Assignment 1 - Association Rules Analysis on the Solar Flare Dataset

Tanushree Kumar

DATA 630 - Summer 2024

Professor Ami Gates

University of Maryland Global Campus

Due: June 4, 2024

Introduction

The objective of this analysis is to perform association rules analysis on the solar flare dataset using the Apriori algorithm to uncover significant relationships between different attributes of solar flares. Specifically, the aim is to identify patterns and associations between various characteristics of solar flares, such as their classification, activity levels, and other attributes that can help in understanding their behavior and potentially improve predictive models.

Solar flares are sudden bursts of energy, characterized by "large eruptions of electromagnetic radiation" (National Oceanic and Atmospheric Administration, n.d.) emitted by the Sun, often associated with sunspots and solar magnetic activity. This burst of energy can be detected as a burst of radiation ranging "across the electromagnetic spectrum, from radio waves to X-rays and gamma rays" (European Space Agency, 2019).

Understanding the characteristics and behavior of solar flares is important as these flares can have a huge impact on space weather forecasting, satellite operations, power grids, and other technological infrastructure on Earth. It's imperative that there is adequate information surrounding this topic in order to mitigate the impacts felt on Earth. For example, the occurrence of solar flares has been linked to increased electromagnetic disturbances, which can disrupt technology and pose risks to astronauts. This dataset contains information about different classes of solar flares and their associated attributes.

The association rules analysis is utilized for identifying hidden relationships within datasets, making it an ideal method for this dataset. By applying association rules mining, frequent itemsets can be discovered and rules can be generated to explain relationships between

the various attributes of solar flares. By setting appropriate support and confidence thresholds, the focus can be shifted to finding significant and reliable patterns.

Analysis

The solar flare dataset, obtained from the UCI Machine Learning Repository (Bradshaw, 1989), contains thirteen attributes describing the characteristics of solar flares:

- Numerical classes of solar flares
    - c-class: Number of C-class flares in the following 24 hours; categorized as common flares
    - m-class: Number of M-class flares in the following 24 hours; categorized as moderate flares
    - x-class: Number of X-class flares in the following 24 hours; categorized as severe flares
- Categorical attributes describing sunspot properties
    - mod_zurich_class: Modified Zurich class of the solar region; denoted as A, B, C, D, E, F, H
    - largest_spot_size: Size of the largest spot in the solar region; denoted as X, R, S, A, H, K
    - spot_distribution: Distribution of the spots in the solar region; denoted as X, O, I, C
- Categorical attributes describing solar activity and history
    - activity: Activity level of the solar region; 1 = reduced, 2 = unchanged
    - evolution: Evolution of the solar region over the past 24 hours; 1 = decay, 2 = no growth, 3 = growth

    - previous_day_activity: Activity level of the solar region in the previous 24 hours; 1 = nothing as big as an M1, 2 = one M1, 3 = more activity than one M1
    - hist_complex: Historical complexity of the solar region; 1 = Yes, 2 = No
    - become_hist_complex: Whether the region has become historically complex; 1 = yes, 2 = no
- Categorical attributes describing the area of sunspots
    - area: Total area of the solar region; 1 = small, 2 = large
    - area_largest: Area of the largest spot in the solar region; 1 = <=5, 2 = >5

The categorical attributes describing sunspot properties are part of a sunspot group classification. Each column is mapped respectively to the “general form 'Zpc', where 'Z' is the modified Zurich Class, 'p' describes the penumbra of the principal spot, and 'c' describes the distribution of spots” (Codes, Terminology and Classifications, n.d.).

It is important to note the difference between the three class flares mentioned. Since the radiation can be emitted in X-rays, solar flares are classified according to their peak brightness in X-ray wavelengths. There are five categories used by scientists, however, the focus will be on the three categories found in this dataset, in the order from most to least intense:

- X-class: big solar flares; cause radio blackouts and radiation storms
- M-class: medium-sized solar flares; cause brief radio blackouts and minor radiation storms
- C-class: small solar flares; 10 times less powerful than an M-class flare (European Space Agency, 2019)

Before developing the model, exploratory data analysis (EDA) was performed to understand the distribution and relationships of the variables. Functions like summary() and str() were used to get an overview of the data.

Based on the EDA, data preprocessing was performed such as handling missing values, encoding categorical variables into factors, and scaling numerical features ensuring the data is in an appropriate format for association rules mining. From the dataset description URL (Bradshaw, 1989), it was mentioned that there were no missing values, thus no preprocessing on that end was not required. The categorical variables were converted to factors to facilitate the mining process. The numerical variables of ‘c-class, ‘m-class’, and ‘x-class’ were scaled and discretized into categories of ‘low, ‘medium’, and ‘high’ respectively to enable the Apriori algorithm to process them.

Summary statistics and the data structure were examined to understand the distribution of values and the relationships between different attributes. The dataset shows a high prevalence of "low" class solar flares and common patterns in sunspot characteristics and activities. These insights lay a solid foundation for applying the Apriori algorithm to uncover significant association rules and enhance our understanding of solar flare activities.

The Apriori algorithm was used for this analysis, which worked by identifying the frequent itemsets in the dataset and then generating association rules based on the identified itemsets. The algorithm relies on three key parameters: support, confidence, and lift. Support is defined as the proportion of transactions that contain the itemset. Confidence is the likelihood that a rule is true given the antecedent. Lift is the ratio of the observed support to that expected if the items were independent.

The modeling process involved converting the dataset into a transaction format that is suitable for the Apriori algorithm. Running the algorithm generated the frequent itemsets and rules. The support and confidence parameters were fine-tuned to extract the rules, which were then evaluated based on the support, confidence, and lift metrics. The Apriori algorithm was applied to the dataset with different support and confidence thresholds to generate association rules. The default parameters generated rules with a minimum support of 0.1 and a confidence of 0.6. The adjusted parameters used a lower support threshold of 0.05 and a higher confidence of 0.7 to identify more specific and reliable rules. It's important to adjust and check these parameters as a high support threshold might miss interesting rules, while a low threshold might produce too many insignificant rules. Similarly, setting an appropriate confidence level helps in filtering out weak associations. The rules were sorted by the highest lift values to prioritize the most significant associations, The redundant rules were then identified and removed to simplify the rule set.

Results

The output of the association rule mining consists of discovered rules describing relationships between different solar flare attributes. Each rule consists of antecedent (if) and consequent (then) parts, where the antecedent represents the conditions under which the rule applies, and the consequent represents the outcome or prediction. For example, a rule may state: that if the flare classification is x-class and the spot size is large, then the activity level is high.

With the default support (0.1) and confidence (0.6) levels, the Apriori algorithm generated a total of 65,633 rules. The average support of the rules found with the default parameters is approximately 0.1669, taken from the summary of the generated rules. With the support set to 0.05 and confidence set to 0.7, the Apriori algorithm generated a total of 144,718

rules. The average confidence of the rules found with the support of 0.05 and confidence of 0.7 is approximately 0.9427, taken from the summary of the generated rules. After sorting by lift and removing redundant rules, the output shows the first 10 pruned (non-redundant) rules. These rules can be seen in Figure 1 in Appendix A.

The pruned association rules show significant relationships between various attributes of the solar flare dataset. For instance, Rule 1 indicates a strong relationship between ‘spot_distribution=C’ and ‘largest_spot_size=K’. Rules 3 and 4 show deterministic relationships for ‘become_hist_complex=1’, indicating specific classifications (‘mod_zurich_class=H’ and ‘spot_distribution=X’).

All the top 10 rules have a high lift value of approximately 4.98, indicating strong associations between the antecedents and consequents. Rules with a confidence of 1.0 are deterministic, meaning the consequent always follows when the antecedent is present.

The rules can be interpreted in relation to the objective. Rule 1 of {spot_distribution=C} -> {largest_spot_size=K} (confidence: 0.9474, lift: 6.3947) with ‘spot distribution C’ referring to the specific distribution pattern of sunspots where C class sunspots are a bipolar sunspot group. The ‘largest spot size’ of K indicates that the largest sunspot size falls into the ‘K’ p-value of sunspots; where K-sized sunspots are large and asymmetric (Codes, Terminology and Classifications, n.d.). This rule indicates that when sunspots exhibit the distribution pattern "C", there is a very high likelihood, 94.74% confidence, that the largest sunspot will be in the "K" size category. The high lift value of 6.39 suggests that this association is much stronger than would be expected by chance. This relationship is critical for predicting the potential size of sunspots based on their distribution, which can be important for anticipating the energy release and potential impacts of solar flares.

Rule 2 of {previous_day_activity=3} => {activity=2} (confidence: 0.7096, lift: 5.1096) with ‘previous day activity’ of 3 indicates a high level of observed solar activity from the previous day. The ‘activity’ of 2 refers to a moderate level of activity on the current day. This rule suggests a relationship where a high level of activity on one day (previous_day_activity = 3) is likely followed by a moderate level of activity the next day (activity = 2). With a confidence of 70.97% and a high lift value, this pattern indicates a potential cycle or decay pattern in solar activity. Understanding such patterns can aid in forecasting daily solar activities, thereby assisting in the prediction of solar flares.

Rule 3 of {become_hist_complex=1} => {mod_zurich_class=H} (confidence: 1.0000, lift: 4.9846) with ‘become historically complex’ value of 1 indicates that the sunspot group has developed into a historically complex configuration. The ‘modified Zurich class’ of H refers to a unipolar group of sunspots. This rule asserts that every time a sunspot group becomes historically complex (become_hist_complex = 1), it will be classified as Zurich class H with absolute certainty (100% confidence). The strong lift value supports the robustness of this rule. This insight can be utilized to classify and predict the behavior of sunspot groups.

Rule 4 of {become_hist_complex=1} => {spot_distribution=X} (confidence: 1.0000, lift: 4.9846) with ‘become historically complex’ value of 1 indicates that the sunspot group has developed into a historically complex configuration. The ‘spot distribution’ of X refers to an undefined distribution of the unipolar sunspots. This rule demonstrates that when sunspot groups become historically complex (become_hist_complex = 1), they always exhibit the spot distribution pattern "X". This suggests a definitive relationship between complexity and the spatial distribution of sunspots.

Rule 5 of {mod_zurich_class=B} => {largest_spot_size=X} (confidence: 1.0000, lift: 4.9846) with ‘modified Zurich class’ of B indicates a bipolar sunspot group with no penumbra for any spots. The ‘largest spot size’ X indicates a sunspot with no penumbra. This rule indicates that all sunspot groups classified as Zurich class B will have the largest spot size in category X. The high confidence and lift values confirm this as a strong, reliable rule. This pattern helps in understanding the characteristics of class B sunspots and predicting their potential size.

Rule 6 of {largest_spot_size=X} => {mod_zurich_class=B} (confidence: 1.0000, lift: 4.9846) is an inverse of Rule 5. This further confirms that all sunspots with the largest spot size classified as X are always classified as Zurich class B. This mutual relationship enhances the understanding and predictability of this particular sunspot group's characteristics.

Rule 7 of {mod_zurich_class=H} => {spot_distribution=X} (confidence: 1.0000, lift: 4.9846) with ‘modified Zurich class’ of H refers to a unipolar group of sunspots. The ‘spot distribution’ of X refers to an undefined distribution of the unipolar sunspots. This rule indicates that all sunspot groups classified as Zurich class H will exhibit the spot distribution pattern X. The strong, definitive association between Zurich class H and spot distribution X highlights a critical characteristic of unipolar sunspot groups.

Rule 8 of {spot_distribution=X} => {mod_zurich_class=H} (confidence: 1.0000, lift: 4.9846) is an inverse of Rule 7, confirming that all sunspots with distribution type X are always classified as Zurich class H.

Rule 9 of {spot_distribution=O, evolution=3, area_largest=1} => {mod_zurich_class=B} (confidence: 0.7742, lift: 3.8591) with ‘spot distribution’ of O represents an open distribution pattern of sunspots. The ‘evolution’ value of 3 indicates a growth in the evolution of sunspots. The ‘largest area’ value of 1 denotes a small area for the largest sunspot. This rule combines

multiple attributes to predict that sunspot groups with distribution type O, at evolution stage 3, and the largest area in category 1 are likely to be classified as Zurich class B. This multifaceted approach offers a deeper understanding of how various factors interact to determine sunspot classification.

Rule 10 of {spot_distribution=I, mod_zurich_class=F} => {mod_zurich_class=B} (confidence: 0.7742, lift: 3.8591) with ‘spot distribution’ of I represents an intermediate distribution of sunspots where the spots lie scattered between the beginning and end of the group. The ‘modified Zurich class’ F is an elongated bipolar sunspot. This rule suggests that sunspot groups with distribution type I and initially classified as Zurich class F are likely to be reclassified as Zurich class B. This transition hints at the dynamic nature of sunspot classification and evolution, providing insights into how sunspots develop and potentially lead to solar flares.

These rules reveal intricate relationships between various characteristics of sunspots and their classifications, which are crucial for understanding and predicting solar flares. The high confidence and lift values across these rules indicate strong, reliable patterns that can be used for forecasting solar activities. By understanding these relationships, scientists can better anticipate solar events, contributing to space weather forecasting and mitigation strategies.

Conclusion

The main findings were that significant relationships were identified between various attributes of solar flares. For instance, the Zurich class 'H' is strongly associated with spot distribution 'X'. The Apriori algorithm successfully discovered meaningful patterns that can potentially aid in predicting solar flare activities.

As with any analysis, there are some limitations to be seen. The discretization process may lead to loss of granularity in the data which can impact the accuracy and usefulness of the

results. The support and confidence thresholds need careful tuning to balance the discovery of significant rules and the exclusion of less relevant ones. The analysis is limited to the associations present in the dataset and may not capture all possible patterns due to the dataset's constraints.

Future work could focus on integrating additional data sources, such as solar imagery and magnetic field measurements, to improve the predictive accuracy of association rules. With the inclusion of more data, different discretization techniques could be applied to test new thresholds to optimize the rule discovery process. Furthermore, exploring alternative algorithms may offer insights into more complex patterns in solar flare behavior.

References

Bradshaw, G. (1989, March). Solar Flare database. Uci.edu. https://archive.ics.uci.edu/ml/machine-learning-databases/solar-flare/flare.names

Codes, Terminology and Classifications. (n.d.). Www.sidc.be; SIDC - Royal Observatory of Belgium. Retrieved June 4, 2024, from https://www.sidc.be/educational/classification.php

European Space Agency. (2019). ESA - What are solar flares? Esa.int. https://www.esa.int/Science_Exploration/Space_Science/What_are_solar_flares

National Oceanic and Atmospheric Administration. (n.d.). Solar Flares (Radio Blackouts) | NOAA / NWS Space Weather Prediction Center. Www.swpc.noaa.gov. Retrieved June 4, 2024, from https://www.swpc.noaa.gov/phenomena/solar-flares-radio-blackouts

Appendix A

All figures mentioned and visualizations produced in the report can be seen below. The R code is attached as a separate file.

Figure 1

Top ten pruned rules evaluated using lift.

```
> #Inspecting the top 10 pruned rules
> inspect(rules_pruned[1:10])
     lhs                                                rhs                    support    confidence coverage   lift     count
[1]  {spot_distribution=C}                           => {largest_spot_size=K} 0.05555556 0.9473684  0.05864198 6.394737 18
[2]  {previous_day_activity=3}                       => {activity=2}          0.06790123 0.7096774  0.09567901 5.109677 22
[3]  {become_hist_complex=1}                         => {mod_zurich_class=H}  0.05246914 1.0000000  0.05246914 4.984615 17
[4]  {become_hist_complex=1}                         => {spot_distribution=X} 0.05246914 1.0000000  0.05246914 4.984615 17
[5]  {mod_zurich_class=B}                            => {largest_spot_size=X} 0.20061728 1.0000000  0.20061728 4.984615 65
[6]  {largest_spot_size=X}                           => {mod_zurich_class=B}  0.20061728 1.0000000  0.20061728 4.984615 65
[7]  {mod_zurich_class=H}                            => {spot_distribution=X} 0.20061728 1.0000000  0.20061728 4.984615 65
[8]  {spot_distribution=X}                           => {mod_zurich_class=H}  0.20061728 1.0000000  0.20061728 4.984615 65
[9]  {spot_distribution=O, evolution=3, area_largest=1} => {mod_zurich_class=B}  0.07407407 0.7741935  0.09567901 3.859057 24
[10] {spot_distribution=O, evolution=3, area_largest=1} => {largest_spot_size=X} 0.07407407 0.7741935  0.09567901 3.859057 24
```

Figure 2

Graph of the rules.

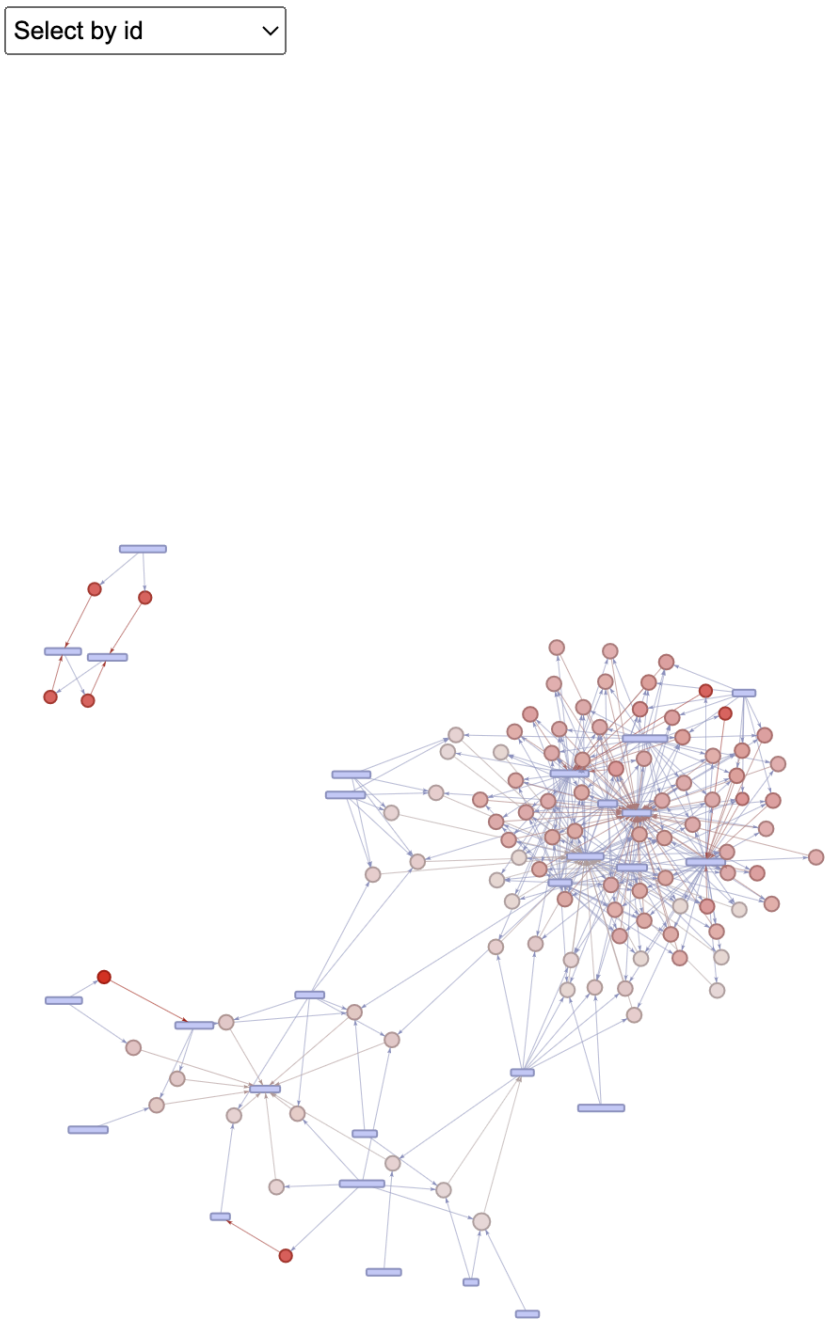

Figure 3

Scatterplot of the rules.

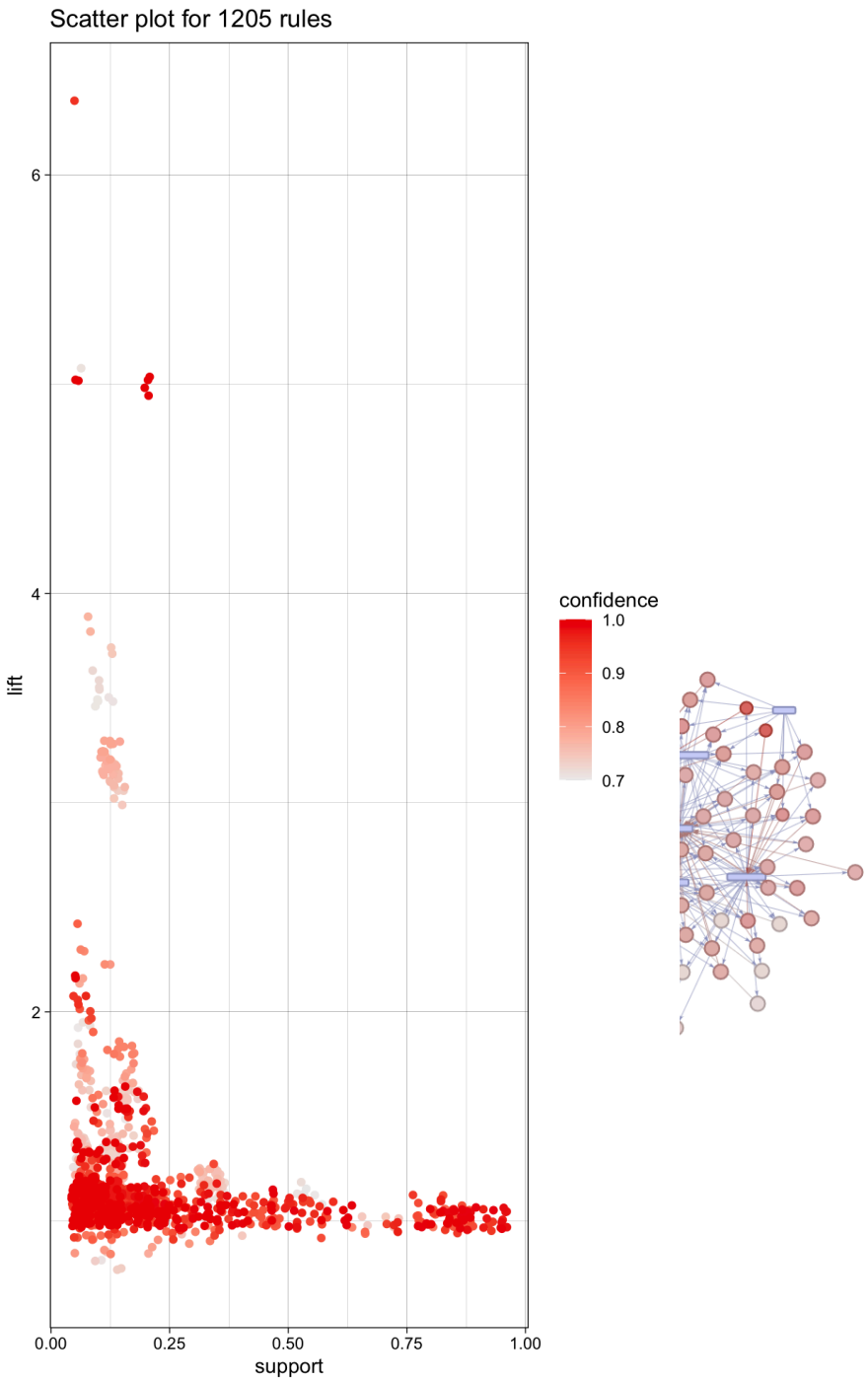

Figure 4

Grouped plot of the rules.