Improve Prediction Accuracy for Clinic Medical Data Analysis

Master Graduate Project
Presenter: Linlin Zhou
Advisor: Dr. Yingshu Li
Outline

- Introduction
- Evaluation of state of art
- High Accurate Decision Tree Pruning & Ensembling Algorithm
  - Decision Tree
  - Pruning Method
  - Ensembling Method
- Result and Evaluation
- Conclusion
Introduction

- Clinic medical data:
  - From practical clinic experiments and treatments
  - Valuable on symptoms description
  - One of major sources to diagnose and develop treatment
  - Costful and difficult to obtain
Data Analysis on Clinic Medical Data

• Help to understand and detect medical diseases symptoms
• Drive new medical discovery
• Cut down on fraud and abuse
• Reduce medical treatment research and development cost
• Improve patient outcome
Challenges on Clinical Medical Data Analysis

• Limited medical data size
• Complicated and redundant medical data features
• High noises
• Unbalanced disease classes
  • Most are healthy or usual classes
  • Small particle but much more meaningful diseases or unusual classes
• High requirement on accuracy
Contributions

• Propose a high accurate clinic medical data classifier algorithm
  • Based on decision tree classification
  • Improved with pruning and ensembling algorithm
  • Presenting significantly higher accuracy than existing classification algorithms

• Implement and evaluate my algorithm on a Psoriasis clinic medical dataset
  • The results show our algorithm overwhelms other existing classification algorithms
  • A list of important features in Psoriasis dataset

• Our algorithm can be implemented on other clinic medical dataset
Psoriasis Clinic Medical Dataset

- Psoriasis is an immune-mediated disease that causes raised, red, scaly patches to appear on the skin
  - Different types of Psoriasis need different treatments.
    - The type of Psoriasis cannot be determined accurately.
  - No standard methods for Psoriasis detection and lacking accuracy of diagnose.
  - Most patients with severe Psoriasis suffer a long time and have no sufficient treatments

- Dataset: 606 patients carrying Psoriasis from a hospital
  - 89 features of specific physical examinations
  - 5 classes representing Psoriasis types
  - Different data type of 89 features
    - Integer, String, float,
    - Unformatted with a lot of typos and unrecognized characters
Evaluation on Existing Classification Algorithms

- Most of existing classification algorithms present low classification accuracy
  - <90%

- Decision tree and Decision tree based algorithms show higher accuracy
  - DT, RT and RT with cross validation
  - >84%

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Parameters</th>
<th>Accuracy Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayesian (NB)</td>
<td>K = 5</td>
<td>0.81</td>
</tr>
<tr>
<td>K Nearest Neighbors (KNN)</td>
<td>Cart</td>
<td>0.79</td>
</tr>
<tr>
<td>Decision tree (DT)</td>
<td>n_components = 9</td>
<td>0.84</td>
</tr>
<tr>
<td>DT with PCA</td>
<td></td>
<td>0.86</td>
</tr>
<tr>
<td>Support Vector Machine (SVM)</td>
<td>n_estimators = 100</td>
<td>0.81</td>
</tr>
<tr>
<td>SVM with AdaBoost Classifier</td>
<td>Max_feature = 9</td>
<td>0.86</td>
</tr>
<tr>
<td>Random Forest (RF) with CCA</td>
<td>Max_feature = 9</td>
<td>0.87</td>
</tr>
<tr>
<td>RF with GridSearch CrossValidation</td>
<td></td>
<td>0.89</td>
</tr>
</tbody>
</table>
High Accurate Decision Tree Pruning & Ensemble (HADPE) Algorithm

- A Decision Tree based high accuracy algorithm
  - Understandable classification rules and outputs
    - Find the influencing features through tree structure
    - Significant for doctors and patient to understand the causes and symptoms
  - Higher accuracy compared to other algorithms
  - Easy to implement and high running efficiency
Decision Tree

- Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training samples are at the root
  - Features are categorical (if continuous-valued, they are discretized)
  - Samples are split recursively based on selected features
  - Features are selected on the basis of a heuristic or statistical measure (e.g., information impurity)
  - Stop when
    - All samples belong to same class
    - Or no remaining features for further splitting

- Predict by finding a path to leaf node that matches the values of features
Decision Tree Example

Three Classes
Two Features

X[0] > 0.7
X[0] ≤ 0.7

X[1] > 0.4
X[1] ≤ 0.4
X[1] > 0.5
X[1] ≤ 0.5

{1,1,2,2} → class1
{2,2} → class2

{1,1,2,3,1,2} → class1
{1,3} → class3

{1} → class1
{3} → class3
Decision Tree Feature Selection

- Used in CART (introduced by Leo Breiman)
- Split tree node to achieve lowest impurity after splitting
  - Impurity measured by Gini index:
    - $Gini(S) = 1 - \sum_{i \in S} p_i^2$
    - $p_i$: probability of class $i$ in set $S$
    - Higher Gini index, higher impurity
- Efficient to split continuous values
- More reliable compared to other criteria
Problems in Decision Tree

- Overfitting: selection of an attribute that is non-optimal for prediction
- Fragmentation: data are fragmented into (too) many small sets
- Caused by:
  - Trying to fit noise
  - Lack of identical features

Prediction accuracy comparison on training set and test set
Pruning: Solve Overfitting in DT

- Remove sections of tree that provide little power to classify observations
  - Set criteria when splitting tree node to subtree nodes
  - Stop splitting tree node when criteria is not satisfied
  - Common criteria:
    - Maximum number of leaf nodes
    - Maximum number of tree layers
    - Minimum impurity
Pruning: Solve Overfitting in DT

- HADPE uses pruning with minimum impurity criteria for splitting
  - Only split a tree node if this tree node has impurity larger than minimum impurity

```
<table>
<thead>
<tr>
<th>max_leaf_node</th>
<th>max_tree_layer pruning criteria</th>
<th>min_impurity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.89</td>
<td>0.92</td>
</tr>
</tbody>
</table>
```

avg_prediction_accuracy
Decision Tree Ensemble Method

- Bagging a strong classifier from a list of weak/sub classifiers
  - Build a list of sub classifiers from random training sets
    - Each sub classifier built by a distinct pruning criteria (min_impurity value)
  - Validate each sub classifier by a validation set
    - Increase the weights the sub classifiers that
      correctly predict the class of the samples in validation set
    - Decrease otherwise
  - To detect and set higher weight to the sub classifiers that
    - Have better pruning criteria
    - Have higher prediction accuracy

\[
\omega_{i(j+1)} = \begin{cases} 
\omega_{ij} \cdot e^{\alpha} & : \text{if classifier i predicts sample (j + 1) correctly} \\
\omega_{ij} \cdot e^{-\alpha} & : \text{otherwise}
\end{cases}
\]

\[0 < \alpha < 1\]
Decision Tree Ensemble Method

- Prediction:
  - After training, predictions for unseen samples (test_set) can be made by taking the majority vote in the case of decision trees.

- Better model performance
Results and Evaluation

- Implement based on Psoriasis clinic medial dataset
- Dataset is cleaned and preprocessed for missing data, unrecognized data, etc.
- Randomly select training set, validation set, testing set
  - 70% training set
  - 30% testing set
  - Validation set: [100, 200, 300]
- Five HADPE configurations on sub DT classifier size:
  - [10, 15, 20, 25, 30]
- Min_impurity criteria randomly chosen from [0, 0.1]
Prediction Accuracy

avg_prediction_accuracy

No of sub classifier in HADPE
Running Time

![Graph showing average running time (s) for different numbers of validation data and classifiers. The x-axis represents the number of validation data (100, 200, 300), and the y-axis represents the average running time (s). The graph includes lines for different numbers of classifiers: 10, 20, and 30.](image)
Feature Importance
Feature Importance

• Most Six Important Features in construction the tree:
  • Column 4: Low Pressure
  • Column 3: High Pressure
  • Column 5: White Blood
  • Column 15: Basophil
  • Column 17: Red Blood
  • Column 40: Total Bilirubin
Conclusion

- Investigate the problem of clinic medical data analysis
  - Meaning and challenges

- Evaluate existing classification algorithms
  - Lower accuracy on the medical data

- Propose our high accuracy classification algorithm
  - Hybrid pruning and ensembling

- Our evaluation shows our algorithm significantly overwhelms existing works

- The algorithm can be used on other clinical dataset
Summary

• Not fitting algorithm:
  • Convolutional Neural Networks (CNN)
  • Deep Neural Networks (DNN)
  • Support Vector Machine
  • KNN
  • ...

• Fitting algorithm:
  • Decision tree (DT), Random Forest (RF)
  • Ensemble Method (Boosting, Bagging)

• Clinical Data Management
  • Small Size -- Ask for as more data as possible
  • Missing value – Imputing method
THANK YOU!

Q&A