Cybercrime Profiling: Decision-Tree Induction, Examining Perceptions of Internet Risk and Cybercrime Victimisation

Ameer Al-Nemrat
ACE, UEL, United Kingdom
Email: ameer@uel.ac.uk

Chafika Benzaïd
Division Sécurité Informatique, CERIST, Algérie
cbenzaid@cerist.dz

***Abstract*—The Internet can be a double-edged sword. While offering a range of benefits, it also provides an opportunity for criminals to extend their work to areas previously unimagined. Every country faces the same challenges regarding the fight against cybercrime and how to effectively promote security for its citizens and organisations. The main aim of this study is to introduce and apply a data-mining technique (decision-tree) to cybercrime profiling. This paper also aims to draw attention to the growing number of cybercrime victims, and the relationship between online behaviour and computer victimisation. This study used secondhand data collected for a study was carried out using Jordan a s a case study to investigate whether or not individuals effectively protect themselves against cybercrime, and to examine how perception of law influences actions towards incidents of cybercrime. In Jordan, cybercafés have become culturally acceptable alternatives for individuals wishing to access the Internet in private, away from the prying eyes of society.**

***Keywords*-Digital forensics; Cybercrime profiling; Data mining; Classification tree.**

I. Introduction

[1] argues that the anonymity of the online world leads to feelings of unconstraint and individuals are more likely to commit crime, as they feel they can deviate with impunity. [2] argues if this were the case, much more deviant activity could be predicted. Either way, [2] maintains that there is one current certainty; cybercrime is becoming increasingly more global."cybercrime is more than a globalised phenomenon that can be committed anywhere on the Internet, from anywhere, at anytime, but is constituted of ideas that transcend cultural and geographical boundaries". Therefore, every country faces the same dilemma of how to fight cybercrime and how to effectively promote security to their citizens and organisations.

In the analysis, Data-mining took place. One of the most popular methods of representing information produced from data-mining is decision-tree induction, as this offers a clear framework for learning and reasoning from feature-based examples. Decision-trees help ascertain which course of action to take or classification to choose, and the rewards and risks of each choice. A decision-tree is considered to be a highly effective machine learning method used to generate classification models. "A decision tree is a flow-chart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution" [3]. It is therefore possible for the outcomes of various options to be explored. In instances such as the research for this paper, where resources are limited, decision-trees provide a useful means of choosing from different options, strategies, or projects. They have been studied and thoroughly tested as a method for presenting information in areas of both pattern recognition and machine learning.

A pattern is identified by breaking down decision-making into stages. It is first of all tested against the characteristic specified at the root point of the decision tree. Once the pattern has been tested against this attribute, or it is apparent that the question asked at the base is answerable using the given pattern, the tree offers several branches corresponding to the value of the attribute in the given example. After moving down the appropriate branch a subtree (or *child node*) is reached, where the initial process must begin again. This process is repeated for all subsequent sub-trees reached until a 'leaf' is encountered, at which point time the pattern is classified within the class as belonging to the class named by that leaf.

This study proposes that the first line of defence against cybercrime is personal awareness and public education of cybercrime methods and measures. Some experts might argue that technology is the first line of defence. This study demonstrates that personal awareness is essential if technology is to be effectively used as a defence tool. Technological advancement and weaknesses in the legal systems render many countries unable to cope with the rapid changes, highlighting the importance of self-protection against cybercrime. Underlying much of the work required to reduce e-crime is the importance of educating individuals regarding potential risk and to highlight personal responsibility for future crime prevention.

The paper is organized as follows. In Sections II and III a detailed description is given of the hypothesis and data collection procedure adopted. Section IV looks at the introduced algorithm used for the imputation of incomplete data. Section V describes the test used to measure the association between variables. The conducted experimental

and the obtained results are presented in Section VI and Section VII, respectively. Section VIII discusses the obtained results. Finally, Section IX concludes the paper.

II. Hypothesis

This paper proposed that Internet users who frequent cybercafs are more inclined to practice risky online behaviour. due to fundamental lack of awareness of computer-associated risks and failure to take precautionary measures, these Internet users consequently more likely to be victims of cybercrime.

III. Data Collection and Survey Procedures

133 participants who frequent cybercafs in Amman (Jordan) were interviewed to determine their views regarding the following:

- Factors which might influence online precautionary behaviour in cybercafs;
- Attitudes and behaviour towards supporting or resisting cybercrime victimisation.

This research was conducted over two, two-week periods in July 2009. cybercafés were selected based upon their central locations and apparent popularity, as they open 24/7. Participants were randomly asked if they like to participate in the survey.

The questions posed in the survey were designed to assess whether or not Internet users consider the various types of prohibited online activity to be acceptable or deviant.

The questionnaire consisted of 34 multiple choice questions (see appendix):

- Questions $(1-3)$ were demographic;
- Questions $(4-9)$ assessed cybercafé behaviour;
- Questions $(10-11)$ and $(17-21)$ assessed the strength of current guardianship tools against risks from the Internet;
- Question (12) recorded what type of cybercrime (if any) had taken place;
- Questions $(13-16)$ recorded responses to having been a victim of cybercrime;
- Questions $(22-26)$ assessed the perceived levels of morality associated with Internet usage;
- Questions $(27-34)$ assessed the perceived level of appropriate punishment for cyber criminals.

The aim of each of the six cybercrime categories listed in Question 12 (including the possibility of no cybercrime having occurred) was to perform statistical association tests between the cybercrimes listed and the data provided by Questions 1 to 34 (excluding Question 12).

Associations found between cybercrime categories and the questions asked according to the classification-tree induction are as follows:

(a) Internet fraud and Questions $(8, 28, 33)$.
(b) Identity theft and Questions $(11, 22, 23, 26)$.
(c) Hacking and Questions $(11, 21, 33)$.
(d) Online stalking and Question (11).
(e) Cyber-harassment and Questions $(5, 21, 34)$.
(f) Other types of cybercrime ¡None¿
(g) Absence of cybercrime and Questions $(2, 22)$.

In this research, in order to better understand the relationship between the $Q12$ (Have you ever been the victim of one of these types of cybercrime?)

- Internet Fraud
- Identity Theft
- Hacking
- Online Stalking
- Viruses
- Phishing

Cybercrime types and the rest of the questionnaire' observed variables classification-tree induction was used.

IV. Handling Incomplete Data

Non-response to questions is a common problem with questionnaires, and the cybercrime questionnaire was no exception. This resulted in the corresponding dataset being incomplete (Table I):

Questionnaire question	data (n=133)
Q1	0
Q2	0
Q3	0
Q4	0
Q5	0
Q6	1
Q7	1
Q8	2
Q9	0
Q10	1
Q11	0
Q13	15
Q14	37
Q15	39
Q16	1
Q17	0
Q18	0

Questionnaire question	data (n=133)
Q19	0
Q20	0
Q21	1
Q22	3
Q23	2
Q24	1
Q25	1
Q26	1
Q27	0
Q28	0
Q29	1
Q30	1
Q31	0
Q32	5
Q33	0
Q34	1

Table I
Frequencies of missing values in the data sets

One approach to handling incomplete data is to only use fully completed cases, but there are two potential problems with doing this, outlined as follows:

- Deletion of cases will reduce the sample size. This, in its turn, will increase the probability of obtaining non-significant results when the null hypothesis is false. In other words, the power of Fisher's exact test will be reduced.
- If absent data is completely down to chance, i.e. the missing answers were excluded for no reason/at random, then forming a sample out of the completed cases alone may still give an accurate representation of the associated population. On the other hand, if missingness (the manner in which data is missing from a sample) is not an entirely random process, then the complete cases can constitute a biased sample [4], [5], [6].

Imputation refers to the replacement of missing values with estimated values. In the case of categorical data, such as the values of the cybercrime datasets, a simple approach is to replace the missing values with either random values or

with the mode corresponding to the given results for each question; however, such an approach would not take into account the possibility that missingness could be dependent on one or more values present in other features, a not uncommon scenario with questionnaires [5]. A number of effective approaches to imputation have been suggested (such as by [6]; and [5]). One of the more recent proposals comes from [7], who suggests using a combination of additive regression [8] and bootstrapping [9]. The following algorithm (Algorithm 1) was used for the imputation of the incomplete data and is adapted from Harrell [7].

Algorithm 1

1: **for each** variable x of S with missing values **do**
2: fill the missing values in x with a random sample without replacement from the non-missing values in x;
3: **end for**
4: $i = 0$;
5: **repeat**
6: **for each** x originally containing missing values **do**
7: horizontally partition S into subsets S_1 (value for x missing) and S_2 (value for x present);
8: draw random sample S_3 from S_2 with replacement such that $|S_3| = |S_2|$;
9: use S_3 to fit a flexible additive regression model M to predict x via the $areg$ function in the R package [10];
10: use M to predict x in S_1;
11: replace predicted value of x with nearest permissible value for x;
12: **end for**
13: $i = i + 1$;
14: **until** $(i > 8)$

This approach to imputation was implemented via the *aregImpute R* algorithm function [11], [7].

V. Test of Associations

A standard test for measuring association between two discrete-valued variables is Pearson's Chi-square test with Yates's correction. However, this test assumes that 80% of the expected values will be greater than 5. Because this condition will not necessarily be fulfilled in every case for the cybercrime data, Fisher's exact test was used.

Fisher's exact test [12] (pp. 188-189) is able to test for association without the restrictions imposed by the chi-squared test. The statistical null hypothesis for the test is that there is no association between two discrete-valued. Because of the complexity of determining the p-values of Fisher's test when the number of rows or columns is greater than 2, the p-values were estimated using 2000 Monte Carlo simulations [13], though the use of the *Fisher.test* function within the *R* (statistical package) [14].

A. Multivariate Analysis Using Classification-Tree Induction

In the univariate analysis used to examine missing data, each test involved only a single predictor variable (covariate). Therefore, a better understanding of the relationships that exist between types of cybercrime and the other variables used in the questionnaire may be achieved using multivariate statistical techniques, such as stepwise logistic regression [15] and classification-tree induction [8], [16], [17]. Given the interpretability of classification-trees, this approach was applied to the data using the Classification And Regression Trees (CART) approach to tree induction [16].

Theory: Let S be a data set of feature vectors x that are vector points within a feature space Φ. Each feature vector is labelled with the class k to which it belongs; consequently, S gives rise to a set of $|S|$ class-labelled point within Φ.

Let R be a region of Φ that contains S. The mixture of class-labelled points in R gives rise to class heterogeneity within R. This heterogeneity can be measured by the *Gini index*,

$$\sum_{k=1}^{K} p_{mk}(1 - p_{mk}) \tag{1}$$

where p_{mk} is the proportion of class k vectors in region R_m. It might be possible to partition R_m into two regions, R_{m1} and R_{m2}, so that the average class heterogeneity in R_{m1} and R_{m2} is less than that in R_m. The optimal partition of R_m is the combination of the feature $x \in x$ and the partition of that feature results in the maximal decrease in class heterogeneity. This process can be repeated for R_{m1} and R_{m2}. By continuing in this manner, R will be recursively portioned into several regions (Figure 1).

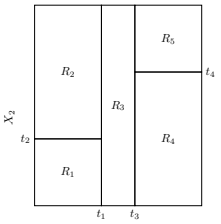

Figure 1. A recursive partitioning of R into five regions based on finding the optimal partition at each step (taken from [8]).

Each partition of a region corresponds to a binary split of the leaf nodes of a tree, with the root node of the tree corresponding to the initially unpartitioned region R. Consequently, the recursive partitioning of R corresponds to the growth of a tree (Figure 2). Such a tree is a classification tree: a new vector $\dot{x}$ is classified by "dropping it down" the tree, seeing which leaf node it reaches, and assigning $\dot{x}$ to the majority class in the region associated with the leaf node.

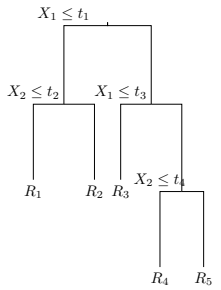

Figure 2. The classification tree corresponding to the partitioning shown in Figure 1 (taken from [8])

A fully grown classification tree T, which has branched

from S, it is unlikely to perform the same upon another data set as it has done upon S because of over-fitting (e.g., [18], pp 6-12). On the other hand, a rooted subtree of T is likely to perform better with respect to another data set. T, therefore, needs to be pruned back.

A common approach to pruning a classification tree is to use the cost-complexity measure of T, $mis_\alpha(T)$, with respect to a complexity parameter α [16]:

$$mis_\alpha(T) = mis(T) + \alpha|\tilde{T}| \quad (2)$$

Here, $mis(T)$ is the misclassification rate for T with respect to S, and $|\tilde{T}|$ is the number of leaf nodes in T (a measure of the size of T). As T increases, $|\tilde{T}|$ increases but $mis(T)$ decreases; therefore, there is a subtree T_α of T that minimises $mis_\alpha(T)$ for a given α. By repeating this process using cross-validation (e.g., [18], pp 32-33) to test successive values of α, the overall optimal pruned tree can be determined.

VI. Experimental

The theory illustrated above was applied to each of the seven cybercrime categories present in the questionnaire set (including the case of 'no cybercrime'). The software used for tree induction was the *rpart* library from the *R* statistical package, which is based on [16]. The default settings of the *rpart* library were used, except that *50-fold* cross-validation was used instead of *10-fold* cross-validation.

VII. Results

The data produced six trees. There were no trees for unspecified cybercrimes extracted from the data. The resulting trees (the output from R) are shown in Figure 3 to Figure 8, and each tree/figure is followed by the corresponding interpretations of it. It is worth mentioning here how each tree represents data. Each node shows the majority class (i.e. 1 if the cybercrime category of interest is present; 0 if it is absent) and the frequency distribution of the classes at the node is shown as: $< frequencyofClass0 > / < frequencyofClass1 >$. It is also important to clarify that Q in all trees refers to the corresponding question from the survey, for example $Q8$ is '*Question 8*'. Below are the trees and their corresponding interpretations:

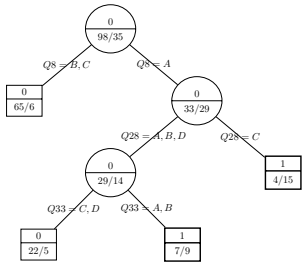

Figure 3. Tree 1: Classification tree for Internet fraud with respect to Jordanian data

Tree 1:
A Jordanian user is most likely to be a victim of Internet fraud;
if(they mostly use the Internet between Monday and Thursday (Q8),
and they believe that the punishment should be only somewhat severe if they are caught with destructive malware (Q28));
or
if(they mostly use the Internet between Monday to Thursday (Q8),
and they do not believe that punishment should be only somewhat severe if they are caught with destructive malware (Q28),
and they agree that the justice system treats computer crime as seriously as street crime (Q33))

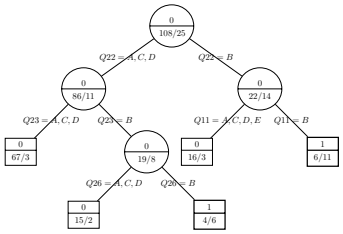

Figure 4. Tree 2: Classification tree for identity theft with respect to Jordanian data

Tree 2:
A Jordanian user is most likely to be a victim of identity theft;
if(they either do approve or strongly disapprove of pirated software (Q22),
and they do sometimes disapprove of destructive malware (Q23),
and they do sometimes disapprove of using a device to obtain free wireless or phone connections (Q26));
or
if(they do sometimes disapprove of pirated software (Q22),
and they mostly fear identity theft (Q11)).

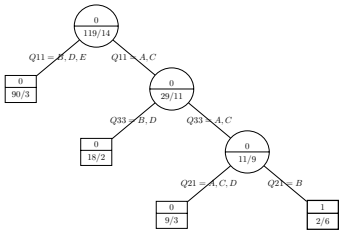

Figure 5. Tree 3: Classification tree for computer penetration with respect to Jordanian data

Tree 3:
A Jordanian user is most likely to be a victim of hacking (and other forms of computer penetration)
if(they mostly fear Internet fraud or computer penetration (Q11),
and they have a strong opinion about whether the justice system treats computer crime as seriously as street crime (Q33),
and they strongly agree that they are more comfortable using an Internet café when visiting unknown websites than when using their own computer (Q21)).

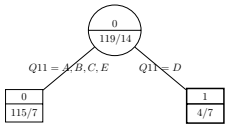

Figure 6. Tree 4: Classification tree for online stalking with respect to Jordanian data

Tree 4:
A Jordanian user is most likely to be a victim of online stalking;
if(they mostly fear online stalking (Q11).

Tree 5:
A Jordanian user is most likely to be a victim of cyber-harassment;

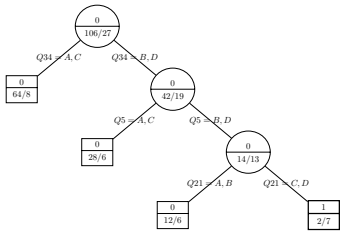

Figure 7. Tree 5: Classification tree for cyber-harassment with respect to Jordanian data

if(they do not have a strong opinion about whether cyber-criminals are as dangerous as street criminals (Q34),
and they use an Internet café once a week or three times a week (Q5),
and they at least agree with the statement that they are more comfortable using an Internet café when visiting unknown websites than when using their own computer (Q21)).

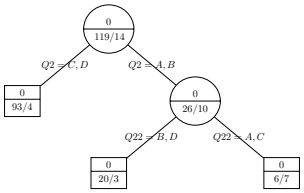

Figure 8. Tree 6: Classification tree for absence of cybercrime with respect to Jordanian data

Tree 6:
A Jordanian user is most likely not to be a victim of cybercrime;
if(they are less than 22 years old (Q2),
and they have a strong opinion about the usage, creation or distribution of pirated software (Q22)textbf).

Result of the imputed data is, as Table II shows, the associations indicated by the imputed data.

	Internet fraud		**Identity theft**		**Computer penetration**		**Online stalking**		**Cyber-harassment**		**Other**		**None**	
	F	T	F	T	F	T	F	T	F	T	F	T	F	T
Demographics	•		•										•	•
Internet café behaviour	•	•	•		•		•		•	•			•	
Guardianship against risks from Internet	•		•	•	•	•	•	•	•					
Response to victimisation										•				
Moral associations with Internet usage	•		•	•									•	•
Level of punishment considered appropriate for cybercrimes	•	•			•	•			•	•	•		•	

Table II
ASSOCIATIONS BETWEEN CYBERCRIME CATEGORIES AND SIX BROAD GROUPS OF USER FEATURES, BASED ON JORDANIAN DATA.
F = SIGNIFICANT WITH RESPECT TO FISHER'S EXACT TEST; T = SIGNIFICANT WITH RESPECT TO CLASSIFICATION-TREE INDUCTION.

There are a number of directions that can be taken with regards to data analysis. Firstly, there is the question of the sensitivity of the results to the imputation used. This could be checked by repeating Algorithm 1 using several different random seeds and observing the variability of the results, a process known as multiple imputation.

CART-based tree induction was used. However, the CART algorithm grows trees in a greedy manner. A second alternative approach is therefore to perform tree induction using a genetic algorithm([19]), permitting a more extensive search for optimal trees.

A third alternative multivariate technique is logistic regression with multiplicative interaction terms([15]):

$$\ln\left(\frac{p(y=1|x_1,\ldots,x_q)}{p(y=0|x_1,\ldots,x_q)}\right) = \beta_0 + \sum_{i=1}^{q}\beta_i x_i + \sum_{i\neq j}\beta_{i,j}x_i x_j \quad (3)$$

In the above equation (Equation(3)), response variable y corresponds to a cybercrime category, $x_1,\ldots,x_q$ are covariates (including binary dummy variables) associated with questionnaire questions Q1 - Q11 and Q16 - Q34, and β_0, β_i, and $\beta_{i,j}$ are regression coefficients. Given that tree induction and logistic regression can complement each other([20]), the use of logistic regression may reveal some new associations.

VIII. DISCUSSIONS

Upon the interpretations of trees 3 to 8 (above), it becomes apparent that cybercafé users who have been victims of different cybercrime types have opposing views about different kinds of online behaviours. For instance, the victims of internet fraud agreed that the punishment should be severe for individuals caught with destructive malware. On the other hand, victims believe that the justice systems treat computer crime as seriously as street crime. However, the agreement to this statement by cybercafé users seems to be inconsistent, because there is no existing law related to cybercrime in Jordan. The only explanation for this finding is that traditional Jordanian law regarding fraud is used in some Internet fraud cases, where no distinction is made between digital and physical fraud. Furthermore, the tree 3 indicates that cybercafé users who have experienced any form of hacking strongly agree that they are more comfortable using an Internet Café when visiting unknown websites than when using their own computer.This is because they might feel safer using a public computer, as having their public computer hacked is less problematic than their own, thus they are more cautious. Therefore, a cybercafé is an environment where a high proportion of "unawareness" issues and behaviours associated with online activities are found. Research also established that in these shared environments, Internet users are likely to be going online to communicate with friends and family, and engaging in online banking and money transfers. This makes Internet Cafés a key target for cyber criminals, who rely on the lack of awareness of the cafés users, especially for client-side caching. *Client-side caching* involves the temporary storage of copies of web pages by web browsing software on the hard drive of a personal computer. All commonly, used web browsers employ this technique, for example, it enables the use of

a browser's "back button". It also saves the return to the source of a previously downloaded web page when the page remains unchanged.A security problem arises when personal information cached by a web browser remains at the end of the user's session. Subsequent users may be able to navigate to pages stored in the browser cache and access this information.

IX. Conclusion

The purpose of this study was to introduce a new approach, Decision Tree, as a data-mining technique to be used in cybercrime profiling. The purpose of this study was to address this gap in the literature, by presenting the first systematic study that questioned cybercrime victimization among cybercafé users. According to results obtained from this study, online users who visit cybercafés are more inclined to engage in risky online behaviour. They lack the level of risk awareness associated with taking precautionary measures, so they are more likely to be victims of cybercrime.These findings were assessed using the data-mining technique of classification trees which, As far as could be ascertained during the research process is the first to be used in cybercrime profiling.

References

[1] K. S. Williams, "Using tittle's control balance theory to understand computer crime and deviance," *International Review of Law Computers and Technology*, vol. 22, no. 1, pp. 145–155, 2008.

[2] D. Wall, *Hunting Shooting, and Phishing: New Cybercrime Challenges for Cybercanadians in The 21st Century*. The ECCLES Centre for American Studies, British Library, 2007.

[3] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

[4] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, Dec. 1976.

[5] ——, *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York, USA, 1987.

[6] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London, 1997.

[7] F. E. Harrell, "Multiple imputation using additive regression, bootstrapping, and predictive mean matching [online]," *Available: http://lib.stat.cmu.edu/S/Harrell/help/Hmisc/html/aregImpute.html [Accessed 7 December 2009].*, 2007.

[8] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York, USA, 2001.

[9] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[10] F. E. Harrell, "Additive regression with optimal transformations on both sides using canonical variates [online]," *Available: http://sekhon.berkeley.edu/library/Hmisc/html/areg.html [Accessed 7 December 2009]*, 2007.

[11] ——, "Impute: A predictive mean matching multiple imputation strategy [online]," *Available: http://lists.utsouthwestern.edu/pipermail/impute/2001-November/000156.html [Accessed 7 December 2009]*, 2001.

[12] W. J. Conover, *Practical Nonparametric Statistics*. John Wiley and Sons, New York, USA, 1999.

[13] W. M. Patefield, "Algorithm as159. an efficient method of generating r x c tables with given row and column totals," *Applied Statistics*, vol. 4, p. 9197, 1981.

[14] R. D. C. Team, *R: A language and environment for statistical computing (Version 2.10.0)*. R Foundation for Statistical Computing, Vienna, Austria, 2009.

[15] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. Wiley Interscience Publication, 1989.

[16] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Chapman and Hall/CRC, New York, USA, 1984.

[17] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[19] D. R. Carvalho and A. A. Freitas, "A hybrid decision tree/genetic algorithm method for data mining," *Information Sciences*, vol. 163, pp. 13–35, 2004.

[20] C. Perlich, F. Provost, and J. S. Simonoff, "Tree induction vs. logistic regression: a learning-curve analysis," *The Journal of Machine Learning Research*, vol. 4, pp. 211–255, 2003.