Designing an Ensemble Classifier over Subspace Classifiers using Iterative Convergence Routine

Bhanukiran Vinzamuri $^1$  Kamalakar Karlapalem $^2$

$^1$vinzamuri@research.iiit.ac.in  
Centre for Data Engineering  
International Institute of Information Technology Hyderabad

$^2$kamal@iiit.ac.in  
Centre for Data Engineering  
International Institute of Information Technology Hyderabad

25 October 2011
Outline of the presentation

1. Introduction
   - Ensemble Classifiers

2. Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3. Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4. Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements
Slide Status

1. Introduction
   - Ensemble Classifiers

2. Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3. Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4. Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements
Utility of Subspace Clusterings

- Many subspace clusterings for a given dataset exist. Each represents a view of the dataset.
- Interpretation and utilization of these clusterings can enhance our knowledge of the data immensely.
- This knowledge can also be used to design a more accurate ensemble classifier!
Knowledge within subspaces

- Negatively related (Kittler et al. 1998) different classifier members creates a diverse ensemble for classification.
- Can several dissimilar subspace clusterings be utilized cumulatively for building a good classifier?
- Can subspace clusterings reveal information on transforming the data to further suit supervised learning?
Overview

- Using decision trees (classifiers) to come up with appropriate subspaces.
- Using a greedy iterative procedure which checks the goodness of decision trees.
- Using a clustering algorithm on identified subspaces to obtain disparate clusterings.
- Generating an *augmented* dataset using clusterings generated for further classification.
Slide Status

1. Introduction
   - Ensemble Classifiers

2. Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3. Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4. Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements
Challenges

- Modeling mutual existence of clustering and classification together in one framework.
- Identification of good subspaces in the dataset which can give disparate clusterings and build a better classifier.
- Quantifying goodness of subspaces and deciding on how many subspaces to evaluate.
1 Introduction
   - Ensemble Classifiers

2 Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3 Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4 Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements
Data Generation Step

Figure: Data generation for the Ensemble Classifier
Illustrative Example

Figure: Illustrative example of working of our ensemble classifier
Slide Status

1. Introduction
   - Ensemble Classifiers

2. Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3. Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4. Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements
**IsConvergent Procedure**

**Procedure1:** *IsConvergent*

**Require:** DecisionTree $DT$, DecisionTree $DT'$, Threshold $Thres$.

1. flag=false
2. if $DT$.accuracy == $DT'$.accuracy then
3.    flag=true
4. else if $DT$.attributes == $DT'$.attributes then
5.    flag=true
6. else if $DT$.accuracy − $DT'$.accuracy > $Thres$ then
7.    flag=true
8. end if
9. return flag
Slide Status

1. Introduction
   - Ensemble Classifiers

2. Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3. Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4. Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements
Importance of augmented attributes (VI)

Without labels

With labels

<table>
<thead>
<tr>
<th>Number of convergence iterations</th>
<th>Clustering Dissimilarity (VI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

- optdigits
- waveform
- dermatology
- factors
- fourier
- pixel

Number of convergence iterations

Performance with induced noise
Importance of augmented attributes (RI)

Without labels

With labels

Number of convergence iterations

Clusterings Dissimilarity (RI)

- optdigits
- waveform
- dermatology
- factors
- fourier
- pixel
Slide Status

1. Introduction
   - Ensemble Classifiers

2. Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3. Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4. Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements
## Performance with induced noise

**Table:** Performance comparison of Our Ensemble Classifier against different classifiers with varying induced noise levels

<table>
<thead>
<tr>
<th>Dataset</th>
<th>J48 10%</th>
<th>J48 20%</th>
<th>J48 50%</th>
<th>Our Ensemble Classifier 10%</th>
<th>Our Ensemble Classifier 20%</th>
<th>Our Ensemble Classifier 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Digits</td>
<td>0.759</td>
<td>0.648</td>
<td>0.27</td>
<td>0.771</td>
<td>0.66</td>
<td>0.2706</td>
</tr>
<tr>
<td>Fourier</td>
<td>0.603</td>
<td>0.531</td>
<td>0.230</td>
<td>0.64</td>
<td>0.549</td>
<td>0.228</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.833</td>
<td>0.719</td>
<td>0.49</td>
<td>0.861</td>
<td>0.728</td>
<td>0.499</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ADABOOST.M1 10%</th>
<th>ADABOOST.M1 20%</th>
<th>ADABOOST.M1 50%</th>
<th>STACKING 10%</th>
<th>STACKING 20%</th>
<th>STACKING 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Digits</td>
<td>0.724</td>
<td>0.597</td>
<td>0.296</td>
<td>0.798</td>
<td>0.714</td>
<td>0.405</td>
</tr>
<tr>
<td>Fourier</td>
<td>0.577</td>
<td>0.487</td>
<td>0.216</td>
<td>0.710</td>
<td>0.595</td>
<td>0.333</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.81</td>
<td>0.689</td>
<td>0.465</td>
<td>0.84</td>
<td>0.693</td>
<td>0.472</td>
</tr>
</tbody>
</table>
Major contributions

- Combining clustering and classification in one framework!
- Evaluating the dissimilarity of subspace clusterings generated.
- Ensemble classification framework with augmented data instead of multiple classifiers!
Slide Status

1. Introduction
   - Ensemble Classifiers

2. Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3. Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4. Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements
Multiple clustering (Muller et al. 2010): Projection onto orthogonal spaces.


Comparing Subspace Clusterings (Patrikainen and Meila 2006): Clustering error, Rand-index and VI.
Slide Status

1. Introduction
   - Ensemble Classifiers

2. Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3. Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4. Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements
Future work

- Extending the data generation framework from using other classifiers apart from decision trees.
- Coming up with a mathematical model explaining the relationship between unsupervised, supervised learning and convergence.


Muller, Emmanuel et al. (2010). “Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data”. In: *Tutorials at ICDM*.

Patrikainen, Anne and Marina Meila (2006). “Comparing Subspace Clusterings”. In: *IEEE Transactions on Knowledge and Data Engineering*.

Slide Status

1. Introduction
   - Ensemble Classifiers

2. Overview of our approach
   - Challenges
   - Illustration
   - Algorithm

3. Experimental Results
   - Importance of augmented attributes
   - Performance with induced noise

4. Major Contributions
   - Related Work
   - Future Work
   - Acknowledgements

Acknowledgements
We would like to thank the LNM Institute of Information Technology Jaipur and Microsoft Research India for providing the necessary travel grant for attending this conference and presenting this paper.