Self Reflection of key concepts

Data Mining, COSC2111
Shonil Dabreo, s3835204
Mark Pereira, s3797413

Understanding of key Concepts

- Data Mining has a variety of algorithms to find “golden nuggets” or for prediction . The choice of algorithms depends on the problem and the structure of the data.
- There are Classification, Clustering, Association Finding, Attribute Selection and Visualization techniques in Weka. It also includes different algorithms for different techniques.
- Explored Neural Networks another algorithm using the javaNNS and MultilayerPerceptron in Weka.
- Data Mining processes such as Data preprocessing, Feature Engineering, Training, validating and testing the model were undertaken.
- Comparison of models done in terms of performance using prediction measures like Relative Absolute Error.

Solution for real-world problem

- There is always a solution to any real-world problem. To solve these problems understanding the data or patterns in the data is very important.
- Sometimes user data can be noisy. The data could have missing values, outliers, mislabelled category values, sanity values or blank spaces. The data needs preprocessing before analyzing the data.
- We need to explore the behavior of the data and visualization graphs are the best way to understand the data.

Solution for real-world problem

- Once data is explored using visualization, building and evaluation of the model is a psyched task.
- Thus, this brings a lot of issues when it comes to modeling. Overfitting, Underfitting, imbalanced accuracy (unequal category distribution), mislabelled errors (False positive or False Negative) are some of the issues that could be solved by appropriately tuning the model with more number of iterations.
- Thus validating tuned models using performance measures could result in an efficient solution to the real –world problem.

Extracting Golden Nuggets

- To find the patterns knowing how to visualize data is very important.
- Visually figuring out the correlation relationship (linear line) between the independent attributes such as for imdb dataset, gross, critical review, user votes and imdb scores. This helps to determine the representation of underlying structure of the data.
- Stating the hypothesis by focusing on the problem and then using visualizations to further check whether its true or not helps to better understand in ways to solve the problem.
- Based on the analyzed results, narrowing the data also helps in improving the performance of the model.

Extracting Golden Nuggets

- This will also help in proper tuning of the model understanding the in-accuracies.
- Analyzing the model outputs aids in understanding how the algorithms work or if it doesn't then why.
- For e.g., Decision Tree works best when the data is binary and it doesn't work when the data contains numbers. Another example is association finding helps to identify association between data.
- These golden nuggets assist in decisions that benefit the organization, governments as well as the customers.

References

- RMIT Data Mining learning materials.
- Assignment 1 and Assignment 2