


<p>Award No: 0453463</p>	
<p>Project Title: REU Site: A Computer Science and Engineering REU Site for Florida, Puerto Rico and Latin America – “A Parallel Feature Selection Algorithm from Random Subsets”</p>	
<p>Investigators: Daniel J. Garcia, Lawrence Hall, Dmitry Goldgof, and Kurt Kramer</p>	
<p>Institution: University of South Florida</p>	
<p>Website: http://figment.csee.usf.edu/~kkramer/sipper/</p>	<p>Description of Graphic Image: Sample images of plankton. These are the images whose features are used by our algorithm in order to classify other types of plankton.</p>
<p>Project Description and Outcome</p>	

Ideas:

In this research a new method of feature selection called *selection from Random Subsets* was developed to classify plankton images in real time. Plankton classification is widely used by marine biologists to identify the location of plankton populations and determine the density of such populations, an important matter, considering the importance on plankton on the marine ecosystem.

The method has two stages. In the first stage, a number of randomly selected feature sets of fixed size are generated from the pool of all features. Then, a 10 fold cross validation is run to determine how well they are able to classify the data. The sets of features are then sorted using a given criteria, such as training time (here) or the number of support vectors generated, and the best of these randomly generated sets are selected for the second stage of the algorithm.

In the second stage of the method we have a number of ranked feature sets. Using these sets, a new set composed of the union of the features found in the selected sets is created. At this point, the classifier is trained using the newly created feature sets, and then it is tested against a previously unseen test set to see how well it performs. The number of feature sets selected for the second stage of the method can vary from 2 to the number of sets generated during the first stage of the process.

The method was evaluated on five different sets of plankton images with 1000 images in each set. On these sets the classification accuracy was comparable with the Wrappers method, which is a feature selection method well-known for its high classification accuracy, while being approximately 10 times faster. In addition to this, the Random Subsets approach can be run on all available processors in parallel, further speeding up the process.

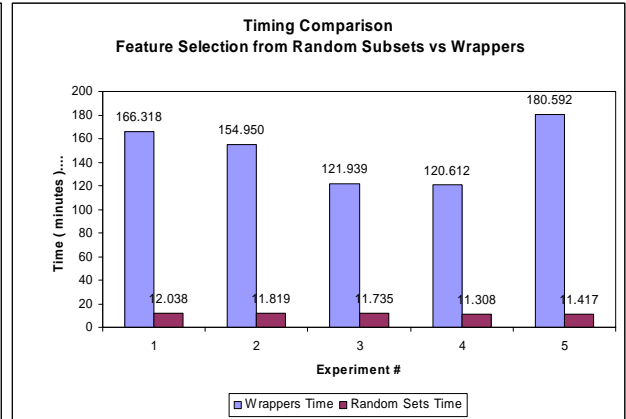
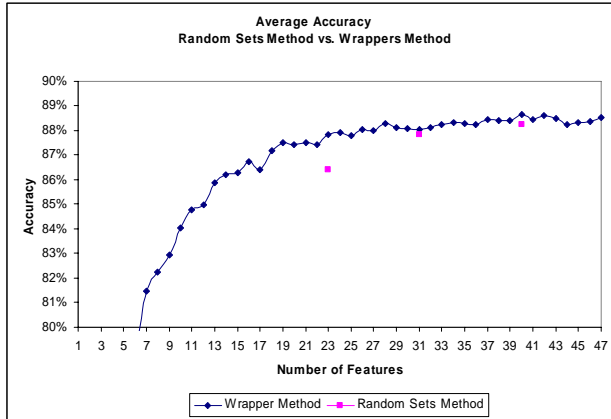
Tools:

N/A

People:

N/A

Additional Graphic Image



Description of Graphic Image:

The graphs above summarize the result from running the feature selection algorithm from random subsets on the plankton data set. As can be seen, the accuracy obtained from the random subsets algorithm is just as good as the result obtained from the well known Wrappers algorithm while requiring much less time.