Interactive Knowledge Frontier Discovery with COBWEB-KFD

Matt Honeycutt, MS and Douglas Talbert, PhD
Department of Computer Science, Tennessee Technological University, Cookeville, TN
Steve Talbert, RN, PhD
University of Central Florida, Orlando, FL

Abstract

Knowledge frontier discovery is a novel technique for identifying interesting subpopulations of a dataset with respect to classification performance. A knowledge frontier is a collection of meaningful groups where any sub-partition with significantly different predictive accuracy is not meaningful. This research introduces knowledge frontiers and knowledge frontier discovery. The first knowledge frontier discovery algorithm, COBWEB-KFD, is also described in detail. Knowledge frontier discovery extends clustering and subgroup discovery to provide information that those techniques cannot.

Introduction

Performance of machine learners can be degraded in domains that deal with noisy data, inconsistently described objects, or where critical features are not available for analysis. Computer-aided trauma triage is a domain where machine learning techniques have not yet achieved acceptable performance. Results presented in (Talbert and Talbert 2007) indicate that state-of-the-art machine learning techniques are unable to improve upon existing triage decision aides. An analysis of the results suggests that there may be subpopulations that the learners work well on, but there also exist subpopulations that the learners perform poorly on. Identifying and understanding these groups could lead to improvements in classifier performance. Our experiments with existing techniques, such as subgroup discovery (Gamberger et al. 2007) and clustering (Fisher 1996), were unable to identify such subpopulations.

This paper presents the COBWEB-KFD *knowledge frontier discovery* algorithm for identifying meaningful subpopulations of patients with similar predictive performance with respect to machine learning classifiers. While motivated by computer-aided trauma triage, we believe knowledge frontier discovery to be a generally applicable and useful technique for assessing classifier behavior.

Knowledge Frontiers and Knowledge Frontier Discovery

Knowledge frontiers are a novel technique for organizing datasets into subgroups that are both conceptually meaningful and statistically consistent with respect to classifier performance. Knowledge frontiers provide valuable insight that cannot be obtained with other similar approaches.

Knowledge frontiers defined

A *knowledge frontier* is a collection of meaningful groups where any sub-partition with significantly different predictive accuracy is not meaningful. A knowledge frontier represents a classifier's performance boundary; partitioning the dataset further provides no new insight into the classifier's performance.

A knowledge frontier consists of one or more subpopulations of the original data where each group is: conceptually meaningful, such as *male patients between the age of 30 and 40 with a pulse less than 40*; and statistically similar with respect to classifier performance, such as *all items were correctly classified between 65% and 70% of the time*. Further, there must not be a way to further partition the items within a knowledge frontier into conceptually meaningful groups that have significantly different performance with respect to the classifier.

Benefits of Knowledge Frontiers

Knowledge frontiers provide new insight that fulfills a vital step in the knowledge discovery from databases process: feedback that can potentially be used to improve classifier performance. Poorly understood frontier groups (those with low average classifier accuracy) can be analyzed further to uncover the causes of poor performance or used to create pre-classification filters which flag items that are similar to poorly-understood groups.

Knowledge Frontier Discovery Process

The process of discovering knowledge frontiers within a dataset is known as *knowledge frontier discovery (KFD)*. The KFD process involves three steps. First, the dataset must be augmented with performance metrics for a

machine learning classifier. Such metrics can be obtained with multiple applications of randomized 10-fold cross validation (Kohavi 1995). The performance for a particular item is simply the percentage of times the item is correctly classified across all trials.

The second step searches for a knowledge frontier satisfying the user-specified constraints on the frontier meaningfulness and similarity. Any technique that discovers an appropriate collection of subpopulations can potentially be used in this step.

The final step in the process is the evaluation of the frontier nodes by the user. If the frontier does not satisfy the user's information needs, the process should be iterated with revised constraints.

COBWEB-KFD

The COBWEB-KFD knowledge frontier discovery algorithm combines the COBWEB clustering algorithm (Fisher 1996) with a pruning technique inspired by cost-complexity pruning (Breiman et al. 1984). Because the algorithm locates a knowledge frontier within the COBWEB cluster hierarchy, the knowledge frontier will be conceptually meaningful. This property has been well established for cluster hierarchies created by COBWEB (Fisher 1996).

COBWEB-KFD first creates a COBWEB cluster hierarchy. Class information and classifier performance information are withheld so that they do not influence the clustering process. The clustering process is guided by COBWEB's *acuity* and *cutoff* parameters.

The second step performs a bottom-up search of the hierarchy and is guided by an *alpha* parameter that allows the user to bias the search for frontier nodes. Smaller alpha values bias the search towards smaller, more consistent nodes, while larger alpha values will bias the search towards nodes that are larger but less consistent.

The search for knowledge frontier nodes is similar to pruning decision trees. Beginning at the leaves of the cluster hierarchy, each node is evaluated with respect to its children. Intuitively, if the performance of the machine learner on the children is significantly more consistent (as determined by the user-specified alpha parameter) than the performance of the learner on the parent, the children become knowledge frontier nodes, and no further pruning is performed along the path from the children to the root of the hierarchy. Otherwise, the children are pruned, and the search continues at the next level up the cluster hierarchy.

The decision to prune utilizes a cost metric inspired by Breiman et al. 1984. The metric incorporates both node impurity and number of children. Node impurity is measured using sum-of-squared error over the accuracy values of items within the node, where T is the node being evaluated:

$$SSE(T) = \sum_{i \in T} (acc_i - acc_{avg})^2$$

When determining whether or not the children of a node can be pruned, two values are computed: the pruned cost and unpruned cost of T:

$$PrunedCost(T) = SSE(T) + \alpha$$

$$UnprunedCost(T) = \sum_{child \in T} SSE(child) + \alpha * |T.Children|$$

If the pruned cost is less than or equal to the unpruned cost, the children are pruned, and the process is repeated at the next level up the hierarchy.

Conclusion and Future Work

Preliminary experiments in the domain of trauma triage have been very positive. A domain expert found that the subgroups discovered by COBWEB-KFD were both interesting and clinically meaningful. This knowledge will hopefully improve the quality of computer-aided trauma triage in the future.

Knowledge frontiers and knowledge frontier discovery are novel contributions to the machine learning and data mining communities. As such, there are numerous avenues to pursue in future work including: development of new metrics, evaluation of alternate clustering and pruning techniques in the KFD process, and the creation of more efficient classifier accuracy estimate techniques. Finally, the utility of COBWEB-KFD and knowledge frontier discovery needs to be assessed beyond the domain of trauma triage.

References

L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Chapman & Hall/CRC, 1984.

D. Fisher. Iterative optimization and simplification of hierarchical clusterings. Journal of Artificial Intelligence Research, 4:147-180, 1996.

D. Gamberger, N. Lavrac, A. Krstacic and G. Krstacic. Clinical data analysis based on iterative subgroup discovery: experiments in brain ischaemia data analysis. Applied Intelligence, 27(3):205-217, 2007.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI, pages 1137-1143. Morgan Kaufmann, 1995.

S. Talbert and D. Talbert. A comparison of a decision tree induction algorithm with the acs guidelines for trauma triage. In AMIA 2007 Symp Proc, p 1127, 2007.