CHAPTER I

INTRODUCTION

1.1 Background

Since 1990s, researchers have been utilizing data mining and machine learning to process large and complex data sets. These two methodologies have been successfully and efficiently used in a variety of fields and industries. From the medical industry to banking, manufacturing and various business strategy applications. There are unlimited potentials that can be unlocked with the study and implementation of these powerful tools.

Due to the vast daily use of the internet by the majority of our society, data mining has grown to be a foundational part of our daily lives, often without us even knowing it. From social media to simple searches, big companies around the world are trying their best to extract the most and best data about each individual as possible. This has been the primary key to success and often sole business model for businesses such as Facebook, Google, YouTube, Amazon and countless others.

One area of focus that has room for implementation of these tools is the fashion industry. Data mining and machine learning can effectively and efficiently help the processing, categorization and cataloguing of product data at scale to comply with the growing demand of the E-commerce world.

Figure 1.1: Category and Attribute Prediction

Recently, Sheeman Jain and Vijay Kumar [1] have done similar research of this study using Naïve Bayes classification, decision tree, random forest and Bayesian forest to train the machine learning model to classify clothes into three different categories and found accuracy of 73 to 90 percent for the respective categories.

Similar to previous study, we are also using the same open source data-set called DeepFashion compiled by Ziwei Liu, Ping Lou, Shi Qiu, Xiaogang Wang, and Xiaoou Tang from Multimedia Laboratory, The Chinese University of Hong Kong. They created FashionNet, a model that learns garment aspects by collectively predicting landmark positions and enormous attributes. These estimated landmarks are then used to pool or gate the learnt feature maps, resulting in robust and discriminative clothing representations. They showed the efficiency of FashionNet and the use of DeepFashion through comprehensive tests [2], which may greatly assist future study as this one.

The intent of this study, is mainly to categorize the clothes from DeepFashion into three different category; Top, Bottom, and Whole similar to Sheeman Jain and Vijay Kumar's work [1]. Unlike theirs, this study will employ only one method, the decision tree, and will use significantly less data. This study will only use 20,000 data, instead of 289,222. We also will be using Fine Annotation instead of Coarse Annotation from DeepFashion: Category and Attribute Prediction Benchmark (Figure 1.1). DeepFashion provided the garment images as well as image annotations. Nonetheless, we still need to combine the necessary data into one spreadsheet and adding the final labels one-by-one according to the image category. Also, another difference between this study and the previous one is that we will utilize two types of parameter tuning techniques to find the best decision tree parameter to increase the performance and comprehensibility of a non-parameter-tuned decision tree.

1.2 Problem Statement

The problem that will be discussed in this study is whether the parameter-tuned decision tree can predict the types of apparels accurately. Specific research questions are as followings.

1. What are the optimal decision tree parameters?
2. What attributes have a substantial impact on the final labels??
3. What improvements does parameter tuning provide to the model?
4. How does the model built with a parameter-tuned decision tree perform?
5. Does the final labels matched accurately with the pictures?

1.3 Objectives

The objective is to develop decision tree model with optimal parameter that can accurately predict the right labels of all the clothes in the data-set. Listed below are our main objectives:

1. identifying the optimal decision tree parameter,
2. determining which attributes have a significant impact on the final labels,
3. observing the model's improvements as a result of parameter tuning,
4. to observe how the model created with a parameterized decision tree performs,
5. to verify if the final labels matched the images accurately.

1.4 Restrictions and Assumptions

1. The sample data is obtained from an open-source data called DeepFashion: Category and Attributes Prediction Benchmark, Fine Annotation, which includes 20,000 clothing images, 50 clothing categories and 26 of clothing attributes.
2. The research is only conducted in short amount of time with only one source of data.
3. The clothing images are already annotated by the source.
4. It is assumed that the data-set is valid.
5. It is assumed that the annotations in the benchmarks are all correct.
6. To develop the final labels, the writer will utilize decision tree method.
7. The program that the writer use to develop the model is Python.

1.5 Benefits

This study is expected to provide benefits to readers, in both practical and theoretical terms.

1.5.1 Practical Benefit

The practical benefit of this research is to show that it is possible to predict clothing category according to their attributes which can be useful for businesses to provide a better user experience to their customers and in turn will increase sales.

1.5.2 Theoritical Benefit

With this study, the researcher hopes to help readers understand the steps and concepts behind categorizing clothes for future research. Also that the analytical framework can be a reference for other studies in fashion category prediction with their choice of datasets.

1.6 Thesis Structure

The structure of this study is as follows.

1. **Chapter I: Introduction**
   This chapter describes the background, problem statement, objectives, restrictions, assumptions, purpose, benefits, and timeline of the study.

2. **Chapter II: Theoretical Framework**
   This chapter describes the theoretical references used in this research. It consists of the definition of machine learning and the theories behind it that will be used in this research.

3. **Chapter III: Methodology**
   This chapter describes the steps taken to reach the objectives stated in Chapter I. It consists of steps and algorithms for data Evaluation, data pre-processing, data processing, model building, and model evaluation.

4. **Chapter IV: Experiment and Results**
   This chapter contains the result obtained from applying the steps and algorithms explained in Chapter III.

5. **Chapter V: Conclusion and Suggestions**
   This chapter summarizes achieved results, draw conclusions and proposes directions for further research.