

# Ontology Based Personalized Modeling for Chronic Disease Risk Evaluation and Knowledge Discovery: An Integrated Approach

Anju Verma

A thesis submitted to Auckland University of Technology in  
fulfillment of the requirements for degree of  
Doctor of Philosophy (PhD)

2009

School of Computer & Information Sciences

Primary supervisor: Prof. Nikola Kasabov

Other supervisors: Dr. Qun Song, Prof. Elaine Rush, Dr. Neil Domigan

## Table of Contents

Attestation of Authorship .....	16
Acknowledgements .....	17
Abstract .....	20
Chapter 1. Introduction .....	22
1.1 Background .....	22
1.2 Goals of the thesis .....	26
1.3 Organisation of the thesis .....	28
1.4 Major contributions of the thesis .....	30
1.5 Resulting publications .....	32
Chapter 2. Methods and Systems for Risk Evaluation in Medical Decision Support Systems .....	36
2.1 Inductive and transductive reasoning.....	37
2.2 Global, local and personalized modeling.....	41
2.3 Weighted K-nearest neighbour method (WKNN) .....	44
2.4 Weighted-weighted K nearest neighbour algorithm for transductive reasoning (WWKNN) and personalized modeling.....	45
2.5 Neuro-Fuzzy Inference Method (NFI) for personalized modeling.....	47
2.6 Transductive neuro-fuzzy inference system with weighted data normalization (TWNFI) for personalized modeling .....	49
2.7 Summary.....	53
Chapter 3. Ontology Systems for Knowledge Engineering: A Review.....	55

3.1 What is Ontology?.....	55
3.2 Applications of ontology .....	56
3.3 Tools for developing an ontology .....	57
3.4 Methods for developing ontology .....	61
3.5 Existing ontologies .....	64
3.6 Conclusion .....	67
 Chapter 4. A Novel Chronic Disease Ontology (CDO) for Information Storage and Knowledge Discovery .....	 68
4.1 Chronic Disease Ontology (CDO) .....	68
4.1.1 Organism Domain.....	69
4.1.2 Molecular Domain .....	70
4.1.3 Medical Domain.....	78
4.1.4 Nutritional Domain .....	79
4.1.5 Biomedical informatics map.....	79
4.1.6 Information retrieval.....	81
4.1.7 Visualization of the ontology .....	82
4.2 Knowledge discovery through the chronic disease ontology (CDO) .....	83
4.3 Summary.....	87
 Chapter 5. An Integrated Framework of Ontology and Personalized Modelling for Knowledge Discovery .....	 88
5.1 Integration framework for ontology and personalized modeling .....	88
5.2 Knowledge discovery through the integration of personalized modeling tools and the chronic disease ontology (CDO).....	95
5.3 Conclusion .....	99

Chapter 6. Cardiovascular Disease Risk Evaluation Based on the Chronic Disease Ontology (CDO).....	100
6.1 Cardiovascular disease, prevalence and description .....	100
6.2 Existing methods for predicting risk of cardiovascular disease .....	101
6.3 Data Exploration .....	104
6.3.1 Description of selected data .....	105
6.3.2 Rationale for selecting variables.....	107
6.3.3 Statistical Analysis.....	110
6.4 Risk prediction and knowledge discovery with the ontology based personalized decision support (OBPDS).....	127
6.5 Integrated framework of ontology based personalized cardiovascular disease risk analysis .....	160
6.6 Examples of integration of the chronic disease ontology and the personalized risk evaluation system for cardiovascular disease .....	162
6.7 Conclusion .....	163
Chapter 7. Type 2 Diabetes and Obesity Risk Evaluation and Knowledge Discovery Based on the Chronic Disease Ontology (CDO).....	166
7.1 Type 2 diabetes, prevalence and description.....	166
7.2 Obesity, prevalence and description .....	168
7.3 Diabetes prediction models.....	171
7.4 Data Exploration .....	173
7.4.1 Description of selected data .....	173
7.4.2 Rationale for selecting variables.....	174
7.4.3 Statistical Analysis.....	179

7.5 Risk prediction method and knowledge discovery .....	192
7.6 Integration framework of ontology and personalized diabetes risk analysis and knowledge discovery .....	213
7.7 Examples for integration of the chronic disease ontology and personalized diabetes risk analysis model.....	214
7.8 Conclusion .....	215
Chapter 8. Conclusions, Discussion and Directions for Future Research ....	218
8.1 Achievements.....	218
8.2 Further developments .....	222
References .....	226
Appendix A WWKNN Algorithm.....	250
Appendix B NFI Learning Algorithm .....	252
Appendix C TWNFI Learning Algorithm.....	258
Appendix D Formulas used to calculate percentages for nutrient variables (Atwater and Bryant, 1900).....	264
Appendix E NeuCom .....	265
Appendix F Siftware .....	269

## List of Figures

<i>Figure 1.1.</i> Venn diagram showing three chronic diseases with overlapping causes.....	23
<i>Figure 1.2.</i> Venn diagram illustration of nutrigenomics as the intersection between health, diet, and genomics (Picture taken from Ruden et al, 2005)..	24
<i>Figure 1.3.</i> Structure and organization of the thesis. ....	29
<i>Figure 2.1.</i> A block diagram of an inductive reasoning system. A global model $M$ is created based on data samples from $D$ and then recalled for every new vector $x_i$ (From: Song and Kasabov, 2004). ....	38
<i>Figure 2.2.</i> A block diagram of a transductive reasoning system. An individual model $M_i$ is trained for every new input vector $x_i$ with data samples $D_i$ selected from a data set $D$ , and data samples $D_{o,i}$ generated from an existing model (formula) $M$ (if such a model exists). Data samples in both $D_i$ and $D_{o,i}$ are similar to the new vector $x_i$ according to a defined similarity criteria (From: Song and Kasabov, 2006).....	40
<i>Figure 2.3.</i> Example of transductive reasoning. In the centre of a transductive reasoning system is the new data vector (here illustrated with two vectors – $x_1$ and $x_2$ ), surrounded by a fixed number of nearest data samples selected from the training data $D$ and/or generated from an existing model $M$ (From: Song and Kasabov, 2006). ....	41
<i>Figure 2.4.</i> A block diagram of the NFI learning algorithm (From: Song and Kasabov, 2004). ....	48
<i>Figure 2.5.</i> A block diagram of the TWNFI algorithm (From: Song and Kasabov, 2006). ....	50

<i>Figure 4.1.</i> The general structure of the organism domain in the chronic disease ontology.....	70
<i>Figure 4.2.</i> General structure of molecular domain in the chronic disease ontology.....	71
<i>Figure 4.3.</i> A screenshot from the chronic disease ontology showing information about the gene ACE.....	78
<i>Figure 4.4.</i> Picture of a disease gene map for type-2 diabetes showing few genes related to type 2 diabetes through various mutations. ....	80
<i>Figure 4.5.</i> A screenshot of an example of the query tool showing a gene list responsible for the regulation of blood pressure and causing cardiovascular disease, obesity and type 2 diabetes by means of insertion (a type of mutation). ....	81
<i>Figure 4.6.</i> Visualization for the structure of the chronic disease ontology using TGViz plug-in.....	82
<i>Figure 4.7.</i> A screenshot of an example of a gene list obtained from the chronic disease ontology at chromosome 2. ....	84
<i>Figure 4.8.</i> A screenshot of a list of genes present on chromosome 2 in the chronic disease ontology which cause disease by dinucleotide repeat mutation.....	85
<i>Figure 4.9.</i> A screenshot of a list of genes involved in blood circulation obtained from the chronic disease ontology. ....	86
<i>Figure 4.10.</i> A screenshot of a list of genes (AGTR1 gene and LPL gene) involved in blood circulation that cause disease by dinucleotide repeat mutation.....	86

<i>Figure 5.1.</i> The ontology-based personalized decision support (OBPDS) framework consisting of three interconnected parts: (1) An ontology/database module; (2) Interface module; (3) A machine learning module. ....	89
<i>Figure 5.2.</i> The general framework for the ontology based personalized risk evaluation system.....	90
<i>Figure 5.3.</i> Example of framework for use of knowledge from the chronic disease ontology (CDO) to personalized model. ....	96
<i>Figure 5.4.</i> An example of utilization of knowledge from the personalized risk evaluation model for cardiovascular disease within the chronic disease ontology (CDO) and reuse for subsequent subjects.....	97
<i>Figure 5.5.</i> An example of use of knowledge from the personalized model for type 2 diabetes within the chronic disease ontology (CDO) and reuse for subsequent subjects.....	98
<i>Figure 6.1.</i> Bar graph of NNS97 data for all subjects with age and risk of cardiovascular disease (n=2,875).....	111
<i>Figure 6.2.</i> Bar graph of NNS97 male data for age and risk of cardiovascular disease (n=1,305).....	112
<i>Figure 6.3.</i> Bar graph of NNS97 female data for age and risk of cardiovascular disease (n=1,570).....	113
<i>Figure 6.4.</i> Bar graph showing variables ranked (highest to lowest) according to signal to noise ratio for the whole data. ....	118
<i>Figure 6.5.</i> Bar graph showing variables ranked (highest to lowest) according to signal to noise ratio for male subjects only. ....	119
<i>Figure 6.6.</i> Bar graph showing variables ranked (highest to lowest) according to signal to noise ratio for female subjects only. ....	120

<i>Figure 6.7.</i> Linear relationship between the variables (listed below) using a correlation coefficient for the whole data. ....	121
<i>Figure 6.8.</i> Linear relationship between the variables (listed below) using a correlation coefficient for male subjects.....	122
<i>Figure 6.9.</i> Linear relationship between the variables (listed below) using a correlation coefficient for female subjects.....	123
<i>Figure 6.10.</i> Illustration of rules extraction from clusters based on nearest subjects. ....	139
<i>Figure 6.11.</i> Example of male Subjects 1 and 2 with cluster centers based on nearest neighbors using principal component analysis (PCA).....	140
<i>Figure 6.12.</i> Example of female Subjects 1 and 2 with cluster centers based on nearest neighbors using principal component analysis (PCA).....	151
<i>Figure 6.13.</i> Integrated framework of ontology based personalized cardiovascular disease risk analysis.....	161
<i>Figure 7.1.</i> Bar graph showing ranked variables (highest to lowest) for whole data using signal to noise ratio for prediction of type 2 diabetes by gene markers. ANGPTL3, AGPT4, TNF genes are ranked at high position.....	181
<i>Figure 7.2.</i> Bar graph showing ranked variables (highest to lowest) for whole data for prediction of type 2 diabetes by gene markers using p-value derived from t-test. The lowest p-value explains the most important gene.....	182
<i>Figure 7.3.</i> Bar graph showing ranked variables (highest to lowest) for male subjects using signal to noise ratio for prediction of type 2 diabetes by gene markers. ANGPTL3 and MMP2 are the most important genes for male subjects and are ranked at highest position. ....	183

*Figure 7.4.* Bar graph showing ranked variables (highest to lowest) for male subjects for prediction of type 2 diabetes by gene markers using p-value derived from t-test. ANGPTL3 and MMP2 are most important genes..... 184

*Figure 7.5.* Bar graph showing ranked variables (highest to lowest) for female subjects using signal to noise ratio for prediction of type 2 diabetes by gene markers. ANGPTL3 and ANGPT 4 are the most important genes for female subjects. .... 185

*Figure 7.6.* Bar graph showing ranked variables (highest to lowest) for female subjects for prediction of type 2 diabetes by gene markers using p-value derived from t-test. ANGPTL3 and ANGPT4 are the most important genes for female subjects..... 186

*Figure 7.7.* Linear relationships between general, clinical and genetic variables (listed below) for whole data using correlation coefficient (Red colour: high positive correlation). .... 190

*Figure 7.8.* Linear relationships between the general, clinical and genetic variables (listed below) for male subjects using correlation coefficient. (Red colour: high positive correlation). .... 191

*Figure 7.9.* Linear relationships between the general, clinical and genetic variables (listed below) for female subjects using correlation coefficient. (Red colour: high positive correlation). .... 192

*Figure 7.10.* Example of male Subjects 1 and 2 with cluster centers based on nearest neighbors using principal component analysis (PCA)..... 198

*Figure 7.11.* Example of female Subjects 1 and 2 with cluster centers based on nearest neighbors using principal component analysis (PCA)..... 207

*Figure 7.12.* Integration framework for chronic disease ontology and personalized risk evaluation of type 2 diabetes. .... 214

*Figure E.1.* Screenshot of the NeuCom environment. .... 265

*Figure F.1.* Screenshot of the Siftware environment..... 269

## List of Tables

Table 3.1 Comparison of ontology development tools.....	60
Table 4.1 General structure of the chronic disease ontology sub-domains.....	69
Table 4.2 List of genes present in the chronic disease ontology (CDO).....	72
Table 6.1 Description of subjects in NNS97 data.....	105
Table 6.2 List of variables from NNS97 data for initial experiments.....	107
Table 6.3 Prevalence of hypertension (risk factor for cardiovascular disease) in 2,875 subjects from the National Nutrition Survey 1997.....	114
Table 6.4 Average, maximum and minimum values of the selected variables in whole, male and female population.....	115
Table 6.5 Results of correlation coefficient for male samples, female samples and the whole dataset.....	125
Table 6.6 Accuracy (%) results comparison of NNS 97 data using 13 variables for male data.....	132
Table 6.7 Accuracy (%) results comparison of NNS 97 data using 13 variables for female data.....	133
Table 6.8 Accuracy (%) results comparison of NNS 97 data using 13 variables for the whole dataset.....	134
Table 6.9 Examples of TWNFI personalized models for two different male subjects; high risk and low risk male subjects; showing different weights for same variables with global weights representing importance of variables.....	137

Table 6.10 Example of TWNFI personalized models for two female subjects; high risk and low risk subjects; showing different weights for same variables with global weights representing importance of variables.....	150
Table 7.1 WHO classification of BMI for Obesity (World Health Organization, 2000).....	170
Table 7.2 Comparison of existing methods to predict risk of type 2 diabetes.....	172
Table 7.3 Distribution of male and female subjects as diagnosed without or with type 2 diabetes.....	174
Table 7.4 List of clinical variables and genes used for personalized risk evaluation and knowledge discovery.....	176
Table 7.5 Comparison of minimum, maximum and average values of clinical variables among male and female subjects.....	180
Table 7.6 List of first six genes for whole data and male, female subjects according to signal to noise ratio and t-test.....	187
Table 7.7 List of genes selected for personalized modeling for male subjects with their description .....	188
Table 7.8 List of genes selected for personalized modeling for female subjects with their description .....	189
Table 7.9 Accuracy (%) comparison of diabetes data using clinical and genetic variables for male subjects.....	194

Table 7.10: Accuracy (%) comparison of diabetes data using clinical and genetic variables for female subjects.....	195
Table 7.11 Examples of TWNFI personalized models for two different male subjects; high risk and low risk; with weight of variables and genes with global weights representing importance of the variables.....	197
Table 7.12 Examples of TWNFI personalized models for two different female subjects; high risk and low risk; with weights of variables and genes with global weights representing importance of the variables.....	206

*dedicated to my hubby.....*

## **Attestation of Authorship**

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in acknowledgements) nor material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institution of higher learning.

---

Anju Verma

## **Acknowledgements**

This thesis arose in part out of research that has been done over the last few years. Over this time, I have worked with a great number of people whose contributions to the research and the development of the thesis deserve special mention. It is my pleasure to convey my gratitude to them all in my acknowledgments.

In the first place I would like to acknowledge my gratitude to Prof. Nikola Kasabov for his supervision, guidance, and advice from the very early stage of this research and for providing me with extraordinary experiences throughout the research. Without his unwavering faith and advice, and stimulating suggestions and discussions, the completion of my study would not have been possible. His experience as a true scientist made him as an oasis of ideas and passion for research, which has inspired and enriched my growth as a student and a researcher. I am indebted to him more than he knows.

I would like to express my sincere gratitude to Brent Ogilvie, who showed faith in me and helped me all the way through my studies. Words fail to express my appreciation for him, as he has always given me moral support and professional guidance with his words of encouragement.

I would like to thank Prof. Elaine Rush for her motivation, invaluable comments, generosity of time and willingness to help. I would also like to thank my other supervisor Dr. Qun Song, who was always there to provide his valuable suggestions. My special thanks to Dr. Neil Domigan, for his constant encouragement to write thesis. I would like to thank Prof. Ajit Narayanan for the help and support. I would also like to thank Russel Pears for his help and assistance during my research.

Armaan, my hubby, the guiding star of my life, deserves all the credit for me finishing my studies as he made room in our lives for me to study, caught the worst of my down at times when this project felt insurmountable, and had the faith that I would get there. Thank you from the bottom of my heart for all the support and sacrifices.

My darling dolls, Monalisa and Alisha deserve a special mention here as they never demanded special time or care though both were going through transition phases of life. Monalisa had a sister and also started primary school during my studies and Alisha even though a small baby never bothered me during my studies and used to play on her own and spent most of her time at day care when she was supposed to be under my care. I appreciate Monalisa's understanding as well all the way during my studies. I would also like to thank 'KinderCare' especially Angela for looking after Alisha so well.

I have reached at this stage with the blessings of my parents. My parents and parents-in law have been an integral part of my research journey through their continuous support and encouragement. My father-in-law always encouraged and showed faith in me and always showered his most precious blessings on me. My acknowledgements would be incomplete if I do not mention my Mum, as she always wanted me to finish my research and during the course of my study she was diagnosed with a serious disease. As she did not want to disturb me during my research, she never informed me about her sickness. I wish everyone could be blessed with a great mum as mine. I thank my mum for her optimism, unfailing love and encouragement.

I am much indebted to Joyce D'Mello, who always encouraged me to finish my research and thesis in time. I would like to thank Peter Hwang for all his

support and encouragement especially with Matlab as well as for our useful discussions during the last phase. I would also like to thank Dr. Ilkka Havukkala for his advice and willingness to share his bright thoughts with me, which was very fruitful in shaping up my ideas and research. I would also like to thank other KEDRI members namely, Paulo, Vishal and Alex for their help.

I would also like to thank Cherry Gordon, who was my first contact person for administration work at AUT. I would also like to thank postgraduate office (Annette, Elena, Martin Wilson) and other AUT admin staff for their help during the period of research. I would also like to acknowledge the help and services provided by the library staff.

This work was funded by FidelityGenetic (an affiliation of Pacific Channel Limited) and the Foundation for Research, Science and Technology of New Zealand under grant number CSHA0401. I would also like to thank the Ministry of Health, New Zealand for providing me with NNS97 data. I would also like to thank Maurizio, PhD. Student from DIMET, University Mediterranea of Reggio Calabria, Italy and Transplant Regional Center of Stem Cells and Cellular Therapy, "A. Neri", Reggio Calabria, Italy for sharing diabetes data.

Finally, I would like to thank everybody who played important role in the successful completion of this thesis, as well as express my apologies that I can not mention each one by name.

And to God, who makes all things possible.

## **Abstract**

Populations are aging and the prevalence of chronic disease, persisting for many years, is increasing. The most common, non-communicable chronic diseases in developed countries are; cardiovascular disease (CVD), type 2 diabetes, obesity, arthritis and specific cancers. Chronic diseases such as cardiovascular disease, type 2 diabetes and obesity have high prevalence and develop over the course of life due to a number of interrelated factors including genetic predisposition, nutrition and lifestyle. With the development and completion of human genome sequencing, we are able to trace genes responsible for proteins and metabolites that are linked with these diseases.

A computerized model focused on organizing knowledge related to genes, nutrition and the three chronic diseases, namely, cardiovascular disease, type 2 diabetes and obesity has been developed for the Ontology-Based Personalized Risk Evaluation for Chronic Disease Project. This model is a Protégé-based ontological representation which has been developed for entering and linking concepts and data for these three chronic diseases. This model facilitates to identify interrelationships between concepts.

The ontological representation provides the framework into which information on individual patients, disease symptoms, gene maps, diet and life history can be input, and risks, profiles, and recommendations derived. Personal genome and health data could provide a guide for designing and building a medical health administration system for taking relevant annual medical tests, e.g. gene expression level changes for health surveillance.

One method, called transductive neuro-fuzzy inference system with weighted data normalization is used to evaluate personalized risk of chronic disease. This

personalized approach has been used for two different chronic diseases, predicting the risk of cardiovascular disease and predicting the risk of type 2 diabetes. For predicting the risk of cardiovascular disease, the National Nutrition Health Survey 97 data from New Zealand population has been used. This data contains clinical, anthropometric and nutritional variables. For predicting risk of type 2 diabetes, data from the Italian population with clinical and genetic variables has been used. It has been discovered that genes responsible for causing type 2 diabetes are different in male and female samples.

A framework to integrate the personalized model and the chronic disease ontology is also developed with the aim of providing support for further discovery through the integration of the ontological representation in order to build an expert system in genes of interest and relevant dietary components.