from being a bad (high risk) candidate to a good (low risk) candidate. Rather as they approach their birthday, they gradually become a better (lower risk) candidate. Fuzzy logic would represent this by them having a low partial membership as a good candidate far away from their 25th birthday, and higher partial membership as their 25th birthday approaches.

When using techniques such as a decision tree to classify data, overtraining of the data is often an issue. Overtraining happens when a classification method uses its training data and creates leaves and nodes so that each of the training instances is correctly classified. The problem with this is when new data is run through, the rules are too specific to the training set. This tends to cause more error in classifying new data, especially when the data contains noise. To avoid this, some decision trees are pruned, which means that some of the nodes are removed so that the tree can better classify new data. Another approach to avoid over training is to use fuzzy logic.

The Iris data set contains 150 instances. Each of the instances is classified into Iris-setosa, Iris-versicolor, and Iris-virginica. There are 50 instances in each class. Each instance has values for four attributes: petal width, petal length, sepal length and sepal width. (Frank & Asuncion, 2010) From previous analysis using correlation to class, it was found that most instances can clearly be separated by class because there is a fairly clear numeric range of petal width values for each class; however there is an overlap region in the versicolor and virginica values. When using a hard classifier, the overlap region is often misclassified. Fuzzy logic, however would show that an instance having partial classification in two classes. This is not the probability of an instance belonging to a class. (Fuzzy logic trees 1.0)

From previous analysis, the Iris data will most likely classify Iris-setosa clearly, and the overlap region for Iris-virginica and Iris-versicolor will give partial classification. The graph [figure 2] below shows the partial classification for an instance with petal width values ranging from 1.4 to 1.8.

value	No. of instances	Versicolor instances	Virginica instances	% versicolor	% virginica
1.4	8	7	1	87.5	12.5
1.5	12	10	2	83.3	16.7
1.6	4	3	1	75	25
1.7	2	1	1	50	50
1.8	12	1	11	8.3	91.7

Figure 1: Iris Partial Membership

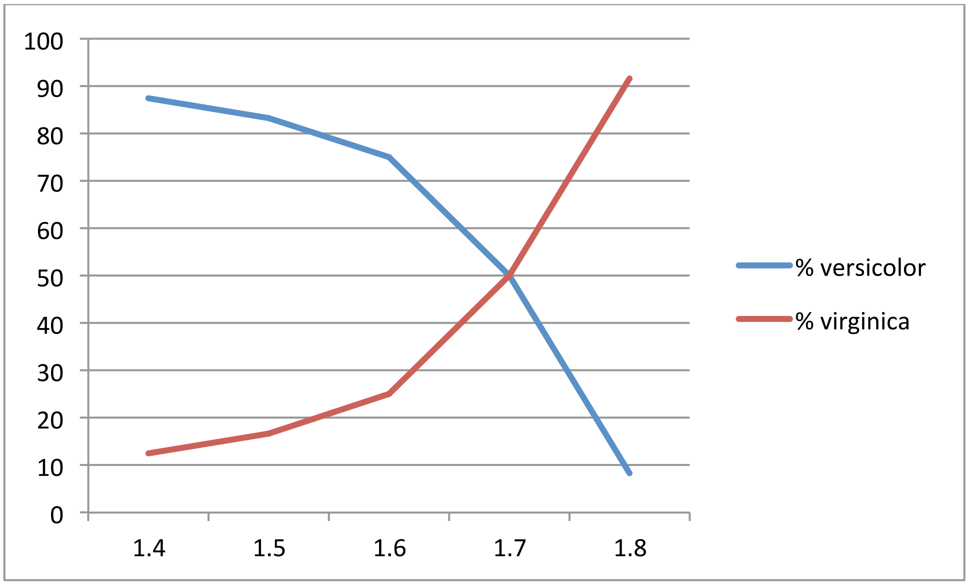

Figure 2: Iris Partial Membership

Therefore if an instance had a petal width of 1.5, it would be classified as 83% versicolor and 17% virginica. When calculating error for an instance, the class that the instance has most membership will be considered its class. The advantage of using fuzzy logic is that if this may not be the class, it could be determined what the true class is based on the partial membership of the other possible class.

The data set used in this analysis is the cluster data set. It is an artificial data set with four attributes, all which have very similar standard deviations. (Aleshunas, 2011) The advantage of using an artificial data set is that it can be manipulated to clearly test the performance of the fuzzy logic algorithm. For the sake of simplicity, Class 1 was removed because the maximum and minimum values of its attributes overlapped the other classes, sometimes almost entirely. Also, only the attribute with the highest correlation to class was used, again to simplify classification. Three extreme outliers were removed, to simplify classification.

The cluster data set was broken up into training and test sets. The training set included approximately 80% of the given data, and the test set included the remaining 20%. This allowed the fuzzy logic algorithm enough data to accurately come up with a rule set while still leaving enough data to test how well the fuzzy rule set classifies. The data was broken up by assigning a random number to each instance and sorting the data by this number. This was done five times to ensure the data was indeed in a random order. The first 74 instances were assigned to be the test data and the last 295 instances were assigned to be training data.

Once the data was separated into a test and training set, the training set will be used to develop a rule set. The correlation between each attribute and class was calculated. The attribute with the highest correlation to class, Attribute C, was used to sort the data. Where numeric ranges give a clear indication of class, a rule was created and the data was removed.

The data was then be manipulated by making the numeric ranges of that Attribute C overlap. This was by calculating the average, maximum and minimum values for Attribute C in each class. The averages were then made closer together by adding or subtracting a value (-22 from values in Class 2 and -20 from Class 4) from each instance. Once the data overlapped sufficiently (about 40%), the data was again analyzed using correlation. Where numeric ranges give a clear indication of class, rules were developed. Class 4 gave a clear indication of class. The following rule was the result:

1. If $C < .-.44$ the instance belongs to Class 4

The numeric range that gave clear indication of belonging to Class 3 was from -0.44 to 0.24. The following rule was the result:

2. If $-.44 \leq C \leq .25$ the instance belongs to Class 3

Lastly, if the instance had a value of 9.43 or above, there was a clear indication of the value belonging to Class 2. The following rule was the result:

3. If $C > 9.41$ the instance belongs to Class 2

Where there was an overlap region, the numbers of instances per value were calculated and rules giving the instance partial classification to class were developed. Since the data was composed of continuous values, the overlap region was divided into discrete ranges. For each range, the number of instances was counted for each class. Once this was done, the number of instances per class was divided by the total number of instances to find the percent membership an instance would have. The data is as follows:

	No. of instances	class 2	class 3	%class 2	%class 3
.25 - 2.54	29	3	26	10	90
2.54 - 4.82	8	4	4	50	50
4.82 - 7.10	27	21	8	78	22
7.10 - 9.41	55	53	2	96	4

Figure 1: Partial Classification Table

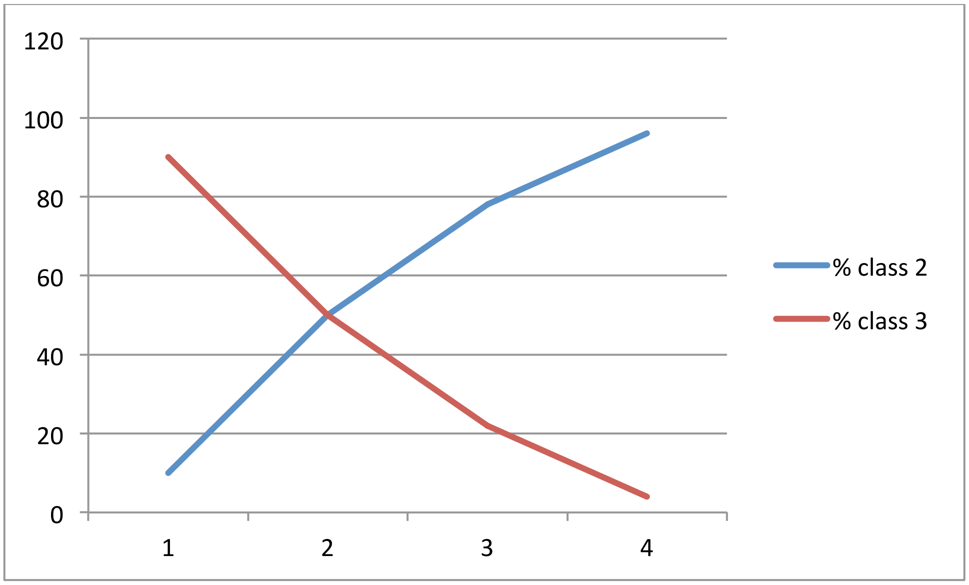

Figure 2: Partial Classification Graph

The following rules were developed from this:

4. If .26 ≤ C ≤ 2.54 instance belongs 10% to Class 2 and 90% to Class 3
5. If 2.54 < C < 4.82 instance belongs 50% to Class 2 and 50% to Class 3
6. If 4.82 ≤ C ≤ 7.10 instance belongs 78 % to Class 2 and 22% to Class 3
7. If 7.10 < C < 9.41 instance belongs 96% to Class 2 and 4% to Class 3

When the test data is run through the rule set, an instance that is classified into a partial classification will be considered part of the class with the highest partial classification value. This is done so that percent error can be calculated. In the above example, if a day has a partial classification of .7 sunny and .3 cloudy, the day will be considered sunny. However, the reason this classification method is more appropriate for certain data is that we also know more information about the day. The partial value shows that the weather is not all sunny or all cloudy, but both.

Lastly, the data that was used to come up with the rule set using fuzzy logic was run through the C4.5 algorithm, which is based on entropy. The result was the following rule set.

Rule 1:
Attribute C <= -17.8466
-> class4 [98.4%]
Rule 7:
Attribute C > 6.28904
-> class2 [95.7%]
Rule 6:
Attribute C > -17.8466
Attribute C <= 6.28904
-> class3 [88.3%]

Default class: class3

Assumptions

This analysis assumes that the following factors are true:

- The training and test sets are representative of the entire population.
- New data will be one of the three classes.

Results

The following rule set was developed by using the fuzzy logic algorithm:

1. If $C < .-.44$ the instance belongs to Class 4
2. If $-.44\leq C \leq .25$ the instance belongs to Class 3
3. If $C > 9.41$ the instance belongs to Class 2
4. If $.26 \leq C \leq 2.54$ instance belongs 10% to Class 2 and 90% to Class 3
5. If $2.54 < C < 4.82$ instance belongs 50% to Class 2 and 50% to Class 3
6. If $4.82 \leq C \leq 7.10$ instance belongs 78 % to Class 2 and 22% to Class 3
7. If $7.10 < C < 9.41$ instance belongs 96% to Class 2 and 4% to Class 3

This rule set gave a 5.4% error rate.

The following rule set was developed using the C4.5 algorithm:

```
Rule 1:
Attribute C <= -17.8466
-> class4 [98.4%]
Rule 7:
Attribute C > 6.28904
-> class2 [95.7%]
Rule 6:
Attribute C > -17.8466
Attribute C <= 6.28904
-> class3 [88.3%]
Default class: class3
```

This rule set also gave a 5.4% error rate.

Although both rule sets gave the same error rate, the fuzzy logic instances that were considered "errors" cannot be considered strictly errors. The instances that gave error had partial membership to their actual class according to the fuzzy logic rules. Therefore, both methods give accurate classification rules; however, the fuzzy logic rules provide a higher level of informational detail about instances that fall into the numeric overlap region.

Issues

Finding an appropriate data set and algorithm to use for the analysis

Appendices

None

References

Aleshunas, J. (2011). *Cluster Set.* Retrieved November 30, 2011, from Mercury: http://mercury.webster.edu/aleshunas/Data%20Sets/Supplemental%20Excel%20Data%20Sets.htm

Chapple, M. (2011). *Classification*. Retrieved November 20, 2011, from About.com: http://databases.about.com/od/datamining/g/classification.htm

Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science