New Classification Models through Evolutionary Algorithms

Alberto Cano
Department of Computer Science and Numerical Analysis
University of Cordoba, Spain
acano@uco.es

ABSTRACT

This Doctoral Thesis presents new computational models on data classification which address new open problems and challenges in data classification by means of evolutionary algorithms. Specifically, we pursue to improve the performance, scalability, interpretability and accuracy of classification models on challenging data. The performance and scalability of evolutionary-based classification models were improved through parallel computation on GPUs, which demonstrated to achieve high efficiency on speeding up classification algorithms. The conflicting problem of the interpretability and accuracy of the classification models was addressed through a highly interpretable classification algorithm which produced very comprehensible classifiers by means of classification rules. Performance on challenging data such as the imbalanced classification was improved by means of a data gravitation classification algorithm which demonstrated to achieve better classification performance both on balanced and imbalanced data. All the methods proposed in this Thesis were evaluated in a proper experimental framework, by using a large number of data sets with diverse dimensionality and by comparing their performance against other state-of-the-art and recently published methods of proved quality. The experimental results obtained have been verified by applying non-parametric statistical tests which support the better performance of the methods proposed.

Full PDF available at http://www.uco.es/users/i52caroa/Thesis%20Alberto%20Cano.pdf

REFERENCES

[1] A. Cano, A. Zafra, and S. Ventura, "Speeding up the evaluation phase of GP classification algorithms on GPUs," *Soft Computing*, vol. 16, no. 2, pp. 187–202, 2012.

[2] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, pp. 37–54, 1996.

[3] A. Cano, A. Zafra, and S. Ventura, "Speeding up multiple instance learning classification rules on GPUs," *Knowledge and Information Systems*, vol. 44, no. 1, pp. 127–145, 2015.

[4] A. K. Jain, R. P. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[5] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley, 2005.

[6] A. Cano, A. Zafra, and S. Ventura, "An interpretable classification rule mining algorithm," *Information Sciences*, vol. 240, pp. 1–20, 2013.

[7] A. Ghosh and L. Jain, Eds., *Evolutionary Computation in Data Mining*, ser. Studies in Fuzziness and Soft Computing. Springer, 2005, vol. 163.

[8] A. Abraham, E. Corchado, and J. M. Corchado, "Hybrid learning machines," *Neurocomputing*, vol. 72, no. 13-15, pp. 2729–2730, 2009.

[9] E. Corchado, M. Graña, and M. Wozniak, "New trends and applications on hybrid artificial intelligence systems," *Neurocomputing*, vol. 75, no. 1, pp. 61–63, 2012.

[10] E. Corchado, A. Abraham, and A. de Carvalho, "Hybrid intelligent algorithms and applications," *Information Sciences*, vol. 180, no. 14, pp. 2633–2634, 2010.

[11] W. Pedrycz and R. A. Aliev, "Logic-oriented neural networks for fuzzy neurocomputing," *Neurocomputing*, vol. 73, no. 1-3, pp. 10–23, 2009.

[12] A. A. Freitas, *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2002.

[13] X. Yu and M. Gen, *Introduction to Evolutionary Algorithms*. Springer, 2010.

[14] P. Espejo, S. Ventura, and F. Herrera, "A Survey on the Application of Genetic Programming to Classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 2, pp. 121–144, 2010.

[15] J. Landry, L. D. Kosta, and T. Bernier, "Discriminant feature selection by genetic programming: Towards a domain independent multi-class object detection system," *Journal of Systemics, Cybernetics and Informatics*, vol. 3, no. 1, 2006.

[16] I. De Falco, A. Della Cioppa, F. Fontanella, and E. Tarantino, "An Innovative Approach to Genetic Programming-based Clustering," in *Proceedings of the 9th Online World Conference on Soft Computing in Industrial Applications*, 2004.

[17] E. Alba and M. Tomassini, "Parallelism and evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 5, pp. 443–462, 2002.

[18] G. Luque and E. Alba, *Parallel Genetic Algorithms: Theory and Real World Applications*, ser. Studies in Computational Intelligence. Springer, 2011.

[19] P. E. Srokosz and C. Tran, "A distributed implementation of parallel genetic algorithm for slope stability evaluation," *Computer Assisted Mechanics and Engineering Sciences*, vol. 17, no. 1, pp. 13–26, 2010.

[20] M. Rodríguez, D. M. Escalante, and A. Peregrín, "Efficient Distributed Genetic Algorithm for Rule extraction," *Applied Soft Computing*, vol. 11, no. 1, pp. 733–743, 2011.

[21] S. Dehuri, A. Ghosh, and R. Mall, "Parallel multi-objective genetic algorithm for classification rule mining," *IETE Journal of Research*, vol. 53, no. 5, pp. 475–483, 2007.

[22] A. Folling, C. Grimme, J. Lepping, and A. Papaspyrou, "Connecting Community-Grids by supporting job negotiation with coevolutionary Fuzzy-Systems," *Soft Computing*, vol. 15, no. 12, pp. 2375–2387, 2011.

[23] P. Switalski and F. Seredynski, "An efficient evolutionary scheduling algorithm for parallel job model in grid environment," *Lecture Notes in Computer Science*, vol. 6873, pp. 347–357, 2011.

[24] NVIDIA Corporation, "NVIDIA CUDA Programming and Best Practices Guide, http://www.nvidia.com/cuda," 2013.

[25] J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Kruger, A. E. Lefohn, and T. J. Purcell, "A survey of general-purpose computation on graphics hardware," *Computer Graphics Forum*, vol. 26, no. 1, pp. 80–113, 2007.

[26] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips, "GPU Computing," *Proceedings of the IEEE*, vol. 96, no. 5, pp. 879–899, 2008.

[27] S. Che, M. Boyer, J. Meng, D. Tarjan, J. W. Sheaffer, and K. Skadron, "A performance study of general-purpose applications on graphics processors using CUDA," *Journal of Parallel and Distributed Computing*, vol. 68, no. 10, pp. 1370–1380, 2008.

[28] D. M. Chitty, "Fast parallel genetic programming: Multi-core CPU versus many-core GPU," *Soft Computing*, vol. 16, no. 10, pp. 1795–1814, 2012.

[29] S. N. Omkar and R. Karanth, "Rule extraction for classification of acoustic emission signals using Ant Colony Optimisation," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 8, pp. 1381–1388, 2008.

[30] K. L. Fok, T. T. Wong, and M. L. Wong, "Evolutionary computing on consumer graphics hardware," *IEEE Intelligent Systems*, vol. 22, no. 2, pp. 69–78, 2007.

[31] L. Jian, C. Wang, Y. Liu, S. Liang, W. Yi, and Y. Shi, "Parallel data mining techniques on Graphics Processing Unit with Compute Unified Device Architecture (CUDA)," *The Journal of Supercomputing*, pp. 1–26, 2011.

[32] A. Cano, A. Zafra, and S. Ventura, "A parallel genetic programming algorithm for classification," in *Proceedings of the 6th International Conference on Hybrid Artificial Intelligent Systems (HAIS). Lecture Notes in Computer Science*, vol. 6678 LNAI, no. PART 1, 2011, pp. 172–181.

[33] M. Paliwal and U. Kumar, "Neural Networks and Statistical Techniques: A Review of Applications," *Expert Systems with Applications*, vol. 36, no. 1, pp. 2–17, 2009.

[34] C. Campbell, "Kernel methods: A survey of current techniques," *Neurocomputing*, vol. 48, pp. 63–84, 2002.

[35] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991.

[36] S. Tsumoto, "Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model," *Information Sciences*, vol. 162, no. 2, pp. 65–80, 2004.

[37] D. Martens, B. Baesens, T. V. Gestel, and J. Vanthienen, "Comprehensible Credit Scoring Models using Rule Extraction from Support Vector Machines," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1466–1476, 2007.

[38] S. Alonso, E. Herrera-Viedma, F. Chiclana, and F. Herrera, "A web based consensus support system for group decision making problems and incomplete preferences," *Information Sciences*, vol. 180, no. 23, pp. 4477–4495, 2010.

[39] N. Xie and Y. Liu, "Review of Decision Trees," in *Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, vol. 5, 2010, pp. 105–109.

[40] D. Richards, "Two Decades of Ripple Down Rules Research," *Knowledge Engineering Review*, vol. 24, no. 2, pp. 159–184, 2009.

[41] J. Cano, F. Herrera, and M. Lozano, "Evolutionary Stratified Training Set Selection for Extracting Classification Rules with trade off Precision-Interpretability," *Data and Knowledge Engineering*, vol. 60, no. 1, pp. 90–108, 2007.

[42] S. García, A. Fernández, J. Luengo, and F. Herrera, "A Study of Statistical Techniques and Performance Measures for Genetics-based Machine Learning: Accuracy and Interpretability," *Soft Computing*, vol. 13, no. 10, pp. 959–977, 2009.

[43] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," *Decision Support Systems*, vol. 51, pp. 141–154, 2011.

[44] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354–2364, 2011.

[45] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[46] H. Kaizhu, Y. Haiqin, K. Irwinng, and M. R. Lyu, "Imbalanced learning with a biased minimax probability machine," *IEEE Transactions on Systems and Man and and Cybernetics and Part B: Cybernetics*, vol. 36, no. 4, pp. 913–923, 2006.

[47] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proceedings of 15th European Conference on Machine Learning*, 2004, pp. 39–50.

[48] S.-H. Wu, K.-P. Lin, H.-H. Chien, C.-M. Chen, and M.-S. Chen, "On generalizable low false-positive learning using asymmetric support vector machines," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 1083–1096, 2013.

[49] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[50] I. Kononenko and M. Kukar, *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Combridge, U.K.: Horwood Publ., 2007.

[51] B. Li, Y. W. Chen, and Y. Q. Chen, "The nearest neighbor algorithm of local probability centers," *IEEE Transactions on Systems and Man and and Cybernetics and Part B: Cybernetics*, vol. 38, no. 1, pp. 141–154, 2008.

[52] L. Peng, B. Peng, Y. Chen, and A. Abraham, "Data gravitation based classification," *Information Sciences*, vol. 179, no. 6, pp. 809–819, 2009.

[53] C. Wang and Y. Q. Chen, "Improving nearest neighbor classification with simulated gravitational collapse," in *Proceedings the International Conference on Computing, Networking and Communications*, vol. 3612, 2005, pp. 845–854.

[54] Y. Zong-Chang, "A vector gravitational force model for classification," *Pattern Analysis and Applications*, vol. 11, no. 2, pp. 169–177, 2008.

[55] D. J. Newman and A. Asuncion, "UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences," 2007. [Online]. Available: http://archive.ics.uci.edu/ml/

[56] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, pp. 255–287, 2011.

[57] S. Ventura, C. Romero, A. Zafra, J. A. Delgado, and C. Hervás, "JCLEC: a Java framework for evolutionary computation," *Soft Computing*, vol. 12, pp. 381–392, 2007.

[58] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemannr, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[59] R. Kohavi, "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the 14th international joint conference on Artificial intelligence (1995)*, vol. 2, 1995, pp. 1137–1143.

[60] T. Wiens, B. Dale, M. Boyce, and G. Kershaw, "Three way k-fold cross-validation of resource selection functions," *Ecological Modelling*, vol. 212, no. 3-4, pp. 244–255, 2008.

[61] O. J. Dunn, "Multiple comparisons among means," *American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.

[62] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[63] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Information Sciences*, vol. 180, no. 10, pp. 2044–2064, 2010.

[64] S. García, D. Molina, M. Lozano, and F. Herrera, "A Study on the use of Non-parametric Tests for Analyzing the Evolutionary Algorithms' Behaviour: A Case Study," *Journal of Heuristics*, vol. 15, pp. 617–644, 2009.

[65] A. Cano, A. Zafra, and S. Ventura, "Solving classification problems using genetic programming algorithms on GPUs," in *Proceedings of the 5th International Conference on Hybrid Artificial Intelligent Systems (HAIS). Lecture Notes in Computer Science*, vol. 6077 LNAI, no. PART 2, 2010, pp. 17–26.

[66] I. De Falco, A. Della Cioppa, and E. Tarantino, "Discovering interesting classification rules with genetic programming," *Applied Soft Computing*, vol. 1, no. 4, pp. 257–269, 2001.

[67] C. Bojarczuk, H. Lopes, A. Freitas, and E. Michalkiewicz, "A Constrained-syntax Genetic Programming System for Discovering Classification Rules: Application to Medical Datasets," *Artificial Intelligence in Medicine*, vol. 30, no. 1, pp. 27–48, 2004.

[68] K. C. Tan, A. Tay, T. H. Lee, and C. M. Heng, "Mining multiple comprehensible classification rules using genetic programming," in *Proceedings of the IEEE Congress on Evolutionary Computation*, vol. 2, 2002, pp. 1302–1307.

[69] M. A. Franco, N. Krasnogor, and J. Bacardit, "Speeding up the evaluation of evolutionary learning systems using GPGPUs," in *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, ser. GECCO '10, 2010, pp. 1039–1046.

[70] A. Cano, A. Zafra, and S. Ventura, "Parallel evaluation of Pittsburgh rule-based classifiers on GPUs," *Neurocomputing*, vol. 126, pp. 45–57, 2014.

[71] A. Zafra and S. Ventura, "G3P-MI: A Genetic Programming Algorithm for Multiple Instance Learning," *Information Sciences*, vol. 180, pp. 4496–4513, 2010.

[72] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, pp. 31–71, 1997.

[73] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *Knowledge Engineering Review*, vol. 25, no. 1, pp. 1–25, 2010.

[74] A. Cano, A. Zafra, and S. Ventura, "An EP algorithm for learning highly interpretable classifiers," in *Proceedings of the 11th International Conference on Intelligent Systems Design and Applications, ISDA'11*, 2011, pp. 325–330.

[75] J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation, Chapter 4: Context-Free Grammars*. Addison-Wesley, 2006.

[76] M. L. Wong and K. S. Leung, *Data Mining Using Grammar Based Genetic Programming and Applications*. Kluwer Academic Publisher, 2000.

[77] J. Quinlan, *C4.5: Programs for Machine Learning*, 1993.

[78] J. Bacardit and N. Krasnogor, "Performance and Efficiency of Memetic Pittsburgh Learning Classifier Systems," *Evolutionary Computation*, vol. 17, no. 3, pp. 307–342, 2009.

[79] C. Márquez-Vera, A. Cano, C. Romero, and S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data," *Applied Intelligence*, vol. 38, no. 3, pp. 315–330, 2013.

[80] A. Cano, A. Zafra, and S. Ventura, "Weighted data gravitation classification for standard and imbalanced data," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1672–1687, 2013.

[81] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

[82] N. Hansen, "The CMA evolution strategy: A comparing review," in *Towards a New Evolutionary Computation. Advances on Estimation of Distribution Algorithms*, J. A. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, Eds. New York: Springer-Verlag, 2006, pp. 75–102.

[83] C. Marquez-Vera, A. Cano, C. Romero, A. Y. Mohammad, H. M. Fardoun, and S. Ventura, "Early dropout prediction using data mining: A case study with high school students," *Expert Systems*, vol. 33, no. 1, pp. 107–124, 2016.

[84] J. Alcalá-Fdez, L. Sánchez, S. García, M. del Jesus, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V. Rivas, J. Fernández, and F. Herrera, "KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems," *Soft Computing*, vol. 13, pp. 307–318, 2009.

[85] A. Cano and S. Ventura, "Gpu-parallel subtree interpreter for genetic programming," in *Proceedings of the Conference on Genetic and Evolutionary Computation*, 2014, pp. 887–894.

[86] A. Cano, J. Luna, E. Gibaja, and S. Ventura, "LAIM discretization for multi-label data," *Information Sciences*, vol. 330, pp. 370–384, 2016.

[87] A. Cano, S. Ventura, and K. Cios, "Multi-objective genetic programming for feature extraction and data visualization," *Soft Computing*, vol. 21, no. 8, pp. 2069–2089, 2017.

[88] A. Ben-David, "Comparison of classification accuracy using Cohen's weighted kappa," *Expert Systems with Applications*, vol. 34, no. 2, pp. 825–832, 2008.
[89] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.
[90] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.
[91] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. London, U.K.: Chapman & Hall/CRC, 2007.
[92] A. Cano, J. M. Luna, A. Zafra, and S. Ventura, "A Classification Module for Genetic Programming Algorithms in JCLEC," *Journal of Machine Learning Research*, vol. 16, pp. 491–494, 2015.
[93] A. Cano, J. Olmo, and S. Ventura, "Parallel multi-objective ant programming for classification using gpus," *Journal of Parallel and Distributed Computing*, vol. 73, no. 6, pp. 713–728, 2013.
[94] A. Cano, J. M. Luna, and S. Ventura, "High performance evaluation of evolutionary-mined association rules on gpus," *Journal of Supercomputing*, vol. 66, no. 3, pp. 1438–1461, 2013.
[95] A. Cano, S. Ventura, and K. Cios, "Scalable CAIM discretization on multiple GPUs using concurrent kernels," *Journal of Supercomputing*, vol. 69, no. 1, pp. 273–292, 2014.
[96] A. Cano, D. T. Nguyen, S. Ventura, and K. J. Cios, "ur-CAIM: improved CAIM discretization for unbalanced and balanced data," *Soft Computing*, vol. 20, no. 1, pp. 173–188, 2015.