

**DPPred: An Effective Prediction Framework with Concise Discriminative Patterns**

**Abstract:**

In the literature, two series of models have been proposed to address prediction problems including classification and regression. Simple models, such as generalized linear models, have ordinary performance but strong interpretability on a set of simple features. The other series, including tree-based models, organize numerical, categorical and high dimensional features into a comprehensive structure with rich interpretable information in the data. In this paper, we propose a novel Discriminative Pattern-based Prediction framework (DPPred) to accomplish the prediction tasks by taking their advantages of both effectiveness and interpretability. Specifically, DPPred adopts the concise discriminative patterns that are on the prefix paths from the root to leaf nodes in the tree-based models. DPPred selects a limited number of the useful discriminative patterns by searching for the most effective pattern combination to fit generalized linear models. Extensive experiments show that in many scenarios, DPPred provides competitive accuracy with the state-of-the-art as well as the valuable interpretability for developers and experts. In particular, taking a clinical application dataset as a case study, our DPPred outperforms the baselines by using only 40 concise discriminative patterns out of a potentially exponentially large set of patterns.

**Existing System:**

Many pattern-based models have been proposed in the last decade to construct high-order patterns from the large set of features, including association rule-based methods on categorical data and frequent pattern-based algorithms on text data and graph. Recently, a novel series of models, the discriminative pattern-based models have demonstrated their advantages over the traditional models.

They prune non-discriminative patterns from the whole set of frequent patterns, however, the number of discriminative patterns used in their classification or regression models is still huge (at the magnitude of thousands). How to select concise discriminative patterns for better interpretability is still an open issue.

**Proposed System:**

We propose a novel discriminative patterns-based learning framework (DPPred) that extracts a concise set of discriminative patterns from high-order interactions among features for accurate classification and regression. In DPPred, first we train tree-based models to generate a large set of high-order patterns. Second, we explore all prefix paths from root nodes to leaf nodes in the tree-based models as our discriminative patterns. Third, we compress the number of discriminative patterns by selecting the most effective pattern combinations that fit into a generalized linear model with high classification accuracy or small regression error. This component of fast and effective pattern extraction enables the strong predictability and interpretability of DPPred.