|  |  |
| --- | --- |
|  | **2017** |
|  | Resultat d'imatges de liu university  Carles Sans Fuentes |

|  |
| --- |
| **[main features affecting earnings on population]** |
| The project tries to identify the main features on the population to earn more than 50K dollars per year. Those features are identified using a very well-known rule based algorithm called a priori algorithm, based on confidence, support and lift |

MOTIVATION

Inequality in salaries due to inequality of opportunities together and discrimination has always been one of the major issues that has concerned the population among different countries. There is always though (unless there is direct evidence from the discrimination) some background to reason upon the election of the best candidate for a position: experience, education and several other aspects (e.g. the physical appearance in an interview) define a lot of times why is one candidate chosen over another. Nevertheless, are there intrinsic variables that can make someone being discriminated indirectly or directly given some prejudices of the population? It is alleged and acknowledged by different countries, for instance, that there is clear evidence of discrimination for the female gender on their jobs; not only about the difference on salaries for a similar or equal position than a man, but rather also about the chances to get to a higher position given its capabilities and hence earning more money in the public and private sector.

In order to evaluate and assess objectively how large these issues are and also several other ones (e.g. gender discrimination by origin or color), the usage of computer and its incremental power and speed during the last decades have been key to be able to analyze the issue. Thus, the evaluation of large amounts of data (called today Big Data) using algorithms that evaluate and assess information efficiently in almost no time has been able to show to the society the dimension of certain problems with clear proof to discuss and argument the subject that matters. Nevertheless, when the analysis of data is done, it is always crucial that when you run algorithms and establish models, all variables available should be included in the first instance for the purpose of seeing the maximum evidence and the true cause of the issue. Thus, it is important to be aware that information can be presented and shown from a lot of perspectives, and that some of them may lack of a good evaluation, the good data or a wrong model. Thus, it is always important to be able to show in a clear matter your findings making them reproducible.

DATASET EXPLANATION

The dataset used for this data mining a project is a well-known dataset called adults[[1]](#footnote-1), implemented in several softwares and evaluated with different algorithms. The dataset, extracted by Barry Becker, contains information about the 1994 Census Boreau database in the US. More specifically, it contains 48842 observations from people with 15 attributes to take into consideration with its following classes:

1. Age: continuous variable related to the observations.
2. Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. Fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: continuous.
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
15. Whether the observations win more than 50000$ or not: >50K, <=50K.

The idea for this data set is to find relations and importance of variables to whether the variable of winning more than 50000$ or not. Having that said, the dataset shows that the Probability for the label '>50K' is 24.78% and Probability for the label '<=50K' : 75.22% (once unknowns are cleaned up on the data base provided[[2]](#footnote-2))

THE A PRIORI ALGORITHM

The apriori algorithm is a frequent items set algorithm used to find association rules over transactional databases. It is really used in areas such as the food industry or the banking industry in order to understand costumer behavior and elaborate bundles and reccomed products to clients. The working procedure of the algorithm is starting with individual item sets and then enlarging them under some constraints, being the most typical ones the follwoing ones:

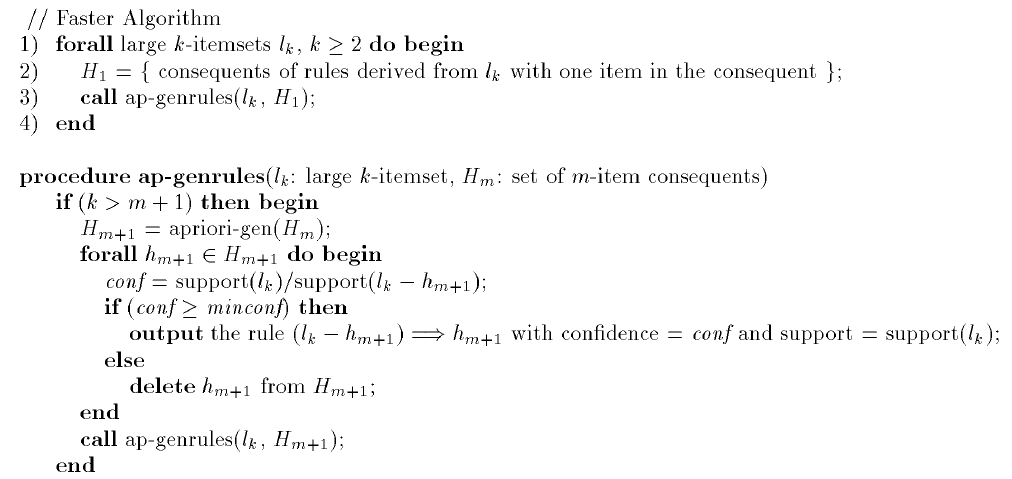
* **Minimum** support: Probability that a transaction contains product X and Y given all transactions.
* **Confidence**: Probablity of Y for all transactions containing X.
* **Lift**: measure that identifies the probbility of a rule given the total possibility of that rule in the model. It is given by the formula lift (X) =Prob(X|Y) /Prob(X)
* **Target** **feature**: It is also typical to use some target features or products contained in the dataset as a constrain in order to assess rules based on this feature or product.

Figure 1 PseudoCode from the apriori algorithm in Han - Data Mining Concepts and Techniques 3rd Edition – 2012, page 253.

The pseudocode for the algorithm can be found in the following picture[[3]](#footnote-3):

One of the limitations of the algorithm is that it requires data to be discrete. Also, it can generate a huge amount of inefficientrules, so good constraints must be designed in order to be optimal.

DATASET CLEANING

TEST RESULTS

test results: describe the results.

discussion: analyze the results, compare with what was done before, did you get good/bad results?, were these results expected? discuss

future: ideas for how to improve? other things that could be done?

1. More information about the data set as well as the possibility of downloading it can be found in <https://archive.ics.uci.edu/ml/datasets/adult> [↑](#footnote-ref-1)
2. Evaluation and information just explained is found in <http://www.cs.toronto.edu/~delve/data/adult/adultDetail.html> [↑](#footnote-ref-2)
3. Since it is an algorithm explained in class, only the main idea of it is given [↑](#footnote-ref-3)