IARS' International Research Journal
p-ISSN 2202-2821 e-ISSN 1839-6518 (Australian ISSN Agency)

Vol. 07 No. 02 2017
82800702201703

# Analysis of Various Data Mining Techniques using Novel Ratings Prediction

Reema Thareja
Department of Computer Science, Shyama Prasad Mukherjee College, University of Delhi, India

Prerna Goel
Department of Computer Science, Shyama Prasad Mukherjee College, University of Delhi, India

***ABSTRACT –*** *Availability of large amount of data in unprocessed form has increased the need of data mining. Various data mining techniques are available for this purpose, but we need to choose one which is more accurate. Looking at the increasing interest of people in novel reading, data collected for prediction is based on novels. People were asked to rate different genre novel s. This study has predicted ratings of dystopian novel based on ratings given to other genre novels by readers using various data mining techniques and calculated their prediction accuracy.*

Key words: Decision Tree; K-Nearest Neighbor; Linear Discriminant Analysis; Scatter Plot Matrix

## I. Introduction

In recent years, a rapid increase in numbers of novel readers has been observed. A large percentage of people miss some essential foundational reading skills and strategies along their journey [1].

The storytelling aspect of a novel is a multi-faceted form of communication that engages a broad range of brain regions. Reading a good novel allows imagination to take flight. Rarely is the movie adaptation of a book ever quite as good as the original novel. Even the most advanced special effects always fall short of the visual power of own imagination [2].

For all avid readers who have been self-medicating with great books their entire lives, reading books can be good for mental health and relationships with others. Some studies have shown that people who read a lot of fiction tend to be better at empathizing with others. Reading literary fiction improves social perception and empathy, which are crucial to "theory of mind": the ability to guess with accuracy what another human being might be thinking or feeling [3].

Reading fiction makes us treat others better; it is a way of treating ourselves better. Reading has been shown to put our brains into a pleasurable trance-like state, similar to meditation, and it brings the same health benefits of deep relaxation and inner calm.

The purpose of the present study is to discover some rating patterns (i.e. ratings given to dystopian novels based on ratings given by them to other genre novels) of readers by data mining techniques. Additionally, an attempt is made to reveal the interest of readers in novels of different genre. Our goal is to find the most accurate technique for prediction and classification.

## II. Data Mining

Data mining is the process of exploration and analysis of large quantities of data in order to discover useful patterns [4]. Task of data mining is the automatic or semi-automatic analysis of data, to extract useful patterns and relationships by using techniques like artificial intelligence, machine learning, statistics, and database systems via advanced data analysis tools [5].

Data mining methods can be classified into two types based on what specific task to achieve: descriptive and predictive. Descriptive methods characterize the general properties of data whereas predictive methods perform inference on the available data set to predict how a new data set will behave [6].

In this study, both of the methods are used by using R Studio. Data mining techniques employed includes Decision tree using rpart function, K-Nearest Neighbour (KNN) and Linear Discriminant Analysis (LDA) to predict or classify rating of dystopian novel by various users.

### A. Methodology

As stated previously, various data mining techniques are employed during the analyses and their prediction performances are compared. Thus, brief information about the techniques used is presented in this section.

The decision tree is a commonly used method in data mining and can handle both categorical and numerical data. The task is to create a classification or regression model that predicts the value of a target variable based on several input variables [7]. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure. This method classifies data into branch-like segments that construct an inverted tree with a root node at the top of the tree, internal nodes, and leaf nodes [8]. Internal nodes, also called decision nodes, have two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor is root node. [9]. Resulted tree can be used to define one or more decision rules that can describe the relationship between input and target variable. These rules used to display the decision tree, which provides a means to visually examine and describe the tree-like network of relationships that characterize the input and target values [10].

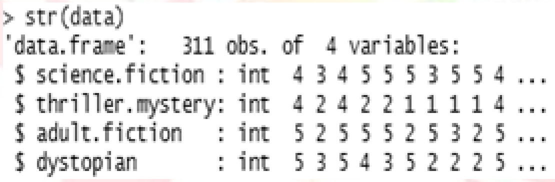
K-Nearest Neighbour (KNN) is a non-parametric, instance-based lazy learning algorithm. Non-parametric means KNN does not make any assumption about the functional form of the problem that needs to be solved. KNN uses raw training instances to make predictions. That is, algorithm doesn't explicitly learn a model. It chooses to memorize the training instances which are used as "knowledge" for the prediction phase [11]. As no learning of the model is required and all of the work happens at the time a prediction is requested, so called lazy learning algorithm. KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry [12]. KNN makes predictions using the training dataset directly. Predictions for a new instance (x) are made by searching through the entire training set for the K most similar instances (the neighbours) and summarizing the output variable for those K instances. For regression, it is the mean output variable. When used for classification, the output can be calculated as the class with the highest frequency from the K-most similar instances. Each instance in essence votes for their class and the class with the maximum votes is taken as the prediction [13].

Linear Discriminant Analysis (LDA) is a technique used in statistics, pattern recognition and machine learning to classify objects into mutually exclusive and exhaustive groups based on a set of measurable object's features. In this, the dependent variable is a group and the independent variables are the object features that describe the group. The dependent variable is always category variable while the independent variables can be any measurement scale. If we assume that the groups can be separated by a linear combination of features that describes the objects, then we can use linear discriminant model. If there are only two features, then the separators between objects group will be lines. If the features are three, the separator is a plane and if more than three, then separators become a hyper-plane [14]. LDA makes predictions by estimating the probability that a new set of inputs belongs to each class. The class that gets the highest probability is the output class and a prediction is made. The model uses Bayes Theorem to estimate the probabilities [15]. It can easily handle the case where the within-class frequencies are unequal and their performance has been examined on randomly generated test data [16].

## III. The Data

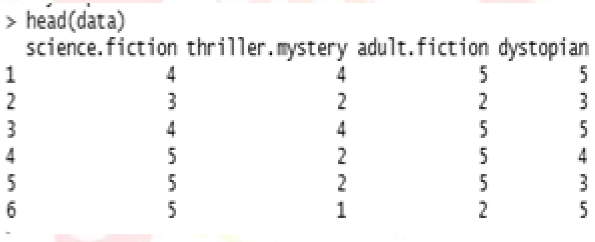
Data was collected from 311 students through survey. Students were asked to rate different genre novels ranging from 1(least liked) to 5(most liked), like "What genre novel do you like most?", "How much rating will you give to science fictions?".

Thus, the final dataset named "data" comprised of 311 rows, each row represents ratings given by a user, and 4 columns, for the ratings of 4 genre novels (science fiction, thriller mystery, adult fiction and dystopian). Furthermore, all the students are at undergraduate level.



```
> str(data)
'data.frame':   311 obs. of  4 variables:
 $ science.fiction : int  4 3 4 5 5 5 3 5 5 4 ...
 $ thriller.mystery: int  4 2 4 2 2 1 1 1 1 4 ...
 $ adult.fiction   : int  5 2 5 5 5 2 5 3 2 5 ...
 $ dystopian       : int  5 3 5 4 3 5 2 2 2 5 ...
```

**Figure 1: Structure of dataset**



```
> head(data)
  science.fiction thriller.mystery adult.fiction dystopian
1               4                4             5         5
2               3                2             2         3
3               4                4             5         5
4               5                2             5         4
5               5                2             5         3
6               5                1             2         5
```

**Figure 2: Head of data**

## IV. Application of Data Mining

Decision tree, Linear Discriminant Analysis and K-Nearest Neighbour Algorithm are applied on data (described in Figure 1) to predict the ratings of dystopian novel based on ratings given to other genre novels. The data is partitioned as 80% training data and 20% test data. Both training and test data are randomly selected. Ratings of dystopian novel are predicted, and then used to calculate accuracy of the prediction.

According to the results (see Table 1), KNN achieves the most accurate predictions for the target variable.

**Table 1: Applied algorithms and prediction results**

| Applied Algorithm | True Classification | False Classification |
|---|---|---|
| **Decision Tree** | 55.55% | 44.45% |
| **K-nearest Neighbour** | 60.31% | 39.69% |
| **Linear Discriminant Analysis** | 54.01% | 45.99% |

We have depicted the relation between adult fiction, thriller mystery and science fiction through scatter plot matrix.

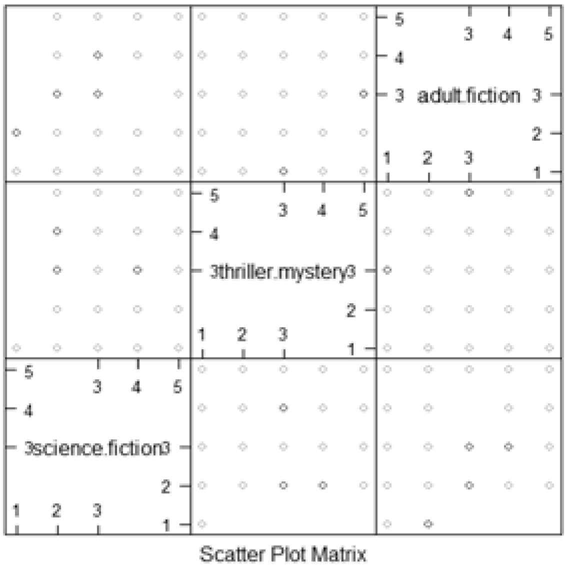
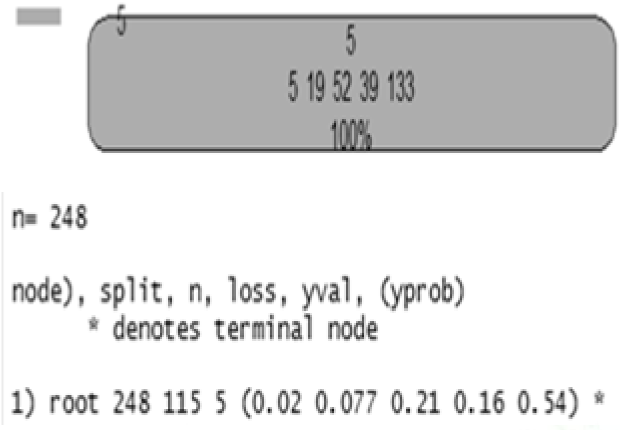


**Figure 3: Scatter Plot Matrix for Predictor Variable**



```
n= 248

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 248 115 5 (0.02 0.077 0.21 0.16 0.54) *
```

**Figure 4: (a) Decision Tree and (b)its Description**

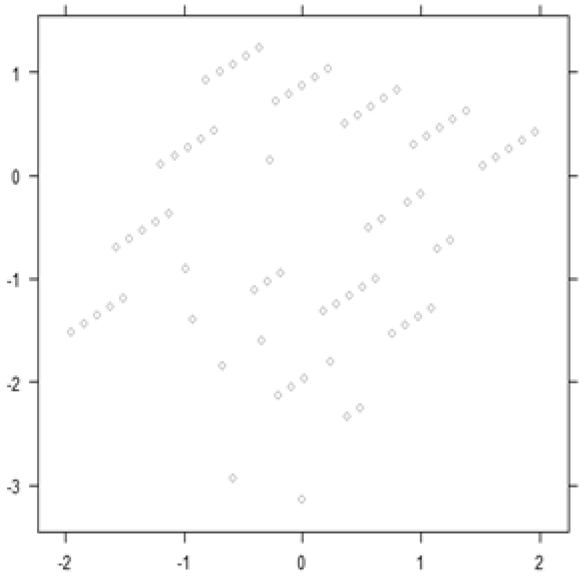


**Figure 5: Graph after LDA prediction**

## V. Discussion and Conclusion

This study tried to predict the ratings of dystopian novel based on ratings given by readers to other genre novels and analysed the accuracy of different data mining techniques. According to the results, K-Nearest Neighbour gives the most accurate results for the dataset at hand.

Many more data mining algorithms can be applied to the dataset to reveal more information from the given dataset. As a result, it can be concluded that data mining techniques can be successfully employed on rating data to predict or classify ratings.

## VI. References

[1]. "Reading Trends." Abrams Learning Trends. Web. 23 June. 2017.

[2]. Bergland, Christopher. "Reading fiction improves brain connectivity and function." Psychology Today. Sussex. Web. 24 June. 2017.

[3]. Dovey, Ceridwen. "Can Reading Make You Happier?" The New Yorker. Condé Nast, 9 June. 2015. Web. 25 June. 2017.

[4]. Berry, M., Linoff, G. "Mastering Data Mining: The Art and Science of Customer Relationship Management." John Wiley & Sons, Chichester (2000). Web. 28 June. 2017.

[5]. "Data Mining." Wikipedia -The Free Encyclopedia. 28 June. 2017. Web. 29 June. 2017.

[6]. "Data Mining Tasks." Wideskills. Web. 29 June. 2017.

[7]. "Decision Tree Learning." Wikipedia The Free Encyclopedia.25 June. 2017. Web. 4 July. 2017.

[8]. LU, Ying. "Decision tree methods: applications for classification and prediction." Shanghai Archives of Psychiatry. Shanghai Municipal Bureau of Publishing, 25 April. 2015. Web. 1 July. 2017.

[9]. Sayad, Saed. "Decision Tree – Classification." Experfy. Web. 4 July. 2017.

[10]. "Decision Trees— What Are They?" SAS. SAS Publisher. Web. 7 July. 2017.

[11]. "A Complete Guide to K-Nearest-Neighbors with Applications in Python and R." Kevin Zakka's Blog. 13 July. 2016. Web. 15 July. 2017.

[12]. Srivastava, Tavish. "Introduction to k-nearest neighbors: Simplified." Analytics Vidhya. 10 October. 2014. Web. 17 July. 2017.

[13]. Brownlee, Jason. "K-Nearest Neighbors for Machine Learning." Machine Learning Mastery. 15 April. 2016. Web. 18 July. 2017.

[14]. Teknomo, Kardi. "Discriminant Analysis Tutorial." Evoledu. Web. 20 July. 2017.

[15]. Brownlee, Jason. "Linear Discriminant Analysis for Machine Learning." Machine Learning Mastery. 6 April. 2016. Web. 21 July. 2017.

[16]. Balakrishnama, S. "Linear Discriminant Analysis - a brief tutorial" Institute for Signal and Information Processing. Web. 25 July. 2017.

[17]. Bozkır, A. Selman, S. Güzin Mazman, Ebru Akçapınar Sezer. "Identification of User Patterns in Social Networks by Data Mining Techniques." Springer-Verlag Berlin Heidelberg (2010): 145-153. Print.