Research of Clothing Sales Prediction and Analysis Based on ID3 Decision Tree Algorithm

SUN Fangshuai[1,a,] LIU Yi[1, b], XURIGAN Saiyin[1,c], ZHANG Qing[2,d]

[1]*Beijing Institute of Fashion Technology, Chaoyang District, Beijing, China*

[2]*China Urban Construction Design & Research institute co.Ltd, Beijing, China*

[a]*765631159@qq.com,* [b]*mableliuyi@163.com*

Abstract

This paper begins with the introduction of the classical algorithm of ID3 decision trees, which is widespread used into data mining. Through a study on the information gain of the noisy data, ID3 algorithm is used to create a corresponding model and an analysis model of prediction that applies to the reality

Key Words: ID3 Algorithm; Decision Tree; Data Mining; Prediction Model

1 Introduction

As a decision support tool, decision tree uses information gain in the theory of information to search for the property field containing the maximum amount of information, establishes a node in a corresponding decision tree, builds the branches according to different values of the property field and iterates lower nodes and branches in the subsets of each branch. The utilization of decision tree visualizes data rules, saving time for the construction process, and makes output results more intelligible and accurate. Moreover, static testing can be used to evaluate the a model as well as measure its reliability.

In modern society, people are more and more particular about fashion; hence multiple factors are influencing apparel sales nowadays, including price, type, and size, besides traditional factors of season and material. Consequently, in order to promote sales and maintain as little inventory as possible, sellers need to identify appropriate factors consistent with the specific condition in their stores. Thus ID3 algorithm can be utilized to create an effective sales model, helping sellers to reduce unnecessary loss.

2 Decision Tree

2.1 An Overview of Decision Tree

Decision Tree is a prediction model representing the mapping relation between object properties and object values. Given the certain probability of all situations, it is specifically used in decision analysis to calculate the probability of zero-or-greater expected value of net present value, assess the project risks and judge whether a project is desirable or not.

2.2 Basic Strategy of ID3 Algorithm

Starting with a single node of the training examples, if the examples are in the same category, the node is identified as a leaf and represented by a corresponding symbol. Otherwise, the algorithm employs the entropy-based measurement of information gain as heuristic information, and selects an attribute that can most satisfactorily classify the examples. The attribute represents "test" or "decide" on the node. In this version of the algorithm, all attributes are classified, i.e. discrete values. For each known value of the test attributes, a branch is created, based on which examples are divided. Following the same process, the decision tree of each division is generated recursively. Once an attribute appears in a node, it should not be considered again in any descendants of the node.

3 Information gain

ID3 uses information gain as its attribute selection measure. This measure is based on pioneering work by Claude Shannon on information theory. Let node N represent or hold the tuples of partitions and reflects the least randomness or “impurity” in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple tree is found. The expected information needed to classify a tuple in D is given by $Info(D) = -\sum_{i=1}^{m} p_i \log_2 \left(p_i \right)$ where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $\frac{\left| C_{i,D} \right|}{\left| D \right|}$. A log function to the base 2 is used since the information is encoded in bits. $\inf o\left(D \right)$ is just the average amount of information needed to identify the class label of a tuple in D.Note that, at this point, the information we have is based solely on the proportions of tuples of each class. $\inf o\left(D \right)$ is also known as the entropy of D

Now, suppose we were to partition the tuples in D on some attribute A having v distinct values, $\left\{ a_1, ..., a_v \right\}$ as observed from the training data. If A is discrete-valued, these values correspond directly to the v outcomes of a test on A.Attribute A can be used to split D into v partitions or subsets, $\left\{ D_1, ..., D_v \right\}$ where D_j contains those tuples in D that have outcome a_j of A.These partitions would correspond to the branches grown from node N. Ideally, we would like this partitioning to produce an exact classification of the tuples. That is, we would like for each partition to be pure. However, it is quite likely that the partitions will be impure. This amount is measured by $Info_A \left(D \right) \sum_{j=1}^{v} \frac{\left| D_j \right|}{\left| D \right|} Info\left(D_j \right)$ The term $\frac{\left| D_j \right|}{\left| D \right|}$ acts as the weight of the jth partition. $Info_A \left(D \right)$ is the expected information required to classify a tuple form D based on the partitioning by A.The smaller the expected information require, the greater the purity of the partitions

4 The Application of ID3 Decision Tree

Data mining with decision tree has been commonly used in financial enterprises, renowned e-commerce sites, weather prediction, medical diagnosis and shopping analysis. By fitting decision tree into the field of apparel sales, this research aims to reduce inventory gluts and boost sales. In spite of a variety of factors influencing sales of apparel, only weather, size and color are considered in this research. Temperatures in a Chinese city (Fig 1) and the sales data of a shirt from a certain brand (Table 1) from Oct. 21 to Oct. 30 are selected as the research data. First of all, relevant data is processed. According to the statistics provided by the weather bureau, the average of minimum temperatures in October is 10.1°C. For simplicity and convenience, temperature above 10°C is considered to be high and below 10°C to be low. In view of the average sales of 37.6 in these 10 days that has been calculated, we define sales greater than 37 as positive while less than or equal to 37 as negative. Then Table 2 is obtained after pre-processing.

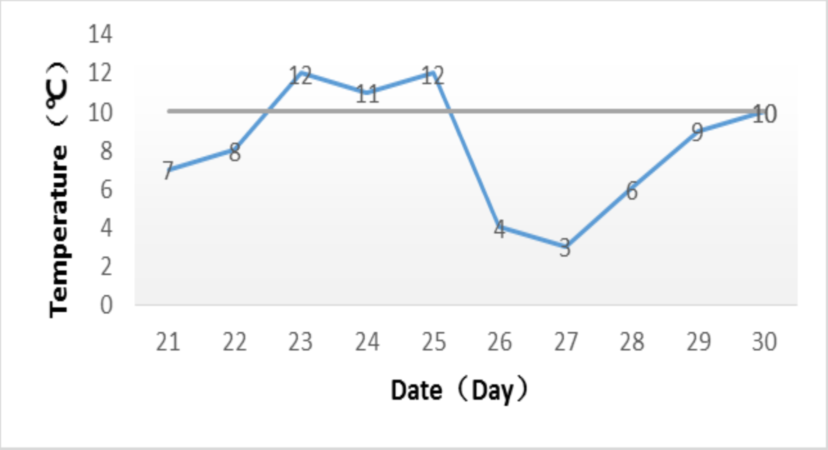
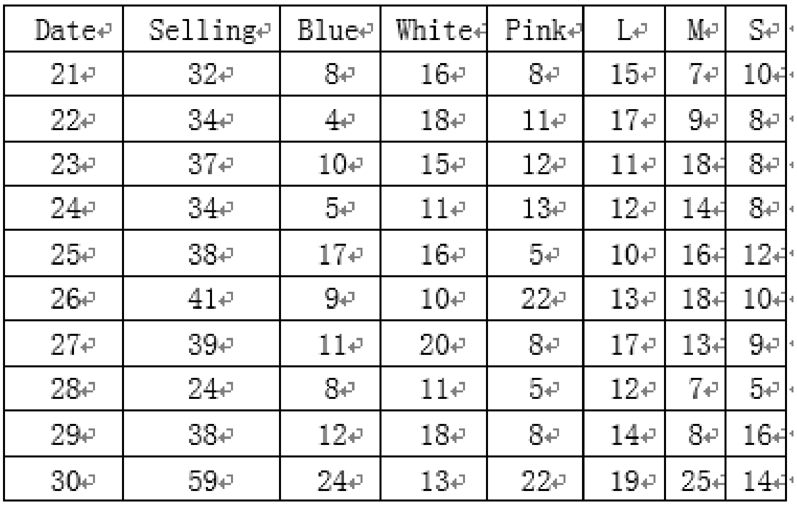


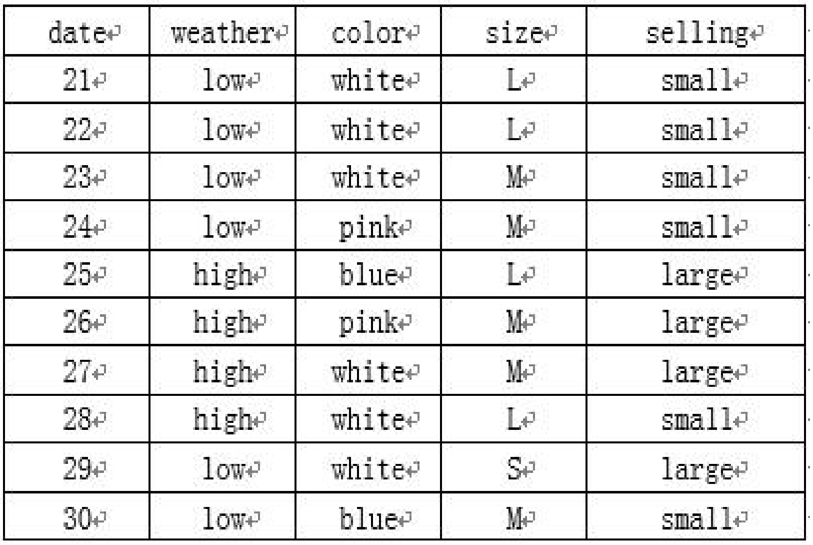
Fig.1 Temperature Changing Tendency

Table 1 Sales Data



Date	Selling	Blue	White	Pink	L	M	S
21	32	8	16	8	15	7	10
22	34	4	18	11	17	9	8
23	37	10	15	12	11	18	8
24	34	5	11	13	12	14	8
25	38	17	16	5	10	16	12
26	41	9	10	22	13	18	10
27	39	11	20	8	17	13	9
28	24	8	11	5	12	7	5
29	38	12	18	8	14	8	16
30	59	24	13	22	19	25	14

Table 2 Sales Amount and Influencing Factor



date	weather	color	size	selling
21	low	white	L	small
22	low	white	L	small
23	low	white	M	small
24	low	pink	M	small
25	high	blue	L	large
26	high	pink	M	large
27	high	white	M	large
28	high	white	L	small
29	low	white	S	large
30	low	blue	M	small

With the help of ID3 algorithm, we calculate that the information gain of the gross sales is 0.97, the information gain of color is 0.267 and entropy is 0.703; the information gain of size is 0.485 and entropy is 0.485; the information gain of weather is 0.650 and entropy is 0.320. It is obvious that the information gain of weather is greater than that of size and color. Therefore, the attribute of weather is selected as the initial node, which is indicated in Fig 2.

Then we iterate the same process to get the information gain of color of 0.540

and that of size 0. After computing again, we know that the information gain of size (the third attribute) is 1, with esults shown in Fig 3

5 Discussion and Conclusion

Situations of Positive Sales

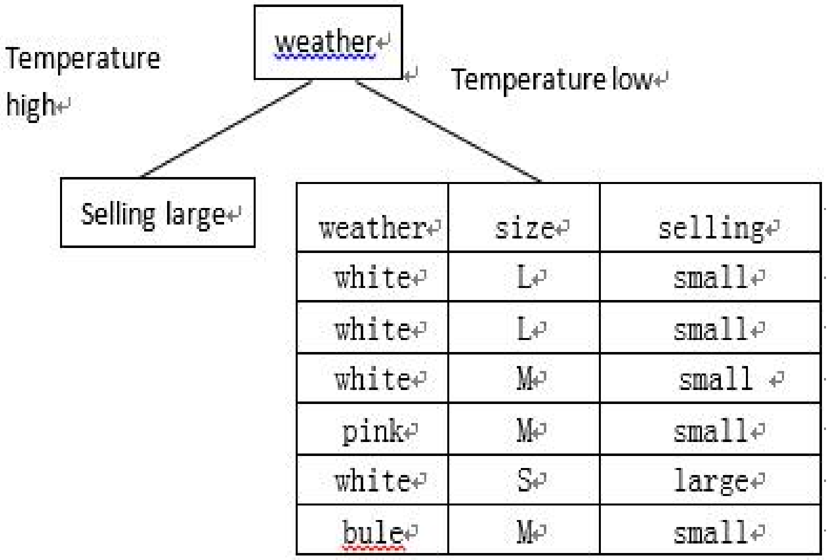
Rule 1: if high temperature and then action= positive sales

Rule 2: if low temperature ∩ white ∩ S and then action= positive sales

Situations of Negative Sales

Rule 1: if low temperature ∩ (pink ∪ blue) and then action=negative sales

Rule 2: if low temperature ∩ white ∪ (L ∪ M) and then action=negative sales



weather	size	selling
white	L	small
white	L	small
white	M	small
pink	M	small
white	S	large
bule	M	small

Fig 2 initial node decision trees

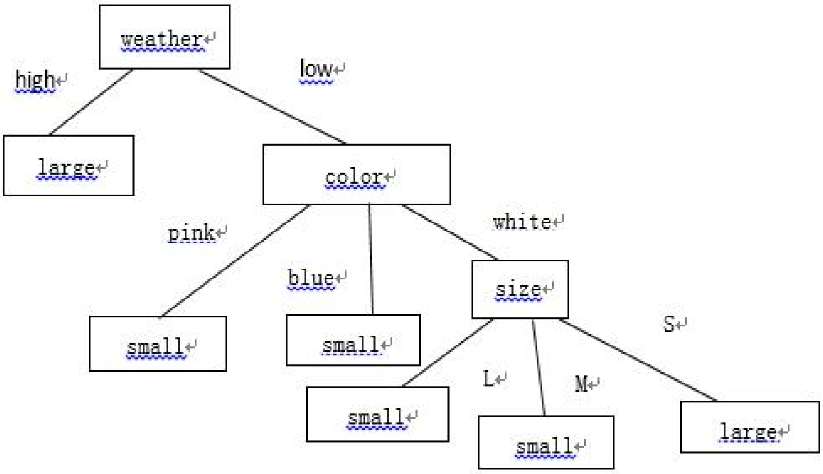


Fig 3 decision tree

Finally it is concluded that the shirt sells well in a higher temperature. In a lower temperature only white color and S size sell best while the case of other colors is not satisfactory. Therefore, it is recommended that the shirt hit the shelves in summer or early autumn days.

6 Summary

By means of ID3 algorithm of decision tree, this paper explores how factors of weather, color and size influence apparel sales and makes predictions on future sales of a commodity. The generalization of this approach to the retail sector like convenience stores, cosmetics shops, furniture stores etc. will help decision makers effectively sort out valuable associated information in the market analysis, improve sales performance and manage the store in an efficient way.

Acknowledgements

This work was financially supported by Beijing College Student Research Training Program (2014,2015), Open Project of Digital and Interactive Media Key Laboratory (KF2013-01， KF2013-13), Beijing Institute of Fashion Technology Teaching reform and innovation team project（JGTD-1404）, PHR(IHLB)-Innovative Research Team PTTBIFT_TD_002.

Beijing education reform project（2013-ms145）

References

[1] J.Han and M.Kamber Data Mining：Concepts and Techniques Morgan Kaufmann Publishers 2011 231-236

[2] Wang Chengjun, Cunstormer Churn Prediction and Analysis Based on Improved ID3 Decision Tree Algorithm, Computer Science, July, 2010

[3] Zhang Qian, Prediction and Analysis of Retail Production Bsed on Data Ming, April ,2008

[4] San Mateo, C4.5 Programs for Machine Learning, Morgan Kaufmann Publishers, 1993

[5] Tian Linqin, Application of Data Mining Technology in Tobacco Industry, Journal of Agricultural Science and Technology, March ，2012

[6] Wang Ying, Li Renwang, Li Bin, Zhang Zhile, Costume Sales Forecasting Model Based on CURE Algorithm and C4.5 Decision Tree, Journal of Textile Research, September, 2008

[7] Zhang Gefu, Ou yang, Hao nan, Xu qi, Application of Decision Tree in Apparel Marketing Based on Appearance of Consumers, Journal of Computer Applications, July,2010

[8] Du liying, the Application of ID3 Decision trees Algorithm, Light Industry Science and Technology, October, 2014