Individual Contribution Report

CSE 578: Data Visualization

Islam Hamed
ihamed@asu.edu
1221217088

- # The Contribution:

➢ Teamwork:

I took the initiative to be the leader of the team. First of all, I created the slack group and started the conversation between team members. Also, I was responsible for suggesting each meeting agenda before the meeting. I managed to have our first meeting to know each other, by responding for small questions (Where are you from? what is your current job? and what is your previous experience and education). In week six, I suggested to have a daily meeting to finish the project, and set our thoughts together.

➢ Technical:

I started with analysis four attributes (Capital-loss, Native-country, Sex, and Occupation), to select some of them to be included in System Documentation Report. In this stage, every attribute's visualization was independently from others. I was exploring whether the attribute's visualization can be helpful in my classification task in general or not, regardless other visualizations.

Capital-loss was the fourth one of the best attributes. As, it is really helpful to classify the income label (>50k, or <=50k). Occupation, and sex attributes were selected to be included in the System Documentation Report, as they have important insights.

One of the most important criteria, that I depended on when analyzing visualization and taking a decision, is if this attribute has values that can classify both income labels (>50k, and <=50k) or not. For instance, if there are two values from this attributes, one classify label >50k by 65%, and another classify label <=50l by 75%. So, if we think of two labels as class 1 (>50k) and class 0 (>=50k), we need to have attributes that able to classify both classes as true positive and true negative.

I. **Capital-loos**: it is a continuous attribute, so scatter plot chart is one of the best charts can plot this attribute. It has range from 0:< 5000. Some of these ranges lies in both class 1 and 0, but we can determine a percentage for each class by the calculate the occurrence of provided values in class 1 verse its occurrence in class 0 in the dataset. On the other hand, there are another ranges, they lay in

only one of the classes such as value 3000 lies in class 1. So, this attributes can classify the input values to class 1, class 0 or percentage of one of the classes.

II. **Occupation**: it is a categorical attribute, so many charts can be used to visualize it (pie chart, bar chart and mosaic chart). Actually, pie chart provides a distribution insights of the categories, so it is not helpful for classification. However, bar chart provides the occupation's value verse income label. At first, I plotted income verse count of record in each category, the range of counts is 0:3263. However, there are some categories that have very small count value, accordingly there is a huge difference in bars. So, I plotted the income verse the percentage of occurrence in each category, so the range is 0:100 and the bars more consistent. All percentages > 50% indicates only class 0. So, this feature cannot classify both income classes (>50k, and <=50k). However, this feature can be used to indicate for class <=50k by percentage. For instance, if the individual's occupation is priv-house-serv, so definitely his/her income is <=50k. According to the dataset, every individual' occupation is priv-house-serv her/his income <=50k, as shown in the chart. Also, we can use it beside other feature as supporting decision by percentage.

III. **Sex**: it is a categorical attribute, so I did the same as occupation in visualization. All percentages > 50% indicates only class 0. So, this feature cannot classify both income classes (>50k, and <=50k). However, we can interpret that male is more likely to have an income >50k than female despite of number of male in the dataset is greater than number of female.

IV. **Native-country**: it is a categorical attribute, so I did the same as occupation and sex in visualization. It was not fit into bar chart with raw count of values, as it has huge variance in occurrence. Also, all percentages > 50% indicates only class 0. So, this feature cannot classify both income classes (>50k, and <=50k). However, this feature can be used to indicate for class <=50k by percentage. For instance, if the individual's native-country is Holand-Netherland or Outlying-US(Guam-USVI-etc), so definitely his/her income is <=50k. According to the dataset, every individual' native country Holand-Netherland or Outlying-US her/his income <=50k, as shown in the chart. Also, we can use it beside other

feature as supporting decision by percentage. For example, for Dominican-Republic, it is 98% the income is <=50k. On the other hand, countries such as Iran, India and France, are little confusing. So, we need to use another feature as primary feature.

• Overview of Team's Work:

First, we needed to select 8 attributes from the datasets. So, we reviewed the description of each attribute. However, we decided to select these attributes based on their visualization. Visualization provides many insights (example: values verse income, count of certain value verse income, and count of value verse counts of another values) that helped us to have a decision. Then we divided the attributes among us randomly. We had a weekly meeting to discuss our progress, and feedback. We selected the best attributes based on our analysis, and discussions. Also, we eliminated some attributes from System Documentation Report, based on our analysis.

• Conclusion:

In machine learning process, data preprocessing is and essential step before using data. One of data preprocessing techniques is feature selections, to avoid the curse of dimensionality. I used to select this features or attributes based on their description. However, after this project and its attributes analysis, I figured out that attribute's description is not always good indicator for which feature will be helpful. On the other hand, feature visualization is a very good indicator. Actually, it is helpful, quick and accurate one. Instead of using some expensive computational techniques such as entropy, data visualization is more efficient and effective in respect of optimization.