Association Rules

* Intro
  + Apriori Algorithm
* Binning
  + Intro

Since most of the explanatory variables are numerical and continuous, data binning is required to divide the numerical values into buckets.

* + Number of bins

To determine the optimal number of bins, Freedman–Diaconis rule [link] is utilized, which is a rule to select the size of bins to be used in a histogram. The equation for the rule is

where IQR(x) is the interquartile range of the data and n is number of observations in sample x. The numbers of bins obtained for every variable using this rule are later used to bin the data. To get the accurate results, two methods, fixed width and adaptive width binning, are used as the binning strategies and the results are compared in the later section.

* + Fixed width Binning

Fixed width binning divides data into bins with intervals of equal length. The advantages of this strategy include that it is simple to implement, and it produces a straight-forward and reasonable abstraction of data. On the other hand, this method is unsupervised, and as a consequence, it is hard to know just from the result of binning that if data are divided in a desired way.

* + Adaptive width Binning

Adaptive width binning divides data into intervals with equal content via quantiles. Depending on the data distribution, this strategy may give a better classification of data. However, there are several pitfalls of this method. Equalizing the interval height may lead to an over-weighting of outliers, and data points with the same value may fall in different groups.

Although Freedman-Diaconis rule is designated for equal width binning, for consistency, the same numbers of bins are used for each variable in both methods.

* Results

Since the dataset only contains 77 samples (77 communities), in order to obtain rules that are more general, a higher support is required compared with bigger datasets. The starting point is set to be 0.5 support and 0.8 confidence. The support value is lowered by step of 0.1 or 0.05 through the analysis process, to help with the selection of proper support value used in the final analysis. The confidence is unchanged through the process since it is decided that 80% is the lowest acceptable confidence in this analysis and the rules with higher confidence would be automatically included.

* + Fixed Width Binning

For the data binned with fix width binning method, every itemset with higher than 0.5 support only has one items. Table [1] shows the result generated by Apriori algorithm with 0.4 support and 0.8 confidence.

Table [1]. Association rules with fixed width binning with 0.4 support and 0.8 confidence

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Antecedents** | **Consequents** | **Support** | **Confidence** |
| 1 | {crimes(105,550]} | {black(-0.097,48.5]} | 0.487 | 0.974 |
| 2 | {white(-0.084,16.8]} | {asian(-0.049,4.9]} | 0.436 | 0.895 |

The first association rules in Table [1] indicates that the communities that have low violent crime rates also have low percentages of black people. The second rule says that the communities with low percentages of white people also have low percentages of Asian people.

As the support value is decreased to 0.3, the following list in Table [2] is appended to the previous list in Table [1].

Table [2]. Association rules with fixed width binning with 0.3 support and 0.8 confidence

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Antecedents** | **Consequents** | **Support** | **Confidence** |
| 1 | {hospital 0,  crimes (105,550]} | {black (-0.097,48.5]} | 0.321 | 0.962 |
| 2 | {black (48.5,97.1]} | {hispanic (0.915,18]} | 0.321 | 0.893 |
| 3 | {black (48.5,97.1]} | {white (-0.084,16.8]} | 0.321 | 0.893 |
| 4 | {hospital 0,  black (-0.097,48.5]} | {crimes (105,550]} | 0.321 | 0.806 |
| 5 | {other 1} | {asian (-0.049,4.9]} | 0.308 | 0.923 |
| 6 | {black (48.5,97.1]} | {asian (-0.049,4.9]} | 0.308 | 0.857 |

The appended list tells the relationship between percentages of races in communities, that communities with high percentages of black people usually have low percentages of Hispanic, white and Asian people. The rules with 0 hospitals in the itemsets do not indicate much since about two thirds of the communities do not have a hospital.

* + Adaptive Width Binning

For the data binned with adaptive width binning method, every itemset with higher than 0.3 support only has one items. Table [3] shows the result using adaptive width binning with 0.2 support and 0.8 confidence.

Table [3]. Association rules with adaptive width binning with 0.2 support and 0.8 confidence

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Antecedents** | **Consequents** | **Support** | **Confidence** |
| 1 | {asian (-0.049,0]} | {black (24,97.1]} | 0.231 | 0.900 |
| 2 | {hispanic (0.915,4]} | {black (24,97.1]} | 0.205 | 0.941 |
| 3 | {white (-0.084,5]} | {black (24,97.1]} | 0.205 | 0.941 |

Unlike the result obtained from fixed width binning section, the association rules with highest support with the adaptive width binned data do not involve the class variable, instead, they reveal a relationship between percentages of races in communities: there is usually a high percentage of black people if there is low percentage of Asian, Hispanic, or white people.

The support is decreased to 0.15 and the following list in Table [4] is appended to the current association rule list.

Table [4]. Association rules with adaptive width binning with 0.15 support and 0.8 confidence

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Antecedents** | **Consequents** | **Support** | **Confidence** |
| 1 | {hispanic (53.8,86.1]} | {black (-0.097,24]} | 0.192 | 0.938 |
| 2 | {white (50.8,84.1]} | {black (-0.097,24]} | 0.192 | 0.938 |
| 3 | {hospital 0,  hispanic (53.8,86.1]} | {black (-0.097,24]} | 0.179 | 1.000 |
| 4 | {black (-0.097,24],  hispanic (53.8,86.1]} | {hospital 0} | 0.179 | 0.933 |
| 5 | {hispanic (53.8,86.1]} | {hospital 0} | 0.179 | 0.875 |
| 6 | {hispanic (53.8,86.1]} | {hospital 0,  black (-0.097,24]} | 0.179 | 0.875 |
| 7 | {white (-0.084,5]} | {asian (-0.049,0]} | 0.179 | 0.824 |
| 8 | {crimes (1560,2760]} | {black (24,97.1]} | 0.167 | 1.000 |
| 9 | {white (13,30.6]} | {hospital 0} | 0.167 | 1.000 |
| 10 | {white (-0.084,5],  asian (-0.049,0]} | {black (24,97.1]} | 0.167 | 0.929 |
| 11 | {hospital 0,  asian (-0.049,0]} | {black (24,97.1]} | 0.167 | 0.867 |
| 12 | {white (-0.084,5],  black (24,97.1]} | {asian (-0.049,0]} | 0.167 | 0.813 |
| 13 | {crimes (105,313]} | {black (-0.097,24]} | 0.154 | 0.923 |
| 14 | {crimes (451,548]} | {black (-0.097,24]} | 0.154 | 0.923 |
| 15 | {other 3} | {black (-0.097,24]} | 0.154 | 0.857 |

As explained in the previous section, association rules with item “hospital 0” can be ignored in this analysis since most of the communities do not have a hospital. Row 8, 13, and 14 in Table [4] indicate a relationship between class variables and percentage of black people. To conclude, communities that have high violent crime rates also have high percentages of black people; communities that have relatively low violent crime rates also have low percentages of black people. The rest of the rules are all about percentage of races, most of which disclose the conflict between percentages of black people and all the other races: when there is a low percentage of Hispanic, Asian, and white people, there is usually a high percentage of black people.

* Conclusion

The fixed width binning provides a better classification in this case, as association rules with higher support are obtained with Apriori algorithm, which means that the rules generated are more generic for the current sample of community data.

Results from both strategies indicate that there is a strong relationship between the class variable and the percentage of black people. Both include rules that where there is low violent crime rate, there is usually a low percentage of black people, while adaptive width binning also gives that where there is a high violent crime rate, there is usually a high percentage of black people. In addition, results from both strategies reveals a complementary relationship between the percentage of black people and percentage of all the other races in a community.

Reference

Freedman, David; Diaconis, Persi (December 1981). "On the histogram as a density estimator: L2 theory" (PDF). Probability Theory and Related Fields. Heidelberg: Springer Berlin. 57 (4): 453–476. doi:10.1007/BF01025868. ISSN 0178-8051. Retrieved 2009-01-06.