Literature Review on the paper “induction of decision trees “

This paper focusses on one microcosm of machine learning and on a family of learning systems that have been used to build knowledge-based systems of a simple kind.It outlines the features of this family and introduces its members. All these systems address the same task of inducing decision trees from examples.It summarizes an approach to synthesizing decision trees that has been used in a variety of systems, and it describes one such system, ID3, in detail. Current commercial systems are powerful tools that have achieved noteworthy successes. The groundwork has been done for advances that will permit such tools to deal even with noisy, incomplete data typical of advanced real-world applications. The paper also reviews the central facet of the induction algorithm and reveals possible improvements in the field.

The practical importance of machine learning of this latter kind has been underlined by he advent of knowledge-based expert systems. People came up with a quite different approach that sees learning as the acquisition of structured knowledge in the form of concepts (Hunt, 1962; Winston,1975),discrimination nets (Feigenbaum and Simon, 1963), or production rules (Buchanan, 1978). The knowledge needed to drive the pioneering expert systems was codified through protracted interaction between a domain specialist and a knowledge engineer. “It is obvious that the interview approach to knowledge acquisition cannot keep pace with the burgeoning demand for expert systems”; Feigen- Baum (1981) terms this the 'bottleneck' problem. While the typical rate of knowledge elucidation by this method is a few rules per man day, an expert system for a complex task may require hundreds or even thousands of such rules. More concretely, learning provides a potential methodology for building high performance systems. This approach, characteristic of a large proportion of the early learning work, produced self-improving programs for playing games, balancing poles, solving problems and many other domains. This perception has stimulated the investigation of machine learning methods as a means of explicating knowledge (Michie, 1983). Carbonell, Michalski and Mitchell (1983) classified machine learning on three principle dimensions.Since artificial intelligence first achieved recognition as a discipline in the mid 1950's, machine learning has been a central research area. At one extreme there are adaptive systems that monitor their own performance and attempt to improve it by adjusting internal parameters.

The paper summarises an approach to synthesizing decision trees that has been used in a variety of systems. In one classification problem studied, this method reduced a totally opaque, large decision tree to a hierarchy of nine small decision trees, each of which 'made sense' to an expert. The task is solved in terms of a collection of notional super-attributes, after which the subtasks of inducing classification rules to find the values of the super-attributes are approached in the same top-down fashion. It is this lack of familiarity (and perhaps an underlying lack of modularity) that is the chief obstacle to the use of induction for building large expert systems.The aim of this paper has been to demonstrate that the technology for building decision trees from examples is fairly robust. While decision trees generated by the above systems are fast to execute and can be very accurate, they leave much to be desired as representations of knowledge. The groundwork has been done for advances that will permit such tools to deal even with noisy, incomplete data typical of advanced real-world applications.

The paper describes several experiments that were carried out to validate their claim of assessing predictive accuracy. In a simpler domain (1,987 objects with a correct decision tree of 48 nodes), randomly selected training sets containing 20% of the objects gave decision trees that correctly classified 98% of the unseen objects. In another version of the same domain, 39 attributes gave 551 distinct objects with a correct decision tree of similar size. Training sets of 20% of these 551 objects gave decision trees of almost identical accuracy. The worth of ID3's attribute-selecting heuristic can be assessed by the simplicity of the resulting decision trees, or, more to the point, by how well those trees express real relationships between class and attributes as demonstrated by the accuracy with which they classify objects other than those in the training set (their predictive accuracy). When training sets containing 20% of these 715 objects were chosen at random, they produced decision trees that correctly classified over 84% of the unseen objects. A straightforward method of assessing this predictive accuracy is to use only a part of the given set of objects as a training set, and to check the resulting decision tree on the remainder. ID3's total computational requirement per iteration is thus proportional to the product of the size of the training set, the number of attributes and the number of non-leaf nodes in the decision tree. This domain is relatively complex since a correct decision tree for all 715 objects contains about 150 nodes. In one domain, 1.4 million Chess positions described in terms of 49 binary valued attributes gave rise to 715 distinct objects divided 65% :35% between the classes.

The paper puts forth several developments that are taking place in the fields of machine learning and a family of learning systems that have been used to build knowledge-based systems while a variety of systems are known, the paper describes the system ID3 in detail, while its mentioned that the ID3 algorithm is widely used to generate a decision tree from a dataset, the paper fails to mention the drawback of the system, ID3 does not always guarantee an optimal solution.It can converge upon local optima. It uses a greedy strategy by selecting the locally best attribute to split the dataset on each iteration.
The optimality of the algorithm can be improved by backtracking during the search for the optimal decision tree at the cost of possibly taking longer. ID3 can overfit the training data, hence small decision trees should be preferred over large ones.This algorithm creates small trees in most cases, however it does not always produce the shortest decision tree feasible. Apart from that, using ID3 on continuous data is more difficult than using it on factored data (factored data has a discrete number of possible values, thus reducing the possible branch points). If the values of an attribute are continuous, there are many more places to split the data on that attribute, and finding the optimum value to split by can take a long time.

While the paper has mentioned various subparts under ID3 wherein developments under Information gain was mentioned, one area where I think there is a decent scope of research is the entropy which is a measure of the amount of uncertainty in the (data) set.For each remaining attribute in ID3, entropy is calculated. On this iteration, the attribute with the lowest entropy is utilised to split the set {\displaystyleS}S. In information theory, entropy is a measure of how much information is expected to be gained while measuring a random variable; it may also be used to quantify how unknown the distribution of the quantity's values is. Because the distribution of a constant amount is precisely known, it has zero entropy. Entropy is maximised when a random variable is uniformly distributed (discretely or continuously uniform). As a result, the higher the entropy at a node, the less information regarding data categorisation at this stage of the tree is known, and thus the larger the opportunity to enhance classification here.