Elitist and Ensemble Strategies for Cascade Generalization

Huimin Zhao, University of Wisconsin - Milwaukee, USA
Atish P. Sinha, University of Wisconsin - Milwaukee, USA
Sudha Ram, University of Arizona, USA

ABSTRACT

Several methods have been proposed for cascading other classification algorithms with decision tree learners to alleviate the representational bias of decision trees and, potentially, to improve classification accuracy. Such cascade generalization of decision trees increases the flexibility of the decision boundaries between classes and promotes better fitting of the training data. However, more flexible models may not necessarily lead to more predictive power. Because of potential overfitting problems, the true classification accuracy on test data may not increase. Recently, a generic method for cascade generalization has been proposed. The method uses a parameter — the maximum cascading depth — to constrain the degree that other classification algorithms are cascaded with decision tree learners. A method for efficiently learning a collection (i.e., a forest) of generalized decision trees, each with other classification algorithms cascaded to a particular depth, also has been developed. In this article, we propose several new strategies, including elitist and ensemble (weighted or unweighted), for using the various decision trees in such a collection in the prediction phase. Our empirical evaluation using 32 data sets in the UCI machine learning repository shows that, on average, the elitist strategy outperforms the weighted full ensemble strategy, which, in turn, outperforms the unweighted full ensemble strategy. However, no strategy is universally superior across all applications. Since the same training process can be used to evaluate the various strategies, we recommend that several promising strategies be evaluated and compared before selecting the one to use for a given application.

Keywords: cascade generalization; data mining; decision tree; elitist strategy; ensemble method; voting method

INTRODUCTION

Modern organizations often rely on data mining techniques to mine their collections of operational data for novel, valuable information that can be used to support decision making (Rajagopalan & Krovi, 2002). An important category of data mining problems is predictive data mining (also called supervised learning), in which a prediction model is built based on historical data, whose outcomes on a dependent variable are already known, and then applied to predict the outcomes of new cases.
Classification is a type of predictive data mining problem in which the dependent variable that needs to be predicted is discrete. Some examples of applications include bankruptcy prediction (Kim & McLeod, 1999; Sung, Chang, & Lee, 1999; Tam & Kiang, 1992; Sarkar & Sriram, 2001), financial performance prediction (Lam, 2004), bond rating analysis (Huang, Chen, Hsu, Chen, & Wu, 2004), credit evaluation (Sinha & May, 2005), credit risk assessment (Doumpos, Kosmidou, Baourakis, & Zopounidis, 2002), and network intrusion detection (Zhu, Premkumar, Zhang, & Chu, 2001).

Decision tree techniques have been used widely in predictive data mining, especially classification applications, largely because they generate models that resemble human reasoning and are easily understood (Weiss & Kulikowski, 1991). Most decision tree algorithms have an inherent representational bias; that is, they do not learn intermediate concepts and use only one of the original features in the branching decision at each intermediate tree node (Weiss & Kulikowski, 1991). Therefore, the trees they learn are called univariate trees. The trees are also called orthogonal trees, because the decision boundaries in the feature space are hyper planes, each of which is restricted to be geometrically orthogonal to a branching feature’s axis. Such representational bias limits the flexibility of decision trees and potentially jeopardizes classification accuracy.

Several methods have been proposed in the past for cascading other classification algorithms (e.g., logistic regression) (Hosmer & Lemeshow, 2000), with decision tree learners to alleviate the representational bias and potentially to improve the classification accuracy. Such a generalization of classification methods is called cascade generalization (Gama & Brazdil, 2000). The branching decision at an intermediate node in a generalized decision tree may be based on a constructed feature (often a linear combination of multiple features) learned by a cascaded classification algorithm. The constructed features represent intermediate concepts, which help to relax the representational bias of the base decision tree learner and allow the decision boundaries to be oblique (rather than axis-orthogonal) hyper planes in the original feature space. Such generalized decision trees have been named multivariate decision trees (Brodley & Utgoff, 1995), oblique decision trees (Heath, Kasif, & Salzberg, 1993; Murthy, Kasif, & Salzberg, 1994), discriminant trees (Gama, 1999) and linear discriminant trees (John, 1996; Yildiz & Alpaydin, 2000). When the branching decision at an intermediate node is based on a linear combination of two features, at most, the generalized decision trees are called bivariate decision trees (Bioch, van der Meer & Pothen, 1997).

Cascade generalization of decision trees increases the flexibility of the decision boundaries that separate different classes so that training data can be fitted better; the apparent accuracy (i.e., accuracy derived based on training data) tends to increase. However, more flexible models may not lead to higher predictive power (i.e., the true classification accuracy on test data may not increase along with flexibility. A model that fits training data well may not predict unseen data well, which is commonly known as the overfitting problem. There is always a trade-off between model flexibility and generalizability (Weiss & Kulikowski, 1991). Recently, a generic method for cascade generalization has been proposed. This method uses a parameter, the maximum cascading depth, to constrain the degree that other classification methods are cascaded with decision tree learners (Zhao & Ram, 2004). Empirical evaluation has demonstrated that the most productive degree of cascading varies across classification problems. To exploit the full potential of cascade generalization, all generalized decision trees under different parameter settings should be generated. An efficient algorithm also has been developed to generate a collection (forest) of all such trees in a single run without repeating any common nodes (Zhao & Sinha, 2005).

In this article, we propose several new strategies that were inspired by ensemble methods (Dietterich, 2000) (also called voting methods) (Bauer & Kohavi, 1999) such as bagging.
Accelerating Large-Scale Genome-Wide Association Studies with Graphics Processors
Mian Lu and Qiong Luo (2014). *Big Data Management, Technologies, and Applications* (pp. 349-380).
[www.igi-global.com/chapter/accelerating-large-scale-genome-wide-association-studies-with-graphics-processors/85463?camid=4v1a](www.igi-global.com/chapter/accelerating-large-scale-genome-wide-association-studies-with-graphics-processors/85463?camid=4v1a)

Graph Applications in Chemoinformatics and Structural Bioinformatics
Eleanor Joyce Gardiner (2012). *Graph Data Management: Techniques and Applications* (pp. 386-420).
[www.igi-global.com/chapter/graph-applications-chemoinformatics-structural-bioinformatics/58620?camid=4v1a](www.igi-global.com/chapter/graph-applications-chemoinformatics-structural-bioinformatics/58620?camid=4v1a)