Rebecca Wessell
IST 565: Data Mining
12/10/12

Final Project

Project Overview and Dataset

Can the characteristics of a country's flag be used to identify the continent from which the flag comes? To support this initial question, I'm also asking: Are there similarities among flags from the same continent? Are there correlations among flag attributes? This project will investigate a dataset that contains certain attributes of a country and its flag and will use various data mining techniques to answer this initial question.

The data set I will be using is the flag dataset from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Flags). It tracks information about the attributes of the flags from each country and information about each country. The flag dataset contains flags for only 194 countries, as it was donated to UCI in 1990. The dataset tracks the following attributes:

1. Geographical attributes: country, continent, area, zone of country in continent
2. Population and demographics: population, language, religion
3. Flag attributes: colors, shapes, images (animate vs. non-animate), text

In total, there are 30 distinct attributes within the dataset, which can be viewed at the dataset's website. To begin with, I will describe the preprocessing steps I undertook in order to analyze the dataset. For classification, I will use the Naïve Bayes, J48 decision tree and Support Vector Machine (using the Sequential Minimal Optimization implementation) methods to predict the continent class label. I will use EM clustering under both the classification and normal setting to see how well it predicts the class label and to see if any interesting correlations appear. Finally, I will use the Apriori associator to generate a list of 20-30 strong rules, and I will take a closer look at approximately 3-5 rules to see what associations or patterns may exist between the attributes of a flag.

Preprocessing

To make my analysis easier, I converted the continent attribute from numeric nominal attributes to text nominal attributes. This made analysis and evaluation of the results easier as I did not need to refer to the legend to identify which number corresponds to which continent.

Comment [PS1]: This is a good strategy to make your results more readable.

In order for the machine learning tasks to work on this dataset, several of the attributes will need to be removed. When attempting to identify the continent, I will keep that attribute as well as the flag attributes, but I will remove the remaining geographic and demographic attributes.

Since some of the attributes are binary, such as the colors and several of the flag attributes (where 1 indicates the item is present, and 0 indicates it is not), I have changed these to nominal attributes to more accurately reflect what the data represents. The

attributes that were nominalized are: red, green, blue, gold, white, black, orange, saltires, crescent, triangle, icon, animate, and text.

Furthermore, some machine learning techniques require nominalized or discretized data; so some of the flag attributes (such as the attribute "stripes" which lists a number for how many stripes a flag has) will need to be converted to nominal attributes for some of the data mining tasks. I will indicate which ones I have changed during each task.

Classification

Algorithm	Accuracy	Parameters
Naïve Bayes	43.29%	Default; 20-fold cross-validation
J48 Decision Tree	44.84%	Binary splits: true, Error reduced pruning: true; 10-fold cross-validation
SMO SVM	45.87%	Filter type: No normalization/standardization, kernel: NormalizedPolyKernel (exponent: 3.0); 10-fold cross-validation
EM Clustering	36.59%	numClusters: 5; Classes to cluster evaluation

None of the supervised learning and clustering to classes results achieved very high accuracies; SMO SVM was the highest with almost 46%. While these accuracies are not high by any standard means, they were higher than I was expecting and thus, these results are still worthwhile to investigate. Before running the classification tasks, I expected accuracies in the 20-30% range (which the EM clustering fell into) as I didn't expect the algorithms to be able to glean any helpful information from the flag attributes to identify continents. However, all of the supervised learning techniques are able to predict close to 1 out of 2 cases, which is mildly surprising given that it is attempting to identify a continent of origin based solely on attributes like color and the presence of shapes and images—there are no "helper" attributes like population, religion or language present.

Furthermore, the methods all did better at classifying flags of African countries; the TP rate of Africa for each supervised learning method were 65.4%, 59.6%, and 73.1%, respectively, with the precision rate being between 50-60% in each case. Moreover, Naïve Bayes and J48 did better at classifying flags of European countries than SMO SVM, with TP rates of 68.6% and 71.4% with precision being around 45% in each case. South America proved the most difficult country to ascribe flags to, as each classifier had a 0% TP rate for that continent. Naïve Bayes also struggled quite a bit with classifying North American flags, as the TP rate as only 9.7%. The rest of the continents had around 30-50% accuracy that varied with each classifier.

However, despite this perspective, these results are still not good enough to justify using any of these methods to accurately predict any continent of origin based on a flag.

Comment [b2]: Yes the upperbound of the accuracy may not be very high for this problem since the association between flag and region may exist but the strength may be limited.

You can try use some reference point, like a random guess or majority vote baseline.

Comment [PS3]: It's good to take a look at only a part of the dataset if it has some special features.

Comment [b4]: I agree with Peiyuan. In the clustering analysis you found green as a color to distinguishing African flags from others. It would be more convincing if your classification result also supports that. My prediction is yes

Thus on the side of classification, the experiment is not sufficient to be able to predict a flag's continent with a necessary degree of accuracy.

Clustering

Centroids of EM Clustering with clusters set to 3 and 66% split

Cluster 0: Africa, 0.9 bars, 1.1 stripes, 3.34 colors, red, green, white, main hue is green, 0.1 circles, 1 sunstar, top left and bottom right are green

Cluster 1: Oceania, 2.51 stripes, 4 colors, red, blue, white, main hue is blue, 0.3 circles, 0.6 crosses, has a saltire, 0.8 quarters, 5.4 sunstars, top left is white, bottom right is blue

Cluster 2: Tie with Asia and Europe, 0.4 bars, 1.5 stripes, 3.2 colors, red, white, main hue is red, 0.2 circles, 0.06 crosses, 0.6 sunstars, top left and bottom right are red

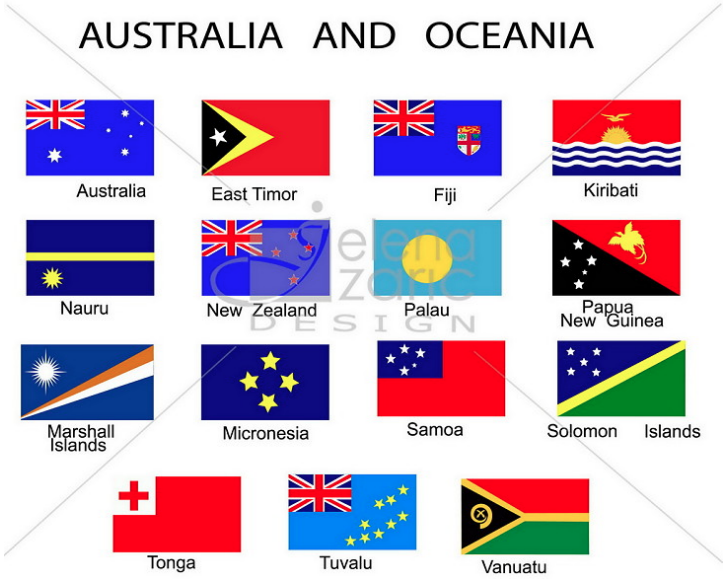
I chose to do 3 clusters for EM instead of letting the algorithm determine the number of clusters. I did this mainly because the algorithm typically chose too many clusters (with many being "junky" or having less than 10% of the results) or the clusters were heavily skewed (with 1 cluster having more than 50% of the results). Having 3 clusters yielded the most balanced results. Moreover, I chose to leave out the attributes that were numerical and had a value of 0 or the attributes that were nominal/binary and were not present when giving the descriptions of the centroids above. For instance, none of the centroids had a majority of orange for a color, so the absence of orange was not included in the centroid description. Overall, red and white seem to be some of the most common colors to appear on flags with many flags having between 3-4 colors, and looking at a chart of the world's flags validates this finding.

Cluster 0 is defined as having flags predominantly from Africa, and the distinguishing feature for this cluster is the presence of the color green. Green is listed as the majority main hue, and both the bottom right and top left corners tend to be green in this cluster. On average, there is also 1 stripe and 1 sunstar per flag, and at least 3 colors. Looking at the 55 flags of Africa below, 43 of them have the color green and 49 have at least 3 colors, and just under half of them have at least 1 sunstar.

Comment [PS5]: If you want to analyze if the flag features can predict which continents they come from. You may try to use the number of the continents as the number clusters.

(Image from depositphotos.com)

Cluster 1 is perhaps the most interesting cluster within this set, and the features indicated reminded me a lot of Australia's flag (shown below). On average, there were 2-3 stripes per flag along with four colors (red, white and blue were majority colors). Furthermore, the majority of the flags had a saltire and around 5 sunstars, with the top left corner being white and the bottom right corner being blue. If we compare these majority attributes with the Australia's flag, the results are surprising. Australia's flag has a saltire, red, white, blue (with the main hue being blue), 1 stripe, and 6 sunstars. Top left corner is mainly white and the bottom right corner is mainly blue. In fact, even looking at the flags that come from Oceania today (of which there are 15 countries included), a majority of them have blue, red and sunstars. Four of them have saltires, including Australia.



(Image from jelenazaric.com)

Cluster 2 is mainly defined by its lack of attributes as it only has 2 majority colors (red and white) and hardly any other discernible attributes. For instance, it only had on average 0.2 circles, 0.06 crosses and 0.6 sunstars, which would mean that a randomly selected flag from that centroid would probably not have a cross, a circle, or a sunstar. Since both Asia and Europe were lumped into this category, it was a detriment to the cluster as no truly defining characteristics emerged.

Association Rules

For the Apriori Associator to function, I changed all the remaining numeric attributes to nominal.

1. IF gold=0 THEN white=1, animate=0 and text=0
Confidence: 85%; Lift: 1.48; Support: 45%

If a flag does not have the color gold, then it is likely the color white is present and that neither animate images (e.g. an image of an animal) nor text are present on the flag. In other words, out of the 103 flags that didn't have gold, 88 of these also had white and did not have animate images or text. Besides having high confidence and support, this rule also has high lift, and it is further supported by the reverse rule that has 79% confidence along with 1.48 lift and 45% support.

2. IF white=1, sunstars=0 and icon=0 THEN gold=0, quarters=0, crescent=0, and text=0
Confidence: 88%; Lift: 2.11; Support: 26%

While the support for this rule is much lower than the first, it's still an interesting rule given its complexity, its high confidence and lift. Essentially the rule states that if a flag has the color white, 0 sunstars and no icon (a non-animate image), then the flag will not have gold, quarters, crescents or text. Out of the 58 flags that had white and no sunstars or icons, 51 of them did not have gold, quarters, crescents or text. The high lift for this rule indicates that it's quite a strong rule in regards to the confidence and the dataset overall. Furthermore, this rule in conjunction with the first rule seem to imply some sort of correlation between having white and not having gold.

3. IF orange=0 and icon=0 THEN animate=0
Confidence: 94%; Lift: 1.18; Support: 65%

This rule doesn't have as high of a lift value as the two aforementioned rules, but its support and confidence are quite high. This rule states that if a flag does not have orange or an icon, then it likely does not have an animate image. This rule held true for 126 flags out of 134 that had neither orange nor an icon.

4. IF crosses=0 animate=0 THEN orange=0 saltires=0
Confidence: 94%; Lift: 1.16; Support: 66%

This rule has the same confidence as above, along with slightly higher support and lower lift. However, it reinforces the association between the absence of an animate image and the absence of the color orange. Out of the 136 flags that didn't have crosses or animate images, 128 did not have orange or saltires.

Many of the rules that Apriori generated had very high support and confidence along with lifts that were greater than 1.1. Furthermore, most of the rules were marked by a correlation between the absences of elements (e.g. if there's no orange and no icon, then there's no animate image). Rather than using the rules to make generalizations about

what elements might appear together in a flag, the rules are helpful for identifying flags that are out of the ordinary. For instance, if a flag has orange and no animate image, this goes against many of the rules. These rules are thus helpful for pinpointing those flags that require closer examination due to their uniqueness.

Conclusion

While the classification methods did not prove completely useful, the clustering and association rules lent some valuable insights and tools into this dataset. The supervised learning methods were decent in identifying flags from Africa and Europe, but they all failed to correctly identify a single flag from South America. The accuracies for the other continents was middling at best with no accuracy being greater than approximately 55%.

The clustering showed that flags from certain continents do share similar attributes, as it clustered flags from Africa and Oceania into two distinct groups. The defining characteristics from each of these groups proved to hold true for a majority of the flags from these continents. For instance, many flags in Africa have the color green, while flags from Oceania tend to have more sunstars than other flags. The clustering thus proved better at finding some correlations between flags from continents and grouping these together successfully than did the supervised learning methods. Furthermore, Apriori rules should a marked correlation between the absence of elements, particularly the color orange, shapes, and images. Since many of the flags lack the same characteristics, the rules can be used to identify flags that are abnormal or different from the norm.

This experiment showed that flags from certain continents do share enough similarities to be grouped together in a meaningful way, and that flags from some continents can be predicted with a degree of accuracy. This was particularly true for Africa across the board with SMO SVM being able to correctly classify almost 3 out of 4 cases and with EM clustering showing one strongly typed African cluster. These results are also strong due to the fact that Africa has the most countries of any continent, thus giving it more instances to cluster or classify. Finally, the Oceania cluster surprised me as Oceania only has 15 countries, being the smallest second to South America. The results were distinct for that cluster, and reviewing the flags of that continent showed that these results did hold true in the real world. The implications from these findings could mean that flags from certain continents share more commonalities than flags from other continents or that these countries are more tightly linked in a sociocultural way thus resulting in more similar flags, but further research would need to be done to test this hypothesis. In the end, the experiment raised more questions than it answered.