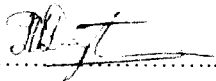


The work included in this report was done by me, and only by me, and the work has not been submitted for any other academic qualification at any institution.



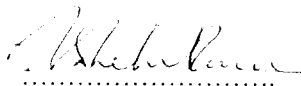
D.G.N. Dayaratne

31st Dec. 2008

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for the MSc Research Project.



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



Dr. Shehan Perera

(Supervisor)

Acknowledgements

First of all, I would like to thank my supervisor Dr. Shehan Perera who initially proposed the idea to develop a Genetic Algorithm optimized K Nearest Neighbor classification framework. Without his guidance and invaluable suggestions this project would not have been a success.

Also I wish to acknowledge Mrs. Vishaka Nanayakkara MSc Coordinator for 2005 MSc batch/ Head of the Dept. of Computer Science & Engineering for her guidance and support. I would like to extend my sincere thanks to Dr. Sanath Jayasena MSc Coordinator for 2007 MSc batch for his invaluable mentoring and guidance to make this project a success.

The authors of the various references used, notably JGAP Java Genetic Algorithm Package are also acknowledged with gratitude.

Also I should mention my family and friends for their support and encouragement extended to me throughout the project. Without their support this would not have been a success.



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

Table of Contents

| | | |
|-----|---|----|
| 1 | Introduction..... | 1 |
| 2 | Literature Review..... | 3 |
| 2.1 | K-Nearest Neighbor Classification | 3 |
| 2.2 | Genetic Algorithms | 8 |
| 2.3 | Optimizing Weighted K-Nearest Neighbor with a Genetic Algorithm | 10 |
| 2.4 | Usage of GA-KNN for Real World Applications..... | 12 |
| 2.5 | GA Frameworks..... | 13 |
| 2.6 | Background to the Existing gaKnn | 15 |
| 2.7 | Framework Designing Aspects | 16 |
| 3 | Methodology..... | 18 |
| 3.1 | Problem Definition..... | 18 |
| 3.2 | Approach..... | 18 |
| 3.3 | Scope..... | 19 |
| 4 | Design and Implementation..... | 24 |
| 4.1 | Overall Design | 24 |
| 4.2 | Module Details..... | 27 |
| 4.3 | Sample Implementations..... | 35 |
| 5 | Results and Evaluation | 41 |
| 5.1 | Goals | 41 |
| 5.2 | Testing Strategy | 41 |
| 5.3 | Test Results | 44 |
| 5.4 | Test Result Analysis..... | 50 |
| 6 | Conclusions and Future Work..... | 55 |
| 6.1 | Conclusions and Contributions | 55 |
| 6.2 | Future Work | 56 |
| 7 | References | 57 |

| | |
|---|----|
| Appendix A: Attribute-Relation File Format (ARFF) | 59 |
| Appendix B: Introduction to JGAP Framework..... | 61 |
| Appendix C: Customizing gaKnn | 63 |



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

List of Figures

| | |
|---|----|
| Figure 2.1: KNN classifier | 3 |
| Figure 2.2: KNN algorithm | 5 |
| Figure 2.3: Effect of scaling attributes | 6 |
| Figure 2.4: Feature selection and instance selection for KNN classifier. | 8 |
| Figure 4.1: Overall design | 24 |
| Figure 4.2: KNN optimization | 25 |
| Figure 4.3: KNN classification | 26 |
| Figure 4.4: Main class diagram for data access | 30 |
| Figure 4.5: Algorithm for chromosome evaluation | 31 |
| Figure 4.6: Main class diagram for KNN optimization | 32 |
| Figure 4.7: Program flow for KNN optimization | 33 |
| Figure 4.8: Main class diagram for KNN classification | 34 |
| Figure 4.9: Program flow for KNN classification | 35 |
| Figure 5.1: GA optimization progress for the abalone data set with Method 1 | 44 |
| Figure 5.2: GA optimization progress for abalone data set with Method 2 | 45 |
| Figure 5.3: Weights for the abalone data set | 46 |
| Figure 5.4: GA optimization progress for adult data set with Method 1 | 47 |
| Figure 5.5: Weights for the adult data set | 48 |
| Figure 5.6: GA optimization progress for gender training data set with Method1 | 48 |
| Figure 6.1: Training time vs file size | 53 |

List of Tables

| | |
|--|----|
| Table 4.1: Data set details of adult data set | 35 |
| Table 4.2: Attribute details of adult data set | 36 |
| Table 4.3: Data set details of abalone data set | 36 |
| Table 4.4: Attribute details of abalone data set | 37 |
| Table 4.5: Data set details of gender training data set | 37 |
| Table 4.6: Attribute details of gender training data set | 38 |
| Table 5.1: Class distribution of abalone data set | 42 |
| Table 5.2: Classes of abalone data set | 42 |
| Table 5.3: Selected attributes of gender training data set | 44 |
| Table 5.4: Optimum parameters of abalone data set with Method 1 | 45 |
| Table 5.5: Optimum parameters of abalone data set with Method 2 | 46 |
| Table 5.6: Optimum parameters of adult data set with Method 2 | 47 |
| Table 5.7: Optimum parameters for gender training data set Method 2 | 49 |



Symbols, Notations, Abbreviations and Acronyms

| | |
|-------|---|
| gaKnn | Genetic Algorithm Optimized Nearest Neighbor classification framework |
| KNN | K Nearest Neighbor |
| GA | Genetic Algorithm |
| WKNN | Weighted K Nearest Neighbor |
| SFS | Sequential Forward Selection |
| SBS | Sequential Backward Selection |
| JGAP | Java Genetic Algorithm Programming |
| ARFF | Attribute-Relation File Format |



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