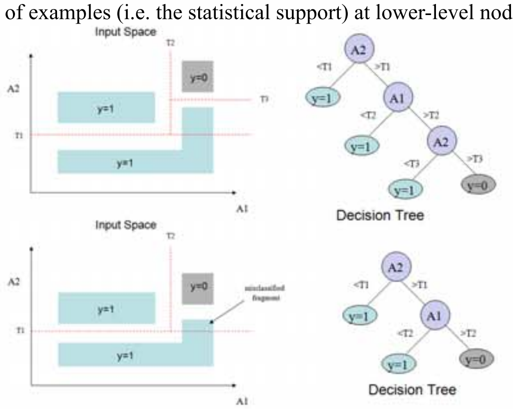
The Effect of the Fragmentation Problem in Decision Tree Learning Applied to the Search for Single Top Quark Production

R Vilalta1, R Valerio2, F Ocegueda-Hernandez1, G Watts3 and M Siller2
1 Department of Computer Science, University of Houston, 4800 Calhoun Rd., Houston Texas 77204-3010, USA
2 Center for Research and Advanced Studies, CINVESTAV, Avenida Cient´ıfica No. 1145, Zapopan Jalisco, 45015, M´exico
3 Department of Physics, University of Washington, Seattle, Washington, USA
E-mail: vilalta@cs.uh.edu, rvalerio@gdl.cinvestav.mx

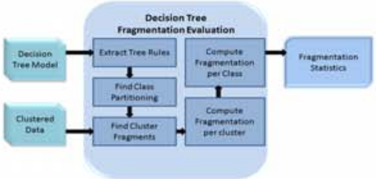
Abstract. Decision tree learning constitutes a suitable approach to classification due to its ability to partition the input (variable) space into regions of class-uniform events, while providing a structure amenable to interpretation (as opposed to other methods such as neural networks). But an inherent limitation of decision tree learning is the progressive lessening of the statistical support of the final classifier as clusters of single-class events are split on every partition, a problem known as the fragmentation problem. We describe a software system called DTFE (Decision Tree Fragmentation Evaluator) that measures the degree of fragmentation caused by a decision tree learner on every event cluster. Clusters are found through a decomposition of the data using a technique known as Spectral Clustering. Each cluster is analyzed in terms of the number and type of partitions induced by the decision tree. Our domain of application lies on the search for single top quark production, a challenging problem due to large backgrounds (similar to W+jets and $t\bar{t}$ events), low energetic signals, and low number of jets. The output of the machine-learning software tool consists of a series of statistics describing the degree of classification error attributed to the fragmentation problem.

Introduction

Decision tree learning algorithms stand as a popular non-parametric approach to classification; the general idea is to have the input or variable space iteratively partitioned into smaller regions until each region exhibits an approximately uniform class distribution. Decision tree learning algorithms have gained increasing acceptance in the physics community because of several factors: the user is relieved from establishing parametric model assumptions; the output is amenable to interpretation; accuracy performance tends to be competitive when compared to other techniques; and CPU cost during training is relatively low. Nevertheless, an inherent limitation exists, also known as the fragmentation problem, in which the continuous partitioning of the training set at every tree node reduces the number of examples (i.e. the statistical support) at lower-level nodes [1].



Decision Tree Fragmentation Evaluator



Experiments on Single Top Quark Production

We performed a pre-processing step before actual analysis. To avoid the class imbalance problem we enforced all signal and background samples to have the same size.

This introduces an assumption of equal priors on the classes that may come unwarranted; nevertheless it helps in avoiding multiple misclassifications on classes represented by small samples.

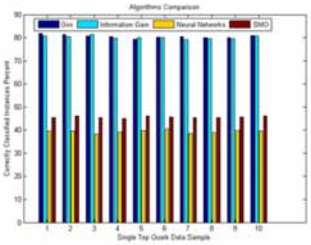
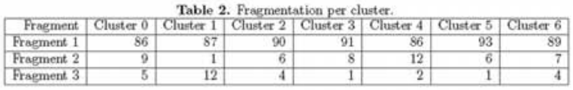


Table 2. Fragmentation per cluster.



Fragment	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Fragment 1	86	87	90	91	86	93	89
Fragment 2	9	1	6	8	12	6	7
Fragment 3	5	12	4	1	2	1	4

Summary and Conclusions

DFTE, Decision Tree Fragmentation Evaluator, is a system that measures the degree of fragmentation exerted by a decision tree classifier on a particular input-output distribution.

The DFTE system outputs a set of statistics that provide useful information on the quality of the decision tree model.Our goal is to understand the behavior of a decision tree classifier under different input-output distributions.

It is important to determine when a classifier is unable to improve on accuracy, either because it is close to Bayes error or because the bias imposed by the classifier is high, and the best chosen model is far from the true concept.

References

[1] Vilalta R, Blix G and Rendell L 1997 Global Data Analysis and the Fragmentation Problem in Decision Tree Induction Proceedings of the 9th European Conference on Machine Learning ECML (Heinderberg: Springer-Verlag) p 312

[2] Li J and Wong L 2002 Solving the Fragmentation Problem of Decision Trees by Discovering Boundary Emerging Patterns Proceedings of the 2002 IEEE International Conference on Data Mining ICDM (Washington DC, USA: IEEE Computer Society) p 653

[3] Liu B, Hu M and Hsu W 2000 Intuitive Representation of Decision Trees Using General Rules and Exceptions Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (AAAI Press; MIT Press) p 615

[4] Hand D J 2006 Classifier Technology and the Illusion of Progress Statistical Science 21 1

[5] Ho K M and Scott P D 1998 Overcoming Fragmentation in Decision Trees Through Attibute Value Grouping Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery (London, UK: Springer-Verlag) p 337

[6] Liu H and Setiono R 1998 Feature Transformation and Multivariate Decision Tree Induction Proceedings of the First International Conference on Discovery Science (London, UK: Springer-Verlag) p 279

[7] DeLisle R K and Dixon S L 2004 Induction of Decision Trees via Evolutionary Programming Journal of Chemical Information and Modeling 44 862