CS 345: Data Mining Names:_____

In-Class Activity 6 _____

Today is about going through a data exploration from start to finish. We will be exploring a ski resort data set, that contains 989 data instances and each instance has 16 attributes. All the data attributes are numeric or nominal. The attributes Rating, Survey, Prize, Punishment represents the overall assessment from a user. The other attributes such as Aspen, Snowmass, ..., Eldora represent the different ski resorts a user rates, so if Aspen=1, then the user is including that resort in their overall rating. The Ski Resort Data Set is taken from a Data Mining Course by Yong Bakos. It is a raw data file in .csv format. Follow the instructions given in Weka ARFF Tutorial document on the course website to first convert skiresortdata.csv file to an .arff file.

Data Preprocessing

While you're still in the Weka ARFF-Viewer window, you should first look at the data objects with missing values and remove the incomplete data from the dataset. Then use Density-Based cluster algorithm to detect whether there are anomalies in the data set. If the cluster algorithm finds any outlier data points far from most other data points, these data points are probably noise data and should be removed from data set as well. Follow the steps below to do so.

Remove Instances Containing Missing Values

The explore tool in WEKA shows excellent statistical results for each data attribute, including the min, max, mean, standard deviation, the number of distinct values and whether there are missing values in the data set. It is straightforward to find out that Loveland and Crested Butte attributes have confused values in certain data objects. Since Loveland and CrestedButte represent the selected ski resort, we can infer that they should be binary attributes just like the other ski resort attributes. Thus the value "1-" and "Q" are very likely errors in the data and should be removed from the data set. Also, the Silverton attribute has 4% missing values of the whole data set (42 total records). Remove the instances containing missing values as well because we do not have any prior knowledge about the incomplete data record.

To remove the instances, first, sort the data by Loveland. Then select all instances that contain a "1-". Delete the instances. Do the same to remove instances containing "Q" for CrestedButte, and finally complete similar steps to delete all instances with missing values for Silverton. You should be left with 862 instances total. Save your file again, and close the ARFF-Viewer window. The ARFF-Viewer makes it difficult to change attribute types, so manually open your .arff file using a Text Editor, and change the

attribute type of Loveland and CrestedButte to be the same as the other ski resorts. Save your changes and close the file.

Detecting Anomalies in the Data

To utilize DBSCAN, you'll first need to install it. Go to Tools-> Package Manager and search for DBSCAN. It should return a Package called optics_dbScan; install it.

Now open the Weka Explorer and open your .arff file. You should be able to see that you have 862 instances and 16 attributes. Go to the Cluster tab and select DBSCAN. Leave all the default settings and click Start. You should find 3 unclustered instances. Scroll up in the results window and find which instances are labeled as NOISE. Notice the similar ranking value for all 3 instances. Remove these 3 instances from the data set; you can do this in the ARFF-Viewer or figure out how to use the Filters in the Pre-process window.

Data Mining Processes

Now that your data has been pre-processed, you're ready to do some knowledge discovery of the data set. We'll apply 4 different strategies to analyze the patterns in the data set. Use K-means to separate the data into groups. Use Naïve Bayes and Decision Tree classifiers to find out the group characteristics of each class. Lastly, use association algorithms to figure out the potential correlation among selected attribute sets.

Simple K-Means Clustering Analysis

Load your data set in Weka Explorer and go to the Cluster tab. Select SimpleKMeans and adjust the number of clusters. Try running the algorithm with 5, 15, 20, and 25 clusters and determine which setting minimizes the within cluster sum of squared errors. Use Training set for the Cluster mode.

Best # of clusters: _____

Within cluster sum of squared errors: _____

How good are the clusters? What % of the data does each cluster cover?

Try to visualize the data in a variety of ways to understand how the clusters are categorized. By looking at the data, which attribute makes the most sense as the class label?

Naïve Bayesian Classifier Analysis

Since the Naïve Bayes classifier requires that attributes are nominal class type, first use the NumericToNominal Filter to transform the numeric attributes into nominal attributes. To do this, go to the Preprocess tab and Choose a Filter. NumericToNominal is located under unsupervised->attribute. Apply the filter to your dataset and check that all the previously numeric attributes are now nominal. Based on your findings from the k-means clustering, Rating seems to be the most likely class label for the data. After running the NumericToNominal filter on your data, the Rating attribute now has 9 distinct values. However, if we use each distinct value as a nominal class label, then we might fail to aggregate the records from different distinct values, which would lower the accuracy of the classifier. Therefore, take a look at the current distribution of the Rating values. The data tend to be scattered when the ranking value is in the 0.15 – 0.85 range, but when the ranking value is greater than 0.9, the instances seem to cluster. Therefore, it seems like a good idea to combine the values 0.9-1, so we can create 5 groups. To do this use the MergeTwoValues filter (unsupervised->attribute). In the options window for the filter, change the attributeIndex to first, firstValueIndex=5, lastValueIndex=6. Apply this filter enough times to get 5 groups. Select the Rating attribute and verify that you now have 5 values.

Now, go to the Classify tab and select the NaiveBayes classifier. Select (Nom)Rating as the class attribute. Evaluate using the training set. Look for interesting trends between the other attributes and the class attribute. In particular, look at the punishment and prize values in relationship with the rating. Note down any trends/relationships you find between these pairs of attributes.

How well did the classifier perform?

Decision Tree Classifier

Since decision tree classifiers also require nominal data, use the same discretized data you used for the Naïve Bayes classifier, and run the J48 decision tree classifier using the training data again.

Are the instances correctly classified?

Visualize the tree. Which attributes were selected for the decision tree?

Now, re-run the J48 classifier using (Nom) Prize as the class attribute. How well does the classifier perform?

Visualize the tree. Using the decision tree, determine the relationship between rating and prize. What trends/relationships do you observe?

Now, re-run the J48 classifier using (Nom) Punishment as the class attribute. How well does the classifier perform?

Visualize the tree. What trends/relationships do you observe?

Now, re-run the J48 classifier using one of the ski resorts as the class attribute. Which one did you select?

By observing the decision tree, what relationships between ski resorts do you observe?

Association Rule Mining

We're going to use association rule mining to discover correlation relationships between ski resorts. To do this, we'll need to first remove the rating, survey, prize and punishment attributes from the dataset. Leave the dataset as you had it for Naïve Bayes and Decision Trees (nominal attributes). Then apply the Remove (unsupervised->attributes) filter with attributeIndices 1-4.

Go to the Associate tab and select Apriori. Run it the lowerBoundMinSupport set to 0.7 and the numRules set to 50. Leave other values as their default. Based on the rules generated, what relationships between ski resorts do you observe?

Now, let's try to understand which ski resorts may have the largest influence on overall rating. Go back to the Preprocess tab and click Undo. Then, rerun the Remove filter with only attributeIndices 2-4 this time. You should now have a version of the data set that only includes ratings and ski resorts. Use the same Apriori settings as before and run Apriori on the updated data. Scroll down in the generated rules until you find ones that deal with Rating. What do these rules tell you about the relationship between ratings and particular ski resorts?

Conclusions: Based on your analysis using clustering, classification, and association rules, what conclusions can you draw about this data set? Write a paragraph about your findings.

Future Work: What other experiments, if any, do you think would be good to run on this data set? What were some of the issues with the above analyses?