REFERENCES

[1] T. Au *et al.*, “Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques”, *Journal of Comparative International Management*, Vol. 6, No: 1, Jan. 2003. ISSN 1718-0864, [Online]. Available: http://www.lib.unb.ca/Texts/JCIM/bin/get.cgi?directory=vol6_1/&filename=li.html. [Accessed: 04-Oct-2011].

[2] Y. Richter *et al.*, “Predicting customer churn in mobile networks through analysis of social groups,” in *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM 2010)*, 2010.

[3] L. Junxiang, “Predicting Customer Churn in the Telecommunications Industry – – An Application of Survival Analysis Modeling Using SAS®,” in *SAS User Group International online proc.*, 2002, vol. Paper No.114–27.

[4] W. Au *et al.*, “A novel evolutionary data mining algorithm with applications to churn prediction,” presented at the IEEE Trans. Evolutionary Computation, 2003, pp. 532–545.

[5] V. Lazarov and M. Capota, “Churn Prediction,” Business Analytics Course, TUM Computer Science, 2007.

[6] S. V. Nath and R. S. Behara, “Customer Churn Analysis in the Wireless Industry: A Data Mining Approach,” in *Annual Meeting of the Decision Sciences Instituite*, 2003, pp. 505–510.

[7] “Eleven Ways to Reduce Telecom Churn.” [Online]. Available: http://www.tmcnet.com/usubmit/2008/01/29/3237095.htm. [Accessed: 04-Oct-2011].

[8] C. P. Parag, “Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services,” *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 6714–6720, Apr. 2009.

[9] “Outsourcing Customer Churn Analysis:: Customer Churn Analysis White Papers.” [Online]. Available: http://www.marketequations.com/white-papers/customer-churn-analysis.html. [Accessed: 31-Oct-2011].

[10] R. J. Jadhav and U. T. Pawar, “Churn Prediction in Telecommunication Using Data Mining Technology,” *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 2, No 2, pp. 17–19, Feb. 2011.

[11] T. Rashid, “Classification of Churn and non-Churn Customers for Telecommunication Companies,” *International Journal of Biometrics and Bioinformatics (IJBB)*, vol. 3, no. 5, p. 82, 2009.

[12] J. Han and M. Kamber, “Classification & Prediction,” in *Data Mining: Concepts & Techniques*, 2nd ed., Morgan Kaufmann Publishers, 2006, pp. 347– 355.

[13] N. Suguna and K. Thanushkodi, “An improved k-nearest neighbor classification using genetic algorithm,” *IJCSI*, p. 18, 2010.

[14] H. Ahn *et al.*, “Global optimization of feature weights and the number of neighbors that combine in a case-based reasoning system,” *Expert Systems*, vol. 23, no. 5, pp. 290–301, 2006.

[15] M. Analoui and M. F. Amiri, “Feature reduction of nearest neighbor classifiers using genetic algorithm,” *Proceedings of world academy of science, engineering and tehchnology*, 2006.

[16] D. G. N. Dayaratne, “GeneticAlgorithm Optimized K-Nearest Neighbor Classification Framework (gaKnn)”, M.Sc. Thesis, Dept. of Computer Science & Engineering, Univ. of Moratuwa, Sri Lanka, 2008.

[17] A. S. Perera *et al.*, “Evolutionary Nearest Neighbour Classification Framework,” presented at the 18th International Conference on Software Engineering and Data Engineering, 2009, pp. 250–255.

[18] “JGAP: Java Genetic Algorithms Package.” [Online]. Available: http://jgap.sourceforge.net/. [Accessed: 12-Jan-2012].

[19] “SGI - MLC++: Datasets from UCI.” [Online]. Available: http://www.sgi.com/tech/mlc/db/. [Accessed: 26-Aug-2013].

[20] Mark Hall *et al.*, "The WEKA Data Mining Software: An Update; SIGKDD Explorations", Volume 11, Issue 1, 2009.

[21] “Mini Lecture: Churn Prediction: Analysis and Applications - YouTube.” [Online]. Available: https://www.youtube.com/watch?v=6yCxbzDjBDc. [Accessed: 22-Feb-2015].

[22] “Turn data into revenue, faster!:Criticality of Predicting Customer Churn.” [Online]. Available: http://www.infosysblogs.com/bigdata/2013/12/_my_personal_experience_of.html. [Accessed: 22-Feb-2015].

[23] B.W. Yap *et al.*, " An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets", *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Herawan,

T;Deris, M.M.; Abawajy, J. (Eds.) 2014,XXI, 730p.235 illus., Hardcover, ISBN: 978-981-4585-17-0

[24] L. Breiman, (1996). "Bagging predictors". Machine Learning 24 (2): 123–140. doi:10.1007/BF00058655. CiteSeerX: 10.1.1.121.7654

[25] W. Liu and S. Chawla, "A Quadratic Mean based Supervised Learning Model for Managing Data Skewness". In: *Proceedings of the Eleventh SIAM International Conference on Data Mining*, pp. 188–198 (2011)

[26] R. Akbani *et al.*, "Applying Support Vector Machines to Imbalanced Data Sets," Lecture Notes in Computer Science, vol. 3201, pp. 39-50, 2004.

[27] W. Liu and S. Chawla "Class confidence weighted knn algorithms for imbalanced data sets", *Proc. 15th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining,* pp.345 -356, 2011

[28] Jason W. Osborne and Amy Overbay (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research & Evaluation, 9(6). [Online]. Available: http://PAREonline.net/getvn.asp?v=9&n=6 [Accessed: 11-Mar-2015].

[29] Prof. Carolina Ruiz, Department of Computer Science, WPI "Illustration of the K2 Algorithm for Learning Bayes Net Structures"

[30] J. Davis and M. Goadrich, "The Relationship between Precision-Recall and ROC Curves," *Proc. Int',l Conf. Machine Learning*, pp. 233-240, 2006

[31] "Lift Charts." [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html. [Accessed: 15-Mar-2015].

[32] Seo, Songwon, "A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets", MSc thesis, Univ. of Pittsburgh, 2006

[33] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions", Computational Statistics and Data Analysis 52 (2008) 5186–5201

[34] G. Brys *et al*., 2004. "A robust measure of skewness," Journal of Computational and Graphical Statistics 13, 996–1017.

[35] Amin, Adnan, et al. "Churn prediction in telecommunication industry using rough set approach." *New Trends in Computational Collective Intelligence*. Springer International Publishing, 2015. 83-95.

[36] Verbeke, Wouter, et al. "Building comprehensible customer churn prediction models with advanced rule induction techniques." *Expert Systems with Applications* 38 (2011): 2354-2364.

[37] “dme.churn/dataset/Balanced/FilledCFS at master · inquire/dme.churn · GitHub.” [Online].Available:https://github.com/inquire/dme.churn/tree/master/dataset/Balanced/FilledCFS. [Accessed: 06-May-2015].

[38] “dme.churn/dataset/Balanced at master · inquire/dme.churn · GitHub.” [Online]. Available: https://github.com/inquire/dme.churn/tree/master/dataset/Balanced. [Accessed: 06-May-2015].

[39] “Bayesian Network Classifiers in Weka for Version 3-5-7.” [Online]. Available: http://www.cs.waikato.ac.nz/~remco/weka_bn/. [Accessed: 12-Oct-2014].

[40] T. Saito and M. Rehmsmeier, (2015). *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*. PLoS ONE, 10(3), e0118432. doi:10.1371/journal.pone.0118432

APPENDIX A – User Manual

gaKnn Churn Prediction Tool Version 1.0

User Manual

gaKnn Churn Prediction Tool is an easy way that you can get to know about the future churners of your product or service. This provides simple and easy to handle, user interfaces to get your work done. The Tool uses the key concepts of `k` nearest neighbor classification and genetic algorithms; supported by the concepts of Naïve Bayesian Weights and Class Confidence Weights (ccw).

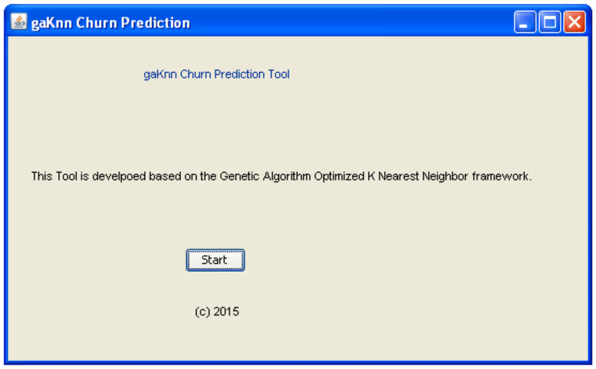
------------------------------------System Requirements---

- `JDK 1.6` or above
- `jgap` library
- `R 3.1.3` or above with `rJava` Package installed
    - Other R libraries required: `foreign, ROCR, gplots, robustbase, infotheo, hexbin, colorspace, ggplot2`
- `opencsv 3.1.2` library (should be compatible with the `jdk` version)

------------------------------------Working with the Tool---------------------------------------

Start the Tool

The Tool starts with the interface Figure A-1. Proceed with the instructions given at each step.



gaKnn Churn Prediction Tool

This Tool is develpoed based on the Genetic Algorithm Optimized K Nearest Neighbor framework.

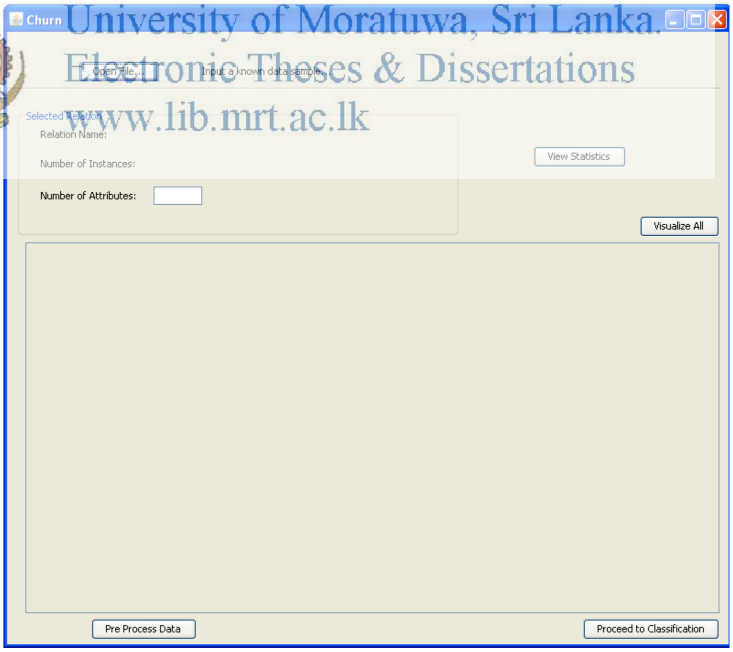
(c) 2015

Figure A-1 First Interface

Selecting the Dataset

Select the dataset through window in Figure A-2. As the dataset is selected it provides a graphical as well as a statistical overview of the selected dataset.

To get the statistical data follow the `'View Statistics'` button.



Relation Name:

Number of Instances:

Number of Attributes:

Figure A-2 Data selection

Data Visualizing

You can get several views about the distribution of data through `'Visualize All'` section.

Scatter plot view

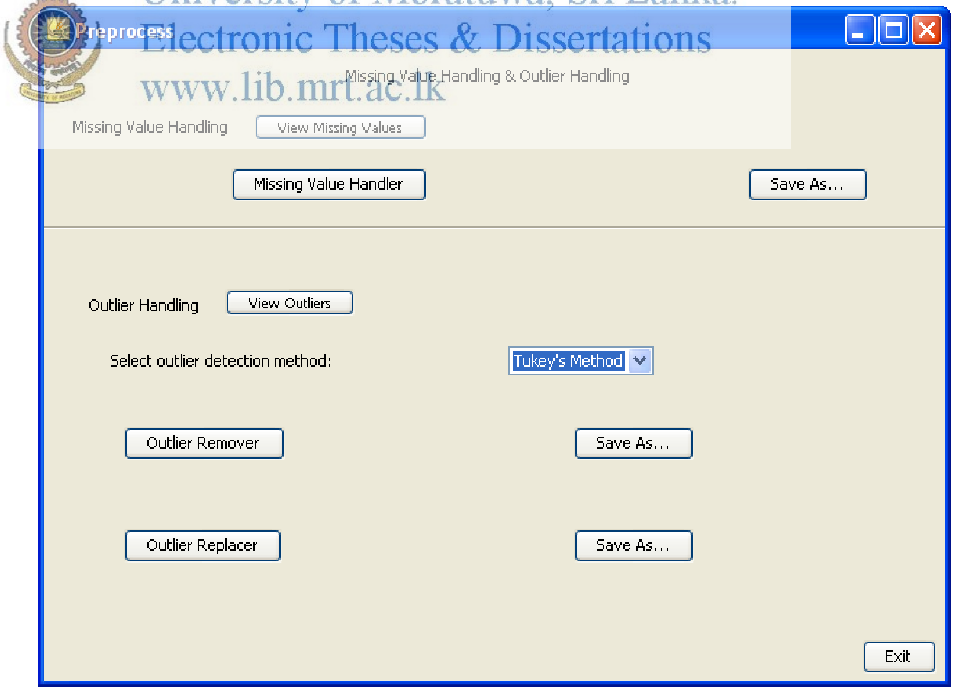
Bar plot view

Data Pre-processing

If you proceeded to the preprocess module via preprocess button in the `'Churn'` window, the window in Figure would be shown.

The `'Preprocess'` window provides you the ability to handle missing values and outliers. Click on the button '`Missing Value Handler'` and it will remove the missing values in the selected data file. The data instances which contain missing values could be viewed by clicking the button '`View Missing Values'`.

If you need to use the missing value handled data file as your training data file, select that data file through the `'Churn'` window.



Select outlier detection method: Tukey's Method

Figure A-3 Data pre-processor

Outliers in the selected data file are visualized as box plots. You can view them through the button ‘`View Outliers`’.

The Tool provides two outlier detection methods;

1. Tukey’s Method
2. MedCouples

Outlier handler provides two handling methods; removal and replacement.

First select the detection method from the dropdown menu and then click on the preferred handling method.

Please note that the visualized outliers are determined by the whiskers extend to the most extreme data point which is no more than range times the inter quartile range from the box.

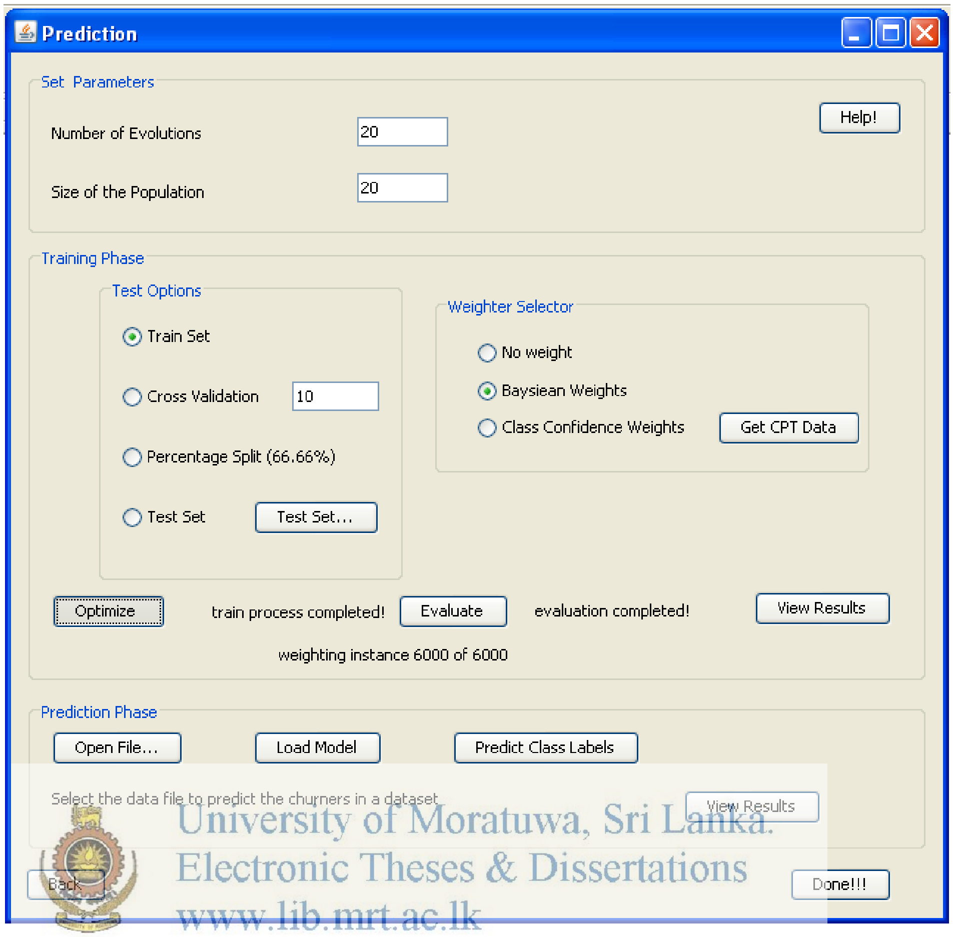
Optimizing the KNN classifier

You can proceed to the optimizing phase via ‘`Proceed to Classification`’ button.

The `Prediction` window allows you to perform both the optimizing and prediction tasks.

The optimizing task involves genetic algorithm techniques. It is by this step the Tool finds an optimum value for `k` and a weight vector for the attributes considering their importance.

The number of times which genetic algorithm is required to iterate is read by the `Number of Evolutions` and the size of the population considered by genetic algorithm is read by `Size of the Population` fields in the `Set Parameters` section.



Prediction

Set Parameters

Number of Evolutions 20

Size of the Population 20

Training Phase

Test Options

● Train Set

○ Cross Validation 10

○ Percentage Split (66.66%)

○ Test Set

Weighter Selector

○ No weight

● Baysiean Weights

○ Class Confidence Weights

train process completed!

evaluation completed!

weighting instance 6000 of 6000

Prediction Phase

Select the data file to predict the churners in a dataset

Figure A-4 Predictor

The population estimated with genetic algorithm is evaluated with KNN.

KNN classification relies on classification of unknown data based on known data. Therefore to test the KNN classifier it requires having two data sets; train set for training and test set for validating. In the `Training Phase` the Tool introduces four options to define the train set and test set; `Train Set`, `Cross Validation`, `Percentage Split` and `Test Set`.

If `Train Set` option is selected the dataset read into the Tool is used as both train set and test set. The `Cross Validation` option partitions the dataset, into specified `k` number of folds which would be used iteratively for `k` times where one fold is taken as the test set while the remaining k-1 folds are used as the train set. The option Percentage Split partitions the dataset into two folds of which one fold contains two-third of the dataset and the other fold contains the remaining one-third. If you use the test set option then you are required to select a separate dataset which is from the same domain as the dataset already read into the Tool. The first dataset is used as the train set while the second fed is used as the test set.

To get a better result on prediction with KNN, instance voting is introduced. Two voting methods were introduced; Bayesian Weights, Class Confidence Weights. You may either perform the classification without either of the two methods by selecting 'No Weight' option. Bayesian Weights add a vote to each prediction based on the Bayesian probability of getting a particular class label. Class Confidence Weights add a vote to each prediction based on the probability of attribute values given the class labels.

After selecting the desired options from Test Options and Weight Selector, click on the Optimize button. It will evaluate the data read into the Tool with genetic algorithm and KNN to produce the optimum value for k as well as a weight vector for the attributes. Results obtained with relevant to the optimization process could be viewed through the View Results button in the Training Phase section.

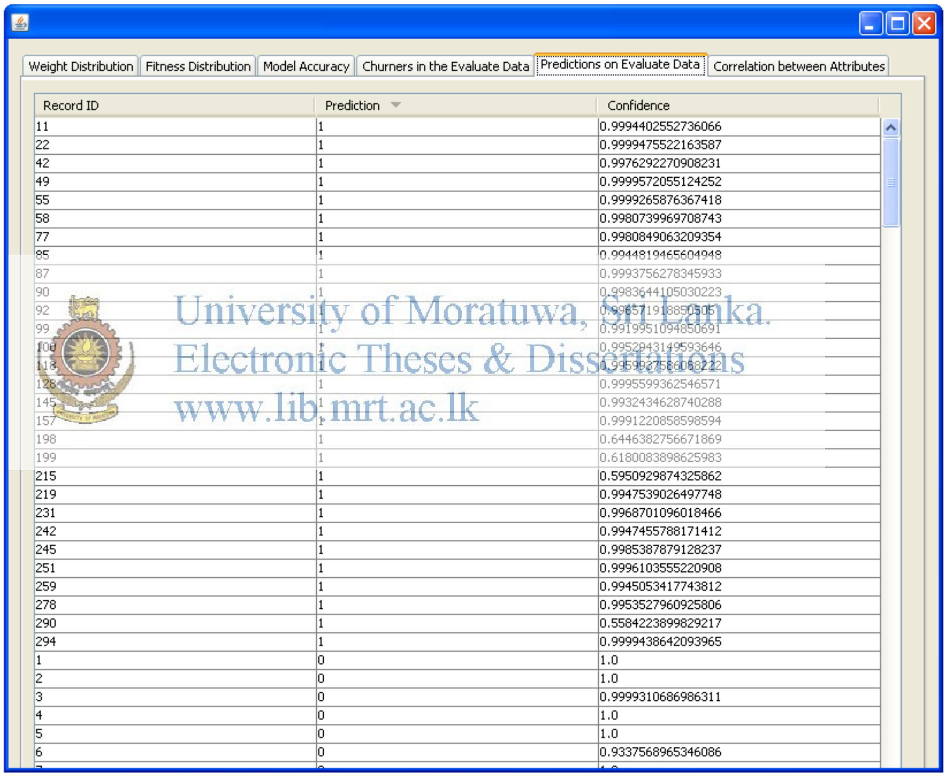
The model is automatically saved as an xml file with the optimized k value and the weights for each attribute. The location of the file is the same as where the training dataset is located and with the same file name as the training dataset with *.prm* extension.

The classifier could be re-evaluated with the Train set or with a separate test set. If you need to re-evaluate the model with the Train set simply click on the Evaluate button without changing any other option. If you need to re-evaluate the model with a separate test set, select the Test set option from the Test Options and choose another known set from the same domain as the initial dataset (Train Set). Results could be viewed through the `View Results` button in the `Training Phase` section.

Prediction

From the `Prediction Phase` section of the `Prediction` window choose the dataset with unknown class labels. You can predict the class labels through `Predict Class Labels` and view the churn results through `View Results` button.

View Results


Weight Distribution
Fitness Distribution
Model Accuracy
Churners in the Evaluate Data
Predictions on Evaluate Data
Correlation between Attributes

Record ID	Prediction	Confidence
11	1	0.9994402552736066
22	1	0.9999475522163587
42	1	0.9976292270908231
49	1	0.9999572055124252
55	1	0.9999265876367418
58	1	0.9980739969708743
77	1	0.9980849063209354
85	1	0.9944819465604948
87	1	0.9993756278345933
90	1	0.9983644105030223
92	1	0.996571918850505
99	1	0.9919951094850691
100	1	0.9952943149593646
118	1	0.9959927586068222
128	1	0.9995599362546571
145	1	0.9932434628740288
157	1	0.9991220858598594
198	1	0.6446382756671869
199	1	0.6180083898625983
215	1	0.5950929874325862
219	1	0.9947539026497748
231	1	0.9968701096018466
242	1	0.9947455788171412
245	1	0.9985387879128237
251	1	0.9996103555220908
259	1	0.9945053417743812
278	1	0.9953527960925806
290	1	0.5584223899829217
294	1	0.9999438642093965
1	0	1.0
2	0	1.0
3	0	0.9999310686986311
4	0	1.0
5	0	1.0
6	0	0.9337568965346086

Figure A-5 Result analyzer

Results could be viewed for different stages of the process. Viewing the results under the `Training Phase` provide the distribution of attribute weights, fitness distribution and evaluated predictions. By selecting the required tab you can easily get the relevant results.

`Weight distribution` shows the importance that has on each attribute when training the classifier. Attributes with high weight values are the most contributing attributes for the final prediction.

`Fitness distribution` shows whether the fitness value is stabilized for the `Number of Evolutions` specified in `Prediction` window. The model provides better results if the fitness value is stabilized at a maximum value.

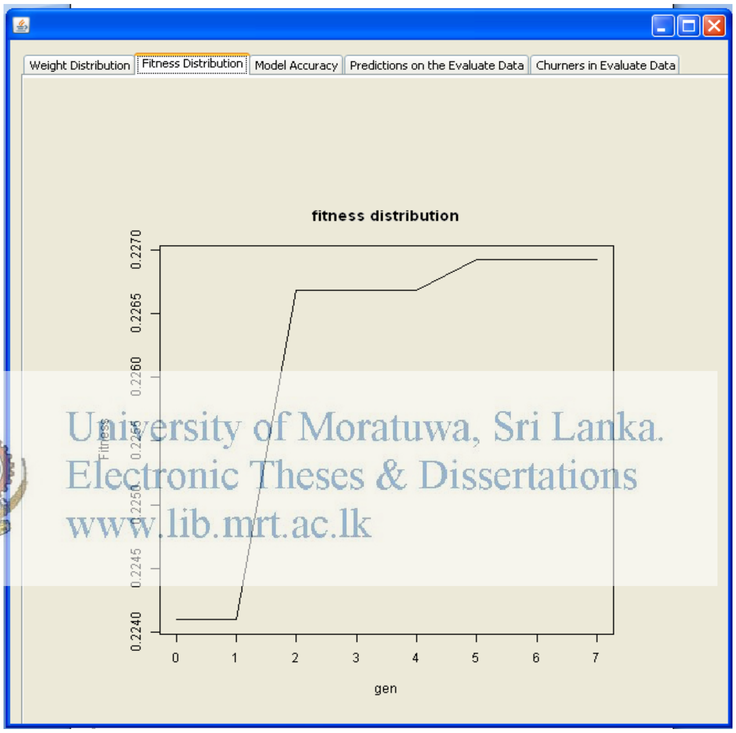


Figure A-6 Fitness distribution

Accuracy of the model could also be viewed in terms of four criteria (Figure A-7).

`'Churners in the evaluated data'` tab shows the predicted churners along with their details which were considered for the prediction. You can by examining the trend in data under '`Visualize All`' section, edit the values of attributes of the churners. Edited file could be saved and re-used for prediction.

`Prediction phase` allows you to load an existing model to predict churners. Select the Load Model button to select the saved model. The model loading should be saved as an xml file in *.prm* extension to be able to use with the Tool.

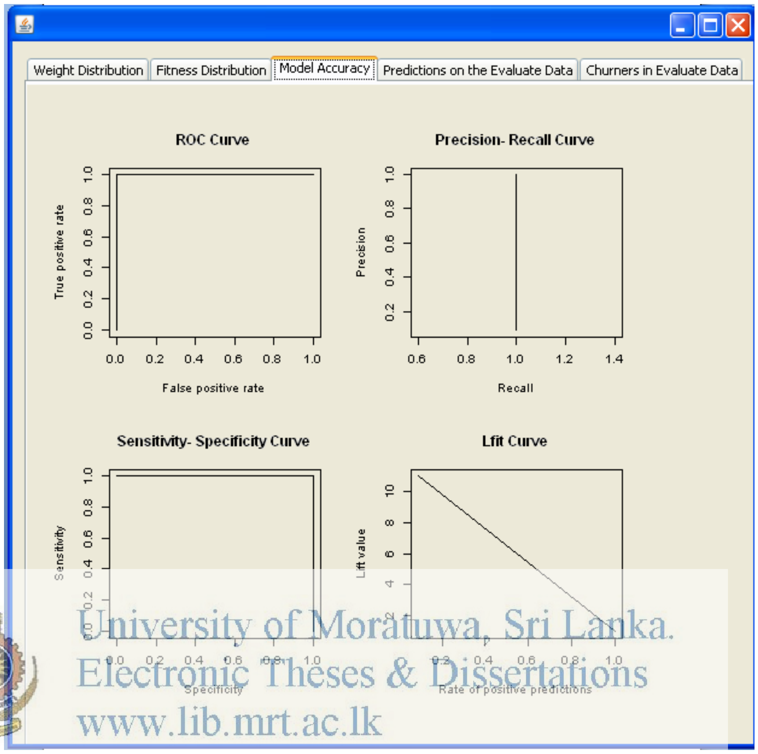


Figure A-7 Model accuracy

APPENDIX B – Source Code

The source code of the Tool could found in the CD attached with the title Appendix B.