Prajwal Kailas
14CO132
9632080195
prajwal967@gmail.com

Feature Selection Inspired Classifier Ensemble Reduction

Directed, Produced and Written by
Ren Diao, Fei Chao, Taoxin Peng, Neal Snooke, and Qiang Shen

Essence :

- The paper is based on ensemble classifiers used the field of machine learning and data mining. Ensemble classifiers are used to improve the accuracy of machine learning applications.
- The task involves the concept of feature selection, which involves selecting the features of a dataset which highly influence the final output.
- The focus is on reducing memory, storage and system runtime overhead. This is particularly helpful when working with large datasets.
- A technique called heuristic harmony search is introduced. Harmony search is a feature selection algorithm used to optimize features of a dataset.
- The performance of the techniques described in the paper are evaluated against high dimensional and large sized benchmarked datasets.

Work Described:

- The paper first focuses on building an ensemble classifier, which can be done by constructing a group of classifiers/models which have different hypothesis followed by combining their predictions to produce the final prediction.
- The paper also dives into classifier ensemble reduction which aims to reduce the amount of redundancy in an ensemble classifier while maintaining the diversity, so that we can deliver the same final results with a reduced subset of classifiers.

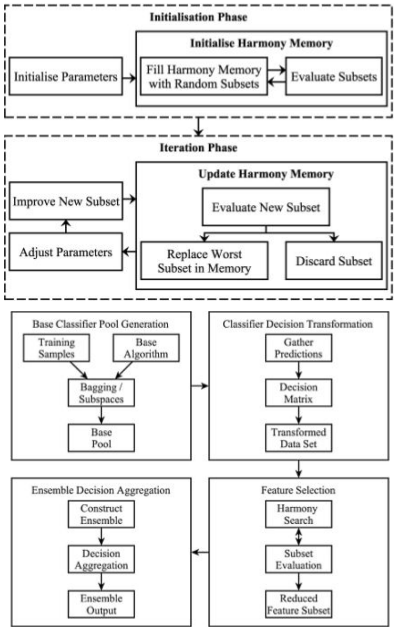
Importance:

- Many fields in computer informatics rely on high accuracy of their systems, by making use of ensemble classifiers the accuracy of a system can be increased, thereby giving better predictions and enabling better use of the systems.
- The complexity in building systems and applications in informatics can be reduced by making use of classifier ensemble reduction.

Methodology:

- Format, clean and analyze the training data. Separate the data in two sets, training data and the test data, in a 70:30 ratio.
- Build an ensemble of classifiers using any number of base machine learning algorithms. Algorithms such as logistic regression, linear regression, random forests, multi layer perceptron, decision trees, gradient boosting can be used.
- A diverse base classifier pool is necessary in building a good ensemble classifier. The ensembling can be done via bagging, stacking, blending etc.
- The ensemble classifier is trained on the training data and their hypotheses are constructed. The hypothesis have different search spaces.
- The predictions of the ensemble classifiers are transformed into artificial features in a newly constructed dataset.
- Discover a subset from the set of ensemble classifier any redundant members and eliminate them while maintaining diversity.
- Feature selection algorithms are used to remove redundant features, maintain the diversity and accuracy of the ensemble classifier

- Harmony search is a meta heuristic feature selection algorithm, which performs feature selection by finding a solution vector, that optimizes the cost function. Harmony search selects a subset of features in such a way that the cost function does not change.
- Harmony search is applied to the newly constructed dataset which contain the artificial features. Harmony search reduces the subset size of th is newly constructed dataset while maintaining the quality and diversity of the ensemble classifiers.
- The final ensemble classifier is a subset of the original ensemble classifier, then new object are classified by this ensemble classifier, and the results of the subset of ensemble classifiers are aggregated to form the final ensemble output.



Pros:

- Large datasets have high memory and processing requirements. Eases working with very large dimensional datasets.
- Reduces memory, storage and system runtime overhead. Feature selection reduces the number of features that are used , thereby decreasing the size and complexity of the dataset.

Cons:

- Diversity of ensembles is reduced slightly. As a subset of the original features is used, some information might be lost.
- Trade off between accuracy and system overhead depending on application.

Future Work:

- Formulation of alternative decision matrix transformation procedures for improving accuracy and diversity of the ensemble classifier.
- Use of other statistical information such as variance while constructing artificial features for the new dataset from the predictions of the base classifiers.

Implementation:

- Using several machine learning algorithms for the construction of ensemble classifiers and training and testing them on a benchmarked training dataset. The predictions of the base classifiers will be obtained from the training data which is followed by constructing the new dataset and then training the ensemble classifier with this newly constructed dataset.
- Make use of logistic regression, linear regression, random forests, multi layer perceptron, decision trees, gradient boosting algorithms as base model classifiers to build the ensemble classifier.