GENERAL APPROACH FOR CHURN PREDICTION WITH GENETIC ALGORITHM OPTIMIZED K-NEAREST NEIGHBOR FRAMEWORK

Tennakoon Mudiyanselage Nipunika Priyadarshani Tennakoon

(118231G)

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science

Department of Computer Science & Engineering

University of Moratuwa
Sri Lanka

June 2015

Declaration

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

Name:

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

The above candidate has carried out research for the Masters Dissertation under my supervision.

Signature of the supervisor: Date:

Name of the supervisor:

Abstract

Customer churn has become one of the most significant topics in today's business. It has become a major challenge for a business with the evolving market and low barriers to switch between the service providers. It has identified that, retaining the old customers is more profitable for a company than acquiring new customers. That motivates business personnel for churn prediction. Service providers can get necessary measures to retain their customers if they could gain prior knowledge on the probable churns in the customer base. But, churn prediction is considered a difficult task.

Various attempts have been made in predicting churn and churn related information. Different data mining techniques had been used in developing churn models. Regression analysis, decision tree based methods and neural network based methods were among the most commonly used techniques. The most successful models suffered from low interpretability which is a main consideration in a churn model while some of the models were domain specific.

K nearest neighbor classifier is one of the best algorithms to be used in classifications. But it has been rarely used in churn prediction. Genetic algorithms are considered an optimization technique which could be used in optimizing performance of other algorithms. Genetic Algorithm Optimized K Nearest Neighbor (gaKnn) is a framework that has tested for its high accuracy. Hence, we developed a Tool based on the gaKnn framework which could be used for churn prediction. We also incorporated two voting mechanisms; Bayesian weights and class confidence weights (*ccw*) to weight the prediction in order to address misclassification issues occur due to class skew.

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Acknowledgements

I sincerely acknowledge the work carried out by the WEKA community and publishing it as an open source code.

I would like to dedicate my sincere thank to my supervisor Dr. Amal Shehan Perera, Senior Lecturer, Department of Computer Science and Engineering for his dedicated support for the success of this research. This would not have become a success without your support from the initial stage of the research.

I would like to thank the entire academic and the non academic staff of the Department of Computer Science and Engineering for their kindness extended to me in every aspect.

Last but not least, I thank my family and all my friends who supported me for the success of this piece of work. Your support was so precious.

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Table of Contents

Declaration ii

Abstract iii

Acknowledgements iv

List of abbreviations... viii

Table of figures ix

List of tables..x

CHAPTER 1 – INTRODUCTION 1

1. Introduction 1

1.1. Background for Churn Prediction 1

1.2. Objectives 2

1.3. Statement of the Problem...3

1.4. About the gaKnn Churn Prediction Tool 4

CHAPTER 2- LITERATURE REVIEW 6

2. Literature Review 6

2.1. Churn Prediction...6

2.1.1. Past attempts on churn prediction 7

2.1.2. Analyzing existing churn models..10

2.2. Analyzing the Data Mining Models Related to the Framework......................12

2.2.1. K nearest neighbor algorithm (KNN) 12

2.2.2. Genetic algorithm (GA) 14

2.2.3. Genetic algorithm as an optimizing technique..................................15

2.2.4. Genetic algorithm optimized k nearest neighbor (gaKnn).........................17

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

2.3. Genetic Algorithm Optimized K-Nearest Neighbor Classification Framework 18

2.4. Data Pre-Processing 21

2.5. Classification vs. Class Skew 22

2.5.1. Mixture models 27

2.5.2. Expectation maximization (EM) algorithm 28

2.5.3. Bayesian networks 28

2.5.4. K2 algorithm 28

2.6. Classifier Evaluation 30

2.6.1. ROC curve/ AUC 31

2.6.2. Lift charts 31

2.6.3. Precision recall (PR) curve...32

CHAPTER 3 – DESIGN 33

3 Design 33

3.1 Overview of Design..33

3.2 Pre-Processing Steps 35

3.2.1. Outlier handling 35

3.2.2. Missing value handling 38

3.2.3. Data discretization..38

3.3 Distance Function and Similarity Calculation...39

3.4 Data Skew..39

3.4.1. Class confidence weights 40

3.4.2. Naïve Bayesian probability weights..43

3.5 Testing and Validation Options..44

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

3.6 Graphical User Interface (GUI) 45

3.7 Saving the Model for Re-Use 47

CHAPTER 4- IMPLEMENTATION 48

4 Implementation 48

4.1 Implementation Environment and Process 48

4.1.1. R language 48

4.1.2. Process implementation 49

CHAPTER 5- TESING AND EVALUATION 56

5 Testing and Evaluation 56

5.1 Testing the Tool 56

5.1.1 Test goals 56

5.1.2 Testing process and results 57

5.2 Evaluation 67

CHAPTER 6- CONCLUSION AND FUTURE WORKS 75

6 Conclusion and Future Works 75

6.1. Conclusion 75

6.2. Future Works 76

REFERENCES 77

APPENDIX A – User Manual 81

APPENDIX B – Source Code 90

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

List of abbreviations

Abbreviation	Description
AUC	Area Under Curve
CCW	Class Confidence Weight
CDR	Call Detail Record
CSV	Comma Separated Values
DMEL	Data Mining by Evolutionary Learning
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
gaKnn	Genetic Algorithm Optimized K nearest neighbor
GUI	Graphical User Interface
JGAP	Java Genetic Algorithms Package
KNN	K nearest Neighbor
NN	Neural Network
PR	Precision- recall
ROC	Receiver Operating Characteristic
TN	True Negative
TP	True Positive

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Table of figures

Figure 2.1-1 Churn prediction process [21] 7

Figure 2.2-1 K nearest neighbor 13

Figure 2.3-1 Overall design of the framework 20

Figure 2.5-1 Pseudo Code for K2 algorithm [29] 29

Figure 3.1-1 Design overview of the Tool 34

Figure 3.6-1Activity flow 46

Figure 4.1-1 Graphical View of Data 51

Figure 4.1-2 Interface to feed data 52

Figure 4.1-3 Data Preprocessing Window 52

Figure 4.1-4 Visualize Results 53

Figure 4.1-5 Prediction window 53

Figure 4.1-6 Churn Results 54

Figure 4.1-7 Accuracy Measurements 54

Figure 5.2-1 AUC for different weighted votes 69

Figure 5.2-2 Change of Lift value 69

Figure 5.2-3 Change of Recall 70

Figure 5.2-4 Performance - Dataset 1 73

Figure 5.2-5 Performance – Dataset 3 73

Figure A-1 First Interface 82

Figure A-2 Data selection 82

Figure A-3 Data pre-processor 83

Figure A-4 Predictor 85

Figure A-5 Result analyzer 87

Figure A-6 Fitness distribution 88

Figure A-7 Model accuracy 89

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

List of tables

Table 5.1-1Tested Data 58

Table 5.1-2 Results for sample drawn from dataset 1 59

Table 5.1-3 Results for sample drawn from dataset 2 59

Table 5.1-4 Results for sample drawn from dataset 3 60

Table 5.1-5 Comparison between different training options 61

Table 5.1-6 Comparison on different test options 62

Table 5.1-7 Comparison outlier detection/handling methods- with Tool 63

Table 5.1-8 Comparison outlier detection/handling methods- with WEKA 64

Table 5.1-9 Results of training phase 64

Table 5.1-10 Results of evaluation with test data 65

Table 5.1-11 Results of training phase 65

Table 5.1-12 Evaluated with test data 66

Table 5.1-13 Results for missing value handled data 66

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk