Feature Selection using Association Rules

Project in Tabular Data Science Couse

Ben Abraham Nageris

March 6, 2023

Abstract

Feature selection is a method of selecting the most relevant features to include in the model construction. Most solutions to feature selection can be split into two main categories: supervised and unsupervised, as each one of the has its advantages and disadvantages. One of the most used techniques to feature selection is based on correlation metrics (usually Chi-Squared and Pearson). This paper presents an alternative approach to performing feature selection using association rules, a commonly used method that originated in data mining. Association Rules is an advanced rule-based machine learning method. This method discovers relations between variables in large datasets and creates if-then rules representation of the relations between a subset of columns to another subset of columns. The new approach presented in this paper leverages "Association Rules" to find and prioritize the different variables related to the target column. This paper presents a comparison between this technique and chi-squared (correlation metric) and evaluated on four datasets from different domains. Empirical results show that this approach performs slightly better than chi-squared.

1 Problem Description

Predictive Model Analytics is a fundamental element in the Data Science pipeline. This step mainly concerns historical data assessment, patterns discovery, trends identification and eventually leveraging this information to predict future trends. Feature selection is one of the most significant problems in predictive model analytics. Selecting the right subset of features to model prediction is necessary to improve data compatibility, simplify the model, shorter training times as well as improve the model explainability. In addition, as data scientists want to create a generalized solution to predict a phenomenon, they would rather have the smallest amount of features (which generalize the problem) and eventually reduce the risk of overfitting.

Solutions to feature selection can be split into two main categories: supervised and unsupervised techniques, and each one has its own advantages and disadvantages. On the one hand, the supervised technique concerns only labeled data and is mainly used in regression and classification models. On the other hand, unsupervised techniques can be used for unlabeled data. This technique includes four types of methods: filter (information gain, Chi-Squared, correlation efficient), wrapper (forward feature selection, recursive feature elimination, exhaustive feature selection), embedded (random forest importance, LASSO regularization), and hybrid methods which combines two or more methods together. Most Data Science pipelines, in practice, use one of the techniques mentioned above. This work focuses on creating an alternative approach to feature selection based on *Association Rules*. Association Rules is an advanced rule-based machine learning method that was originally made for data mining. This method discovers relations between variables in large datasets by searching for frequent if-then patterns and by using three metrics to determine the importance of each pattern. The three metrics used in association rules are: support, confidence and lift.

The first two metrics are:

- Support of an itemset (IS) is the ratio of the occurrence of this itemset in the database.
- Confidence is defined by how many times the if-then statements are found true.

$$support(IS) = \frac{|\{t \in T : IS \subseteq t\}|}{|T|} \text{ , } confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)}.$$

Beyond these two essential metrics, the third metric is *lift*. Lift is defined by the ratio between the expected confidence and the number of if-then statements that are expected to be found true (compares the ratio between the expected Confidence and the actual Confidence).

$$lift(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X) \cdot support(Y)}$$

There are many advantages of using *Association rules* like distinguishing patterns in the dataset as well as understanding correlations and co-occurrences between data sets. In the medical world, association rules are commonly used to help patients diagnosis because many diseases share a subset of the symptoms [9]. Using Association Rules in this domain helps the doctor to find the diagnosis distribution over the subset of symptoms.

This work intends to find the relation between the target column and a subset of features related to this target column by exploiting the different metrics used in *Association Rules* (support, confidence and lift values). In addition, this work leverages association rules' advantages to suggest an automatic pipeline to feature selection using association rules. This work uses a different pattern mining technique generally used to find the influence of the different features and includes them as the model features to predict the target column phenomenon, as this work suggests a new metric discussed later in the paper instead of correlation.

2 Solution overview

2.1 Solution General Approach

The project idea is to make it easy for a data scientist to decide the features in the model's presentation of the problem as well as get an immediate evaluation of this subset of feature selection. The vision behind this proposed automatic pipeline is to computerize and accelerate the feature selection process done usually by data scientists manually or semi-manually nowadays. The solution idea is to leverage association rules into feature selection due to the success of association rules in finding relations between different features in the database. The solution exploits association rules to distinguish and prioritize the most relevant set of features to include in the model's generalized representation via a feature occurrence counter. The ambition is to create an alternative approach to perform feature selection and to suggest a new metric.

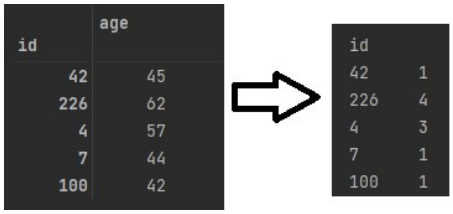
The configuration of the tool is made from a few elements:

- Id column name
- Target column name
- Minimum support
- Minimum confidence
- Requested number of features

These five parameters are the only parameters of the pipeline making the proposed solution easy to adjust to a new dataset. It's worth mentioning that besides these parameters specifications, all steps (preprocessing, feature selection, prioritization and evaluation) happen automatically, therefore these parameters affect directly the model's representation complexability and indirectly its performance and eventually choose the right set of parameters to reach the best fit with their computation and data representation limit.

2.2 Solution Steps

1. **Data Processing** is the first step in the pipeline. This step is responsible for splitting the data into 3 main types (categorical, very numerical and ordinals). Very numerical refers to columns that have more than 20 unique values. In this step, very numerical values are split into bins - binning (Figure 3) and null values get the column's average score. In addition, columns that have more than 70% of null values are being deleted. Beyond that, one-hot-column or label encoding is executed on the categorical columns (depending on the number of different values).



id	age
42	45
226	62
4	57
7	44
100	42

id	
42	1
226	4
4	3
7	1
100	1

Figure 1: Binning: transforming original value to given small interval (bins)

2. After the preprocessing stage the **Apriori Algorithm** [1] (Figure 2) is executed. The first stage in this part is transforming the processed dataset into transactions. Those transactions are the input to the *apriori algorithm* with the minimum support confidence (depending on the dataset) and output if-then rules which suffice the confidence and support input value. Those rules are the building blocks to the feature selector.

	lhs		rhs	support	confidence	lift	count
[1]	{}	=>	{whole milk}	0.25551601	0.2555160	1.000000	2513
[2]	{hard cheese}	=>	{whole milk}	0.01006609	0.4107884	1.607682	99
[3]	{butter milk}	=>	{other vegetables}	0.01037112	0.3709091	1.916916	102
[4]	{butter milk}	=>	{whole milk}	0.01159126	0.4145455	1.622385	114
[5]	{ham}	=>	{whole milk}	0.01148958	0.4414062	1.727509	113
[6]	{sliced cheese}	=>	{whole milk}	0.01077783	0.4398340	1.721356	106
[7]	{oil}	=>	{whole milk}	0.01128622	0.4021739	1.573968	111
[8]	{onions}	=>	{other vegetables}	0.01423488	0.4590164	2.372268	140
[9]	{onions}	=>	{whole milk}	0.01209964	0.3901639	1.526965	119
[10]	{berries}	=>	{yogurt}	0.01057448	0.3180428	2.279848	104

Figure 2: Example for Apriori rules output

3. **Filter** the relevant rules to the predicted target column. This step's input is the apriori algorithm output which is constructed of numerous if-then rules. This step is responsible for filtering out the rules whose then-clause (rhs column in Figure 2) doesn't contain only the target column.

4. In order to decide which feature to include first in the model we have to **prioritize the features**. In this part, the proposed solution saves a frequency structure to count each feature's number of occurrences in the if-clause (lhs column in Figure 2) and sorts in descending order. In addition, one of the pipeline's inputs is the number of features desired to be in the model ($k \in \mathbb{N}$), therefore this step takes only the top-k most prioritized features to the target column.

$$priority(f \in features) = count(f \in aprioriOutputRules)$$

5. In order to test the solution we have to **evaluate** the prediction. This part executes a training on the training data and eventually evaluation on the test data and reports the prediction success ratio over the test dataset. In the attached jupyter notebook the prediction model is the basic decision tree classifier but it can be replaced by any other learning model (depends on the domain settings and prediction requirements).

$$evaluation(festureSelector, Test) = \frac{count(correctPredictions)}{|Test|}$$

6. The solution conducts a **Comparison** between the proposed solution and the chi-squared correlation-based feature selector. This part executes stages 4-5 again but with a different feature prioritizing algorithm. In this step, the feature importance prioritizing is calculated with *chi-squared* which is a very common correlation metric to find relations between two columns.

   Like before, the pipeline chooses the top-k (k is the number of features to include in the model) features which score the lowest p-value (in the *chi-squared* calculation). Then, the pipeline compares the two model's performance on the test dataset and reports their success ratio.

	col1	col2	score	p_val
6	thall	output	63.719537	1.457063e-14
9	cp	output	51.077915	4.708626e-11
12	caa	output	49.332283	4.977274e-10
7	exng	output	27.401796	1.652777e-07
8	slp	output	30.476974	2.409957e-07
11	sex	output	9.367343	2.208854e-03
10	restecg	output	11.675697	2.915108e-03
4	oldpeak	output	53.901137	2.795754e-02

Figure 3: Example of the chi-squared values returned in the dataset "heart Attack" 3.1

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Figure 4: Chi Squared formula

3 Experimental evaluation

3.1 Datasets

The proposed solution was tested and evaluated on the following four datasets taken from Kaggle:

1. **Heart Attack** [11] - This dataset is constructed of 13 columns representing the patient's health status and a column in which the patient has more or less chance of heart attack, 70% of the columns have 4 unique values or less.

2. **Airlines Delay** [10] - This dataset is constructed of 8 columns and the task is to predict whether a given flight will be delayed, given the information of the scheduled departure.

3. **Mobile Price Range** [12] - This dataset is constructed of 21 columns representing the cell-phone's technical specifications (support for wifi, dual sim, ram size, etc.) and the problem is to predict the price category this phone is in where 0 is the cheapest ad 3 is the highest price.

4. **Home Loan Approval** [7] - This dataset is constructed of 12 columns representing the loan requestor details, for example: education, sex, loan term, and property location. The problem is to automate the loan eligibility process (real-time) based on customer detail and therefore return whether this customer is a eligible for loan or not.

All these datasets represent different domain knowledge as well as different prediction problem and data types. Taking these datasets, from a variety of domains, were taken to be a fair comparison between the proposed solution and the state-of-the-art.

3.2 Evaluation

To evaluate the model prediction ratio, after each feature selector (Association-based or chi-squared-based) outputs its top-k most crucial features. Each k-features is sliced from the same dataset and trained using the naive decision tree classifier. Afterward, each trained decision tree is evaluated independently on the same test data, and the prediction success rate is being calculated. To compare the two approaches to feature selection we compare both success rates (the rates are written down in the table 3.3).

$$evaluation(festureSelector, Test) = \frac{count(correctPredictions)}{|Test|}$$

3.3 Results

Database	Association Rules as Feature Selector	Chi-Squared as Feature Selector	number of features (k)
Heart Attack	**0.835**	0.810	8
Home Loan Approval	**0.847**	**0.847**	4
Airlines Delay	0.6	**0.65**	3
Mobile Price Range	**0.245**	0.235	6

The table mentioned above presents the performance comparison between the two feature selectors, chi-squared and association rules-based, over the databases mentioned in 3.1.

The Heart attack [11] database had the best absolute advantage in feature selecting of the proposed solution, while Mobile Price range dataset [12] had the most relative improvement (4.2%).
In other words, empirical results show that the new approach to performing feature selection (feature selection using association rules) exceeded expectations and actually got a higher success prediction rate than the chi-squared based. In one database, Home loan Approval[7], both feature selectors mentioned prioritized the same features and therefore got the same in the evaluation phase. Nevertheless, in the Airline Delay dataset chi-squared actually scored better and achieved a more successful prediction rate. The main reason why both metrics (*chi-squared* and *association-rules based metric*) got around 60% correct prediction is that 7 out of 8 of these columns are categorical more than 1000 unique values. Error analysis showed that columns that had fewer unique values tended to occur more in the rules and eventually if the dataset is mostly constructed of these kinds of columns to even get a better prediction. This conclusion rose up, especially in the heart attack prediction dataset where 70% of the features have 4 or fewer unique values. To sum up, surprisingly, association-rules-based feature selector scored the same or better in three out of the four datasets.

4 Related Work

Feature Selection refers to distinguishing the most related columns to the target in order to find a subset of features that best represents the problem. Solutions to feature selection can be split into two main categories: supervised and unsupervised techniques [8], and each one has its own advantages and disadvantages. On the one hand, the supervised technique concerns only labeled data and mainly used in regression and classification models. On the other hand, unsupervised techniques can be used for unlabeled data. This technique includes four types of methods: filter (information gain, Chi-Squared, correlation efficient), wrapper (forward feature selection, recursive feature elimination, exhaustive feature selection), embedded (random forest importance, LASSO regularization), and hybrid methods which combines two or more methods together [2].

One of the most frequent solutions to this feature selection is using correlation [5, 6, 4] as most people refer to correlation as the obvious "go-to" to perform feature selection. All those works use correlation as the metric to quantify the relationship between each column and the target column. One of the problems of this problem is that sometimes people revolt against statistical tools as it comprises tough mathematics and knowledge.

Moreover, recent feature selection research abandoned those conventional metrics and rather use neural networks instead [13]. Verikas et al. created a new feature selection algorithm for classification that is based on a neural network instead, denoting its flexibility and determining the data's inner pattern and behavior.

Another approach to perform feature selection is using association rules, Chawla et al. [3] created an association rule-based feature selection algorithm. Chawla et al. algorithm creates a relations network of features based on association rules and eventually extracts the important features out of the network's links.

This project further explored the usage of feature selection using association rules and intimidate the obvious step of using correlation by using a different association rules-based computation to calculate the relation between the features and the target column.

5 Conclusion

In this project, a new approach to feature selection and automatic tool implementation is proposed. This work represents an alternative approach to performing feature selection since it doesn't use conventional metrics and created an alternative approach based on Association Rules. The proposed solution is very easy to use and adjusts easily to a new dataset and domain. In addition, this alternative method over-performed the chi-squared in those 4 datasets which, in practice, indicates that feature selection using Association Rules can be useful in real-life scenarios. This project draws attention to the fact that simple mathematical calculations (support, confidence and lift) sometimes succeed in the performance of well-complicated statistical tools (chi-squared) and work efficiently. This work emphasizes the benefits of importing techniques from other fields of interest. Many of nowadays breakthroughs in science are related to the integration of fields and this work is a living example.

5.1 Things I learned

Since I am new to data science and research in particular, a lot of the time spent on the project was invested in reading the articles and the different approaches to feature selection and getting inspiration. After I finished reading, many of the difficulties were to define the pipeline's design and determine clear steps for the research and implementation. While working on the data, I saw the importance of correct and solid pre-processing, because when the binning was too rough, the Apriori algorithm didn't find a sufficient set of rules. Afterward, I saw the importance of selecting the right support and confidence values to create enough sufficient rules. Planning the research beforehand and having an abstract version of the research (the research proposal) really focused the work on the project's objectives and goals. Moreover, thanks to the project, I learned to use decision trees in practice for the first time and had practical and theoretical work.

5.2 Future Steps

In the future, people can further explore the advantages of importing techniques from different fields of interest.

In addition, although evaluating this solution on 4 different datasets from different domains it's recommended to further explore the advantages of using this technique as well as the types of data that suit this technique the best.

Moreover, researching different prediction models and pre-processing techniques can benefit this research and further elevate the use of these kinds of techniques in data science pipelines.

Another further research is to gradient decent the success rate of the model (using different support and confidence values) in order to find the local optimum of the feature selection.

References

[1] Rakesh Agarwal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, volume 487, page 499, 1994.

[2] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[3] Sanjay Chawla. Feature selection, association rules network and theory building. In *Feature Selection in Data Mining*, pages 14–21. PMLR, 2010.

[4] N Gopika and A Meena Kowshalaya ME. Correlation based feature selection algorithm for machine learning. In *2018 3rd international conference on communication and electronics systems (ICCES)*, pages 692–695. IEEE, 2018.

[5] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[6] Mark A Hall. Correlation-based feature selection of discrete and numeric class machine learning. 2000.

[7] Rushikesh Konapure. Home Loan Prediction. https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval, 2022.

[8] Vipin Kumar and Sonajharia Minz. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014.

[9] Carlos Ordonez, Cesar A Santana, and Levien De Braal. Discovering interesting association rules in medical data. In *ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, pages 78–85, 2000.

[10] Ulrik Thyge Pedersen. Can you predict when a flight in delayed? https://www.kaggle.com/datasets/ulrikthygepedersen/airlines-delay/, 2023.

[11] Rashik Rahman. Heart Attack Analysis & Prediction Dataset. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/, 2021.

[12] Abhishek Sharma. Predict the mobile price range. https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification/, 2018.

[13] Antanas Verikas and Marija Bacauskiene. Feature selection with neural networks. *Pattern recognition letters*, 23(11):1323–1335, 2002.