1st Paper : Speech Emotion Features Selection Based on BBO-SVM

Speech is one of the most direct and natural ways of human communication. Along with complex semantic information, voice also contains Emotional Information. It is easier to capture voice signal than other physiological signals. The basis of emotional recognition is to extract useful features from speech signals. The redundant characteristics due to increasing speech feature dimension affects the result of recognition.

The objective of the paper is

- to find a set of low dimensional and sufficient features to characterize the emotion contained in speech signal
- to reduce the processing time and improve the efficiency of recognition.

An optimization method based on BBO-SVM for invariant elements of set is proposed. It utilizes BBO (Biogeography Based Optimization) algorithm to select and optimize the original feature set. Cross-validation result of SVM is used as a standard for evaluating subset. Iterative optimization of the original high-dimensional feature set is done to get the compressed feature subset.

Biogeography is a study of the distribution of species and ecosystems in geographic space and through geological time.
Habitat Suitability Index (HSI) is the capacity of a habitat to support a selected species. It indicates the appropriate species survival index. In general, HSI and number of species in that habitat are directly related.
Habitability is related to features such as rainfall, topography, soil quality, temperature, etc. These are called Suitability Index Variables (SIVs).
Species Emigration rate is high and species immigration rate is low in the habitats with high HSI while Low HSI Habitats have lower species emigration rate and higher immigration rate. The HSI of the habitat can be improved in low HSI Habitats because of the influence of species migration on habitat environment.

BBO Algorithm is based on population. It regards each solution of the population as a habitat, the goodness of the solution as the HSI of the habitat, and each component of the solution as an SIV. It can solve the optimization problem by stimulating the effect of migration and mutation in biogeography on population.

The purpose of the Migration operation is to transfer information between different solutions, i.e. from Optimal Solution to other solutions and from other solutions to the poor solution. In this process, a habitat i is selected with probability proportional to λ_i and for each i, a habitat j is selected with probability proportional to μ_j and then the current component of habitat i is replaced with the corresponding component of habitat j. The algorithm then compares the fitness of Hi and the new solution Hi' and then keep a solution with higher fitness.

Unexpected events can change HSIs of natural habitats. This situation is modeled as SIV mutation in the BBO algorithm. The probability of species quantity is higher when number of species is medium. Each solution H_i of the population is given an associated probability number of species P_i. Based on this, each solution will have a probability π_i to mutate. The value range of the D dimension of this problem is $[l_d , u_d]$, assuming it to be a continuous optimization problem.

The goal is to select a subset of features that makes the emotion classification rate higher. This can be transformed into an optimization problem. We need to maximize the result of the cross validation of SVM. SVM is used because it has excellent learning performance, that is suitable for many classification problems.

Finite iterative calculation gives optimal solution of the problem which is also the optimal subset of emotion feature set. During the iteration the results of the SVM cross validation for each subset correspond to the HSI of the geographical environment which also serves as the criteria to judge the subset of emotional features. We want to remove the redundant feature information and retain the feature subset which greatly contributes to emotion recognition. In this process the specific values of the eigenvector should not be modified, So this process has retained the migration operation of the BBO Algorithm and removed the mutation operation as the mutation operation will change the SIV itself.

Berlin Emotional speech database was used. It has audio data of 7 types of emotions. The experiment chooses 40 audio files for each kind of emotion. openSMILE was used for feature extraction which generated 1582 dimensional speech features and SVM was used to recognize emotions.

The paper sets population size, n as 14 and the number of feature vectors in each population, m as 113. 50 is set as the maximum iteration number. The max emigration and immigration rate were set as 1. For comparison purposes genetic algorithm under the same parameter settings was selected.

As per the obtained results, initially, after the random allocation of the original feature subset, the recognition rate is below 65%. But, the BBO-SVM crosses 80% just after 5 iterations and in about 25 iterations the recognition rate for BBO-SVM reaches 90%.

It can be concluded that, The BBO-SVM algorithm can effectively reduce the dimensions of the original high dimension collection and improve the accuracy of classification. BBO-SVM appears to be more effective than ga-svm and the recognition rate can reach 90.4%. But, there is certain probability of eliminating an excellent feature in this method. This loss can be serious. A Dynamic generation of subset with unfixed length might give better results.

2nd Paper : Biogeography-Based Informative Gene Selection and Cancer Classification Using SVM and Random Forests

Microarray is a technique that helps to measure the expression levels of thousands of genes simultaneously. Genes from two different samples are taken, labeled with a different fluorescent tags (usually red and green) and hybridized onto a microarray plate. Colour and intensity give expression level of each gene on a relative scale.

Microarrays generate a lot of complex data. The number of genes is usually much more than number of samples in microarray gene expression dataset. This leads to increased difficulty in solving the classification problem by machine learning. Most genes do not contribute to the classification process.

To overcome this problem, one way is to select a small subset of informative genes from the data, known as gene selection or feature selection, which helps in tackling overfitting by getting rid of noisy genes, reducing the computational load and increasing the overall performance of the learning models.

Gene selection algorithms are mainly categorized as:

Wrappers: Make use of learning algorithms to estimate the quality or suitability of genes to the modeling problem.

Filters: Evaluate the genes considering their inherent characteristics without making use of a learning algorithm.

Methodology:

BBO has already been covered in detail in the previous paper.

The migration procedure of the original BBO algorithm is retained.

Information Gain heuristics are used as the additional information during mutation. The information gain of a gene stores the ‘information content’ of a gene with respect to the problem under consideration. We partition the informative and non-informative genes into separate sets. We want to focus on only those genes which have non-zero IG values, i.e., the genes that are informative.

Exploration or Exploitation - In mutation, the algorithm either randomly explores newer genes or exploits from the available set of genes with non-zero IG values, based on a user-defined exploration probability.

The two classification algorithms that are used are:

SVM: Employs a max margin linear hyperplane for binary linear classification problems. For non-linearly separable problems, it firstly transforms the data into higher dimensional feature and then employs a linear hyperplane. SVM with recursive feature elimination (RFE) for gene selection has shown high accuracy levels.

Random Forest: It is an ensemble of randomly constructed independent decision trees. Separate test data is not required for checking the overall accuracy of the forest. A majority vote is taken to decide on the class label for each case.

The three datasets used are:

Kent Ridge Biomedical dataset repository: The Colon Cancer dataset, 62 instances representing cell samples taken from colon cancer patients.

DUKE Breast Cancer SPORE frozen tissue bank: The breast cancer dataset, 44 samples.
Kent Ridge Biomedical dataset repository: Leukemia dataset, 72 samples.

Results:

BBO has been used with and without heuristics (Information Gain of Gene). Cross Validation with SVM and Random Forests was done. Better performance than other EA has been observed. Population average in BBO with Information gain heuristics converged to higher accuracy in lesser generations as compared to simple BBO.

BBO with SVM has outperformed BBO with Random Forests. BBO-SVM and BBO-RF have better accuracies compared to the previously best performing algorithms namely SVMRFERG , Fisher-RG-SVMRFE, ACO-AM (Ant Colony Optimization–Ant Miner) and ACO-RF and other algorithms for all 3 datasets.

Conclusion:

The hybrid BBO-SVM and BBO-RF techniques have shown consistently good results when compared against the highest accuracies with accuracies as high as 99.60%. BBO-SVM is found to be better than BBO-RF. A significant speedup in the algorithm may be achieved by parallel implementations where the classification accuracies for individual candidate solutions may be computed in parallel. It is simple to implement, robust and flexible since we can have various possible alternatives as suited to the problem and domain constraints.