Performance Evaluation of Evolutionary and Decision Tree Based Classifiers in Diversity of Datasets

Pardeep Kumar[1], Vivek Kumar Sehgal[1], Nitin[1], and Durg Singh Chauhan[2]

[1] Department of Computer Science & Engineering,
Jaypee University of Information Technology, Waknaghat, Solan (H.P), India
[2] Department of Computer Science & Engineering, Institute of Technology, Banaras Hindu University, Banaras(U.P), India. Currently with Uttrakhand Technical University, Dehradun (UK), India
pardeepkumarkhokhar@gmail.com,
{vivekseh,delnitin}@ieee.org, pdschauhan@acm.org

Abstract. The large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls and many more applications generate streams of digital records archived in huge databases, sometimes in so-called data warehouses A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining. Data mining plays an important role to discover important information to help in decision making of a decision support system. It has been the active area of research in the last decade. The classification is one of the important tasks of data mining. Different kind of classifiers have been suggested and tested to predict the future events based on unseen data. This paper compares the performance evaluation of evolutionary based genetic algorithm and decision tree based classifiers in diversity of datasets. The performance evaluation metrics are predictive accuracy, training time and comprehensibility. Evolutionary based classifier shows better comprehensibility over decision tree based classifiers. These classifiers show almost same predictive accuracy. Experimental results demonstrate that evolutionary approach based classifiers are slower than decision tree based classifiers. This research is helpful for organizations to select the classifiers as information generator for their decision support systems to make future policies.

Keywords: Knowledge Discovery in Databases, Evolutionary Computation, Information Gain, Classification.

1 Introduction

Information plays a vital role in business organizations. Today's business is information hungry. Information can be used by the top level management for

V.V. Das and J. Stephen (Eds.): CNC 2012, LNICST 108, pp. 309–312, 2012.

decision making to make future policies. Due to increasing size of organizations data rapidly, manual interpretation of data for information discovery is not feasible.

Over the last three decades, data mining has been growing on the map of computer science. It deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases. Data mining is regarded as the key element of a much more elaborate process called **Knowledge Discovery in Databases (KDD)** which is defined as the non – trivial process of identifying valid, novel, and ultimately understandable patterns in large databases [1]. One of the important tasks of data mining is classification. The conventional classifiers used for classification are decision trees, neural network, statistical and clustering techniques. There is lot of research going in the machine learning and statistics communities on classifiers for classification. In the recent past, there has been an increasing interest in applying evolutionary methods to Knowledge Discovery in Databases (KDD) and a number of successful applications of Genetic Algorithms (GA) and Genetic Programming (GP) to KDD have been demonstrated.

The STATLOG Project[2] finds that no classifier is uniformly most accurate over the datasets studied and many classifiers possess comparable accuracy. Earlier comparative studies put emphasis on the predictive accuracy of classifiers; other factors like comprehensibility and classification index are also becoming important. Breslow and Aha have surveyed methods of decision tree simplification to improve their comprehensibility [3]. Brodley and Utgoff , Brown, Corruble, and Pittard, Curram and Mingers, and Shavlik, Mooney and Towell have also done comparative studies in the domain of classifiers[4-7]. Saroj and K.K Bhardwaj have done excellent work on GA's ability to discover production rules and censor based production rules [8]. No single method has been found to be superior over all others for all datasets. Issues such as accuracy, training time, robustness and scalability must be considered and can involve tradeoffs, further complicating the quest for an overall superior method.

This paper compares evolutionary approach based genetic algorithm and decision tree based classifiers (CHAID, QUEST and C4.5) on four datasets (Mushroom,Vote,Nursery and Credit) that are taken from the University of California, Irvine, Repository of Machine Learning Databases (UCI) [9].

2 The Classifiers

CHAID, QUEST and C5.0 are decision tree based classifiers [10-12]. Genetic algorithm is the evolutionary approach based classifier [13-17].

3 Experimental Setup

There are four datasets (Mushroom, Vote, Nursery and Credit) used in this research work from real domain. These datasets are available from UCI machine learning repository [9]. Predictive accuracy, training time and comprehensibility are the parameters used for performance evaluation of the underlying classifiers [1, 10-11]. Decision tree based classifiers have been tested using Clementine 10.1 with window XP platform. GA has been tested using GALIB 245 simulator on Linux Platform.

4 Results

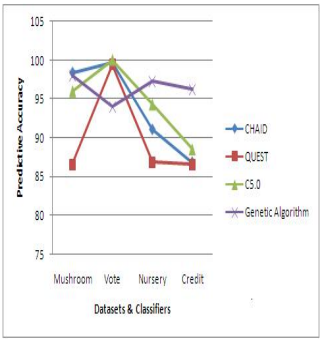

Fig. 1. PA

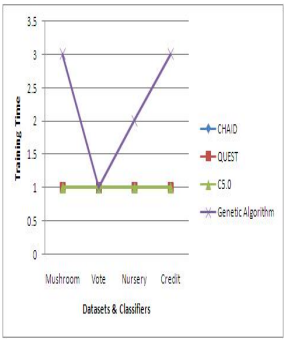

Fig. 2. TT

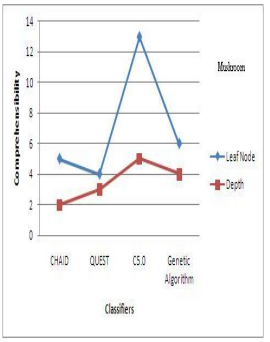

Fig. 3. CM

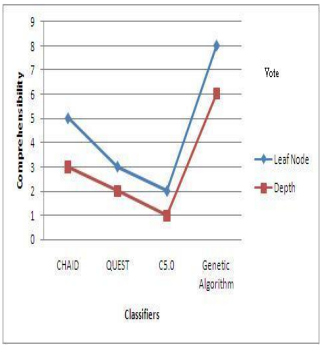

Fig. 4. CV

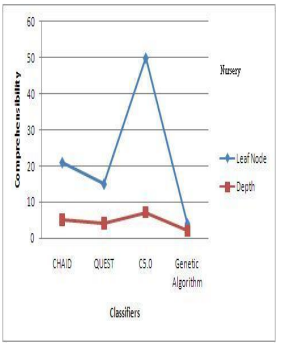

Fig. 5. CN

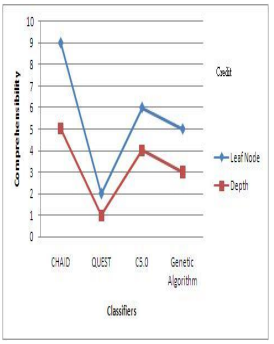

Fig. 6. CC

PA-Predictive Accuracy TT-Training Time CM-Mushroom Comprehensibility CV-Vote Comprehensibility CN-Nursery Comprehensibility CC-Credit Comprehensibility

5 Conclusion

Experimental results given in result section demonstrate that evolutionary approach based genetic algorithm classifier will remain the first choice when predictive accuracy be the selection criteria as it is independent from the domain and size of the datasets. Organizations can rely on genetic algorithm and QUEST when comprehensibility be the selection criteria. Evolutionary approach based classifier scored poor in context of speed as the selection criteria. So, this paper is helpful for organizations to select data mining product for their decision support systems to make policies for future.

References

1. Smith, U.M.F., Shapiro, G.P., Smyth, P.: The KDD process for extracting useful knowledge from volumes from data. Communication of ACM 39(11), 27–34 (1996)
2. King, R.D., Feng, C., Sutherland, A.: STATLOG. Comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence 9(3), 289–333 (1995)
3. Breslow, L.A., Aha, D.W.: Simplifying decision trees: A survey. Knowledge Engineering Review 12, 1–40 (1997)
4. Brodley, C.E., Utgoff, P.E.: Multivariate versus univariate decision trees, Department of Computer Science, University of Massachusetts, Amherst, MA. Technical Report 92-8 (1992)
5. Brown, D.E., Corruble, V., Pittard, C.L.: A comparison of decision tree classifiers with back propagation neural networks for multimodal classification problems. Pattern Recognition 26, 953–961 (1993)
6. Curram, S.P., Mingers, J.: Neural networks, decision tree induction and discriminant analysis: An empirical comparison. Journal of the Operational Research Society 45, 440–450 (1994)
7. Shavlik, J.W., Mooney, R.J., Towell, G.G.: Symbolic and neural learning algorithms: an empirical comparison. Machine Learning 6, 111–144 (1991)
8. Saroj, Bhardwaj, K.K.: A parallel genetic algorithm approach for automated discovery of censored production rules. In: AIAP 2007 Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications, pp. 435–441 (2007)
9. UCI Repository of Machine Learning Databases. Department of Information and Computer Science University of California (1994), `http://www.ics.uci.edu/~mlearn/MLRepositry.html`
10. Han, J., Kamber, M.: Data mining: concepts and techniques: Book (Illustrated), 550 pages (January 2001) ISBN-10: 1558604898, ISBN-13: 9781558604896
11. Quinlan, J.R.: Induction in decision trees. Journal of Machine Learning 1(1), 81–106 (2003)
12. Lim, T.S., Loh, W.Y., Shih, Y.S.: A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Journal of Machine Learning 40, 203–228 (2000)
13. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley (1989)
14. Deb, K.: Genetic Algorithm in search and optimization: The techniques and Applications. In: Proceeding of Advanced Study Institute on Computational Methods for Engineering Analysis and Design, pp. 12.1–12.25 (1993)
15. Saroj: Genetic Algorithm: A technique to search complex space. In: Proceedings of National Seminar on Emerging Dimension in Information Technology, August 10-11, pp. 100–105 (2002)
16. Frietas, A.A.: A survey of evolutionary algorithms for data mining and knowledge discovery, pp. 819–845. Springer, New York (2003)
17. Frietas, A.A.: Data mining and knowledge discovery with evolutionary algorithms, 265 pages (2002) ISBN: 978-3-540-43331-6