Detecting outlying patterns in categorical variables

Its always been amazing to play with data. Most specifically the continuous data that I get for my analysis. However, its a pain when I get to analyse the insights of the categorical data. Ofcourse, there are different ways to handle the categorical data such as converting them to binary form, creating dummy columns for them or by factoring them and giving them all a separate new numeric number. I wonder if its only me, but these methods have really never helped me when I get real data with many categorical variables and many categories in them.

Real data in work mostly comes with categorical variables and sometimes even with more than 100 categories.

For example, say we work for an organisation which handles transactions among worldwide. We are given we a problem to find which are the most uncommon country and currency combination in the system.

| Table 1 | |
| --- | --- |
| AFGHANISTAN | AFG |
| ALBANIA | ALB |
| ALGERIA | DZA |
| AMERICAN SAMOA | ASM |
| ANDORRA | AND |

| Table 1-1 | |
| --- | --- |
| ANGOLA | AGO |
| ANGUILLA | AIA |
| BAHAMAS | BHS |
| BAHRAIN | BHR |
| BANGLADESH | BGD |
| BARBADOS | BRB |

| Table 1-2 | |
| --- | --- |
| AFGHANISTAN | AFG |
| ALBANIA | ALB |
| ALGERIA | DZA |
| AMERICAN SAMOA | ASM |
| ANDORRA | AND |

Recently I was working on a outlier detection problem where I had to find the outlying patterns of the data. The current outlying detection problems are powerful in handling the continuous data. But, when it comes to get the uncommon patterns of the categorical variables, I never got a desired result out of the current outlier detection algorithms. I spent many days spending in utilising the SVM one class, kmeans algorithms to find the outliers of my data. As a result, I was able to get the outlying continuous data. But, was not able to get the uncommon patterns in the categorical variables of my data.

Then, I decided to write my own code from scratch to get the most uncommon patterns in my data.

To explain my program, let have look at my data.

I have a CSV file of country and country codes.

For example:

In the data, the below patterns has occurred twice.

And the below pattern has occurred each once. We need to find these patterns with the least frequency in the give data.

Lets visualize the data to see the pattern of data.

I have attached the data source to the article. From a sample of data, the below bar chart is made.

The plot is the frequency of country and countryside pattern.

From the plot, we can see that pattern like ‘COMOROS-COM’ , ‘CAMBODIA-KHM’ has a frequency of 2 where as patterns like ‘BARBADOS-BRB’ ‘AFGHANISTAN-AFG’ has a frequency of 1. As the graph was drawn on the sample data, there might have been some data loss. But, its fine. The graph is just to illustrate what the requirement is.

| Table 1-3 | |
| --- | --- |
| AFGHANISTAN | AFG |
| ALBANIA | ALB |
| ALGERIA | DZA |
| AMERICAN SAMOA | ASM |

| Table 1-4 | |
| --- | --- |
| ANDORRA | AND |
| ANGOLA | AGO |
| ANGUILLA | AIA |
| BAHAMAS | BHS |
| BAHRAIN | BHR |

Our requirement is to find the most uncommon patterns i.e. the patterns with low frequency.

This problem is entirely based on the categorical data.

So, now here comes my solution. I have done a ready made module which takes all the categorical variables that we feed it and then its going to give a result as -1 for for the least frequency patterns and 1 for the rest all.

