Mining Smartphones sales record

**Group Members**

**Duo zhang, Guanqi Guo**

## Abstract

As we know, data mining is a step of “Knowledge-Discovery in Databases” (KDD), and we can search the necessary hidden information from a mass of data through data mining process. [1] We have three major goals in this project: First, find the most related factor that influences the price most. Second, find the relationship between popular value, price, and monthly sale. Third, prove recommendations to consumers and producers by using the analysis result.

## Introduction

***Background of dataset:***

The “smartphones sales record.xlsx” was coming from the most popular online shopping website of China, called: “TianMao”. It is the official online store of “Taobao” which is belongs to Alibaba Group. Alibaba Group is a Chinese E-commerce company that provides consumer-to-consumer, business-to-consumer and business-to-business sales services via web portals. [1] Since, it has the good reputation in the international community. Therefore, we think the dataset is trustworthy.

***Reason for choosing projects:***

With the development of consumer electronics industry, a variety of consumer electronics products went into our field of vision. Due to the different types and configurations of the smartphones, consumers and producers are all facing with big challenges. Moreover, because of our team members are interested in digital products. Therefore, we try to use data mining method to analysis the smart phone sells data, by using this, to give producers and consumers some useful recommendations. [4]

***Main questions aimed to answer:***

1. Find the most related attribute to the price.
2. Find the relationship between popular value, price, and monthly sale.
3. Making recommendations to customers and producers.

## Methods

***Mode*** means the value which appears mostly in the data. It can express the majority level of the data.

***Average number*** means the trend of the data. It shows the whole level of the data.

***Information Gain*** can explain the ability of an attribute to separate the data. That means, if the value is bigger, it can better represent the whole data. It also has much more connection to the class attribute.

***Lift*** can express the relationship between two groups. If lift=1, that means these two groups are independent. If the lift value is bigger than 1, these two groups have positive correlation. If the lift value is smaller than 1, these two groups have negative correlation.

***Leverage*** can also express the relationship between two groups. If leverage =0, that means these two groups are independent. Moreover, the bigger the value, the more deep positive correlation they have.

***Apriori Algorithm*** is created as a seminal algorithm for mining frequent item sets about Boolean association. It uses prior knowledge, so it is called Apriori. The algorithm does iteration for every layer. We can use this method to find the confidence and support.

## Results & Discussion

Data preprocessing

1. ***Filling missing value***

In this step our team chooses to fill the missing value manually, because we think that fill the missing value manually can improve the credibility and accuracy of the dataset. By searching on the Internet, we found a pattern can replace the missing value of the smart phone effectively and accurately. Through the survey, we know that the hardware configurations of smart phones are often closely related to the brand, means generally the phones which are in same brand; their hardware configurations are similar in a high degree. Therefore, we use the mode value or mean value of an attribute which is in the same brand to replace the missing value. For example, in Figure 1.1, there are some missing values in attribute storage. First, we find which brand is this attribute belongs to, and then we find all the values for storage which belongs to the same brand. At last we use the 8MB (the mode value) to replace the missing values. We use the same method to replace the missing values in operating system, battery type, Storage, Ram, and screen type.



Figure 1.1

1. ***Data Transformation***

To make the results of data mining become more credible, and make the data mining process more efficient, and the patterns found may be easier to understand. We transform some of the values into some specific types.

1. ***Operating system attributes***

As we know that, as the smart phone becomes more and more popular in people's lives, the mobile phone manufacturers are also increasing. Most of the manufacturers prefer to develop their own operating system to suit their products. Therefore, the type of operating system has become very rich. However, most of the operating system is based on the secondary development of Android. When consumers choose mobile phones, they will not care about these details, they only care about the phone is IOS or Android or some other well-known operating system. Therefore, through searching on the official website of the phone, we make all Android-based operating system renamed as “Based on Android”. Like Figure 2.1.

Figure 2.1

1. ***Battery Type attributes***

Also, for mobile phone battery model, we are only concerned about whether it is removable, because most of the users only care about whether their phones are integral or removable. Like Figure 2.2.

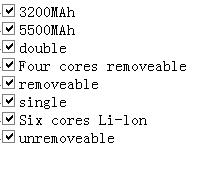
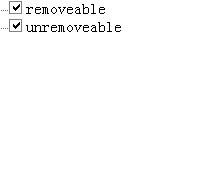
 

Figure 2.2

1. ***Size attributes***

For size attribute, some of the attribute was recorded in different form. As we know according to international practice, the size of mobile phone usually identified with a diagonal length, and the unit is feet. Therefore, we use “Trigonometric formulas” to transform the attributes into the usual format. For example, in Figure 2.3, the first attribute was recorded as length\* width\* thickness. Therefore, we just ignore the thickness, and calculate the diagonal= ≈156. 47 mm, then transform the unit into inches, 156.47mm = 6.14 inches. Unified data format, make data mining becomes easy.

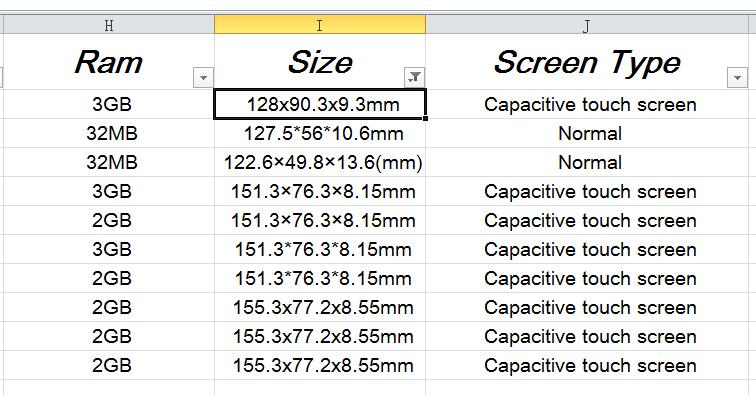


Figure 2.3

1. ***Screen Resolution attributes***

Screen resolution means the number of pixels per unit area on the screen. Commonly used data mining software can’t mine information by formula. Therefore, we use the “Custom Functions” of excel to calculate all the formula, such as Figure 2.4.

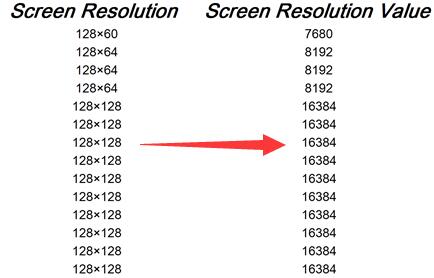


Figure 2.4

1. ***The color attributes***

In order to attract consumers to buy, most of the manufacturers try to use different kinds of qualifier to modify the color of their products. Such as: Mouse, chestnut, sapwood, pink (a kind of flower), withered grass, celadon china, bamboo tea, turtledove feather. However, this makes data mining a lot of obstacles. Therefore, through analysis, we directly transform these adjectives into color, like: gray, red, pink, blue, brown, green, yellow, and purple.

1. ***Screen type attribute***

The 21st century has entered the touchscreen phone era, but there are some smart phone manufacturers choosing to go different ways, they also produced some traditional screen phones which input with keyboard. There are different kinds of screen type in this attribute, but consumers only care about whether their phones’ screen supports multi-touch or whether they are sensitive or not. Therefore, we fail them into three categories: first, Resistive touch screen; second, capacitive touch screen; thyroid, normal. For example, Figure 2.5.

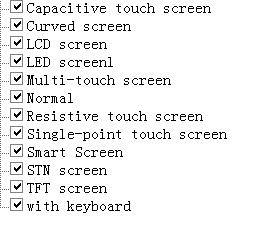
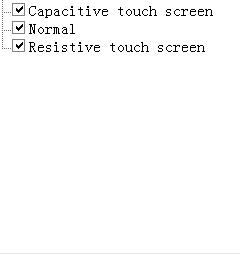
 

Figure 2.5

1. ***Data Smooth (smooth noisy data)***
2. ***Screen resolution attribute***

Due to the special nature of the mobile phone resolution. According to the calculation results of screen resolution, we cluster the screen resolution attribute by the 4 different technology generations of mobile phone screen resolution. They are: first generation, 800\*600; second generation, 1280\*720 to 1366\*168 (720P); third generation, 1920\*1080 (HD1080P); fourth generation, 2560\*1440 and above (2K). Processes as Figure 3.1.

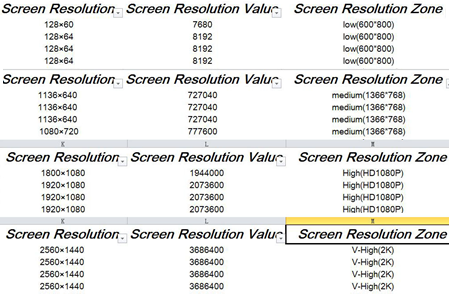


Figure 3.1

1. ***Size attribute***

Similarly, we also work with the numeric type of attribute-the size of the phone. According to the general knowledge of phone size, we divide the size of phone into three groups: Small (below 4.3), Medium (4.3-5.0), and Big (5.5 and above). For example, Figure 3.2.

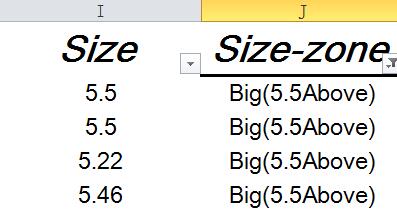
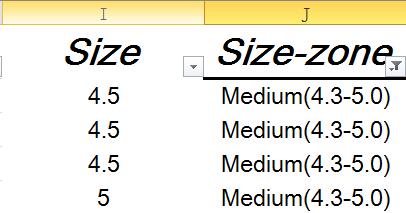
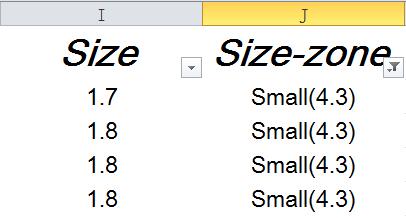


Figure 3.2

1. ***Price attribute***

Because the price is an unstable value, it is difficult to find some patterns according to the attributes like this. Therefore, we divide the price into five groups: v-low, low, medium, high, v-high. For example, Figure 3.3.

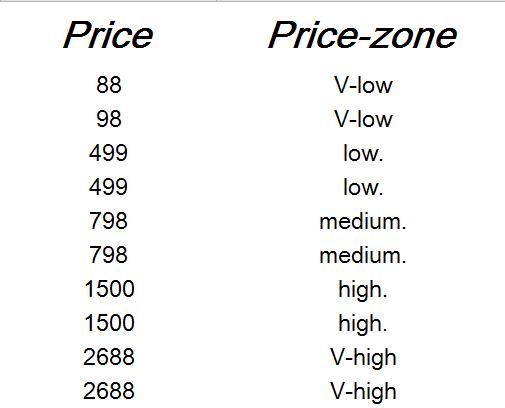
******

Figure 3.3

1. ***Popular values attribute***

Because the ranking percentage does not decide whether a smart phone is popular or not, we split the popular vales into three groups depending on the value of numbers. They are low (<1000), medium (1000-10000), and high (>10000). Moreover, only nominal value can be used in Apriori algorithm. As Figure 3.4



Figure 3.4

1. ***Monthly Sales attribute***

For the same reason and purpose, we separate the Monthly Sales attribute into three groups. They are few (<100), many (100-1000), mass (>1000). As Figure 3.5

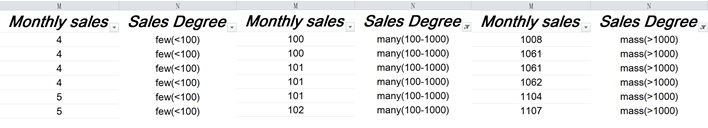


Figure 3.5

1. ***WEKA Analyzing***

In data processing, we use Weka to process our data. Before using weak, we change the data type to numeric and nominal because most of the algorithms cannot be used if the data type does not match. As Figure 4.1.

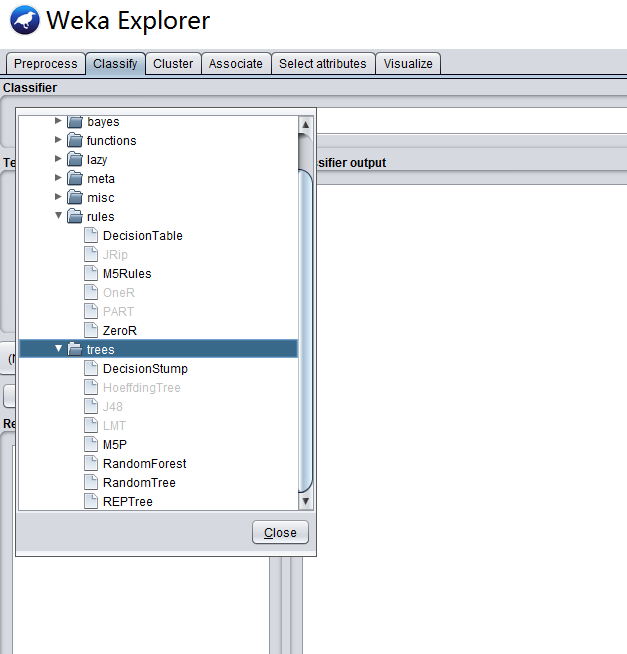


Figure 4.1

For some algorithms, only nominal type can be accepted. For example, when we use J48 (C4.5) to check the data, the class must be nominal type. Moreover, when we use Apriori algorithm to associate the attributes, the error tells us that none of the data types can be numeric. As Figure 4.2.

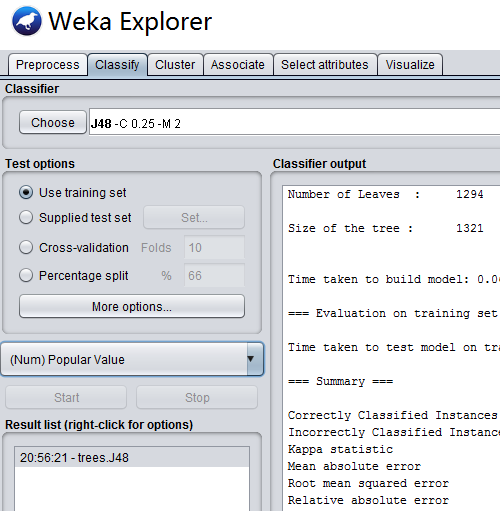


Figure 4.2-1

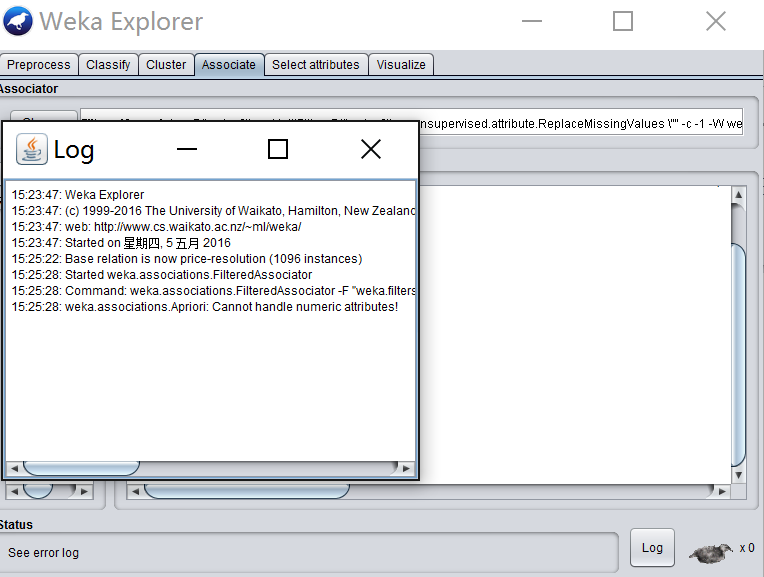


Figure 4.2-2

Use J48 (C4.5) to calculate if the value after smoothed is effective. Take the Price-zone as the class. Then, we find that the average recall is about 70%. Therefore, the data after smoothed is effective. As Figure 4.3.

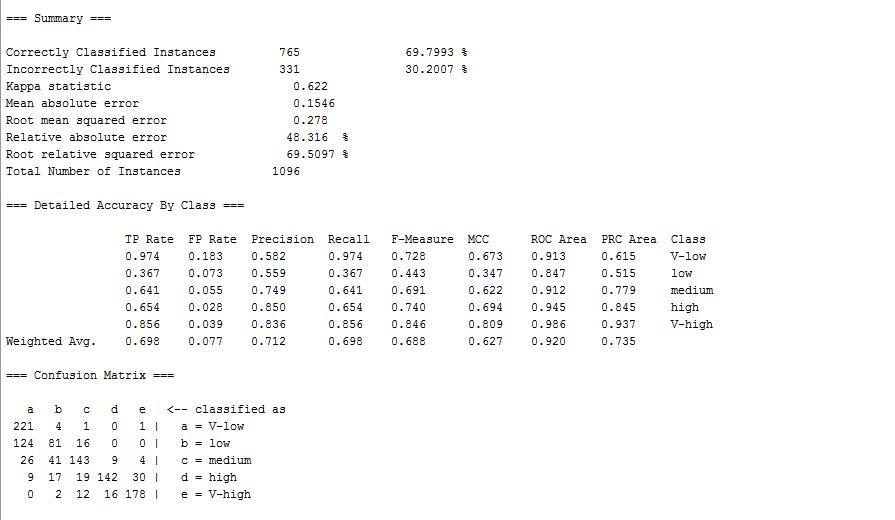


Figure 4.3

Use Information Gain to find the most related attribute to the Price-zone. Take Price-zone as the class. Then we find that the Brand is the most related attribute to the price. After that, Screen Resolution Value attribute is the most related attributes of the parameters of the smartphones. Moreover, we find that the popular degree and sales degree have very low value. Therefore, we can consider that these two attribute do not have much connection to price attribute. As Figure 4.4.

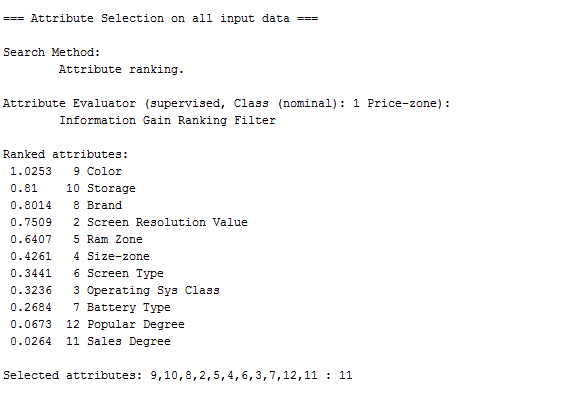


Figure 4.4

1. To find the relationships between price, monthly sales, and popular value, we take sales degree and popular degree as the class. Then we find that monthly sales and popular value relate to each other deeply. After color attribute, we can find that the Brand influences the Sales Degree and Popular Degree most. After Brand, it is Size-zone. As Figure 4.5.

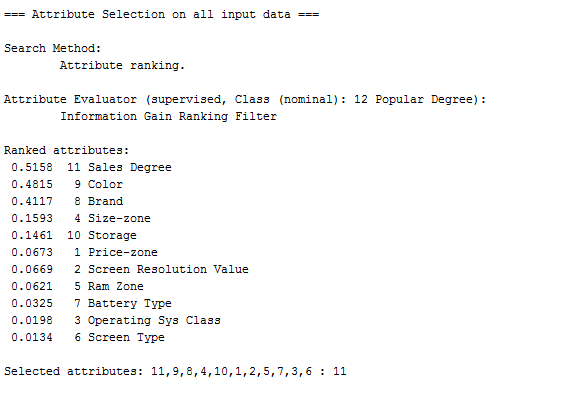
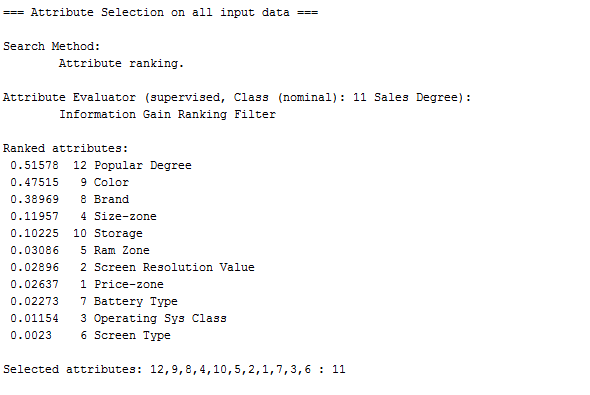


Figure 4.5

To research the connection more among each group of price, monthly sales, and popular value, we use Apriori algorithm to associate the data. Then, we find that only when the price is low and sales degree is few, the popular degree is low; and when price is medium and sales degree is few, the popular degree is low. As Figure 4.6.

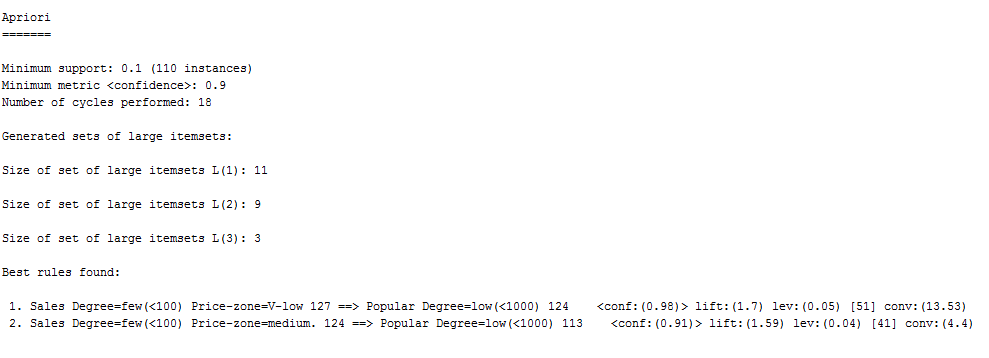


Figure 4.6

## Conclusion

After processing the data, we find many valuable results.

The most related attribute to the price is Brand. That means most of the smartphones are expensive or cheap because of their brands, but not parameters. Therefore, if people want to buy smartphones, which have a high cost performance rate, they should choose the brand carefully.

Except the brand attribute, the screen resolution value influences the price most. We can imagine that the screen resolution influences the cost much.

The relationships between price, monthly sales, and popular values show that only monthly sales and popular values influence each other. Moreover, brands attribute is the most important attribute which relates to both monthly sales and popular values. However, brands are not parameters of smartphones. It tells us that people buy smartphones and evaluate them subjectively and prejudicially.

If some manufacturers want their products have hot sales, they should pay attention to the size of the smartphones because the second related attribute to popular values and monthly sales is Size-zone. They can also decrease the Screen resolution values to reduce the cost for this attribute does not influence the popular value and monthly sales much.

There are also some problems. For example, we cannot process the data which have more than one value in a cell, such as color attribute and storage attribute. Now, we try to count each color for one time, but it influences the sum of the values. In the future, we would find a method to solve this problem.

## References

Use MLA method

|  |
| --- |
|  |
| [1] ["E-commerce in China: The Alibaba phenomenon"](http://www.economist.com/news/leaders/21573981-chinas-e-commerce-giant-could-generate-enormous-wealthprovided-countrys-rulers-leave-it). [The Economist](https://en.wikipedia.org/wiki/The_Economist). 23 March 2013.  [2] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Elsevier, 2011.  [3] Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.  [4] Sodhi, M. S., and Seongha Lee. "An analysis of sources of risk in the consumer electronics industry." Journal of the Operational Research Society (2007): 1430-1439. |

## Author contributions

Guanqi Guo 1/2 Preprocessing Duo Zhang 1/2 Preprocessing

Guanqi Guo 1/2 Translation Duo Zhang 1/2 Translation

Guanqi Guo 1/2 Paper Writing Duo Zhang 1/2 Paper Writing

Guanqi Guo 1/2 WEKA analyzing Duo Zhang 1/2 WEKA analyzing