

MINING GENERALIZED CUSTOMER PROFILES

Fatma E. Giha¹, Y.P. Singh², H.T. Ewe³

¹ Computer Science Department, Computer Man College
P.O. Box 10553, Khartoum, Sudan

Tel: +249-9-12155422, Email: fgiha@yahoo.com

² CDAC (Center For Development of Advanced Computing),
Anusandhan Bhawan, C-56/1, Institutional Area ,Sector-62, Noida – 201307, India.

Fax: +91 120 240 2569, E-mail: ypsingh@cdacnoida.com

³ Faculty of Information Technology

Multimedia University Jalan Multimedia, 63100, Cyberjaya, Selangor, Malaysia

Fax: +60-3-83125264, E-mail: htewe@mmu.edu.my

Abstract

Modeling the customer behavior becomes a key issue in customer retention, direct marketing and product promotion. Many data mining techniques have been designed to address the problem of capturing the true behavior of customers for various businesses applications, in which the data can be represented in form of demographic and transactional datasets. In this paper, we propose to use association rules mining technique for modeling customer behavior in form of constructing *generalized profile association rules*. Our proposed scheme based on transactional databases that consist of demographic data (i.e., personal information), and transactional data where items bought in a customer transaction can come from any level of hierarchy/ taxonomy representation (i.e., computer *is-a* Hardware). The constructed profiles can be used for segmenting the customers into groups, and identifying the potential ones for targeted marketing and Customer retention. We implement pruning techniques and interestingness measures to select the relevant and interesting rules for profiling process. Experimental results on a synthetic dataset are provided to show the effectiveness of the proposed scheme for generalized customer profiling.

Keywords: *Generalized customer profiles, association rules mining, customer modeling*

1. Introduction

Data mining is the process of extracting valid, useful, previously unknown, and ultimately comprehensible knowledge from large databases [1]. Recently, data mining has attracted great attention in the database and machine learning research area, since it has financial and commercial usefulness. Data mining is considered as a step in the whole process of knowledge discovery problem statement [2].

Various data mining techniques with different working principles have evolved to accentuate discovering of knowledge, in form of association rules, characteristic

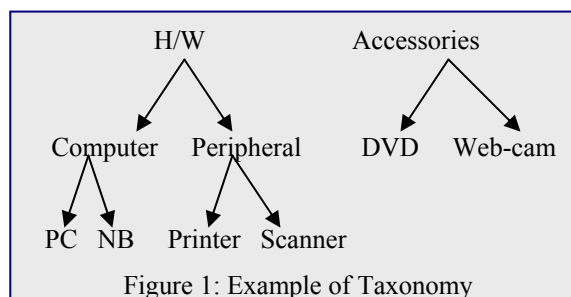


Figure 1: Example of Taxonomy

rules, classification rules, discriminative rules, sequential patterns, clustering rules, neural networks and many others. Data mining techniques can be used efficiently at any businesses application that involves data, such as:

- Increasing business unit and overall profitability
- Understanding customer desires and needs

- Identifying profitable customers and acquiring new ones
- Retaining customers and increasing loyalty
- Cross-selling and up-selling
- Detecting fraud, waste and abuse
- Determining credit risks
- Increasing web site profitability
- Monitoring business performance

Association rules discovery is one of the most powerful techniques in data mining that is used effectively to capture the true behavior of customers in various businesses applications. It searches for the items, which are bought frequently together by the customers.

In this paper, we are proposing association rules discovery for constructing generalized customer profiles, from a transactional database that consist of factual data records (i.e., age and income) as well as transactional data records (i.e., items bought). The transactional data can be represented in a form of hierarchy or taxonomy, which classifies the items into product groups, brands, or categories, etc... Figure 1 represents taxonomy T of transactional dataset for set of items $I = \{I_1, I_2, \dots, I_m\}$. Note that, NB is an abbreviation for Notebook, PC for Personal Computer, and H/W for Hardware.

An edge in T represents an (*is-a*) relationship. If there is an edge in T from p to c (p represents a generalization of c), we call p a parent of c and c a child of p [3]. Association rules are represented in a form $LHS \Rightarrow RHS$, in which we are considering the LHS to be any number of factual or demographic data, while the RHS is devoted for one behavioral or transactional attribute obtained from any level of hierarchy.

Based on our earlier work [4] & [5], we found that the evolution of data mining and knowledge discovery technology provides smarter and more focused customer profiling to increase business returns, because mass marketing is expensive and the returns on investment are frequently questioned.

Our contribution differs from the earlier work of Srikant and Agrawal [3] in that; they were restricted to the transactional data in a form of hierarchy, while we combine the hierarchical transactional dataset along with demographic data in form of customer profiles. In this paper, the generalized profile association rules are proposed to model the customers' behavior that in turn can be used as a powerful tool for retaining customers. Since the number of discovered association rules might be very large, we implement pruning techniques and interestingness measures to select the relevant and interesting rules for profiling process. Experimental results on a synthetic dataset are provided to show the effectiveness of the proposed scheme for customer profiling.

This paper is organized as follows: Section 2 provides problem statement along with definition of the profile association rules and the interestingness measures. The concepts of association rules mining, profile association rules, generalized association rules, and generalized profile association rules are presented in section 3. Pruning techniques and interestingness measures are presented in section 4. Section 5 is devoted for modeling customer profiles and analysis of the experimental results. Section 6 is a conclusion.

2. Problem Statement

1. Given a database of transactions D over a given itemset I , in which data can come from any level of taxonomy, as shown in Figure 1, we propose to investigate applications of *Apriori* algorithm [6] to discover generalized profile association rules of the form:

Factual attributes \Rightarrow Product from any level of taxonomy

We refer to this mining problem as *generalized profile association rules* for building customer profiles. For the given database of transactions D , we add all ancestors of each item to form set of extended transactions. Table 2 represents extended set of transactions of Table 1, where product's Accessories has been added to the product Webcam. Then we can apply *Apriori* to this set of extended transactions to generate generalized association rules.

2. The number of discovered association rules using *Apriori* is very large and many of the discovered rules do not have much meaning. Therefore, it is important to implement pruning and interestingness measurements so as to obtain the most interesting generalized profile association rules. In this paper we implement template-based pruning technique and R -interesting measure to select the most interesting profile rules among set of generalized rules.

3. Mining Generalized Profile Association Rules

Often there exists structures for the data, i.e. datasets have some hierarchy that describes the relation among different attributes for the data. This can be used as prior information before applying association rules discovery algorithms. For example, the hierarchy representation of Figure 1.

We propose to combine the concepts of data generalization and customer profiling to discover generalized profile association rules. In the next section we will discuss the association rules mining technique and its applications to the profile

association rules and generalized profile association rules.

3.1 Association Rules Discovery Technique

The problem of discovering association rules is to discover all rules that above user specified *minsupport* and *minconfidence* [6]. For a simple database of transactions *Apriori* algorithm can be used successfully to discover set of association rules. Moving towards the problem of discovering generalized association rules, given a database of transactions and a set of taxonomies, the problem of discovering generalized association rules is defined as the process of finding association rules above user specified *minsupport* and *minconfidence*, considering that items discovered can be from any level of taxonomy (i.e., leaf-level *itemsets* as well as parent *itemsets*).

A generalized association rule is an implication of the form $X \Rightarrow Y$, where X and Y can have three possibilities:

- x is a leaf-level *itemset*, while y' is an ancestor *itemset* (parent *itemset*)
- x' and y' are both ancestor *itemsets*
- x' is an ancestor *itemset*, while y is a leaf-level *itemset*

Consider the rule “item-Anc=Computer \Rightarrow item-Anc=H/W” with 60% support and 100% confidence, where H/W is an ancestor of the item NB and PC. If a rule “item =NB \Rightarrow item-Anc=H/W” with support 35% and confidence 100%, exists in the set of discovered association rules, it indicates more than half of the H/W sales are Notebooks. The latter rule is considered to be redundant since it is less general than the former one and doesn't convey any additional information. Given that $A, \hat{A} \subseteq I$, we call \hat{A} is an ancestor of A if we can get \hat{A} from A by replacing one or more items in A with their ancestors [3].

3.2 Generalized Profile Association Rules

In this paper we propose the *generalized profile association rules*, in which we tried to construct set of general behavioral rules that model the customer behavior and build profiles for the most profitable customers, rather than just finding the items that are frequently bought together. A generalized profile association rule is an implication of the form $X \Rightarrow \hat{Y}$, where X is a demographic / factual attribute, Y is an item that can come from any level of taxonomy. Consider the sample dataset shown in Table 2, which represents an extended dataset of Table 1; attributes age and income are considered as demographic data, while product (represented in taxonomy of Figure 1) is transactional attribute. To discover generalized

profile association rules, we consider that the data can come from any level of taxonomy for the RHS, while the LHS can contain a simple data of demographic attributes.

Discovering rules from different levels of taxonomy is important, because many interesting rules have not been discovered when we have restricted the rules to the items at leaf-level of taxonomy. For example, given a taxonomy that says; Notebook *is-a* computer *is-a* Hardware. We may discover a profile rule that young customers tend to buy computer, even though rules of “young customers tend to buy Notebook” and “young customers tend to buy Hardware may not reach the minimum constraints to be held.

When we apply *Apriori* to discover set of association rules, large number of rules has been generated. For that reason, we apply some pruning techniques to find the interesting rules among the redundant and insignificant rules.

ID	Age	Income	Item
1	Old	High	NB
2	Old	Average	PC
3	Old	High	NB
4	Young	High	DVD
5	Old	Low	PC
6	Old	Average	PC
7	Young	High	Printer
8	Young	High	NB
9	Young	Average	Web-cam
10	Young	High	NB
11	Young	Average	NB
12	Old	Average	PC
13	Young	High	NB
14	Young	High	DVD
15	Old	Average	PC
16	Young	High	Printer
17	Old	High	Printer
18	Young	High	NB
19	Young	Average	DVD
20	Young	High	DVD

Table 1: Dataset of transactions, where items come from leaf-level of taxonomy

ID	Age	Income	Item	Item-ancestor	Item-ancestor
1	Old	High	NB	Computer	H/W
2	Old	High	PC	Computer	H/W
3	Old	High	NB	Computer	H/W
4	Young	High	DVD	Accessories	-
5	Old	Low	PC	Computer	H/W
6	Old	Average	NB	Computer	H/W
7	Young	High	Printer	Peripheral	H/W

8	Young	High	NB	Computer	H/W
9	Young	Average	Web-cam	Accessories	-
10	Young	High	NB	Computer	H/W
11	Young	Average	NB	Computer	H/W
12	Old	Average	PC	Computer	H/W
13	Young	High	NB	Computer	H/W
14	Young	High	DVD	Accessories	-
15	Old	Average	PC	Computer	H/W
16	Young	High	Printer	Peripheral	H/W
17	Old	High	Printer	Peripheral	H/W
18	Young	High	NB	Computer	H/W
19	Young	Average	DVD	Accessories	-
20	Young	High	DVD	Accessories	-

Table 2: Dataset of extended transactions for Table 1

4. Pruning Techniques

The association rules mining technique often generates huge number of association rules, which makes it very difficult to be analyzed in order to find interesting patterns / rules. Even though some of discovered rules have high support and confidence, but are not interesting. This is particularly true for the datasets whose attributes are highly correlated [7]. Therefore many pruning techniques have been developed and implemented to overcome this problem. The pruning of association rules is due to the presence of spurious and redundant rules that do not convey any additional information. Another reason for pruning is that the generated rules might be very specific (i.e., rules with many conditions) which tend to over fit the data without much predictive power compared to the general rules (i.e., rules with less number of conditions).

In this paper, we provide experimental results using some pruning techniques and statistical measurements that are template-based pruning, Chi square test, and R-interesting measure, to derive the most interesting generalized profile associations.

4.1 Template-based Pruning

Some of the discovered rules may hold uninteresting *itemsets*, or uninteresting combination of *itemsets*. One way of selecting interested rules is to specify template information in which we can explicitly specify what is interesting and what is not [8]. We implement the template-based pruning to

extract relevant set of association rules, because not all rules that passed the *minsupport* and *minconfidence* values are interesting. This technique filter the discovered association rules and selects the ones that match our template criteria, while rejecting the other rules.

We have specified some template information to extract the most interesting rules. Consider the following template, *Accept* “*antecedent = age or income*” and “*consequent = item*”, which means that accept all rules that have age or income attribute in their antecedent part, and item bought in their consequent part. Other rules that do not have this structure will be rejected.

4.2 Chi-square Test

Chi-Square is a well-known data measurement in the study of statistics [7] & [9]. We use chi-square (χ^2) test to measure the degree of independence between two different attributes by comparing their observed patterns of occurrence (actual support) with the expected pattern of occurrence (expected support). (χ^2) value is calculated as follows:

$$\chi^2 = \sum (f_o - f_e)^2 / f_e$$

Where f_o represents an observed frequency (actual support) and f_e is an expected frequency (expected support). The expected support is that all expected frequencies of the presence or absence of an item will be equal in a category or not in a category. For example, in a dataset of 20 customers who bought H/W or Accessories products, we discovered an association rule of “age=old 8 ==> item-Anc=H/W 8”, that is, out of 20 customers, 8 of old age customers tend to buy H/W products. From this information, we can calculate the expected support, as shown in Table 4, using a contingency table of the observed support illustrated in Table 3.

	item-Anc=H/W	item-Anc=accessories	Total
age=old	8	0	8
age=young	7	5	12
Total	15	5	20

Table 3: Contingency table for the observed support

The expected support is calculated as a product of corresponding row and column totals divided by the total number of elements in the lower right cell of the observed support table, that is $(8 * 15 / 20) = 6$. The expected support calculations are illustrated in Table 4.

	item-Anc= H/W	item-Anc= accessories	Total
age=old	6	2	8
age=young	9	3	12
Total	15	5	20

Table 4: Contingency table for expected support

We use the same technique explained above to find (χ^2) value for the 3 rules represented in Table 5. Rule with higher (χ^2) value is more interested, because it shows a higher correlation and dependency between its attributes.

4.3 R-interesting Measure

This technique uses the information provided in taxonomies to find the interesting rules among its ancestors, based on the assumption of statistical independence and strength of the latter rule. R is a threshold value specified by the user, a rule $X \Rightarrow Y$ is interesting if it passed the R -specified threshold with respect to its ancestors. This technique is based on the idea implemented by Srikanth in [3], and states that:

A rule ($X \Rightarrow Y$) is considered to be interesting in a given set of rules, if it has no ancestors or it is R -interesting with respect to its ancestor $X \Rightarrow \hat{Y}$. We call a rule $X \Rightarrow Y$ R -interesting with respect to an ancestor $X \Rightarrow \hat{Y}$, if the support of the rule $X \Rightarrow Y$ is R times the expected support based on $X \Rightarrow \hat{Y}$, or the confidence is R times the expected confidence based on $X \Rightarrow \hat{Y}$.

To explain the idea of R -interesting more clearly, consider the set of discovered association rules represented in Table 5, in which rule # 1 represents an ancestor of rule # (2 & 3), rule 2 represents an ancestor of rule # 3, while rule 3 is just a simple leaf-level rule. With R threshold ≥ 1.1 , Rule # 1 is

considered to be interesting compared to Rules 2 & 3 since it has no ancestors.

Rule 2 is also interesting, because its support and confidence were 1.17 times the expected support and expected confidence of rule 1. Rule 3 is considered to be redundant since its support and confidence were less than 1.1 times expected support and expected confidence of rule 2. Based on the information provided in Table 5, R -interesting measure prunes down much less interesting rules.

5. Generalized Association Rules Customer Model

Most of the data mining techniques are at the customer level, because customers are the main concern of any business. Instead of targeting all customers equally or providing the same incentive offers to every one, a firm can select only those customers who meet certain profitability criteria based on their needs and buying patterns. The key issue for building profiles for the customers or segmenting them is to obtain targeted customers and maximize their long-term revenue and loyalty [10].

Marketing managers are interested in rules like: 80% of male customers with high income tend to buy computer. These kinds of descriptions give them clear customer profiles on which to target their marketing actions. So, generalized profile association rules can provide a basis for better communicate with existing customers in order to offer better and tailored services. We build generalized profiles in set of conjunctive association rules. The conjunctive representation of the association rules is of the form $(A \wedge B \wedge C \Rightarrow W)$, where A, B, C are demographic attributes and $W \subseteq I$ (item bought).

No.	Rule	Support	Expected Support	Confidence	Expected confidence	Chi-square	$R \geq 1.1$
1	age=old 8 ==> item-Anc=H/W	8	6	1	0.75	4.4	-
2	age=old 8 ==> item- Anc=Computer	7	4.8	0.88	0.6	4.2	1.17
3	age=old 8 ==> item=PC	5	2	0.63	0.25	10	1.04

Table 5: Applying interestingness measures to select interesting rules

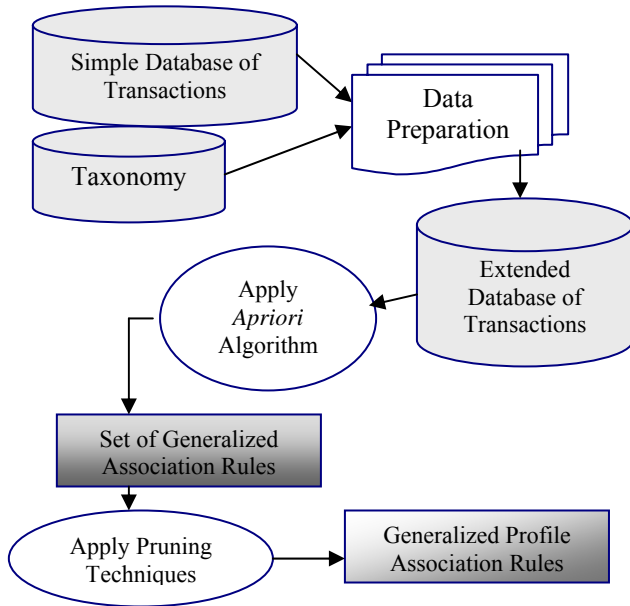


Figure 2: Conceptual flow of the process of discovering generalized profile association rules

Figure 2 shows the conceptual flow of our proposed model of generalized profile association rules, which works as follow:

1. Identify the variables for inclusion in the profile, such as, name, address, products bought, etc.
2. Prepare the dataset for mining process, by adding all ancestors of each item in a simple transaction to the transaction, and call this extended dataset.
3. Apply *Apriori* algorithm to generate all association rules that have *minsupport* and *minconfidence*.
4. Apply template-based and taxonomy-based pruning techniques to select the suitable general rules that describe the behavior of customers.
5. Using Chi-square statistical test to measure the degree of independence of between the attributes that forming the rule.
6. A significance test of *R*-interesting measure is used to select the most interesting rules among its ancestors.
7. The most interesting rules are used to construct generalized customer profiles, where the buying behavior is stored in the database along with the customer's factual / demographic data. Consequently this profiling information can be used to segmenting customers into groups based on their buying behavior and preferences.

5.1 Experimental Results and Discussion

Experiments are done with *Apriori* algorithm that is implemented by WEKA software [11]. WEKA

software uses a specific data format for handling datasets, that's ARFF (Attribute-Relation File Format). Experiments are performed on a small synthetic dataset of customer's transactions database (as appeared in Table 1 along with the taxonomy of Figure 1).

The data in Table 1 has been converted to an ARFF as follows:

```

@Relation simple;
@Attribute age {young, old}
@Attribute income {high, average, low}
@Attribute item {PC, NB, Printer, Scanner, DVD, Web-cam}
  
```

```

@Data
old, high, NB
old, average, PC
old, high, NB
young, high, DVD
old, low, PC
old, average, PC
young, high, Printer
young, high, NB
young, average, Web-cam
young, high, NB
young, average, NB
old, average, PC
young, high, NB
young, high, DVD
old, average, PC
young, high, Printer
...
  
```

Apriori Algorithm has been applied on the small datasets of Table 1 with *minsupport* threshold $\geq 10\%$ and *minconfidence* threshold $\geq 50\%$. When we applied the template-based pruning to select the interesting rules among the large number of discovered rules, such following rules have been obtained:

- age=old income=average 5 ==> product=PC 5 conf:(1)
- age=old 9 ==> product=PC 6 conf:(0.67)
- age=old income=high 3 ==> product=Printer 2 conf:(0.67)
- age=young income=high 10 ==> product=NB 6 conf:(0.6)
- age=young 11 ==> product=NB 6 conf:(0.55)

When we applied *Apriori* with the same support and confidence threshold values, on the extended dataset of Table 2, extra number of rules has been obtained which consists of the rules obtained from Table 1 plus more generalized rules as shown below:

- age=old 8 ==> item-Anc=H/W 8 conf:(1)
- age=old income=average 4 ==> item-Anc=Computer 4 conf:(1)

To select set of interesting rules among its generalized rules, we use the *R*-interesting

technique along with the fact that rules that do not reach the R value are considered less general with respect to its ancestor and hence redundant. From the taxonomy angle, a general rule is a rule that doesn't have any ancestors. This idea was explained clearly in section 4.

As a result of using template-based pruning technique along with the interestingness measures, we reduced the large amounts of the discovered rules to a small number of interested rules considered as generalized profiles for group of customers.

Examples of the generalized customer profiles that we have constructed are as follows:

- $\text{age}=\text{old} \wedge \text{income}=\text{high} \implies \text{product}=\text{H/W}$ (30%, 100%)
- $\text{income} = \text{average} \implies \text{item-Anc} = \text{Computer}$ (25%, 71%)

The above rules can be considered as generalized profiles for group of customers, where their buying behavior will be stored in the database along with their factual / demographic data. The LHS of each rule can be looked as group of customers (i.e., old age with high income customers, or customers with average income); these groups of customers can be used for targeted marketing campaigns or customer retention.

We can conclude that discovering rules from different levels of taxonomy is important, because when we restricted ourselves to the items at leaves of taxonomy, many rules have been undiscovered. The significant set of rules can be considered as generalized profiles for the customers, because they possess the *minsupport* and *minconfidence*, as well as they have no ancestors, and they passed the significant test value.

6. Conclusion

The concept of generalized profile association rules has been proposed so as to construct customer profiles, which are used to determine customer characteristics for segmenting them into various groups for targeted marketing and customer retention. We have constructed the generalized profile by combining flat dataset of customer's demographic data and hierarchical dataset of transactional data. Using the association rules discovery as one of the powerful data mining techniques, *Apriori* algorithm has been applied to discover set of association rules in the form of demographic data in the LHS, while the transactional hierarchical data in the RHS.

Pruning techniques and interestingness measures are used to reduce the number of the discovered rules and select the most interesting rules for

customer modeling, that is, the generalized profile association rules. Experiments are conducted for a small synthetic dataset, in order to show the effectiveness of the proposed scheme.

7. References

- [1] M. Chen, J. Han, and P. Yu. Data Mining: An Overview from Database Perspective. IEE Transactions on Knowledge and Data Engineering, Dec 1996, pp. 866-883.
- [2] H. Mannila. Data Mining: machine learning, statistics, and databases. IEEE Transactions on Knowledge and Data Engineering, June 1996, pp. 2-9.
- [3] R. Srikant and R. Agrawal. Mining Generalized Association Rules. In Proc. of the 21st Int Conference on Very Large Databases. Switzerland. Sep 1995, vol 13(2-3), pp. 161-180.
- [4] F. E. Giha, Y. P. Singh, & H. T. Ewe. Customer Profiling using Data Mining Techniques for e-Commerce Applications. In Proceedings of the National Seminar on E-Commerce, UNITAR, Malaysia, 2003a, pp. 16 - 21.
- [5] F. E. Giha, Y. P. Singh, & H. T. Ewe. Customer Profiling and Segmentation based on Association Rules Mining Technique. In Proceedings of the 7th IASTED International Conference on Software Engineering and Applications, USA, 2003b, pp. 37-42.
- [6] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In Proceedings of 20th International Conference on Very Large Databases, 1994, pp. 487 - 499.
- [7] B. Liu, W. Hsu, and Y. Ma. Pruning and Summarizing the Discovered Associations. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD-99), Aug 1999, San Diego, USA, pp. 125-134.
- [8] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila. Pruning and Grouping of Discovered Association Rules. ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases, April 1995, pp. 47-52.
- [9] C. Silverstein, S. Brin, & R. Motwani. Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. Data Mining and Knowledge Discovery, 1998, vol. 2(1), 39-68.
- [10] C. Bounsaythip and E. Rinta-Runsala. Overview of Data Mining for Customer Behavior Modeling. VTT Information Technology, June 2001.
- [11] L. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Academic Press, 2000.