MASTER'S THESIS | LUND UNIVERSITY 2015

The Business Insight Index – Evaluating Customer Insights through Hybrid Models

Jonathan Bratel

Department of Computer Science
Faculty of Engineering LTH

ISSN 1650-2884
LU-CS-EX 2015-19

The Business Insight Index

(Evaluating Customer Insights through Hybrid Models)

Jonathan Bratel
jonathan.bratel@gmail.com

June 10, 2015

Master's thesis work carried out at Narrative AB.

Supervisors: Pierre Nugues, Department of Computer Science, Faculty of Engineering,
Lund University Pierre.Nugues@cs.lth.se
Erik Söderberg, Head of Business Management, Narrative AB
erik@narrativeteam.com

Examiner: Jacek Malec, jacek.malec@cs.lth.se

Abstract

Customer segmentation and target analysis are two essential tasks when identifying a company's customers. To perform these tasks, this thesis develops and applies hybrid data-mining models, integrating clustering and decision trees. The hybrid models are applied to the life-logging camera company Narrative, in order to gain insights into their customer data. From previous research, we found that these hybrid models lacked means for evaluating the amount of insights proposed to decision makers.

For this reason, we created, tested, and validated a new evaluation measure – the *Description Tree Index*. Through experiments on five separate datasets, we conclude that the measure enables decision makers to evaluate the insights gained through the hybrid model. In each case, the index generates the best results for the expected number of segments.

We then integrated the Description Tree Index with existing evaluation models to form a *Business Insight Index*. This index evaluates customer segmentation and target analysis from both a business and data-mining perspective. By applying the index to the Narrative data, we found four customer segments to present the most insights.

Keywords: Data mining, customer insights, cluster analysis, decision trees, evaluation measures

Acknowledgements

I would like to thank my supervisor Pierre Nugues for his invaluable support and feedback regarding the thesis. Without him, this report would not have been possible to finish. Furthermore, I want to thank Erik Söderberg and Dzana Dedovic at Narrative for guiding me through this project. They have provided me with the support and information needed to produce this thesis. For that, I am truly grateful.

Contents

Acronyms

AGR	Average Gain Ratio
ALC	Average-Linkage Criterion
BII	Business Insight Index
CH Index	Calinski-Harabasz Index
CLV	Customer Lifetime Value
DTI	Description Tree Index
HCA	Hierarchical Cluster Analysis
IGR	Information Gain Ratio
LOF	Local Outlier Factor
MLS	Minimum Leaf Size
RFM	Recency, Frequency, Monetary Value
WAVC	Weighted Attribute-Values per Cluster

Chapter 1
Introduction

This chapter provides the context, background, and purpose of the thesis. The chapter aims to create an understanding of the work undertaken to produce the results. Furthermore, the chapter explains the objectives, limitations, and outline of this report.

1.1 Evaluating Customer Insights

Through the recent evolution of data-mining techniques, information can now be retrieved from large amounts of data to improve decision making. However, in the areas of customer segmentation and target analysis, there is a lack of measures to evaluate the insights discovered and communicated to decision makers (Estivill-Castro, 2002).

In order to create competitive advantages, the ability to interpret and evaluate insights in customer data has become increasingly important for today's businesses. This is the case for Narrative, a life-logging camera company. Narrative wishes to gain more insights into its current customers, so as to better target the profitable ones and develop its products (Söderberg, 2015).

As data mining is defined as the process of finding structural patterns in data, these patterns may generate an understanding of the data itself without having any practical business implications (Witten et al., 2011). Since different segmentation algorithms produce different results, it is hard to compare them based on the amount of insights they generate. There is also the problem regarding what counts as insights. Customer segmentation and target analysis are evaluated from a business perspective, and satisfactory data-mining results may lack in business relevancy.

When performing customer segmentation through data-mining techniques, the evaluation is often based on the internal quality of the segments. By measuring the similarity within each segment and the difference between segments, the internal structures reflect the validity of a specific segmentation. The major disadvantage of this method is that it does not measure the amount of information that may be retrieved from the segmentation.

This means that a data-mining algorithm may produce segments with good internal results, but with little or no insights to be retrieved (Lapczynski and Jefmanski, 2013).

In this thesis, we propose, test, and validate a new evaluation measure: the Business Insight Index (BII). This index integrates the RFM model, the Calinski-Harabasz (CH) index, and an original Description Tree Index (DTI). The DTI evaluates how much customer information is communicated to decision makers. By applying a hybrid approach, where customers are first segmented, the index computes both the quality and the quantity of the segmentation through a decision tree. The results show that the DTI is able to determine which customer segmentation generates the most insights.

1.2 Background and Related Work

The two most common data-mining types for customer segmentation and target analysis are clustering and classification. Clustering is the task of forming groups of objects, where the objects in each group are more similar to each other than objects in other groups. Classification builds predicting models in order to target and understand profitable customers. By combining these two areas of data mining into hybrid models, customer segmentation and target analysis can be performed iteratively (Ngai et al., 2008).

When using a hybrid model for customer segmentation and target analysis, one approach is to first cluster the customers based on a variable describing each customer's lifetime value (CLV). Then, insights can be extracted from the segments using classification. Nimbalkar and Shah (2013) used such a hybrid model consisting of *k*-means clustering and decision trees to allocate marketing and advertising resources in businesses. Dhiman et al. (2013) also used a hybrid model, but with a variety of clustering algorithms, in order to extract valuable information from tax audit datasets. Lapczynski and Jefmanski (2014) integrated clustering and decision tree algorithms and concluded that hybrid models improve the performance of predictive models. They further evaluated the clustering by different internal evaluation measures. However, these measures showed different results for different datasets and no measure correctly computed the expected number of clusters for all datasets.

There are multiple classification techniques in data mining. Methods such as linear regression, Naive Bayes, and neural networks have been developed over the years, and are frequently used. However, for this thesis, we apply decision tree modeling. There are many reasons for choosing decision trees for classification. They implicitly perform feature selection and require little or no effort in terms of data preparation. Furthermore, the decision trees do not require linear relationships between attributes in the data, making them ideal when there are no such relationships. Last but not least, they are easy to interpret and explain to decision makers, which is highly useful in business situations (Rokach and Maimon, 2008). For these reasons, decision trees are used when evaluating Narrative's customers.

Figure 1.1: The Narrative Clip. After Narrative (2015)

1.3 Narrative

Narrative is a Swedish start-up based in Linköping with offices in Stockholm, Lund, and San Francisco. Its main product is the life-logging camera, the Narrative Clip 1 (see Figure 1.1). The product consists of both a hardware camera and a cloud-based software service, where pictures can be uploaded. The camera automatically takes a picture every 30 seconds. The pictures are uploaded to the company servers when the customer connects the camera to a computer. They undergo a selection process, where an algorithm separates the high-quality pictures from the ones of lesser image quality. The selected pictures are then uploaded to the Narrative app, where users can look through the pictures of specific days (Narrative, 2015).

The company was founded in 2012, and launched a funding campaign that raised 11 times the requested capital. Narrative shipped the first version of its camera in 2014, and has now announced a second version, which will ship in 2015. To increase sales of future product versions, the company is interested in examining the characteristics of its customers. By studying areas such as behavior, demographics, geography, and lifestyles of its current customers, Narrative's objective is to gain effective insights from customer data through segmentation. These insights are then meant to be used for targeting the customer segments more effectively in the marketing efforts. It is also meant to be used in product development (Söderberg, 2015).

1.4 Purpose

By generalizing the specific assignment mentioned above, the thesis provides an answer to the following question:

> How can companies apply and evaluate the hybrid clustering and decision tree model in order to segment and target their customers, with the aim of finding effective insights?

We define effective insights as a set of business and data-mining objectives, which should be met in order to ensure the quality of the insights.

1.5 Business Objectives

The customer segmentation and target analysis are the primary business objectives we aim to achieve in this thesis. To perform these, certain criteria are necessary to confirm the business relevancy of the insights gained. In order to ensure the business quality of the segments created through the hybrid models, we used the RFM model. This model takes into account the recency, frequency, and monetary value as the three criteria for evaluating each CLV (Nimbalkar and Shah, 2013). This thesis extends the model to Narrative's customers, which is further explained in Sect. 3.2.

Furthermore, from a business perspective, it is important to evaluate the segment partitioning by the homogeneity within each segment and the heterogeneity across segments. In order to measure the intra-cluster similarity in relation to the inter-cluster dissimilarity, we used the Calinski-Harabasz (CH) index. This index is explained in Sect. 3.3, and is applied in combination with the proposed Description Tree Index (DTI) to form the Business Insight Index (BII). The purpose of the BII is to produce robust results from both a business and data-mining standpoint.

1.6 Data-Mining Objectives

The main data-mining objective is to find a validation measure to evaluate the technical insight value from each hybrid model iteration. The measure, referred to as the Description Tree Index (DTI), evaluates the description level obtained from different numbers of clusters. The DTI also assesses the overall quality of one clustering algorithm in comparison with other algorithms. In order to validate the quality of the index, we tested it on different datasets with different clustering algorithms. If the evaluation measure is qualitative, it should be able to predict the expected number of clusters. Moreover, it should also reflect the proportion of correctly clustered instances for the specific clustering algorithm. This is based on the assumption that correctly clustered instances generate more insights than incorrectly clustered ones.

By combining these business and data-mining objectives, we define effective insights through the BII.

1.7 Limitations

This project aims at investigating the properties of existing customers, where extensive data exist. Although it would be of scientific interest to investigate and compare customers with non-customers, we considered it outside the scope of this thesis. Furthermore, even though many different clustering and classification algorithms exist, we only use the ones described below. Other algorithms are proposed in Sect. 4.1.4, but due to the nature of the data, only three algorithms are included; two clustering algorithms and one decision tree algorithm.

1.8 Structure

This report is structured as follows:

Chapter 2 provides a comprehensive review of the data-mining algorithms used in this project. The chapter reviews the theory behind these techniques.

Chapter 3 describes the evaluation measures used to assess the results. Here, we introduce and thoroughly explain the BII and its components.

Chapter 4 clarifies the approach for this project. The chapter includes a detailed description of the process used, as well as each of its phases.

Chapter 5 reports the results of the project and provides an in-depth analysis.

Chapter 6 describes the findings related to the thesis's original purpose. This chapter provides a discussion regarding the validity of the BII.

Chapter 7 contains a final discussion and the conclusions of this work.

Chapter 2
Algorithms

This chapter describes the theory behind the clustering and decision tree algorithms used for this thesis project. We explain the k-means and HCA algorithms, as well as the decision tree algorithm C4.5.

2.1 Clustering Taxonomy

Cluster analysis is the task of transforming a set of instances into groups of objects. The instances in each groups should be more similar to each other than the objects in other groups. Figure 2.1 shows an example of it.

There are different clustering algorithms in data mining for this purpose. Figure 2.2 illustrates a taxonomy of the algorithms, proposed by Jain et al. (1999). The taxonomy shows a first division between hierarchical clustering analysis (HCA) and partitioned clustering. Partitioning methods generally aim to divide a dataset into a set of clusters with each instance belonging to one cluster. HCA, on the other hand, builds a cluster hierarchy where instances are assigned to a cluster by moving up or down a tree structure of instances (Dhiman et al., 2013).

In this thesis, we chose one technique from each part of the taxonomy based on the advantages and disadvantages over other algorithms: the k-means and HCA algorithms described.

2.1.1 K-means Clustering

K-means clustering divides a population of instances into k clusters. The algorithm bases the partitioning on each observation's position in an N-dimensional space, where N is the number of features. Today, it is one of the most widely used clustering algorithms in data mining due to its simple approach and efficient run time. K-means clustering aims to minimize the within-cluster sum of square Euclidean distances (WCSS) between each

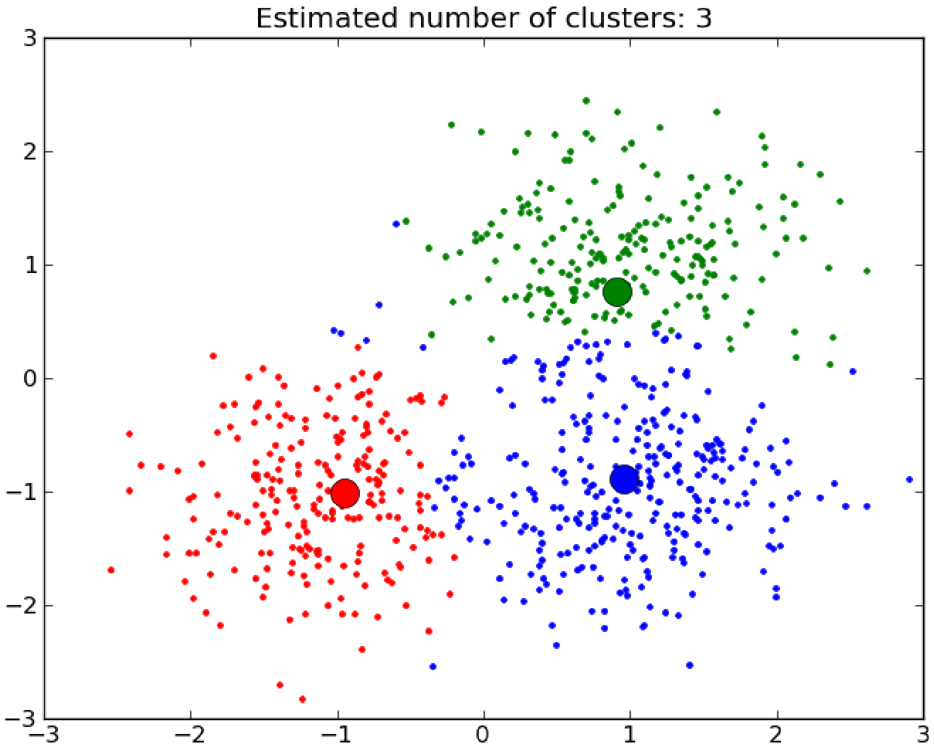


Figure 2.1: An example of clustering. After Scikit-learn (2014)

instance and its closest cluster center. The algorithm partitions the sample of n instances $X = \{x_1, x_2, ..., x_n\}$ into k clusters, with k cluster centers $C = \{c_1, c_2, ..., c_k\}$. Eq. 2.1 describes this (Arthur and Vassilvitskii, 2007):

$$WCSS = \sum_{x \in X} \min_{c \in C} \|x - c\|^2 \tag{2.1}$$

The algorithm itself consists of four separate steps (see Figure 2.3) explained below:

1. Place k cluster centers $C = \{c_1, c_2, ..., c_k\}$ randomly at the instances in the data set;

2. Create k clusters where each cluster $\mathbf{C_i}$ consists of the instances in X that are closer to c_i than other cluster centers for each $i \in \{1, ..., k\}$;

3. Assign the centroid of each cluster $\mathbf{C_i}$ to be the new cluster center c_i;

4. Repeat steps 2 and 3 until C is stable.

This method for finding the k clusters, often called Lloyd's algorithm, may however only find local optimums depending on the initial placement of C. For this thesis project, we use the k-means++ initialization method proposed by Arthur and Vassilvitskii (2007). This method is similar to Lloyd's algorithm, but with better performance in accuracy and speed. It consists of four steps explained below:

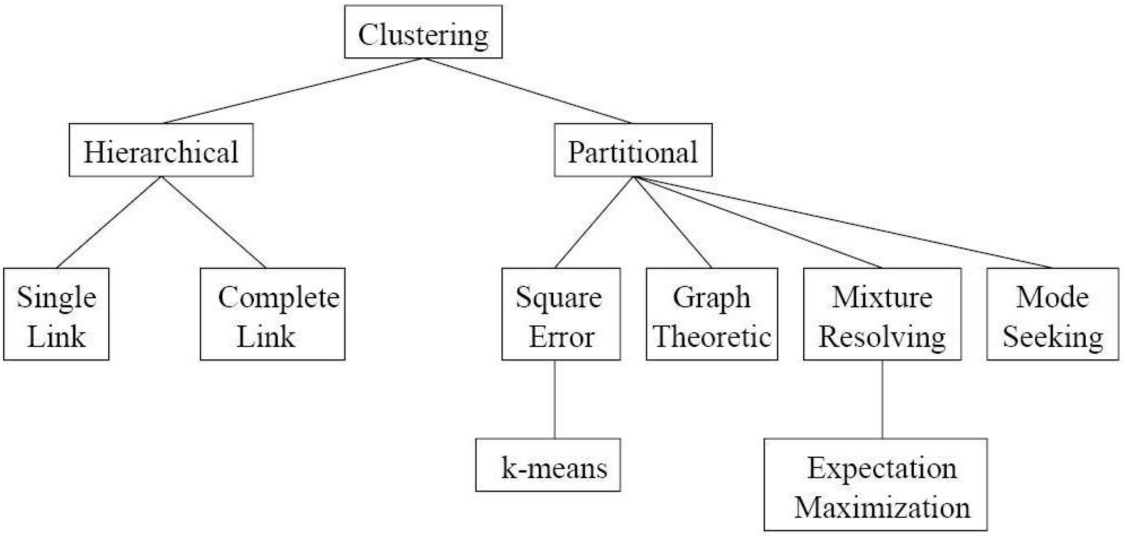


Figure 2.2: A taxonomy of the clustering algorithms. After Jain et al. (1999)

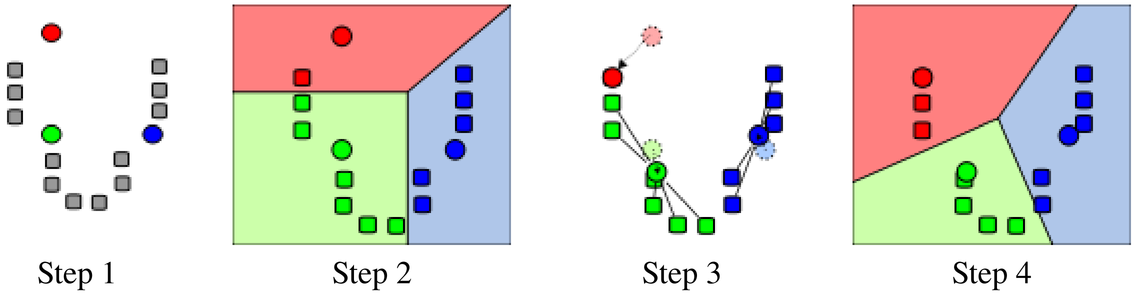


Figure 2.3: The steps of the k-means algorithm. After Amherst (2015)

1. Place cluster center c_1 at an instance chosen randomly from X;
2. Place the next cluster center c_i at instance $x \in X$ with probability $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$ where $D(x)$ is the shortest distance from an instance to the closest center already assigned;
3. Repeat step 2 until k cluster centers are chosen;
4. Follow step 2-4 from Lloyd's algorithm.

Step 2 increases the probability that new cluster centers will be placed as far as possible from the previous ones, which decreases the risk of suboptimal clusters (Arthur and Vassilvitskii, 2007).

2.1.2 Hierarchical Cluster Analysis

HCA aims to structure the instances in a dataset into a hierarchy of clusters (Figure 2.4). There are two types of strategies to achieve this (Sayad, 2015):

Agglomerative: This strategy applies a "bottom up" method by initially assigning each instance as a separate cluster. The clusters are then merged in order to move up the hierarchy.

Divisive: As opposed to the agglomerative strategy, the divisive method uses a "top down" approach where all instances start in a single cluster, which is then split in order to move down the hierarchy.

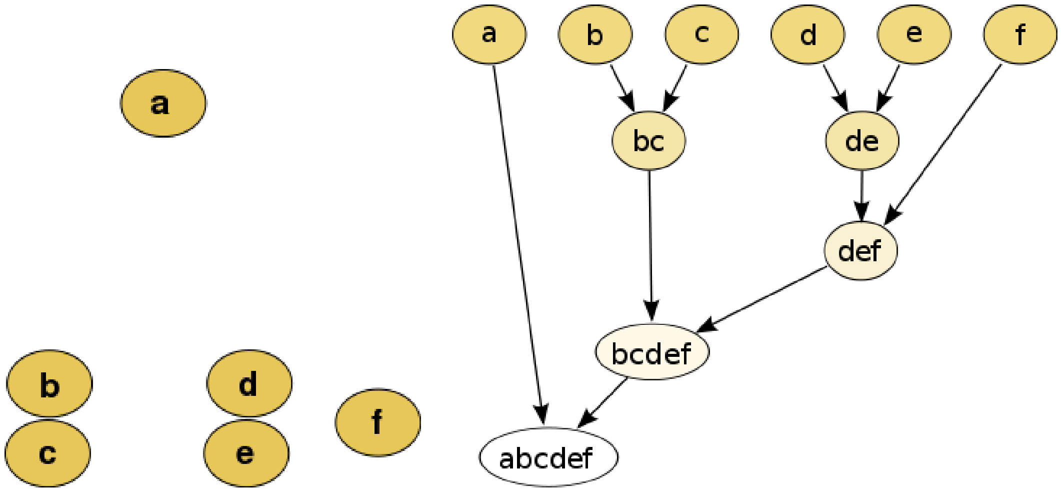


Figure 2.4: An example of hierarchical clustering. Depending on the number of clusters, the algorithm moves up or down the hierarchy to select the clusters. After Sayad (2015)

The HCA algorithm merges or splits clusters based on a measure of dissimilarity between the clusters. This measure, called the linkage criterion, specifies the distance between two clusters and form clusters subsequently. Two of the most common ones are the single-linkage (equation 2.2) and complete-linkage (equation 2.3) criteria. The single-linkage algorithm measures the distance between the two closest instances from two separate clusters, which makes the algorithm sensitive to outliers. The complete-linkage measure on the other hand, measures the maximum distance between the clusters as a linkage criteria. However, this may also result in clusters consisting of outliers (Sayad, 2015).

These linkage criteria require less computation than other criteria, but with the disadvantage that they only take one distance into account. This problem is overcome by the average-linkage criterion (ALC). The ALC (equation 2.4) uses the average distance between each object in one cluster to every object in the other cluster (Figure 2.5). According to Milligan and Isaac (1980), the ALC ranks the first of all linkage criteria. However, depending on the nature of the instances, other criteria may perform better (Milligan and Isaac, 1980).

$$L(r,s) = min(D(x_{ri}, x_{si})) \tag{2.2}$$

$$L(r,s) = max(D(x_{ri}, x_{si})) \tag{2.3}$$

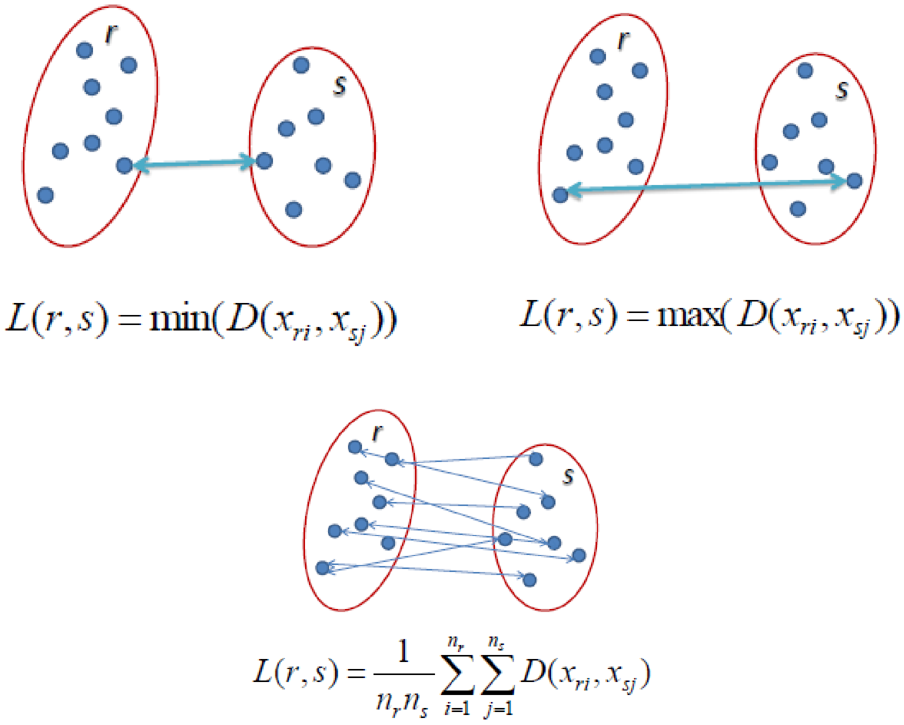

r
s
$L(r,s) = \min(D(x_{ri}, x_{sj}))$
r
s
$L(r,s) = \max(D(x_{ri}, x_{sj}))$
r
s
$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$

Figure 2.5: The single-, complete-, and average-linkage criteria between two clusters r and s. After Sayad (2015)

$$L(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{i=1}^{n_s} D(x_{ri}, x_{si})) \tag{2.4}$$

2.2 Decision Trees

Decision tree learning is an area in data mining, focused on predictive modeling. Decision trees are often used for decision making and visualization of decisions since they use a white box model where every step is observable. The goal of the decision tree is to predict the values of the observations in a dataset, based on a model created from earlier examples. Figure 2.6 illustrates an example of a decision tree. The box nodes represent decision points corresponding to input variables in the dataset, and the arrows represent possible values. The final points in the decision tree are the leaves. These leaves present to what class an observation belongs, based on its values (Rokach and Maimon, 2008).

2.2.1 C4.5 Algorithm

There are various decision tree algorithms. One of the most popular is the C4.5 algorithm developed by Ross Quinlan. The algorithm builds decision trees using a set of training data $S = s_1, s_2, ...$ of already classified examples. Each example s_i consists of the attributes used to describe the example, as well as the class label. The algorithm first selects the

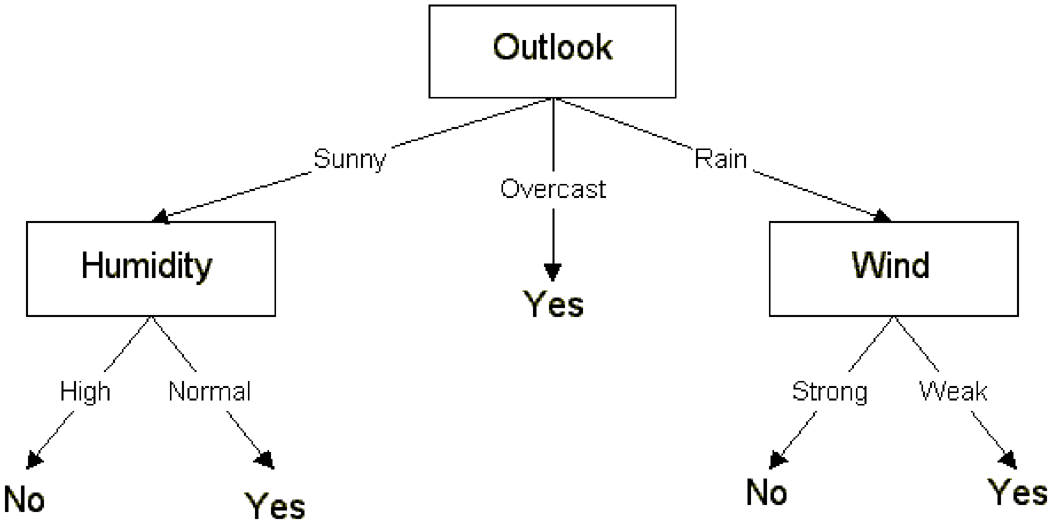


Figure 2.6: A decision tree for deciding whether to play outside or not. After Vidal (2009)

root node based on the attribute that most effectively splits the training set into subsets. The actual splitting point of the attribute is determined by the normalized information gain ratio (IGR) explained below. The algorithm then continues through the subsets until every example is classified (Quinlan, 1986).

The algorithm consists of three base cases:

- All the examples used to build the tree model belong to the same class. If this is the case, the algorithm creates a single leaf node with that class.
- None of the attributes provide any information gain. This forces the algorithm to create a decision node higher up the tree with the most frequent occurring class.
- An example with a previously unknown class value is encountered. As before, the algorithm creates a decision node higher up the tree with the most frequent occurring class.

In pseudo code, the algorithm proceeds as follows (Quinlan, 1986):

1. Check for base cases
2. For each attribute *a*:
   (a) Calculate the normalized IGR from splitting on *a*
3. Use the attribute with the highest IGR to create a decision node (*a*-best)
4. Continue with the subsets created by splitting on *a*-best, and add those attributes as children of the previous node

Information Gain Ratio

As mentioned above, each splitting point in the tree is based on the IGR of an attribute. The IGR builds upon the concept of entropy and information gain, used in information theory. Quinlan (1986) defines the expected information gain as the change in information entropy H from an earlier state to a state that takes into account some new information:

$$IG(Ex, a) = H(Ex) - H(Ex \mid a), \tag{2.5}$$

where Ex is a set of training examples and a is the value of the ath attribute. More precise, equation 2.6 describes the information gain:

$$IG(Ex, a) = H(Ex) - \sum_{v \in values(a)} \left(\frac{|\{x \in Ex | value(x, a) = v\}|}{|Ex|} \times H(\{x \in Ex | value(x, a) = v\}) \right), \tag{2.6}$$

where $value(x, a) = v$ is the specific value of example x for attribute a. In practice, the information gain for an attribute increases if each attribute value contributes to an unique classification. Even though the information gain is a sufficient measure for determining the significance of an attribute, it has several drawbacks. The most notable problem follows when information gain is applied to attributes with a high proportion of unique values, but with low relevance for the decision tree. For example, a decision tree might be used to categorize customers with a number of unique attributes such as email-address, social security number, credit card number etc. These attributes have a high information gain since they effectively identify each customer, but with little or no relevance in practice (Quinlan, 1986).

The IGR solves this problem by dividing the information gain by the intrinsic value of the dataset. Equation 2.7 defines the intrinsic value as:

$$IV(Ex, a) = -\sum_{v \in values(a)} \left(\frac{|\{x \in Ex \mid value(x, a) = v\}|}{|Ex|} \times \log_2 \left(\frac{|\{x \in Ex | value(x, a) = v\}|}{|Ex|}\right) \right) \tag{2.7}$$

The intrinsic value represents the potential information produced by splitting the examples into v subgroups, where each subgroup corresponds to v values on attribute a. A high intrinsic value implies that the subgroups have, more or less, the same size. Opposite, a low intrinsic value means that only a few of the subgroups hold most of the examples.

In general, the IGR for each attribute assumes values between 0 and 1, where a value of 1 indicates that the attribute is crucial to the partitioning of the dataset. Subsequently, a value of 0 implies that the attribute does not contribute to a partitioning at all (Quinlan, 1986).

Accuracy

One of the most important evaluation measures in classification overall, and decision tree modeling in specific, is the accuracy. To calculate the accuracy of any given classification model, the number of correctly classified examples are divided by the total number of examples. This reflects the overall correctness of the algorithm. In order to compute the accuracy in a decision tree, the data is usually divided into a training set and a test set. The decision tree uses the training set to build a model. Then, the algorithm applies the test set to calculate the accuracy (Witten et al., 2011).

A common way to compute the accuracy of a decision tree is through k-fold cross-validation. This model validation measure arbitrary parts the data into k equal subsets. Then, one of the subsets is used for testing the model, while the other $k - 1$ subsets are used to build the model. This is repeated k times, until every subset has been used for testing. The model then computes the accuracy as the average accuracy over all k subsets. The advantage of this method is that every instance is used for both training and testing, which reduces the risk of overfitting. There is no standard value for k, but one of the most commonly used validation methods is the 10-fold cross-validation. According to Witten et al. (2011), this number has proven to be the best during extensive testing with different numbers of k, and various datasets.

Minimum Leaf Size

For many decision tree model algorithms, there is a parameter for determining the minimum number of instances at each leaf. This parameter determines the complexity of the tree, and is referred to as the minimum leaf size (MLS). As the size of the tree is decided by the MLS, this parameter is important. For datasets with different numbers of examples, the MLS is usually adjusted to avoid overly complex or menial trees. When increasing the MLS, the accuracy of the tree often decreases. This is because the instances that would have been assigned to a separate leaf are now placed in a leaf where they do not belong. On the other hand, when decreasing the MLS too much, the resulting tree becomes complex and hard to interpret (Witten et al., 2011).

Choosing a MLS is more a rule of thumb than a standard process, and it depends on the context of the problem. For the purpose of extracting insights from a decision tree, the MLS is often set to a certain percentage of the examples included in the dataset (Witten et al., 2011).

Chapter 3
Evaluation Measures

To evaluate the insights from hybrid models, we applied multiple evaluation measures. In this chapter, we propose and explain the components of the Business Insight Index (BII). These include the Description Tree Index (DTI), the RFM model, and the Calinski-Harabasz (CH) index.

3.1 Description Tree Index

We created the DTI to evaluate how well the decision tree models describe the clusters generated from the cluster analysis. As the information extracted from a decision tree is only as good as the decision maker's ability to understand it, we prioritized the comprehensibility of the tree. The DTI is based on three aspects of the decision tree:

Accuracy: The overall accuracy of the decision tree model indicates how well the model describes the clusters created.

Average gain ratio: The average gain ratio (AGR) of the attributes reflects the average descriptive quality of each decision node in the decision tree.

Weighted attribute-values per cluster: The weighted attribute-values per cluster (WAVC) of the tree represents the amount of information each cluster contributes to the insight level of the clustering.

By taking these three aspects into consideration, we define the DTI as:

$$\text{DTI} = Accuracy \times AGR \times WAVC. \tag{3.1}$$

3.1.1 Accuracy

As explained in Sect. 2.2.1, the accuracy is the number of correctly classified examples divided by the total number of examples. Usually it is one of the final validation measures for a classification model. However, since the purpose of this thesis is to gain insights rather than create an accurate predictive model, we modify the use of accuracy in this thesis. For example, when clustering a dataset into two classes, there is a possibility that the resulting decision tree will only have one internal decision node and an accuracy of 100 % (see Figure 3.1). This means that the decision tree effectively splits the dataset based on a single attribute. While this is desirable from a predictive perspective, it does not result in a lot of insights about the data. Instead, the accuracy is used as factor in the DTI, indicating how well the clustering can be described in a decision tree. A low accuracy means that the clustering is more complex in nature and hard to describe, while a high accuracy indicates a simpler clustering.

3.1.2 Average Gain Ratio

As explained in Sect. 2.2.1, the information gain ratio (IGR) indicates how important an attribute is in creating classes in a decision tree. An attribute with a high IGR will be placed higher up in the decision tree and highly contribute to the partitioning. Even though most of the attributes in a dataset, more or less, contribute to the partitioning of the tree, some are more important than others. By taking into account the AGR, the information contribution of all attributes is reflected, even if not all attributes are present in the decision tree. This also implies that a higher AGR can be achieved by reducing the dimensionality of a dataset, removing attributes with low IGRs.

3.1.3 Weighted Attribute-Values per Cluster

The WAVC measures of how much information is made available from a decision tree by clustering a dataset into k clusters. The WAVC reflects that the decision nodes traversed when reaching a leaf generate specific insights about that leaf. Furthermore, the more instances placed at a leaf, the more important that leaf is to the characterization of the respective cluster. We define the WAVC as:

$$WAVC = \frac{l}{c} \times \frac{1}{l} \sum_{i=1}^{l} \left(\frac{k_i}{k_{Tot}} \times N_i \right) = \frac{1}{c} \sum_{i=1}^{l} \left(\frac{k_i}{k_{Tot}} \times N_i \right), \tag{3.2}$$

where l is the total number of leaves in the decision tree, and c the number of clusters. k_i and k_{Tot} are the number of instances at leaf i respective the total number of instances used to build the model. N_i is the number of decision nodes traversed to reach leaf i. By calculating the average weighted attribute-values to reach each leaf and multiplying with the number of leaves per cluster, we derive the WAVC. In practice, the WAVC estimates the weighted number of attribute-values needed to describe the average cluster (compare Figure 3.1 and Figure 3.2). The tree in Figure 3.1 requires only one decision to describe two segments. The tree in Figure 3.2 however, requires four decision nodes to describe three segments. The increased number of decision nodes generates a higher WAVC, and

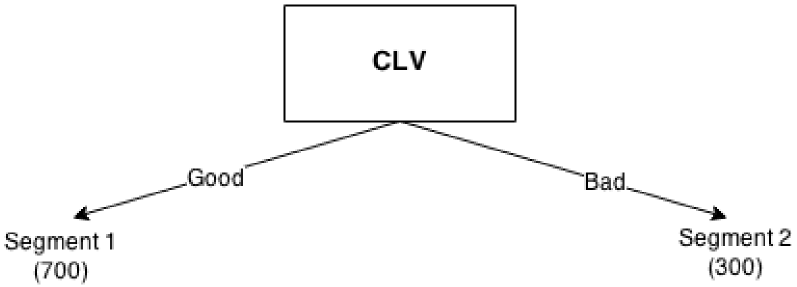


Figure 3.1: An example of a decision tree with two segments and a WAVC of 0.50. The numbers in parentheses are the customers in each leaf

more insights, as the tree require more attribute-values to describe the average segment in Figure 3.2.

When increasing the number of clusters, the total insights gained from all leaves should increase in order to justify the creation of a new cluster. This leads to a cost-benefit model, where the goal is to optimize the WAVC for a given number of clusters.

3.2 The RFM Model

As described in Sect. 1.5, the segments created should be qualitative from a business perspective. The use of the RFM model enables an assessment of the customer lifetime value (CLV) for each customer by measuring recency, frequency, and monetary value (RFM) of individual customers' purchases (Nimbalkar and Shah, 2013). We applied this model to the Narrative dataset by the creation of the following attributes:

Recency of last photo taken (R): This measure represents the relative recency ratio between the latest photo uploaded and the customer's account age:

$$\text{Recency Rate} = \frac{\text{Days between signup and last photo created}}{\text{Account age}}.$$

Frequency of photos (F): The frequency of photos represents the number of days a customer uploads photos divided by the associated account age:

$$\text{Frequency Rate} = \frac{\text{Number of active days}}{\text{Account age}}.$$

Monetary value of a customer (M): As a measure of each customer's monetary value, we chose the average number of photos per day:

$$\text{Monetary value} = \frac{\text{Number of photos}}{\text{Account age}}.$$

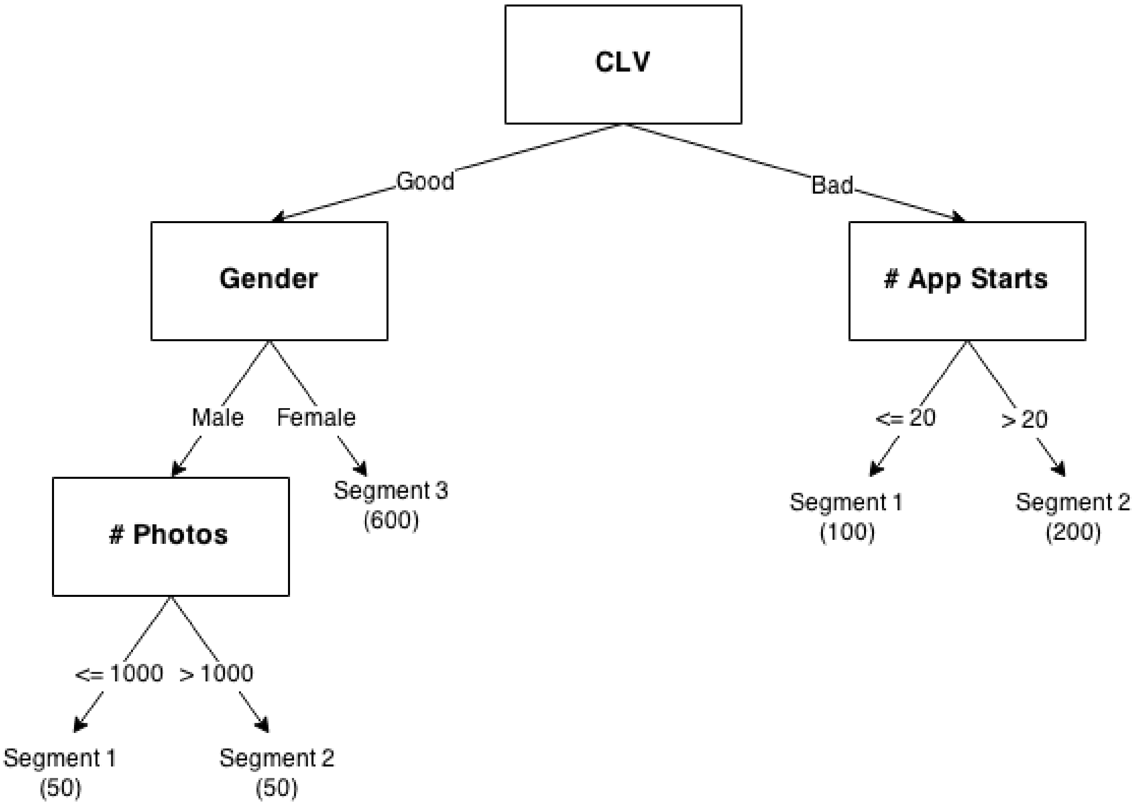


Figure 3.2: An example of a decision tree with three segments and a WAVC of 0.70. The numbers in parentheses are the customers in each leaf

We calculated the CLV by multiplying the RFM values for each customer, using the values as equal weights. By clustering the customers based on their CLV, the validity of the clustering from a business perspective increases. It also enables marketers and analysts to perform target customer analysis, since the RFM values make a ranking of the segments possible.

3.3 The Calinski-Harabasz Index

As mentioned in the introduction, the CH index is a measure for estimating the intra-cluster similarity in relation to the inter-cluster dissimilarity for a given clustering partitioning. Liu et al. (2010) defines the CH index as:

$$CH_k = \frac{SS_B}{SS_W} \times \frac{N-k}{k-1} \tag{3.3}$$

where SS_B is the overall variance between clusters, and SS_W is the overall variance within each cluster. N is the total number of instances and k is the number of clusters. A high variance between clusters (indicating a high heterogeneity) and a low variance within each cluster (indicating a high homogeneity) result in a better CH index. While the SS_B is essentially the same as the within sum of Euclidean square distances explained in Sect. 2.1.1, Liu et al. (2010) calculates the SS_B by the following equation:

$$SS_B = \sum_{i=1}^{k} n_i \|m_i - m\|^2, \tag{3.4}$$

where m_i is the cluster center of cluster i, and m is the mean of all the instances in the dataset. n_i represents the size of cluster i in proportion to the entire dataset. The optimal cluster partitioning, in terms of homo- and heterogeneity, is the one with the highest CH index.

The CH index is based on k-means clustering, which makes it a good choice when evaluating partitions generated by the algorithm (Liu et al., 2010).

3.4 The Business Insight Index

By integrating the DTI with the CH index, after utilizing the RFM model to the data, we enable a clearer overall picture of both the robustness and information available from a specific cluster partitioning. This new measure, proposed as the Business Insight Index (BII), works as a heuristic for the purpose of evaluating customer segments. We define it as:

$$BII_k = \mathrm{DTI}_k \times \text{Normalized } \mathrm{CH}_k \tag{3.5}$$

where k is the number of clusters. We normalize the CH index in order to make it comparable with other datasets.

Chapter 4
Approach

This chapter describes the working process for the thesis project. It defines the phases, as well as the tools and theory applied to produce the results explained in chapter 5.

4.1 CRISP-DM

This project is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM), created and developed by the CRISP-DM consortium (see Figure 4.1). The process consist of six iterative phases and we chose it because it is the most frequently used procedure, and the de facto standard for data-mining projects (Marbán et al., 2009).

4.1.1 Business Understanding

During the initial phase of the project, we defined the purpose and objectives of the thesis together with representatives from Narrative. The objectives included both business and technical goals in order to broaden the perspectives.

In consultation with Narrative, we executed a situation assessment, where resources, risks, and contingencies were evaluated as well as requirements, assumptions, and constraints regarding the project.

We designed a preliminary project plan, taking into account the time, tools and techniques necessary to complete the project. We decided to use the open source software Weka 3.7.11 for the data management. Weka is a Java-based machine learning tool including most of the algorithms necessary to perform a variety of data-mining tasks.

4.1.2 Data Understanding

The purpose of the data understanding phase is to collect the initial data from its sources. Then, the project participants are to describe, explore, and finally verify the data quality.

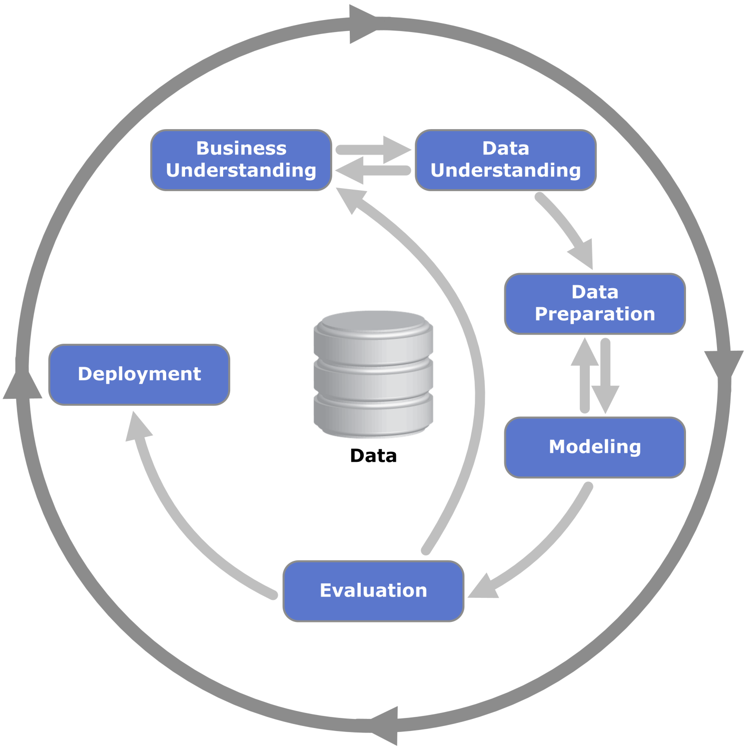


Figure 4.1: Phases of the CRISP-DM

They should describe aspects such as data format, quantity and identity of attributes while correlations, analysis, and hypotheses are explored. Finally, the project members assess the completion, error rate, and missing values in order to evaluate the data quality (Chapman et al., 2000).

Initial Data Collection

In order to validate the Description Tree Index (DTI) before applying the measure to the Narrative dataset, we tested the measure on five separate datasets described below. We collected the data from the UC Irvine Machine Learning Repository. We considered the data to be unbiased enough to provide a basis for testing. Table 4.1 shows a description of all the datasets.

We collected the Narrative dataset from Narrative's business intelligence software Looker. We selected the data based on its business importance to the company, related to the business objective stated in Sect. 1.5.

Data Description

The initial Narrative dataset consisted of 11,952 instances and 148 attributes whereof 3 nominal and 145 numeric ones. We examined the dataset from both a volumetric and statistical viewpoint by computing the basic statistics for each attribute. This included measures such as mean/mode, maximum, minimum, coverage, standard deviation, and distribution.

Dataset	Attribute Types	# Instances	# Attributes	Average LOF	Number of classes
Soybean	Categorical	47	35	1.046	4
Iris	Real	150	4	1.142	3
Seeds	Real	210	7	1.098	3
Glass	Real	214	10	1.419	6
UKM	Integer	403	5	1.091	4
Narrative	Categorical, integer	11,952	148	1.421	?

Table 4.1: Dataset description

Data Quality Verification

We performed an evaluation of the quality of the datasets as we detected missing values and outliers, verifying their meaning. We detected the outliers by using the local outlier factor (LOF) (Schubert et al., 2012). The LOF measures the density of each point in a dataset in comparison with its closest neighbors' densities. The algorithm calculates the density as the distance at which an object is reached from its neighbors (see Figure 4.2).The LOF assigns a value to each point based on its local density. A value of 1 or less indicates a clear inlier. However, there is no standard to what is considered an outlier score, and a score of 1.1 may indicate an outlier in some datasets, while being an inlier in others. Therefore, we used the average LOF as a measure of the overall quality of the datasets in terms of outliers. Table 4.1 shows that the Narrative and Glass datasets have less dense data in comparison with the other datasets (Schubert et al., 2012).

Regarding missing values, it was only the Narrative dataset which contained missing values. We worked out a plan to deal with the missing values, described in Sect. 4.1.3.

4.1.3 Data Preparation

The data preparation step aims to cover all activities necessary to produce the final dataset used for modeling. This phase is usually performed in iterations together with the modeling phase. Since new information may come up during the modeling phase, it may be necessary to re-prepare the data. Tasks such as selecting the final attributes, transformation, and cleaning of the data are performed during this phase (Chapman et al., 2000).

Data Selection

While we considered the five test datasets complete as is, we decided to reduce the dimensionality of the original Narrative dataset. Out of the initial 148 attributes, we selected 21 (three nominal and 18 numeric) for further modeling due to their business importance and high quality. These were the attributes with the highest business impact, while the other attributes lacked of business significance, quality, or correlation independence.

Data Cleaning

The data cleaning step included dealing with missing values and outliers in order to raise the quality of the data. We decided to deal with the Narrative dataset, which contained

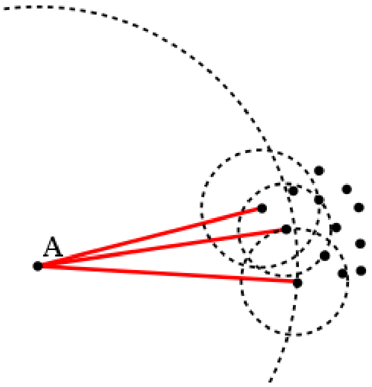


Figure 4.2: LOF compares the local density of point A to the density of its neighbors. In the figure, A has a lower density than its neighbors and is therefore considered an outlier.

both missing values and a relatively high average LOF. We replaced the missing values based on their business context since their meaning were apparent in many cases.

Concerning the outliers, we decided to evaluate the Narrative dataset both with and without outliers, to see if the results would improve. When removing the outliers, we considered every instance with an LOF score higher than the average an outlier. Even if this did not guarantee that all outliers were removed, it increased the quality of the dataset. The removal affected 3,500 (29 %) instances, which were removed on the assumption that it would not degrade the validity of the insights gained from the remaining customers.

Data Construction and Formatting

During the data construction step, we derived and added 11 new attributes to the Narrative dataset. We did this to better describe the relative customer activity, as the former dataset only described the absolute activity of each customer. These included attributes such as app starts per day, uploaded photos per day etc. The new attributes also included the RFM values for each customer in order to evaluate their customer lifetime value (CLV). We discretized the CLVs into two groups, one for customers with a high CLV and one for customers with a low CLV. By doing so, we facilitated further customer segmentation and target analysis, as the new attribute generated the highest information gain ratio (IGR). This meant that the attribute effectively split the dataset into subgroups in the decision tree.

Moreover, we normalized the data in Weka before modeling. This eased the distance measuring process since the attributes exhibited varying unit sizes.

4.1.4 Modeling

The modeling phase includes a selection of the modeling techniques to be used, and the generation of a test design to assess the evaluation measure proposed (Chapman et al., 2000).

Modeling Techniques

When deciding what modeling techniques to use, there are a variety of types to choose from. Especially for the clustering algorithms there were many choices, each with their own advantages and disadvantages. Table 4.2 presents a summary of the pros and cons of four of the most distinct clustering algorithms. We selected techniques based on the limitations in data characteristics and quality for the datasets, particularly for the Narrative dataset. Below, we summarize the limiting features of the Narrative dataset:

Size: The size of the dataset, with over 10,000 instances and 30 attributes, put demands on the clustering algorithm to be computationally fast with low time complexity.

Distribution: The majority of the attributes demonstrated a negative exponential distribution rather than a normal distribution.

Density: Because of the negative exponential distribution the density of the instances was high for low values, while decreasing with increasing values. This meant large differences in densities in the dataset.

Outliers: The dataset exhibited a large proportion of outliers according to the LOF, indicating instances with significantly distant attribute-values.

Comprehensibility: Due to the importance of the dataset from a business perspective, it was important that the resulting clusters would be comprehensible for decision makers.

Due to these constraints we decided to apply only the *k*-means++ algorithm to the Narrative dataset, while also including the HCA algorithm when validating the DTI. Because of the size of the Narrative dataset, it was not practical to apply the HCA algorithm.

Test Design

We tested the quality and validity of the hybrid model through several evaluation measures (see Figure 4.3). In order to assess the accuracy of the decision tree model, we applied 10-fold cross validation. Furthermore, we ran each decision tree as a binary tree. While the numerical attributes always result in a binary splitting point, this is not the case with the nominal attributes. For example, an attribute describing the customers' nationality usually include many countries. Instead of displaying each country in the decision tree, a binary splitting point is created, splitting the dataset between one country and the remainder (e.g. Sweden or not Sweden). By doing this, we avoided overly complex tree structures as a few of the nominal attributes in the Narrative dataset included more than 100 unique values. By selecting a binary tree, unimportant values were excluded, making the understanding and evaluation of the trees easier. By the same token, the minimum leaf size (MLS) was set to 1 % of the examples in the datasets. This in order to keep the decision tree readable and comprehensible.

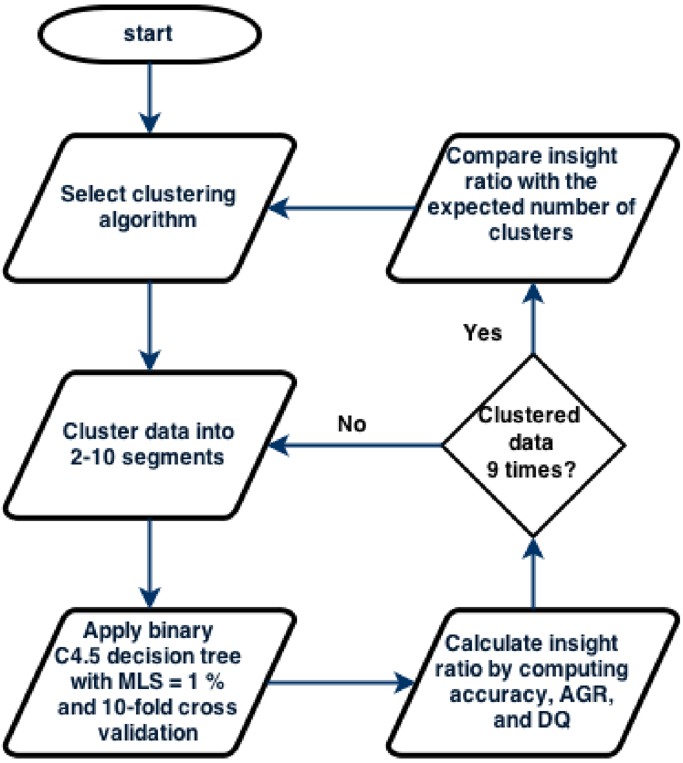

start
Select clustering algorithm
Compare insight ratio with the expected number of clusters
Yes
Cluster data into 2-10 segments
No
Clustered data 9 times?
Apply binary C4.5 decision tree with MLS = 1 % and 10-fold cross validation
Calculate insight ratio by computing accuracy, AGR, and DQ

Figure 4.3: Test design for the DTI

Evaluation Measures Assessment

In order to evaluate the proposed DTI, we compared the results from both the *k*-means++ and HCA algorithms to the expected number of clusters for the test datasets. We expected the measure to generate the highest DTI for the projected number of clusters. However, we anticipated that the algorithms would exhibit differences in suitability for the different datasets. Due to the advantages and disadvantages of each algorithm, one of them would correctly cluster a higher proportion of instances than the other. When this was the case, the resulting DTI would correspond with a higher value.

4.1.5 Evaluation

The previous assessment is intended to verify that the model and evaluation measures meet the data mining objectives from a technical aspect. The evaluation phase is intended to assess the final model's validity in relation to the business objectives. If there is a reason not to apply the resulting model, there is a discussion during this phase. During this step, the final model is evaluated, and the findings are reconnected to the purpose of the project. Chapter 6 describes the evaluation in detail (Chapman et al., 2000).

4.1.6 Deployment

The deployment phase includes the production of the final report as well as a final presentation. This phase also involves a review of the process, evaluating what went right and wrong, and what needs to be improved (Chapman et al., 2000).

Algorithm	Advantages	Disadvantages
K-means++ (centroid model)	Low time complexity Works well even for large datasets Easy to implement	Assumes globular clusters all the time, does not handle skew data well Does not recognize noise Different random settings result in different clusters
EM (distribution model)	Fastest algorithm for learning mixture models No predetermined number of *k* is required Does not bias the cluster sizes to have specific structures	Assumes normal distribution of data, even when this is not the case Always uses all the instances it has access to, leading to complex criteria for deciding how many components to apply Does not recognize noise
DBSCAN (density model)	No predetermined number of *k* is required Can find arbitrarily shaped clusters Performs well with outliers	Does not work well with large differences in densities Border points reachable from multiple clusters may fall into either cluster without a deterministic selection
HCA (connectivity model)	No predetermined number of *k* is required Easy to implement Builds dendrograms using all instances	High time complexity, best case $O(n^2)$ Does not work well for large datasets Does not recognize noise

Table 4.2: Advantages and disadvantages of four common clustering algorithms

Chapter 5
Results

This chapter presents the performance of the Description Tree Index (DTI) for the different datasets, as well as the application of the Business Insight Index (BII) for the Narrative data.

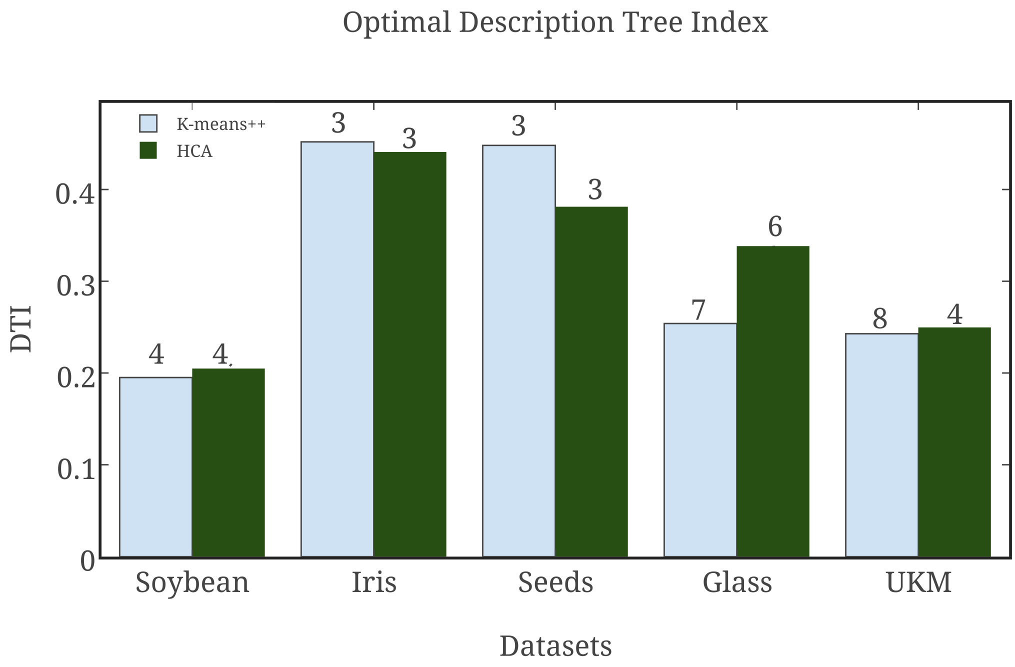
5.1 Overall Performance

Figure 5.1 shows the results of the DTI validation. For the Soybean, Iris, and Seeds datasets, both the *k*-means++ and HCA algorithms generated the optimal DTI for the expected number of clusters (compare with Table 4.1 in Sect. 4.1.2). However, only the HCA algorithm produced the highest DTI for the expected number of clusters when it came to the Glass and UKM datasets

Furthermore, the DTI enables a comparison between clustering algorithms for the same dataset. Regarding the Soybean, Iris, and UKM datasets, the algorithms perform almost equivalent, while the *k*-means++ algorithm outperforms HCA for the Seeds dataset and vice versa for the Glass dataset. Figure 5.2 further reflects this difference between the clustering algorithms. The figure illustrates the proportion of correctly classified instances for the expected number of clusters. This indicates how well the algorithms cluster the instances into the expected segments. In general, the algorithm with the highest proportion of correctly classified instances also generates the highest DTI. The sections below go into detail on how the results are structured.

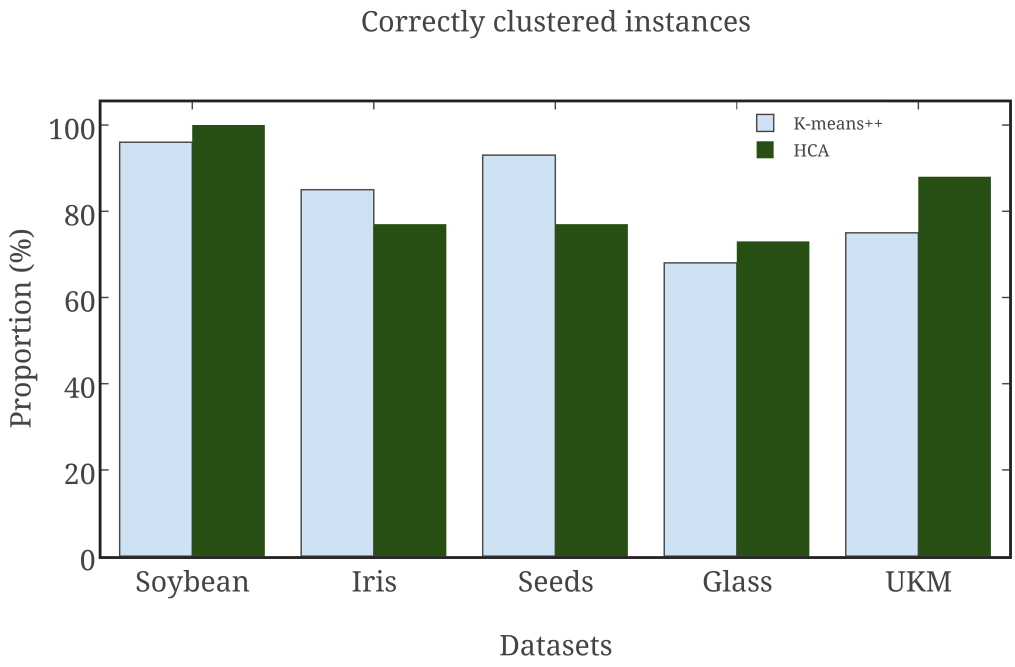
5.2 Description Tree Indexes

Figures 5.3-5.7 show the DTI for the different datasets. As can be seen, the results differ from each other from dataset to dataset. In the Soybean dataset, both the *k*-means++ and the HCA algorithms generate the highest DTI for 4 clusters with a clear peak. After the



Optimal Description Tree Index

Figure 5.1: Optimal DTI for the datasets. The number above each bar indicates for which cluster partitioning the optimal DTI was achieved



Correctly clustered instances

Figure 5.2: Proportion of correctly clustered instances for the datasets

peak, the graphs are declining as the number of clusters increase. The DTI for the Iris data projects a similar pattern, with a clear peak for 3 clusters.

For the Seeds data, the *k*-means++ algorithm performs significantly better than the HCA algorithm, which Figure 5.5 reflects as *k*-means++ generates a higher proportion of correctly clustered instances. However, after 4 clusters, the HCA algorithm produces a higher DTI. This indicates that the HCA algorithm extracts more insights than the *k*-means++ algorithm for these cluster partitions. For the Glass and UKM datasets, the HCA algorithm performs the best, with clear peaks at the expected number of clusters. This also corresponds to the higher percentage of correctly clustered instances. Yet, for the UKM dataset the *k*-means++ algorithm almost reach as high DTI as the HCA algorithm, but for 8 clusters. This might be due to the fact that the initial placement of the first cluster center affects the *k*-means++ clustering. Since the initial placement is done arbitrarily, different random settings result in different clusters, and thus different DTIs.

Moreover, there are significant differences in DTIs between the datasets. For the Iris and Seeds data, the DTI exceeds 0.40, while the DTI for the Soybean data barely reaches 0.20. This is most likely due to a higher average gain ratio (AGR) for these datasets, described below.

To understand what the differences in DTIs depend on, it is necessary to closer study the components of the DTI: the accuracy, AGR, and weighted attribute-values per cluster (WAVC) of the datasets.

5.2.1 Accuracy

Figures 5.8-5.12 display various differences in accuracy between the algorithms. For all datasets, the HCA algorithm produces a better accuracy than the *k*-means++ algorithm for 2-4 clusters, and a higher accuracy in general. This might be because the HCA algorithm forms clusters by building hierarchies including all instances.*k*-means++ on the other hand, is a centroid based algorithm, where each instance is assigned to the closest cluster center. In general, the two algorithms generate accuracies above 95 % for 2 cluster. The accuracies do however decline as the data is divided into more clusters. This is natural, as the increased number of clusters also increase the probability that the C4.5 algorithm will incorrectly classify instances.

Except for the UKM dataset, the accuracies do not peak at the expected number of clusters. This reinforces the notion that accuracy alone is not a good measure for determining the optimal number of clusters.

5.2.2 Average Gain Ratio

The AGRs of the datasets are in many respects similar to each other. Especially for the Soybean, Iris, and Seeds datasets, the AGRs for the two algorithms exhibit only small differences. Instead, it is the relative AGR levels, which differ considerably between the data. The Iris and Seeds datasets have an AGR level that exceeds 60 % for many cluster divisions. This is correlated with the overall high DTIs for these datasets. Moreover, it implies that a majority of the attributes contribute to the cluster classification. With only 4 attributes for the Iris data and 7 attributes for the Seeds data, the low dimensionality level results in attributes that highly contribute to the cluster partitioning. As the UKM dataset

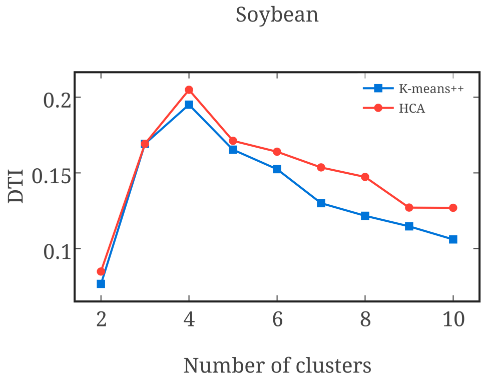


Figure 5.3: DTI for the Soybean data

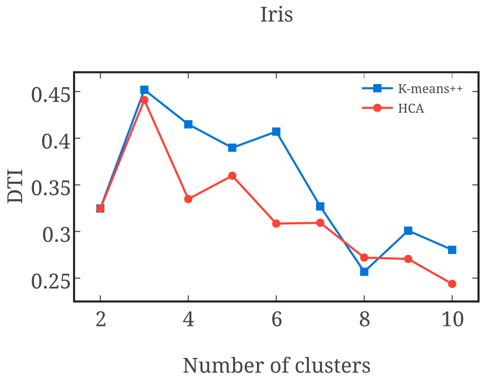


Figure 5.4: DTI for the Iris data

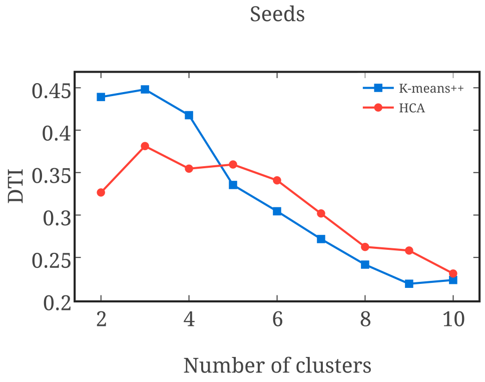


Figure 5.5: DTI for the Seeds data

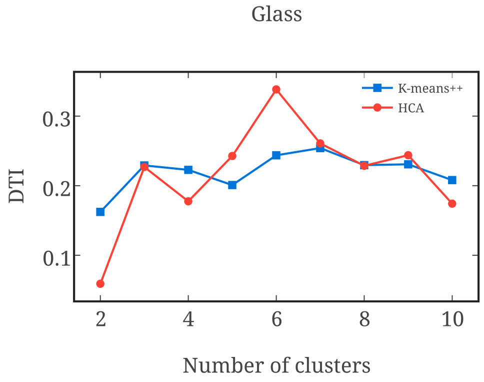


Figure 5.6: DTI for the Glass data

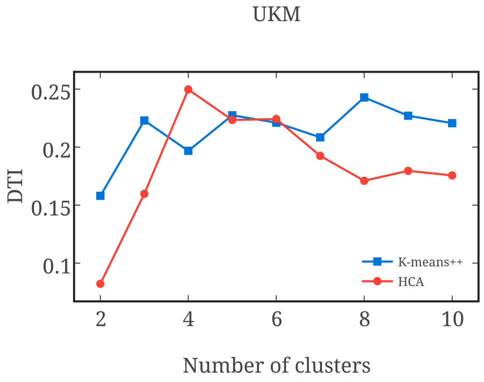


Figure 5.7: DTI for the UKM data

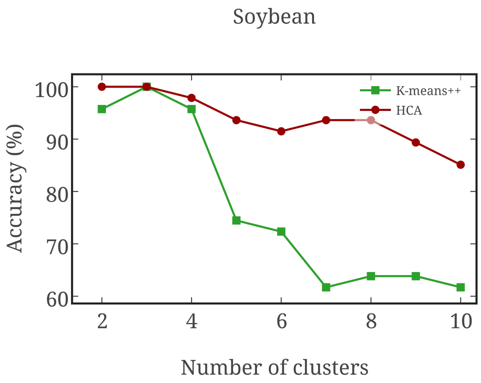


Figure 5.8: Accuracy for the Soybean data

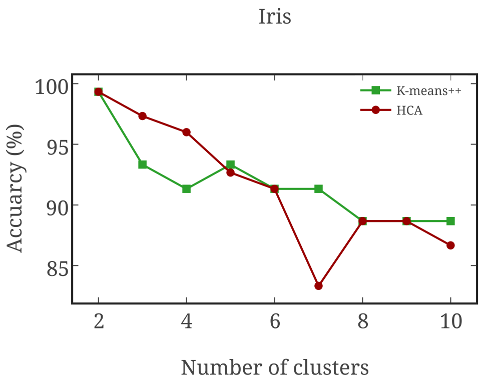


Figure 5.9: Accuracy for the Iris dataset

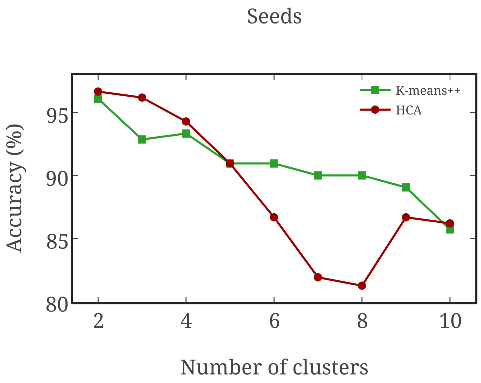


Figure 5.10: Accuracy for the Seeds data

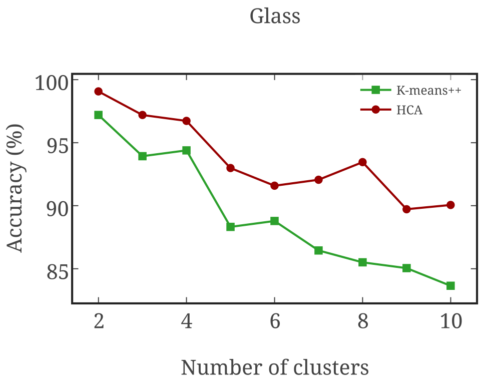


Figure 5.11: Accuracy for the Glass data

only has 5 attributes, the reason for the initial low AGR is related to the low information gain ratio (IGR) of a few of the attributes. These attributes do not contribute at all to the cluster identification, which lowers the AGR. A low AGR reflects a poor choice of attributes, and attribute selection techniques might be desirable.

Even though the HCA algorithm was able to correctly cluster all instances in the Soybean dataset, there was a large proportion of attributes with an IGR of 0. This lowered the AGR and thus the overall DTI.

Additionally, Figures 5.13-5.17 illustrate how the AGRs increase with a growing number of clusters. This is explained by the increased impact of each attribute when creating new clusters. As the number of clusters grows, they end up closer and closer to each other, meaning it will require more attributes to separate them. Thus the AGR increases.

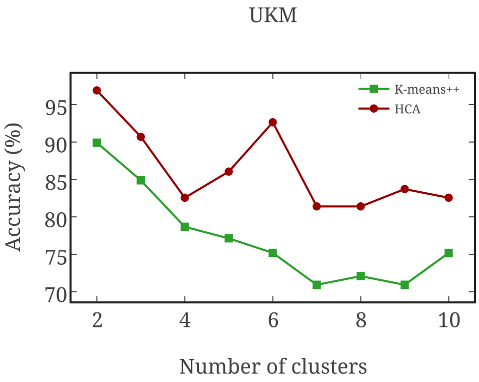


Figure 5.12: Accuracy for the UKM data

5.2.3 Weighted Attribute-Values per Cluster

The WAVC is closely related to the DTI. For many datasets, the peak of the WAVC coincides with the peak of the DTI. This is the case for the Soybean, Iris, Seeds, and Glass data. However, the DTI cannot only be determined by the WAVC, which the UKM dataset reflects. Here, the WAVC peaks at 3 clusters, which is inaccurate for deciding the expected number of clusters. This might depend on the nature of the dataset. It is possible that the data lacks the internal patterns necessary to perform a qualitative clustering. In the previous section, we observed that the UKM dataset obtained a relatively low AGR even though it only had 5 attributes. This raises suspicions regarding the dataset's quality.

Of interest is also the Glass dataset, where there is peak at 3 clusters for both algorithms. For the HCA algorithm, it is however only a local optimum as the WAVC is higher at 6 clusters, which is the expected number of clusters. The *k*-means++ algorithm, on the other hand, achieves its best WAVC at 3 clusters.

Furhtermore, for most of the datasets, the WAVC reaches almost 0.80. This is not the case for the Soybean dataset, where the algorithms only achieve a maximum WAVC of 0.60. This relates to the size of the Soybean dataset. With only 47 instances and a low AGR, the amount of insights to be retrieved is limited.

Moreover, Figures 5.18-5.22 show a pattern in the WAVC of the algorithms. The graphs start to overlap as the number of clusters increases. With an increasing number of clusters, the WAVC decreases until every instance becomes a separate cluster.

5.3 The Narrative Dataset

Figure 5.23 visualizes the resulting DTI for the Narrative dataset. As can be seen, reducing the number of outliers increases the DTI. The data with fewer outliers achieved a clear peak at 4 clusters, as compared to 5 clusters with the original data. The increase in DTI depends on the fact that the *k*-means++ algorithm does not recognize outliers, and the clusters distort when including them. By removing outliers, the clusters become more compact, and more insights are gained through the decision tree. Figures 5.24-5.26 illustrate the compo-

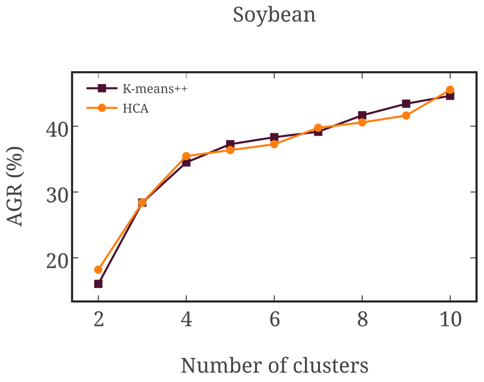


Figure 5.13: AGR for the Soybean data

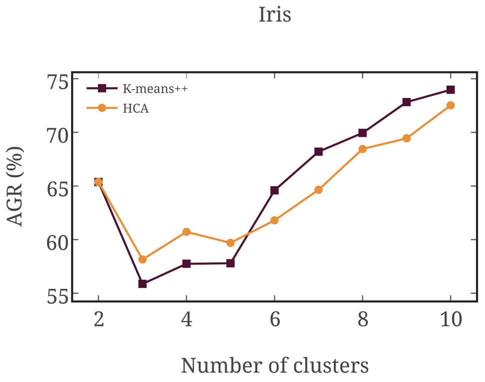


Figure 5.14: AGR for the Iris data

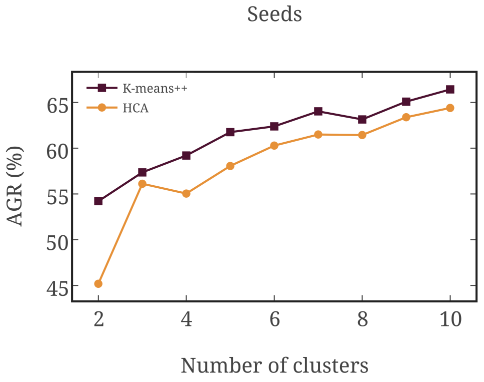


Figure 5.15: AGR for the Seeds data

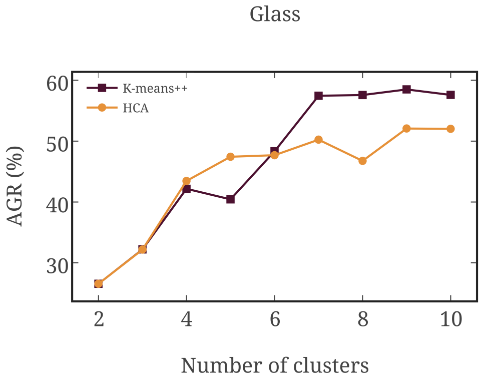


Figure 5.16: AGR for the Glass data

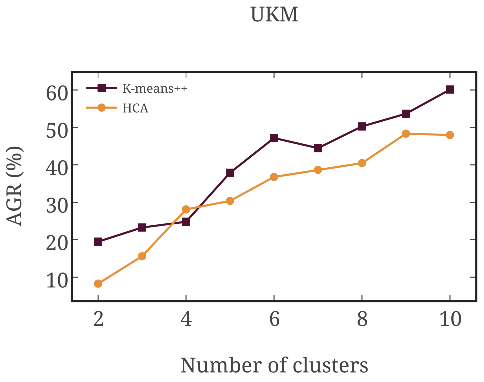


Figure 5.17: AGR for the UKM data

nents of the DTI, with and without outliers. Although there are minor changes in accuracy and AGR, the only component that changes significantly is the WAVC. The WAVC now peaks at 4 clusters, and thus generates the optimal DTI for this cluster partitioning.

However, when studying the normalized Calinski-Harabasz (CH) index in Figure 5.27, it advocates 2 clusters for the original data, and 3 clusters with a reduced number of outliers. This indicates a higher proportion of homo- and heterogeneity for these cluster partitions. Although, when combining the normalized CH and the DTI into the Business Insight Index (BII), the measure indicates 3 clusters for the original data, and 4 clusters with fewer outliers (see Figure 5.28).

Figure 5.29 illustrates an example of the resulting the decision tree for the Narrative data with a reduced number of outliers. Due to the sensitive nature of the information, the attributes and values are replaced with fictive values. The tree provides insight into the 4 segments created from the clustering. As the figure shows, the segments usually have more than one leaf describing the customers. This means that the segments are fragmented internally, and it is not enough with one set of attribute-values to describe the entire segment.

By also studying the cluster centers of the segments, we created a description for the 4 types of company customers. Here, we provide a fictitious example of how the segments could look like:

Young, high-using iPhone owners who cares about photo quality: These users are under 30 years old, have an iPhone, and take photos with a high average quality score. They are mostly males from the U.S., and they take a lot of photos. This segment (segment 1) consists of 1600 (19 %) customers, and has a high average CLV. These customers are highly valuable to Narrative as they are frequent users, and most likely to buy future products.

High-using females who share their photos on social media: This segment consists of females, mostly under 30 years, who take many photos, and share them on social media. The segment (segment 2) consists of 1100 (13 %) customers, and has the highest frequency rating. This means that these customers use the camera more days than other customers. To further market the product to these customers, Narrative should investigate how to better integrate the camera with social media platforms.

Low-using, slow-starting customers with low understanding: This is the largest segment (segment 3) with 3200 (38 %) customers. These users have a low understanding of the product, as evidenced by the many days between sign-up and their first photo. Their lack of understanding of the product result in that the photos they take have a low average photo score, and hence they take fewer photos. To engage these customers, Narrative has to provide better information about the product, as well as reminding the customers to use it.

Early users who loose interest in the product: These customers start using the camera shortly after they get it, but stop after a couple of weeks. This may be because they forget to charge it or are not satisfied with the photo quality. The segment (segment 4) include 2552 (30 %) customers who have a low recency rating compared to the average customer. For this segment, it is important to examine the reasons why these customers loose interest in the product after a while.

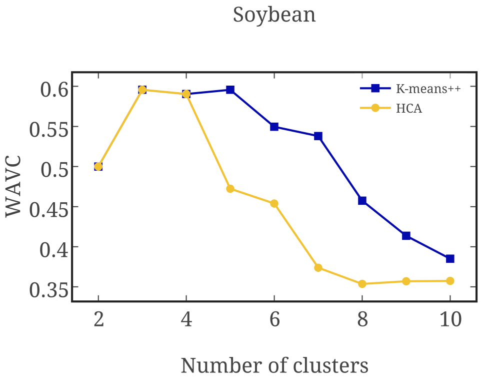


Figure 5.18: WAVC for the Soybean data

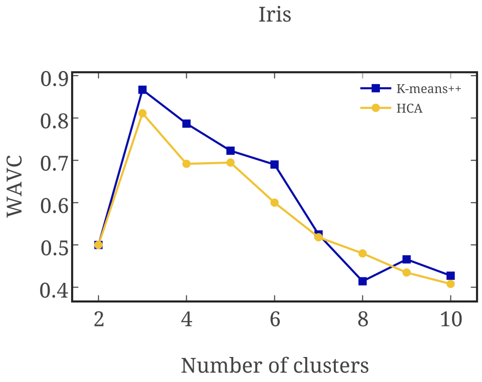


Figure 5.19: WAVC for the Iris data

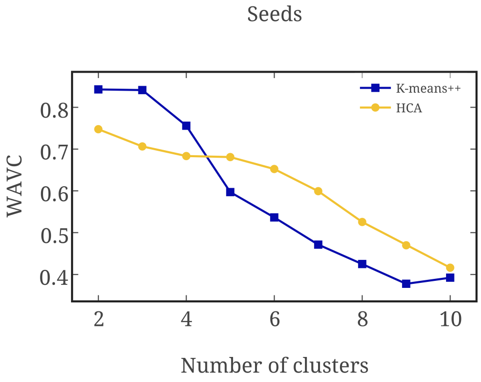


Figure 5.20: WAVC for the Seeds data

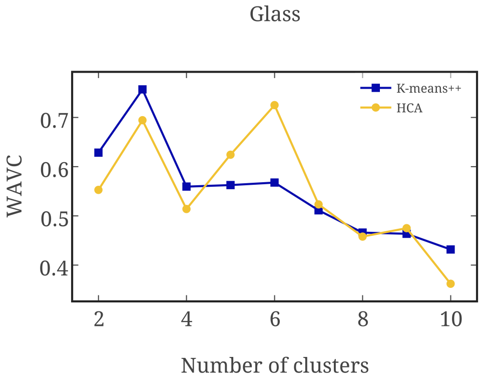


Figure 5.21: WAVC for the Glass data

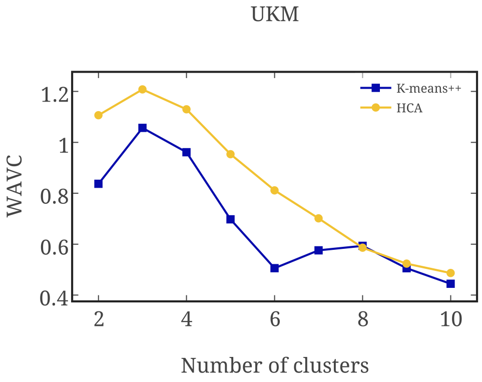


Figure 5.22: WAVC for the UKM dataset

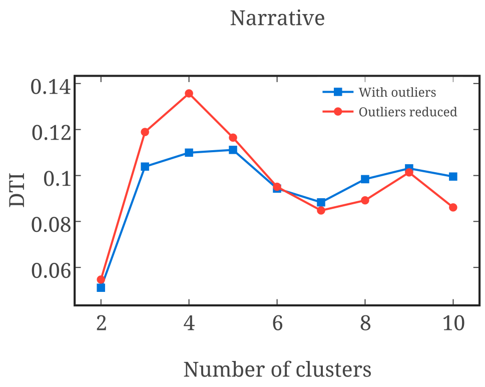


Figure 5.23: DTI for the Narrative dataset

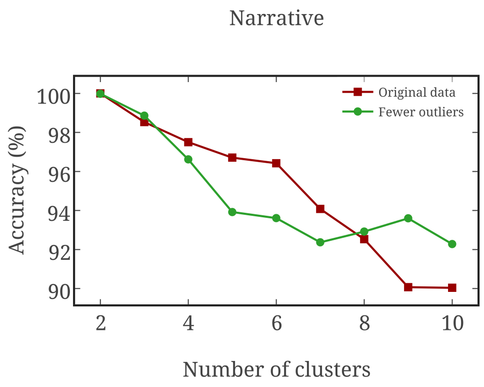


Figure 5.24: Accuracy for the Narrative data

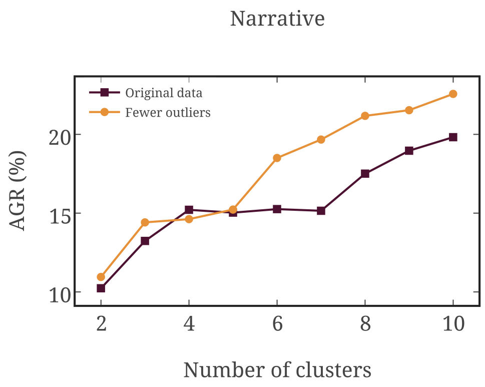


Figure 5.25: AGR for the Narrative data

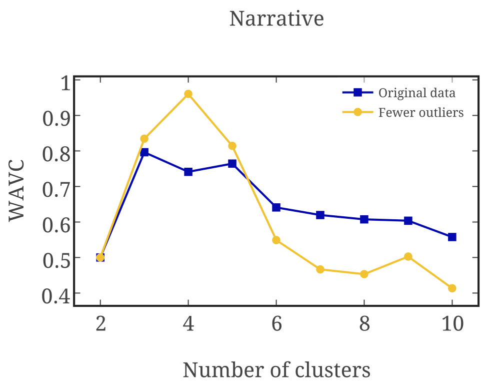


Figure 5.26: WAVC for the Narrative data

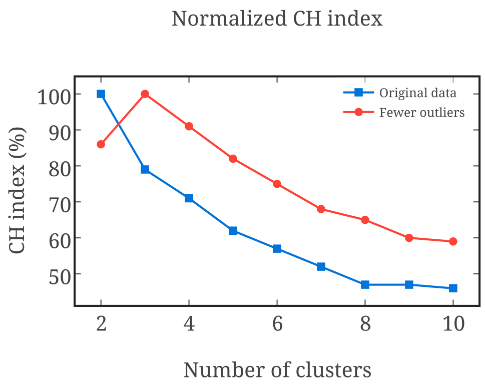


Figure 5.27: CH for the Narrative data

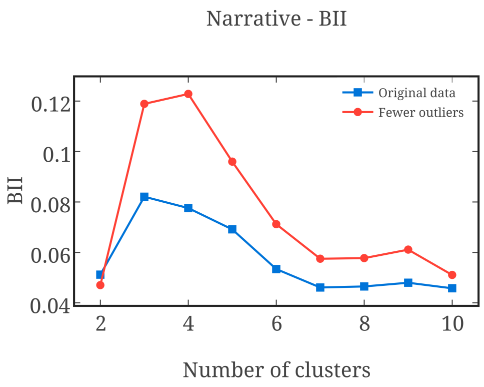


Figure 5.28: BII for the Narrative data

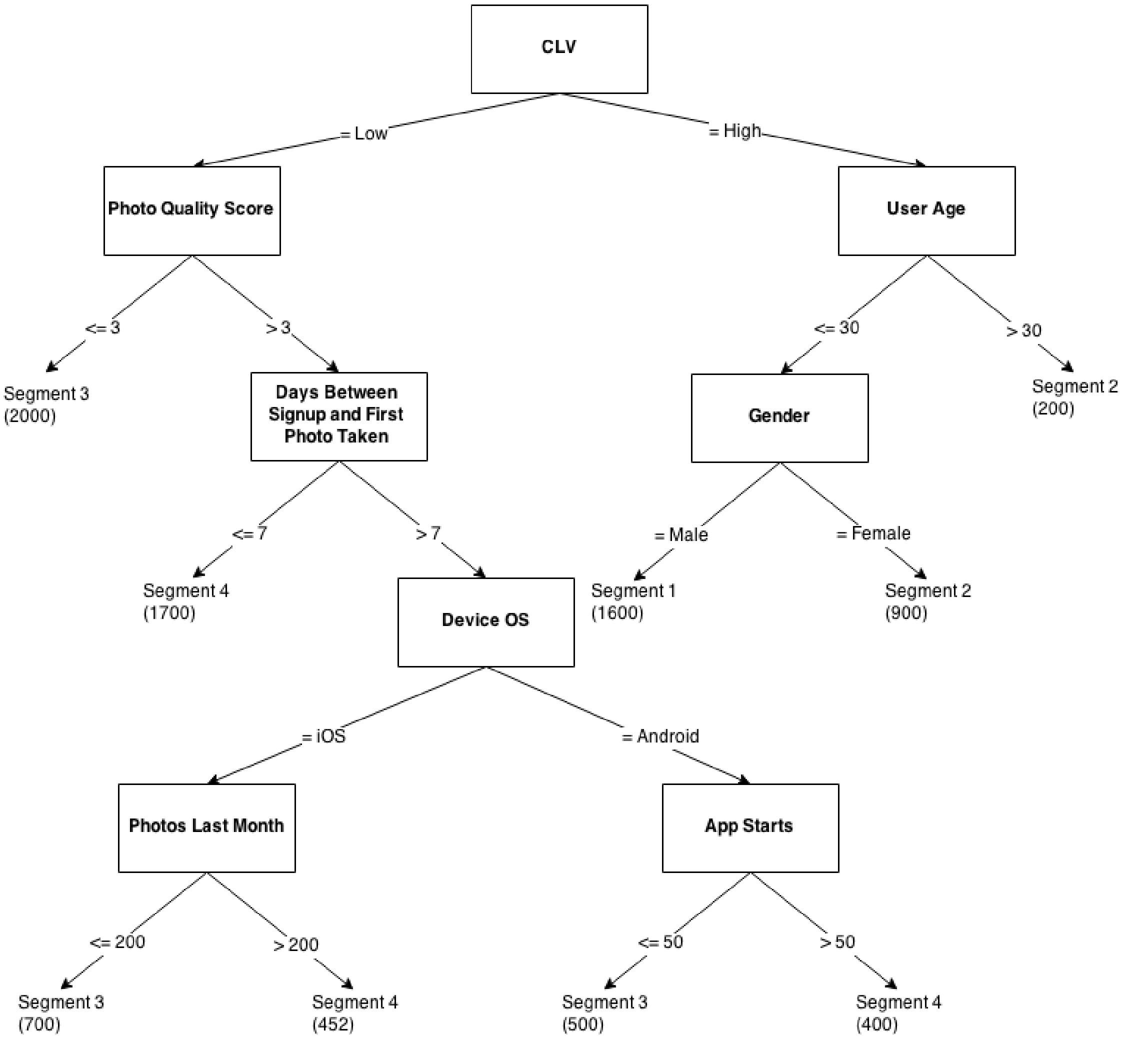


Figure 5.29: An example of the resulting decision tree for the Narrative dataset. Here, the original attributes and values are replaced to protect the company insights

Chapter 6

Discussion

In this chapter, we hold a discussion regarding how the results reconnect to the purpose and objectives presented in the introduction, and why this is the case. We discuss how the application and evaluation of hybrid models for customer segmentation and target analysis affect the results.

6.1 Application

To successfully use hybrid models for customer segmentation and target analysis, several factors are important in order to validate the results. First of all, the selection of customer data is essential to the quality of the analysis. For the Narrative data, even though we selected the attributes with the highest business relevancy, many of them lacked in significance from a data-mining perspective. This became evident when we examined the average gain ratio (AGR) of the attributes. The AGR was considerably lower than the AGR for the test data. The low AGR indicates that only a few of the attributes have an impact on the resulting segmentation. To ensure the descriptive quality in the decision tree, some attributes might have to be removed or transformed to achieve better results.

The use of 32 attributes in the Narrative dataset may have influenced factors such as the Euclidean distance measuring, outlier detection, and the final clustering. By further reducing the dimensionality of the data, other insights might have been presented. On the other hand, from a business aspect it may be as important to find out which attributes are technically insignificant as the attributes that are essential. By using selection criteria, considering technical significance and correlation, as well as business importance, the quality of the attributes increases. Also, the the choice of attributes greatly impact the number of outliers in the data.

By applying the local outlier factor (LOF) to the datasets, we enabled an evaluation of their densities. This is important from a business perspective, as objects closer to each other have a greater probability to respond similarly to specific marketing mixes. Even though our application of the LOF to the Narrative data did not guarantee that all outliers were detected, we detected and removed the ones with the highest LOF score. While it might be questionable to remove outliers from a data-mining perspective, it may be valuable from a business standpoint. By removing the outliers, the quantitative properties of the results deteriorate, while the qualitative properties often improve. This might increase the validity of the results, as the insights become clearer. However, the results might not be applicable in a larger context, as fewer objects are represented. For this reason, it is important to set a limit on how many customers are required to draw both quantitative and qualitative conclusions.

The selection of clustering and decision tree algorithms is another factor with great significance to the results. When selecting techniques, the appropriateness, availability, and constraints of existing algorithms are essential. Unless the structures of a dataset are evident, there is usually more than one way to cluster and classify the data. Especially concerning customer data, a segmentation that is qualitative from a data-mining perspective, may lack business implications in reality. There is no guarantee that the algorithms selected in this project are the optimal ones. We choose these due to their advantages and disadvantages compared to other algorithms. As there are various established and proposed techniques, the importance of trying several algorithms rather than focusing on one increases.

Finally, the minimum leaf size (MLS) of the decision trees is factor with great importance to the complexity and comprehensibility of the resulting insights. As a lower MLS leads to bigger trees in terms of decision nodes and leaves, the accuracy increases. However, the visualization and understanding of the tree decreases. Although this parameter was not up for review in this thesis, its significance cannot be denied.

6.2 Evaluation

In the introduction we described the lack of measures which evaluate the insights presented to decision makers. Through the creation of the Description Tree Index (DTI) and Business Insight Index (BII), we hope to provide a broader understanding, and better evaluation of customer segmentation and target analysis in companies such as Narrative.

6.2.1 The Description Tree Index

The results presented in chapter 5 indicate that the DTI works well as a cluster validation measure. The measure generated the best results for the expected number of clusters in each dataset, at least for one of the algorithms used. Also, the better DTI produced from the two algorithms in each dataset seems to correspond to the proportion of correctly clustered instances.

Even though the optimal DTI in many cases coincides with the expected number of clusters, the original purpose of the DTI is to evaluate the information made available in a decision tree. This means that the use of clustering algorithms that are not compatible with a specific dataset may generate misleading results. In adjacent clusters where instances are close to each other, the k-means++ algorithm has major problems with splitting clusters. This might have been the case of the Glass and UKM datasets where the k-means++ algorithm performed poorly. However, the *k*-means++ algorithm extracted almost as much information as the HCA algorithm, but for 8 clusters instead of 4 in the UKM case. This does not mean that 8 clusters is an optimal choice. Instead, it indicates that the *k*-means++ algorithm generates more insights for 8 clusters rather than 4. For this reason, it is important to select the clustering algorithm best fit for a specific dataset. This argument is further strengthened by the fact that the HCA algorithm produced better results where it generated a higher proportion of correctly clustered instances.

6.2.2 The Business Insight Index

As opposed to internal clustering validity measures, the DTI does not take the separation and compactness of clusters into consideration. This is perhaps the measure's greatest weakness. By integrating the DTI with an internal validity measure, the resulting Business Insight Index (BII) evaluates both the insights, and the homo- and heterogeneity of a particular customer segmentation. Sect. 5.3 shows that the DTI and the Calinski-Harabasz (CH) index favored two different cluster divisions, both for the original data and the dataset with a reduced number of outliers. However, when reducing the number of outliers, the results of the CH index and DTI exhibited greater similarities, as they advocated 3 and 4 clusters respectively. For the original data, the CH index peaked at 2 clusters, while the DTI peaked at 5 clusters. These results indicate that when increasing the quality of the data, the two measures will coincide.

Yet, as the DTI and CH index assess different aspects of the cluster partitioning, they might give different answers. Instead of regarding them as opposites, the measures are integrated in order to evaluate customer segmentation from several aspects. By doing so, the risk of suboptimal results, in comparison with the separate evaluation measures, decreases. To conclude, the BII comes with the gain of a broader picture, considering both the data-mining and business standpoints of the segmentation.

Moreover, the CH index is not the only internal cluster validation measure. There are various assessment methods, each with its own advantages and disadvantages. While the CH index was used for this thesis project, other measures could be applied to enable a comparison of the results. As the measures are based on different aspects of the cluster compactness and separation, they exhibit different results depending on the data characteristics. Since data properties such as monotonicity, outliers, density differences, and skewness impact the measures, the choice of measure should be carefully investigated.

Finally, the use of the RFM model further ensures the business quality of the BII. By integrating the model, marketers, sales-teams, and product developers are able allocate time and resources to the different segments. In this thesis, the recency, frequency, and monetary value were weighted relativistically when deciding each customer lifetime value (CLV). Depending on the business context, different weights might be desirable. Furthermore, we divided the CLV attribute into two discrete groups, one with a low CLV and one with a high CLV. This is however more of a heuristic choice than a standard, as the choice of two CLV groups generated the highest AGR. Other groupings might have produced different results, for better or worse. Depending on the distribution of CLVs, the specific groupings might have to be adjusted.

6.2.3 Narrative

By applying the BII to the Narrative dataset, we segmented and analyzed the customers. The resulting segments present valuable insights in customer behaviour and characteristics. Especially, the value drivers of each segment are identified (e.g. social media access, photo quality etc). These drivers enable a mapping of the customers' present and future potential. The value drivers are then to be used when developing the product, and to improve sales through different marketing mixes.

The decision tree in Figure 5.29 provides a description of the segments, but also a map for segmenting new customers. In the fictitious tree, the CLV created by the RFM model has the highest information gain ratio (IGR), and the attribute thus become the root node. Even though there are other discrete attributes with only two values, such as the gender attribute, the CLV effectively splits the data. This increases the validity of the results, as the customers clearly differ from each other in CLV. The CLVs for the segments enable a ranking and comparison of the segments, in order to target the profitable ones.

Chapter 7
Conclusions and Future Work

The purpose of this thesis is to perform and evaluate customer segmentation and target analysis by means of hybrid models. The creation of the Business Insight Index (BII), and its application to the Narrative data, proved helpful for evaluating customer insights. The measure contributed new insights which can be used for marketing and product development. However, the use of the BII still requires an extensive data understanding and thorough data preparation. To investigate the data properties before applying specific data-mining algorithms is a critical step. Depending on factors such as monotonicity, outliers, density differences, and skewness, the data should be prepared correctly to yield qualitative results. These factors greatly affect the performance of the Description Tree Index (DTI) and the Calinski-Harabasz (CH) index.

By testing the DTI, we conclude that the measure successfully evaluates the amount of insights in a decision tree. However, it does not measure the homogeneity within clusters nor the heterogeneity between them. For this reason, the measure is integrated with the CH index to form the BII. This measure evaluates the insights presented through the decision tree, as well as the compactness and distinction of the customer segments. Moreover, the RFM model enables a qualitative comparison of the segments in order to perform a target customer analysis.

Additionally, there is room for improvement regarding the measures created. Even though the DTI generated the best results for the expected number of clusters, further testing is required to improve and validate the evaluation measure. This includes testing more algorithms on both artificial and real data. Also, we advise experiments with weighted clustering and decision trees in order to better reflect the hierarchical importance of attributes, which often exists in companies. Moreover, we believe that the minimum leaf size (MLS) should be investigated closer. This component could be included in the DTI as a measure for the generated complexity cost of the decision tree.

Finally, as mentioned in Sect. 6.2.2, the CH index is not the only internal cluster validation measure. We recommend comparing different validation measures and their contribution, if they are integrated in the BII.

Bibliography

Amherst, U. (2015). Bioinformatics and genomics. `https://bcrc.bio.umass.edu/courses/spring2012/micbio/micbio660/content/mev-microarray-analysis`. Accessed: 2015-06-08.

Arthur, D. and Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, Chicago. Society for Industrial and Applied Mathematics.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Dhiman, R., Vashisht, S., and Sharma, K. (2013). A cluster analysis and decision tree hybrid approach in data mining to describing tax audit. *International Journal of Computers & Technology*, 4(1):114–119.

Estivill-Castro, V. (2002). Why so many clustering algorithms - a position paper. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75.

Jain, A., Murty, M., and Flynn, P. (1999). Data clustering: A review. *ACM Computer Survey*, 31(3):264–323.

Lapczynski, M. and Jefmanski, B. (2013). Impact of cluster validity measures on performance of hybrid models based on k-means and decision trees. *Advances in Data Mining*, pages 153–162.

Lapczynski, M. and Jefmanski, B. (2014). Number of clusters and the quality of hybrid predictive models in analytical crm. *Studies in Logic, Grammar and Rhetoric*, 37(50):141–157.

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures. In *IEEE International Conference on Data Mining*, pages 911–916.

Marbán, Ó., Gonzalo, M., and Segovia, J. (2009). *A Data Mining & Knowledge Discovery Process Model*. I-Tech, Vienna.

Milligan, G. and Isaac, P. (1980). The validation of four ultrametric clustering algorithms. *Pattern Recognition*, 12:41–50.

Narrative (2015). A new kind of photographic memory. `http://getnarrative.com/narrative-clip-1`. Accessed: 2015-06-08.

Ngai, E., Xiu, L., and Chau, D. (2008). Application of data mining techniques in customer relationship management. *Expert Systems with Applications*, 36:2592–2602.

Nimbalkar, D. and Shah, P. (2013). Data mining using rfm analysis. *International Journal of Scientific & Engineering Research*, 4(12):940–943.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Rokach, L. and Maimon, O. (2008). *Data Mining with Decision Trees: Theory and Applications*. World Scientific Pub Co Inc., Tel-Aviv.

Sayad, S. (2015). Hierarchical clustering. `http://www.saedsayad.com/clustering_hierarchical.htm`. Accessed: 2015-06-08.

Schubert, E., Zimek, A., and Kriegel, H. P. (2012). Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 28(1):190–237.

Scikit-learn (2014). A demo of the mean-shift clustering algorithm. `http://scikit-learn.org/stable/auto_examples/cluster/plot_mean_shift.html`. Accessed: 2015-06-08.

Söderberg, E. (2015). Interview.

Vidal, J. M. (2009). Decision tree learning. `http://jmvidal.cse.sc.edu/talks/decisiontrees/allslides.html`. Accessed: 2015-06-08.

Witten, I., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Inc, Burlington, 3rd edition.

INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTATIONSDAG 2015-06-04

EXAMENSARBETE The Business Insight Index – Evaluating Customer Insights through Hybrid Models
STUDENT Jonathan Bratel
HANDLEDARE Pierre Nugues (LTH), Erik Söderberg (Narrative)
EXAMINATOR Jacek Malec (LTH)

Kundinsikter genom data mining

POPULÄRVETENSKAPLIG SAMMANFATTNING **Jonathan Bratel**

Att öka förståelsen av sina kunder har utvecklats till en allt viktigare uppgift för dagens företag. För att få bättre kunduppfattning utvecklas i detta examensarbete ett nytt mått som bedömer kundinformation.

Ett allt vanligare problem för affärsverksamheter är att försäljnings- och marknadsföringskostnader ökar. Detta beror på att dagens kunder har alltmer olikartade köpbeteenden. För att kunna prioritera de mest värdeskapande kunderna krävs numera att en segmentering genomförs. Att segmentera kunder innebär att dela upp en marknad i mindre delar utefter olika kundegenskaper, exempelvis ålder, kön eller inkomstnivå. Att segmentera kunder har dock blivit en allt mer komplex uppgift, då informationen om kundernas egenskaper och beteende ökat lavinartat de senaste åren. Bara på det sociala mediet Facebook laddas över 500 terabyte data upp varje dag. Med hjälp av data mining kan segmentering och prioritering utföras även på stora mängder data. Data mining består av verktyg och tekniker för att hitta mönster, samband och trender i data. Dessa insikter kan sedan utnyttjas av beslutsfattare för att skapa konkurrensfördelar.

The Business Insight Index

För att kunna utvärdera kundinsikterna skapas i examensarbetet ett nytt mått – Business Insight Index (BII). Detta mått kan användas för att avgöra om en kundsegmentering är mer kvalitativ än annan. Genom att utvärdera mängden information som görs tillgänglig till beslutsfattare kan måttet förbättra kundsegmenteringsprocessen. BII uppvisar goda resultat vid test på fem datafiler där segmenteringen är känd sedan tidigare. För varje datafil genererar måttet bäst resultat för de förväntade segmenten.

Traditionella evalueringsmetoder håller inte måttet

Vanligtvis utvärderas segmentering inom data mining genom att mäta hur inbördes lika segmenten är i förhållande till hur olika de är sinsemellan. Dessa mått tar dock inte hänsyn till mängden information som förmedlas till säljare och marknadsförare. Bättre kundinformation kan på sikt leda till konkurrensfördelar och ökad försäljning. Därför är det viktigt att dels utvärdera segmentens kvalitativa egenskaper, men även till vilken grad dessa kan förstås och kommuniceras.

Narrative

För att hjälpa företaget Narrative att segmentera sina kunder utnyttjas BII. Narrative är ett Linköpingsbaserat företag som marknadsför lifelogging-kameror. Var 30:e sekund tar dessa bilder, vilka kan laddas upp till företagets servrar. Kunder kan sedan komma åt korten via företagets mobil-app. Genom att dela in företagets kunder i segment och sedan utvärdera dessa, får Narrative information om vilka värdedrivare kunderna ser i produkten. Är exempelvis hög bildkvalitet viktigare än anpassningsmöjligheter till sociala medier? Eller är bildfrekvensen den viktigaste faktorn? Då företaget identifierar kamerans värdedrivare kan produkten utvecklas och marknadsföras till de olika segmenten.

Integrering av segmentering och beslutsträd

Genom att analysera segmenten i beslutsträd framkommer vilka egenskaper som är utmärkande för kunderna. Beslutsträdet förutsäger vilka värden som kommer krävas för att en kund ska placeras i ett specifikt segment. Detta verktyg är fördelaktigt då det möjliggör en visualisering av kunderna som enkelt kan förstås av förklaras för beslutsfattare.