Chapter 5.

5 Analysis

5.1 The basic algorithm

The trees created by the Genetic Algorithm were very accurate for their cost range, and the pareto front of decision trees was able to evolve to reach satisfactory levels in very short amounts of time. The structure and thresholds of the individual decision trees were shown to parallel current medical knowledge, confirming that data mining is indeed an exciting and promising field that has the possibility to reveal new and important medical knowledge when run on clinical databases.

5.2 Elitism

For smaller population sizes, the addition of the elite set improved the AUC measures, and for the most part ensured a steady improvement in the fitness of the population over the generations. The populations without the elite set would find improved solutions, but then lose them in later generations. As population size increased, though, the addition of the elite members eventually no longer made a significant difference.

Though genetic programming is a global search, and may avoid becoming trapped at local maximum as the way greedy heuristics do, the tradeoff is that they are computationally intensive.[35] This multi-objective tree algorithm in particular must calculate the fitness values by comparing every tree to every other tree in the population, an operation with an order of N^2. The time and resources needed will only increase when this process is applied to large medical databases. Therefore any techniques that could possibly improve the speed or lower the necessary resources are a valuable addition. The inclusion of an elite set permits the number of trees to be cut to much smaller levels, and still allows solutions with reasonable AUC to be found.

5.3 Mutation

Both hard and soft mutations had similar values (though soft mutation was a bit more stable, and had slightly higher AUC estimates). Adding slight amounts of mutation from the 0% starting point increased the AUC for both, and they remained stable despite increases in the amount of mutation done, until a

mutation rate of 75% was reached. At that point both strategies experienced a sharp drop as their error thresholds were reached [37]. It seems that, whether you are replacing a constant within a node's function, or replacing that entire node and its subtree, the disturbance is enough to increase the AUC to a certain extent. After that point is reached, the results stay within the same area and are not greatly affected by changes in the mutation rate until the error threshold is reached.

5.4 Linear Decision Trees

It was shown that it was possible to create and evolve a pareto front of linear decision trees. However, adding linear capabilities to a decision tree created in this way does not guarantee better classification of the data. It is interesting to note that adding linear capabilities to the decision trees did not improve the AUC for the populations, or decrease the number of nodes necessary to obtain the optimal solutions. In the case of the PIMA diabetes dataset, the result is slightly worse than the univariate. Longer runs for the number of generations, to allow more time for the linear nodes to find the optimal constants, did not greatly improve their performance, either.

Chapter 6.

6.1 Future Work

The trees created by the genetic program at approximately the 1 to 1 cost ratio had very good accuracies, but would in general be very large, with hundreds of nodes at times. The time it takes to run the entire dataset through every tree to discover their confusion matrices (and therefore be able to determine the dominance of a particular tree or not) goes up fast with the size of the trees. Though the inclusion of multiple copies of the same decision patterns inside a single individual is actual desirable during evolution, ensuring that good patterns are not lost or destroyed during the evolution process(30), the pruning of useless nodes or branches would certainly lead to vast improvements in run time. It would be interesting to try pruning during run time, and see what performance gains it may make, how it might affect the final AUC values, and whether it would be cost effective to do so.

The addition of linear capabilities to the decision tree did not improve its capabilities on the UCI datasets used. Genetic Programming has been found to be efficient at optimizing the structure of a tree, but struggles with finding the optimum coefficients for trees, since it uses combinations of random ones.(24) To improve the performance of linear trees created with Genetic Programming, some researchers have used Quasi-Newton optimization techniques to obtain the coefficients.(23) Further study could be done by using this technique to obtain the coefficients, instead of randomly generating and mutating them.

Another possibility would be to run this algorithm on a larger more complicated dataset, that contains more attributes. In a simpler dataset, it is possible that the axis-parallel splits done by the univariate decision trees are optimal, or just as effective as the oblique splits done by the linear trees. With the addition of more attributes, some of which may have values related to and dependent on each other, the greater flexibility of the linear decision tree may make a greater difference.

[1] http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/
[2] http://www.glucobase.com/
[3] Lowrance, William W. Privacy and Health Research: A Report to the U.S. Secretary of Health and Human Services. 1997 May.
[4] http://www.ispor.org/DigestOfIntDB/Default.aspx?rcd=411
[5] Data mining textbook
[6] David J. Hand: Statistics and Data Mining: Intersecting Disciplines. SIGKDD Explorations 1(1): 16-19 (1999)
[7] Lane, T. In Maloof, M., ed., *Machine learning and data mining for computer security: Methods and applications*. London: Springer-Verlag. 2006.
[8] Jason Tsong-Li Wang, Mohammed Javeed Zaki, Hannu Toivonen, Dennis Shasha (Eds.): Data Mining in Bioinformatics. Springer 2005, ISBN 1-85233-671-4
[9] Web Data Mining and Applications in Business Intelligence and Counter-Terrorism by Bhavani Thuraisingham Auerbach Publications © 2003
[10] J. C. Prather, D.F. Lobach, L.K. Goodwin, J. W. Hales, M. L. Hage, W. Edward Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse", 1997
[11] Bartha JL, Martinez-Del-Fresno P, Comino-Delgado R. "Early diagnosis of gestational diabetes mellitus and prevention of diabetes-related complications." European Journal of Obstetrics and Gynecology and Reproductive Biology, Volume 109, Number 1, 1 July 2003, pp. 41-44(4)
[12] Expert Knowledge and Its Role in Learning Bayesian Networks in Medicine: An Appraisal Lecture Notes in Computer Science, Springer Berlin / Heidelberg
[13] , Allan, Veronica Vinciottia, Xiaohui Liua, David Garway-Heathb Artif, "A spatio-temporal Bayesian network classifier for understanding visual field deterioration", Intell Med. 2005 Jun;34(2):163-77
[14] Blanco R, Inza I, Menino M, Quiroga J, Larrañaga P. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. J Biomed Inform 2005;38(5):376-88.
[15] Veropoulos K., Cristianini N., Campbell C., 1999, The Application of Support Vector Machines to Medical Decision Support : A Case Study. Department of Engineering Mathematics, Bristol University, United Kingdom
[16] **A support vector machine approach for detection of microcalcifications** El-Naqa, I.; Yongyi Yang; Wernick, M.N.; Galatsanos, N.P.; Nishikawa, R.M. Medical Imaging, IEEE Transactions on
Volume 21, Issue 12, Dec 2002 Page(s): 1552 – 1563
[17] Osmar R. Zaïane, Maria-Luiza Antonie, Alexandru Coman: Mammography Classification By an Association Rule-based Classifier. MDM/KDD 2002: 62-69
[18] AI book
[19] http://en.wikipedia.org/wiki/Neuron
[20] http://en.wikipedia.org/wiki/Tournament_selection
[20] Ganesh Venayagamoorthy, Viresh Moonasar, Kumbes Sandrasegaran, "Voice Recognition Using Neural Networks" : Communications and Signal Processing, 1998. COMSIG '98. Proceedings of the 1998 South African Symposium on 7-8 Sep 1998
[21] Neural Networks for Prediction and Classification, Encyclopedia of Data Warehousing and Mining, Volume II, I-Z by John Wang (ed)

[22] Support Vector Machines, Encyclopedia of Data Warehousing and Mining, Volume II, I-Z by John Wang (ed)
[23] Text Categorization with Support Vector Machines: Learning with Many Relevant Features (1997) Thorsten Joachims, Proceedings of ECML-98, 10th European Conference on Machine Learning
[24] The Limitations of Decision Trees and Automatic Learning in Real World Medical Decision Making, M Zorman, MM Štiglic, P Kokol, I Malčić - Journal of Medical Systems, 1997 – Springer
[25] http://www.genetic-programming.com/humancompetitive.html
[26] "Neuroimaging of gender differences in alcoholism: Are women more vulnerable?," Alcoholism: Clinical & Experimental Research (ACER), May 2005 ??? (I think) Klaus Ackermann, Bernhard Croissant, Helmut Nakovics, and Alexander Diehl
[27] http://en.wikipedia.org/wiki/Receiver_operating_characteristic
[28] Thomas G. Tape, Interpreting Diagnostic Tests, http://gim.unmc.edu/dxtests/Default.htm
[29] Edwin D. de Jong: The Incremental Pareto-Coevolution Archive. GECCO (1) 2004: 525-536
[30] Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.
[31]
[32] McCance DR, Hanson RL, Charles MA, Jacobsson LTH, Pettitt DJ, Bennett PH, et al. Comparison of tests for glycated haemoglobin and fasting and two hour plasma glucose concentrations as diagnostic methods for diabetes. BMJ 1994;308:1323-8. (21 May.)
[33] http://www.americanheart.org/presenter.jhtml?identifier=4489
[34] http://www.pennhealth.com/ency/article/003482.htm
[35] A Genome Compiler for High Performance Genetic Programming (1998) Alex Fukunaga, Andre Stechert, Darren Mutz
Genetic Programming 1998: Proceedings of the Third Annual Conferencexsz
[35] Defining the Relationship Between Plasma Glucose and HbA1c
Analysis of glucose profiles and HbA1c in the Diabetes Control and Complications Trial Curt L. Rohlfing, BES, Hsiao-Mei Wiedmeyer, MS, Randie R. Little, PHD, Jack D. England, Alethea Tennill, MS and David E. Goldstein, MD
[36] Is fasting blood glucose a reliable parameter for screening for diabetes in hypertension? Andreas Bura, Harald Herknera, Christian Woisetschlägera, Marianne Vlceka, Ulla Derhaschniga and Michael M. Hirschl
[37]'Life course determinants of insulin secretion and sensitivity at age 50 years: the Newcastle Thousand Families Study'. Pearce, M.S et al
[38]
[39] A comparative assessment of classification methods, Kiang M.Y.1. Decision Support Systems, Volume 35, Number 4, July 2003, pp. 441-454
[40] Data mining cardiovascular Bayesian networks Charles R. Twardy, Ann E. Nicholson, Kevin B. Korb, John McNeil
[41] Mining Medical Records for Computer Aided Diagnosis, R. Bharat Rao, Romer Rosales, Stefan Niculescu, Sriram Krishnan Luca Bogoni, Xiang S. Zhou, Balaji Krishnapuram
[43] Mining Genetic Epidemiology Data with Bayesian Networks Application to *APOE* Gene Variation and Plasma Lipid Levels, ANDREI RODIN,[1] THOMAS H. MOSLEY, JR.,[2] ANDREW G. CLARK,[3] CHARLES F. SING,[4] and ERIC BOERWINKLE[1,5]
[44] Learning Bayesian Networks is NP-Hard (1994) David Chickering, Dan Geiger, David Heckerman

[45] Multiobjective optimization and multiple constraint handling withevolutionary algorithms. I. A unified formulation Fonseca, C.M. Fleming, P.J. Dept. of Autom. Control & Syst. Eng., Sheffield Univ.

[46] Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.