Introduction

- "Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules" — Berry and Linoff 1997

- The aim of data mining is to
  - "be able to respond to the patterns, to act on them, ultimately turning the data into information, the information into action, and the action into value" — Berry and Linoff 1997

The task is either classification or regression

- When performing classification the target value must be any of a pre-defined set of values
- For regression, the target value is a continuous value

The normal procedure is to use historical data with known target values to build models that could later be used for prediction
A decision tree

 iris-virginica

 A neural net

 JChipper rules:

 Number of Conditions : 4
 Number of Rules      : 3

 DEFAULT: Iris-virginica [50/2]
 IF ( petalwidth <= 1.7 ) AND ( petallength <= 5.0 ) THEN Iris-versicolor [50/2]
 IF ( petalwidth <= 0.6 ) THEN Iris-setosa [50/0]

 Ensembles

 • An ensemble is a composite model, aggregating multiple base models into one predictive model
   – An ensemble prediction, consequently, is a function of all included base models
 • Both theory and a wealth of empirical studies have established that ensembles are generally more accurate than single predictive models

 Diversity

 • For the ensemble approach to work, the ensemble must contain diversity
   – There would be no point in combining only models that always
     • Make the same mistakes
     • Add the same information
 • We want models that perform well individually and complement each other
Learning\textsuperscript{T2}(* = T\textsuperscript{2}h)\textsuperscript{S}\textsuperscript{3}

Unfortunately (Krogh and Vedelsby, 1995) diversity are highly correlated average error and diversity is well defined for regression problems in classification context have been proposed. Measure all members together when creating the models to combine

\begin{align*}
E = E - A
\end{align*}

**Information Fusion**

- Information fusion is the research about how to aid decision makers with different tasks, by combining data and information from various sources
- It is characterized by the necessity to gather data about objects or situations from multiple sources and combine them to enable effective decision support, often under severe time and resource constraints
- Each source can only provide information from its specific point of view and often only about some specific feature.

**Ensembles in Information Fusion**

- One of the characteristics of information fusion is the need to combine data from several sources
  - To understand the whole picture from all the various fractions of data that is gathered
- Obviously, the use of ensembles is a very natural framework for information fusion
  - New base models can be added when new sources are added
  - Old models can be updated or dropped when they become too faulty or sources are removed or lost

**Diversity and Information Fusion**

- Diversity in ensembles is achieved by dividing datasets into:
  - Different feature sets
  - Different subsets of data
  - Measurements of the problem from different perspectives
- The data used in Information Fusion often come:
  - from different kinds of sensors
  - with different intervals
  - from sensors at different positions

**Problem Statement**

- The main problem: How should ensembles be created to maximize predictive performance?
- The problem statement: How could measurements of diversity and predictive performance on available data be used when combining or selecting base classifiers in order to maximize ensemble predictive performance on unseen data?
- The final goal when building predictive models is to achieve as high predictive performance as possible, this is inherent in the need of a predictive model
- An ensemble can be formed either by simply combining available base classifiers, or by selecting a subset of base classifiers
  - This means that diversity and performance measures can be used either to guide the selection or as an implicit goal when creating the models to combine
The problem statement can be further specified through the following more specific research questions:

1. How do different means of achieving implicit diversity among base classifiers affect the performance of, and diversity in, the ensemble?
2. Can ensemble predictive performance on novel data be estimated from results on available data?
3. Is there an optimization criterion based on an existing measure on available data that is best for the purpose of developing ensembles that maximize predictive performance?
4. Are combinations of single measures a good solution for the purpose of developing ensembles that maximize predictive performance?

Evaluation of techniques

- Dividing training data
  - by features
  - by instances (bootstrapping)
- Varying ANN architecture

The purpose of several studies was to evaluate whether it is possible to estimate the performance of ensembles based on available data.

Previous studies have shown that no diversity measure is well correlated with ensemble accuracy under specific circumstances.

- Is this true in general?
- Is any of the evaluated diversity measures better correlated?
- Are performance measures significantly different than diversity measures?

The overall purpose of the study on implicit diversity was to empirically evaluate some standard techniques for introducing implicit diversity in neural network ensembles.

- Evaluates all combinations of techniques, resulting in 12 different combinations
- Most important criterion is of course generalization accuracy
- But the study also analyses:
  - the levels of diversity produced by the different methods
  - how diversity and generalization accuracy co-vary, depending on the technique used to introduce the diversity

<table>
<thead>
<tr>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four experiments</td>
</tr>
<tr>
<td>- Measuring diversity and performance on either training or validation data</td>
</tr>
<tr>
<td>- Using either</td>
</tr>
<tr>
<td>- All ensembles of a fixed size (enumerated)</td>
</tr>
<tr>
<td>- Randomly drawn ensembles with varying sizes (random)</td>
</tr>
<tr>
<td>- Mean correlation over 11 datasets</td>
</tr>
</tbody>
</table>
Correlation between measures on available data and ensemble test accuracy is generally very low.

Enumerated Train that was proposed

Mean ranks over 30 data sets • • All the complex optimization criteria were comparably good

DI Could they compete with single measures as optimization criterion?

ensemble accuracy Twotypes of base models is comparably good as selection criterion for • • as optimization criterion that combination of measures could outperform single measures – complex optimization criteria are competitive

The great problem is

0.10

0.20

0.30

0.40

0.50

2009-03-24

are, in most cases, ANN ensemble accuracy clearly better than • • using only – indicating its relative importance The method searches for a set of weights optimal for each data set and each specific set of models • • The solution is intended to be used as an optimization criterion when selecting models to include in the ensemble – The method utilizes some rather complex techniques • • It produces a series of solutions, whereas the results for three of them are presented

Five types of base models • • Neural networks • • Decision trees

Study: Combining Measures

No single measure was a good estimator of ensemble performance • • Could combinations of measures work better? – Could they compete with single measures as optimization criterion?

– Is it possible to determine which measures to include and what importance to give them?

Combining accuracy measures with diversity measures fits very well into the basic theory about ensembles • • The ensembles should consist of accurate models that complement each others predictions

Initial results give a strong indication that combination of measures could outperform single measures • • The great problem is how to know exactly which combination to use

– The combined optimization criterion is compared to using Accuracy or Difficulty as optimization criterion

Two types of base models – Neural networks – Decision trees

Results

The complex optimization criteria are, in most cases, clearly better than using only ensemble accuracy or Difficulty as selection criteria

The opposite is true for ensemble accuracy

– Difficulty does not work at all for neural networks – It is significantly worse than all other selection criteria

– Only the complex optimization criteria are competitive regardless of which set of base classifiers that are used

– All the complex optimization criteria were comparably good

Estimating Ensemble Performance

Results showed that all diversity measures evaluated show low or very low correlation with ensemble accuracy

The initial studies also showed correlations between training or validation accuracies and test accuracy to be remarkably low – Most ensembles tend to have very similar training or validation accuracy

– Validation sets are often rather small, so confidence intervals for true error rates when estimated using validation data become quite large

Estimating Ensemble Performance

It could be argued that the main issue is whether an ensemble ranked ahead of another on some measure retains this advantage on predictive performance on test data

The overall purpose of one study was to investigate how well ensemble rankings produced from different measures on available data agree with predictive performance on test data • • Results showed that:

Many ensembles get exactly the same test set accuracies • • Marginal difference in test set performance between ensembles ranked high and low

Study: Combining Measures

No single measure was a good estimator of ensemble performance • • Could combinations of measures work better? – Could they compete with single measures as optimization criterion?

– Is it possible to determine which measures to include and what importance to give them?

Combining accuracy measures with diversity measures fits very well into the basic theory about ensembles • • The ensembles should consist of accurate models that complement each others predictions

Initial results give a strong indication that combination of measures could outperform single measures • • The great problem is how to know exactly which combination to use

– The combined optimization criterion is compared to using Accuracy or Difficulty as optimization criterion

Two types of base models – Neural networks – Decision trees

Results

The complex optimization criteria are, in most cases, clearly better than using only ensemble accuracy or Difficulty as selection criteria

The opposite is true for ensemble accuracy

– Difficulty does not work at all for neural networks – It is significantly worse than all other selection criteria

– Only the complex optimization criteria are competitive regardless of which set of base classifiers that are used

– All the complex optimization criteria were comparably good
Conclusions: Implicit Diversity

• How do different means of achieving implicit diversity among base classifiers affect the performance of, and diversity in, the ensemble?

• The ensemble performance is positively effected by implicit diversity

• Using heterogeneous ensembles, with varied neural network architectures, was clearly beneficial

• Resampling of features was beneficial – Bootstrapping reduced the positive effect

Conclusions: Estimating Ensemble Performance

• Can ensemble predictive performance on novel data be estimated from results on available data?

• No results on available data were strongly correlated with ensemble performance on novel data – many measures were in general almost non-correlated or even negatively correlated with ensemble accuracy on the test set

• Some measures had, in comparison, constantly higher correlation
  – Ensemble accuracy, difficulty, double fault and to some extent base classifier accuracy
  – Double fault and base classifier accuracy were negatively affected when allowing small ensembles

• The ranking study showed that it is very difficult to estimate predictive performance based on available data

Conclusions: Optimization Criterion

• Is there an optimization criterion based on an existing measure on available data that is best for the purpose of developing ensembles that maximize predictive performance?

• Only two measures have proven to constantly perform rather well as optimization criteria:
  – ensemble accuracy and the diversity measure difficulty

• Other measures, like base classifier accuracy and double fault, did under some circumstances perform comparably well:
  – The problem with both these latter measures was their tendency to prefer the smallest ensembles when ensembles of varied size were considered

• Difficulty was significantly better as optimization criterion than ensemble accuracy when the base models were decision trees
  – Significantly worse when the base models were neural networks
  – Suggests that using diversity measures as (part of) an optimization criterion is possible and perhaps also feasible
  – The problem is how to know in advance when to prefer one measure over another

Conclusions: Combining Measures

• Are combinations of single measures a good solution for the purpose of developing ensembles that maximize predictive performance?

• The conclusion regarding combined measures from all but the last study was that combined measures resulted more often than not in better performance
  – Exactly how to combine measures was still an open question.

• A method for optimizing a combined measure was proposed
  – The proposed method was at least as good as the best single optimization criterion regardless of which base models that are used
  – It was shown to be significantly better than using ensemble accuracy as optimization criterion, when using decision trees
  – It was also significantly better than using difficulty as optimization criterion, when using neural networks

Main Conclusions

How could measurements of diversity and predictive performance on available data be used when combining or selecting base classifiers in order to maximize ensemble predictive performance on unseen data?

• Best to somehow combine information about both accuracy and diversity
  – All experiments involving combined measures showed that even straightforward linear combinations were generally better as optimization criterion than using even the best single measure
  – The problem when using straightforward linear combinations is knowing which measures to include

• A method aimed at optimizing a combined optimization criterion was proposed
  – The results indicate that a strong argument for using such optimized combinations is their robustness

Discussion

• There is of course an analogy to ensembles of classifiers when considering combined measures, since they could be viewed as ensembles of measures

• Just as with ensembles of classifiers, it is reasonable to assume that measures can be combined and optimized in many different ways
  – Therefore, further research about combining multiple measures is suggested

• One important question that must be addressed in future work is where the limit is reached when the extra effort of finding a suitable combination of measures is no longer worthwhile
Discussion

- The **standard procedure** when trying to select base classifiers for an ensemble has been to **optimize some performance measure**, usually accuracy.

- Results show that **other measures might be better** as optimization criteria in some cases:
  - **Difficulty** was superior to **ensemble accuracy** when selecting from a pool of decision trees.

- An interesting question is whether it is possible to find rules that can work as **guidelines for when to use which optimization criterion**:
  - Using meta learning.