Use of Bayesian Network in Income Distribution Study

Pacák Vojtěch, Repko Šimon, Sudora Simon

March 27, 2018

1 Introduction & Dataset

We decided to learn Bayesian Network from dataset provided by U.S. Census (archive.ics.uci.edu), which contains information about U.S. citizens' households across different jobs, educational levels, age categories, races, etc. To keep work on "discrete level", we either got rid of continuous variables, or transformed them to discrete ones (e.g. age) by splitting values into multiple levels. Each level is size of 5 years. The main variable of interest is annual income of U.S. citizen as factor of 2 levels (more than $50K or less than $50K). The discretized adjusted dataset contains following information about U.S. Citizens:

- Age category
  - (15,20], (20,25], (25,30]...
- Sex
  - male/female
- Race
  - Amer-Indian-Eskimo, Asian-Pac-Islander, Black, White, Other
- Relationship
  - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- Marital status
  - Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-marriedSeparated, Widowed
- Native country
  - Scotland, South, Taiwan, Thailand, Trinadad&Tobago, United-States ...
- Education
  - Bachelors, Doctorate, Masters, Preschool, Some-college...
- Occupation
  - Tech-support, Craft-repair, Other-service, Sales, Exec-managerial...
- Workclass
  - Private, Self-emp-inc, Federal-gov, State-gov, Without-pay, Never-worked...

The whole dataset has 32561 examples. Some of the variables in the dataset have also missing values. native country has 536(1.6%), occupation 1843(5.7%) and work class 1836(5.6%).

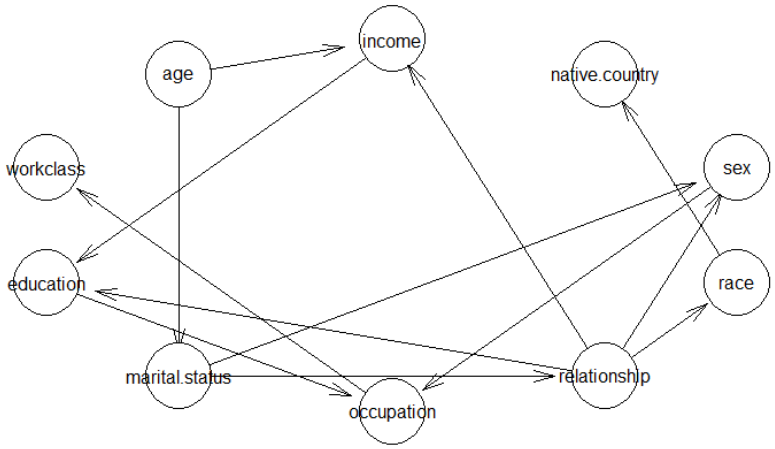

Figure 1: Qualitative BN network model created by EM structure algorithm

2 Qualitative model

To create qualitative model, we had 2 options. Create the qualitative model on our explicit domain knowledge or let the Bayesian network do work for us and learn structure of network from the data. Because of our smaller knowledge about the domain we decided to use the help of algorithm to find the proper qualitative model. After algorithm suggested possible structure of network we evaluated the network based on our common sense and did few adjustments.

Since we want to learn the structure and the dataset has also some missing values we used structural EM algorithm and its implementation in R package „bnlearn“.

On the Figure 1 you can see learnt structure. Our variable of main interest - income is directly connected only with age and relationship. Dependency of income on age is logical. With more years spent working individual gains more experience and thus his income should raise too. However, connection between income and relationship is less obvious. Common sense suggests that this arc should have opposite orientation because people usually wait until their income is high enough to take care of family. But in this case information that individual has a family and kids can be an indicator of higher income.

Variable marital status provides very similar information as relationship variable. They are closely bounded together and thus it doesn't give any new information which can be used for determining income category.

Another obvious fact that has impact on income is field in which individual works. This is represented by variables occupation and work class. Even though this connection wasn't identified by the algorithm, we decided to add dependency between income occupation and work class.

When it comes to size of income education is one of the first thing that comes to our mind. This dependency is also present in the created network. However, it has different orientation and for the purpose of prediction of income is more useful orientation from education node to income node.

Sex is naturally connected with occupation. Since some of the professions are typically more popular among men. Unfortunately, in everyday life we can also hear about sexual discrimination and unequal salary conditions for women. So, in this case directed arc from sex to income would make sense too.

Lastly, variables race, and native country should not have an impact on income. In some cases native country might have an impact on access to education, but when we looked on the data there is only 8.6% of individuals that are not native Americans so in most of the cases this won't have that big impact on the income category.

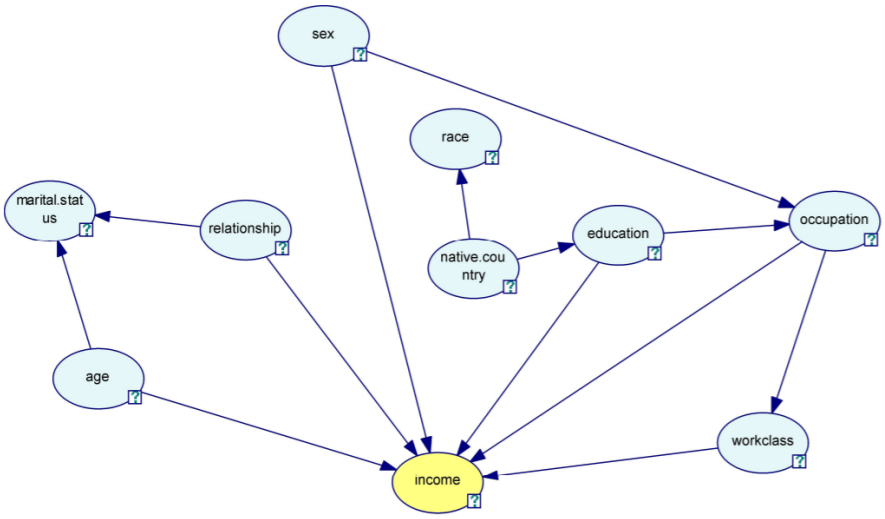

Figure 2: Qualitative BN network model created by EM structure algorithm

We did all the described changes and we implemented adjusted qualitative model Genie software, resulting in Figure 2 structure.

3 Quantitative model

Package „bnlearn“ in R as well as software Genie allow to extract conditional probabilities throughout our BN model. As income is our main point of interest, we picked to highlight several conditional probabilities of Income level with respect to different attributes, as can be seen in Table 1. The first two rows show that there is higher probability for man to receive higher salary then i tis for woman.

Rows 3, 4, and 5 support expectation about positive correlation of education and salary, as additional level of reached education reflects higher probability to earn additional money.

Finally, last row of our table shows that we can obtain more complicated probabilities if interested. As person occupied as Prof-specialty who is self-employed has probability 44% to reach $50K annual salary.

P(Event| Evidence)	Event	Evidence
0.29	Income > $50K	Sex: Male
0.14	Income > $50K	Sex: Female
0.39	Income > $50K	Education: Bachelors
0.56	Income > $50K	Education: Masters
0.73	Income > $50K	Education: Doctorates
0.28	Income > $50K	Occup.: Prof-specialty and workclass: Self-emp-inc

Table 1: Conditional probabilities for prediction of the event Income > $50K

Table 2 reveals conditional probability with respect to Age group of worker. Its obvious and logical that older people with enough experience and expertize are much more likely to have higher salary than fresh graduates in their 20s.
The cond. probability has pattern of parabolic curve, as it firstly sharply increases with additional age, reaching its peak in 45-55 age range and then slowly descending back, as elderly approach retirements.

P(Income >$50K)	Age group
0.002	(15,20]
0.033	(20,25]
0.131	(25,30]
0.235	(30,35]
0.331	(35,40]
0.346	(40,45]
0.391	(45,50]
0.417	(50,55]
0.351	(55,60]
0.266	(60,65]
0.28	(65,70]
0.25	(70,75]
0.25	(75,80]
0.056	(80,85]
0.141	(85,90]

Table 2: Conditional Probabilities with respect to age

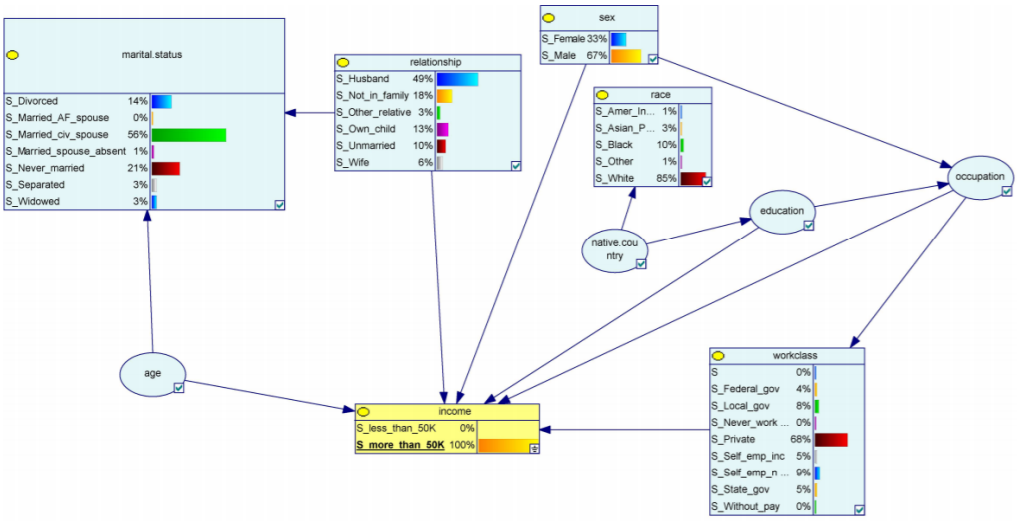

Figure 3: Conditional probabilities for diagnosis when Income > $50K

After prediction of Income, we did also tests with diagnosis when income is higher than $50K. In the Figure 3 are shown bar charts only for the variables with significantly higher evidence of one of its value. Base on the Figure 3 we can say if the individual has income higher than $50K then there is a high chance that he is married, male, white race and he works in private sector.

4 Conclusion

In our short paper, we investigated impact of several personal characteristics on US citizens' income by using Bayesian Network approach. The structure of Network was learned from adjusted dataset, and then slightly modified to accurately, at least according to us, reflect true-world situation. Finally, with help of statistical software, we estimated specific conditional probabilities and highlighted some of them in quantitative part of this paper. We recognized some patterns of income disparities that we kind of expected to be present (based on gender, age, ...), as results mostly met our prior expectations proving this model to be useful at least to some extent.