International Conference on Innovative Research in Engineering, Management and Sciences
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 4 | Issue 9 | ISSN : 2456-3307

Fashion sales prediction using Data Mining

T. Gayathri
Asst Professor, New Horizon College of Engineering, Outer Ring Road, Marathalli, Bengaluru, Karnataka, India

ABSTRACT

Online shopping has widened the sales of attires. Wide range of fashion outfits are made available to the customers at much cheaper rate. Merchandiser has reduction in cost because, it is not essential for him to have a showroom or sale staffs. Even a naïve fashion designer can sell their products through shopping websites. Online shopping sites also provide a platform to understand the fashion market. Data mining can be used to understand the fashion market by predicting the customer mindset. This paper attempts to create a learned model which would predict if the dress designed would be sold or not.

Keywords : Classification, Fashion Sale Prediction, Online Shopping

I. INTRODUCTION

Online shopping provides a comfort for the customer, they can do shopping from any place at any time. It gives an extensive array of products for the customers which is not possible to bring under one roof. Customer also can cancel and return the products whenever their desire. Online shopping also grasps the interest of customer by the huge discount that they provide.

There are also some cons in Online shopping, there is no personnel touch in shopping. The customers do not see the product before shopping. There can be a difference between how the dress looks in real and in photo. A prediction of whether dress can be bought or not would help the customers to make an intelligent decision.

Fashion in general is influenced by the Film industry, but there are certain unknown factors which influences the fashion industry as well. Fashion prediction is vital for many reasons, if the merchandiser can know what kind of goods will be sold, then he can have a strong hold over the market. The merchandiser can influence the customers to buy their product thus increasing their profit. Internet and online shopping play an important role in fashion.

Online shopping site has huge amount of information on the kind of dress people like to buy. When data is available in huge amount it is possible mine pattern and make prediction using data mining. Our age is data age, the presence of huge data makes our life easy. From huge amount of weather data, it is possible to predict day today weather, medical data of patients enable us to predict disease, sales data enables us to predict the mood of the customers.

Data mining is the extraction of information or knowledge from huge volume of data. Data mining involves classification, clustering, association mining. Classification and Clustering techniques create models that predicts the category to which a given data belongs to. In classification learning happens over labelled data. In clustering learning happens over unlabelled data. Association mining is a technique through which the model suggest that a collection of data occur together frequently.

In this paper we attempt to create a learned model using WEKA. This model will be trained using a dress dataset downloaded from UCI machine learning repository [1]. This dataset was created in 2014. It contains 501 instances and 13 attributes.

LITERATURE SURVEY

Few research papers have tried to predict Fashion using data mining. In [2] the author gathers details from social networking site, used natural language processing to extract information and create a decision support model for fashion prediction. There some papers which suggest similar predictions based on books movies etc.

In [3] the data is extracted from fashion trends web pages. Features of previous seasons fashion trend and its corresponding sales value is used to forecast whether new fashion trends features would be hit in the market. After feature extraction Artificial neural network, fuzzy logic is used to create a working model. Coefficient of determination is used to assess the quality of model proposed.

[4] states the difficulty of fashion forecasting. It is a challenge because fashion grows in a non linear fashion, when season and dress attributes are considered. In this paper the author suggests a two-stage prediction model, a short term and long-term prediction with Artificial Neural networks.

[5] -this paper creates an intelligent system to find combination for outfits. Combination for jewelry dress etc. This done by deep learning of meta data of fashion sites. Customer buying pattern is analyzed generally using collaborative filtering, but there is a disadvantage of collaborative filtering. Collaborative recommendation worls on static data hence does not keep up with the change needs of customer. In Fashion industry needs change at a rapid rate. Therefore, in [6] suggest a Collaborative filtering that works on dynamic data.

In [7] has a prototype for a model, this would predict the items in shopping list of customers and provide personalized interaction with customer to improve his experience. [8] attempts to use text mining in extraction of fashion based information from renowned fashion blogs this done to keep up with dynamic Fashion industry. Korean fashion blog is used to do the analyses.

From literature survey we are able to understand that the UCI machine learning dataset are not used by the research. The research paper predominantly uses text mining to extract information and use Neural network to do the prediction. In this research paper we attempt to use dress dataset and explore the various machine learning algorithms to arrive at a consensus.

DATA MINING DEFINITIONS

Data mining is extraction of useful information from data. Data mining is a four steps process which involves: Data collection, Data preprocessing, Machine learning, Pattern Evaluation.

3.1 Data Collection

Collecting or downloading data suitable to problem domain. With the improvement of data mining research this stage has predominantly become finding a suitable source of data from machine learning repository.

3.2 Data Preprocessing

Data cannot be fed as such to machine learning algorithm. Data has to transformed or reduced based on requirement.

Data is usually noisy, contain unnecessary information which might result in less accuracy. Data preprocessing is cleaning of data before machine learning. [9] suggests the use of data pre-processing to improve machine learning. Classification and clustering accuracy is predominantly dependent on the proper representation of data.

3.3 Machine Learning

Machine learning literally means, make the machine learn, machine learns by processing the data with various machine learning algorithm [9]. There is no fixed algorithm to provide high accuracy this is called No Free lunch theorem [10], however deep learning provides a better accuracy in most cases.

For any application it is important to apply few machine learning algorithms to find out the best suited model. Machine learning algorithms can be grouped under Bayes, Rule Based, Neural network and Decision tree.

3.3.1 Naïve Bayes: Naïve Bayes theorem is the best machine learning algorithm to use when the features are independent of one another [12]. Each instance is considered as a vector. The posterior probability of a class given a predictor is found with

P(h|d) = (P(d|h) * P(h)) / P(d)

P(d|h) - the posterior probability of class given a predictor

P(h) - Prior probability of a class

P(d) - Prior probability of a predictor

3.3.2 Decision Tree: Decision tree is arrived at by finding the optimum way to arrange the various nodes. There are two ways to identify the best partition of dataset at node, information gain or gain ratio. The decision tree model which uses information gain is ID3 and gain ratio is J48 [13]

3.3.3 Multilayer Perceptron: Multilayer perceptron contains large number of nodes called as neurons, joined together so that they for input layer hidden layer and output layer. The instances are supplied though the input layer, bias and weight are added at the hidden layer and supplies the class in output layer [14].

3.4 Pattern Evaluation

After machine learning more than model will result. Evaluation of which model is a better model, is performed.

EXPERIMENTATION AND RESULTS

4.1. Dataset

Dress dataset was downloaded from UCI repository. The dataset contained 501 instance and 27 attributes. The attribute of dresses were: style, price, rating, size, season, neckline, sleeve length, waist line, material, fabric type, decoration, pattern type and recommendation. Each instance is information of a type of dress and whether customer buy the dress.

Following model used to identify weak students and propose improvement strategy for them (see Fig.1):

1. Association mining to find strong association rules
2. Feature Selection
3. Resampling
4. Classification

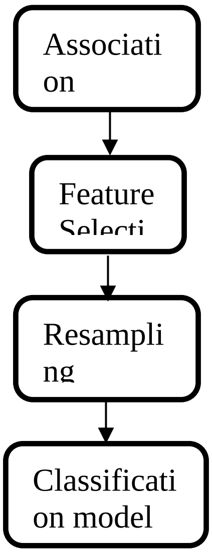

Fig.1 Block diagram of proposed model

4.2. Data preprocessing

All the attributes of dataset are nominal except ratings, the rating can also be converted to nominal dataset using Discretization process. Equal width binning is used, the ratings are mapped among five bins. The rating takes a value from 0 to 5, therefore 5 bins are used.

4.3. Feature Extraction

Dataset under consideration is a nominal dataset. Association mining is used to find the relation among various attributes. When association mining was applied, it observed no rules were mined. Association rules which were extracted had a confidence of 80% or less.

Information Gain Attribute Evaluation was done along with Ranker, even this resulted in equal weightage for all the attributes. So, all the attributes of dress dataset were considered important though the algorithms.

Through Domain based analysis it was found that size of the garment does not contribute to fashion trend, so the attribute was removed.

4.4. Resampling

The dataset has equal proportion of sales and no sales data. It is not necessary to implement resampling to solve Class Imbalance. But improve the classification model resampling is applied. Genetic algorithm based SMOTE (Synthetic minority Oversampling technique) is used. SMOTE algorithm achieves resampling by creating samples which are altered version of minority class instances. Further the algorithm we have applied is Genetic algorithm. Here a crossover and mutation of original sample is done, so the resamples are not exact match of original samples. A main disadvantage of sampling is overfitting. Since samples are duplicate of already existing instance, the classification model tends to be overfitted. This is avoided in our paper using GASMOTE.

4.5. Classification

Dress prediction dataset is used to classify the dataset into whether the dress would be sold or not. This learned system would be able predict the sales of a dress given a set of dress features. The model is created with an interest that, if a fashion designer or a merchandiser provides the idea of a dress then it can predict if the dress is of interest to customers.

Many classification algorithms are available to do supervised learning. According to No Free Lunch theorem [15], it is not possible assure that one algorithm is better than another. So, we are analyzing the dataset through various classification algorithms like, Multi-layer perceptron, Random forest, Random tree, J48, Naïve Bayes, BayesNet and SMO.

Preprocessed and sampled dataset was run through the classification algorithm using 10 folds cross validation. The order in which the data are supplied to the algorithm can bias the output and alter the end results. To avoid this, 10 folds cross validation is used. In k fold cross validation, dataset is divided into k folds and each fold is used for testing at some part of the learning.

Table I. Accuracy metrics for various classifiers

Classifiers	**Accuracy**
Multilayer perceptron	83.6%
Random Forest	83%
Random tree	80%
SMO	69.2%
J48	63.4%
Simple KMeans	70%
Naïve Bayes	65.8
Bayes Net	68%

When we look at the accuracy Multilayer perceptron and Random forest is better model to classify fashion data.

CONCLUSION AND FUTURE WORK

The accuracy of various classification algorithm is analyzed, and a working prediction model is created. This research can be expanded in future. Through this research we can predict sales. Using cognitive data mining, it possible to identify if the features of the dress which when altered can increase the sales of a dress or influence a person to buy the dress.

ACKNOWLEDGMENT

I thank my college New Horizon college of engineering for providing support and tools for this research. I thank Head of Department, Dr.B.Rajalakshmi for her support and guidance.

REFERENCES

1. https://archive.ics.uci.edu/ml/datasets/Dresses_ Attribute_Sales
2. Dang, Nhan Cach, et al. "Framework for retrieving relevant contents related to fashion from online social network data." International Conference on Practical Applications of Agents and Multi-Agent Systems. Springer, Cham, 2016.
3. da Silva Alves, Nelson. "Predicting product sales in fashion retailing: a data analytics approach." (2017).
4. Ni, Yanrong, and Feiya Fan. "A two-stage dynamic sales forecasting model for the fashion retail." Expert Systems with Applications 38.3 (2011): 1529-1536.
5. Li, Yuncheng, et al. "Mining fashion outfit composition using an end-to-end deep learning approach on set data." IEEE Transactions on Multimedia 19.8 (2017): 1946-1955.
6. Cho, Yeong Bin, Yoon Ho Cho, and Soung Hie Kim. "Mining changes in customer buying behavior for collaborative recommendations." Expert Systems with Applications 28.2 (2005): 359-369.
7. Cumby, Chad, et al. "Predicting customer shopping lists from point-of-sale purchase data." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.
8. Anna Rickman, Tracy, and Robert M. Cosenza. "The changing digital dynamics of multichannel marketing: The feasibility of the weblog: text mining approach for fast fashion trending." Journal of Fashion Marketing and Management: An International Journal 11.4 (2007): 604-621.
9. D. H. Deshmukh, T. Ghorpade, and P. Padiya, “Improving classification using preprocessing and machine learning algorithms on nslkdd dataset,” in Communication, Information & Computing Technology (ICCICT), 2015 International Conference on. IEEE, 2015, pp. 1–6
10. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.
11. Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." IEEE transactions on evolutionary computation 1.1 (1997): 67-82.
12. Lewis, David D. "Naive (Bayes) at forty: The independence assumption in information retrieval." European conference on machine learning. Springer, Berlin, Heidelberg, 1998.
13. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993
14. Goodman, Rodney M., and Zheng Zeng. "A learning algorithm for multi-layer perceptrons with hard-limiting threshold units." Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Workshop. IEEE, 1994.
15. Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." IEEE transactions on evolutionary computation 1.1 (1997): 67-82.
16. T, Gayathri., Solution based mining of Students Academic performance,International Journal of Research and Analytical Reviews,6,2,189-192,2019,International Journal of Research and Analytical Reviews
17. Wang, Haixun, et al. "Mining concept-drifting data streams using ensemble classifiers." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. AcM, 2003.
18. Lin, Feng Yu, and Sally McClean. "A data mining approach to the prediction of corporate failure." Applications and Innovations in Intelligent Systems VIII. Springer, London, 2001. 93-106.