Bayesian Network Classifiers versus k-NN Classifier using Sequential Feature Selection

Franz Pernkopf
pernkopf@tugraz.at

University of Washington, Department of Electrical Engineering
M254 EE/CSE Building, Box 352500, Seattle, WA, USA

1 Technical Updates

This paper aims to provide various Technical Updates to the paper "Bayesian Network Classifiers versus k-NN Classifier using Sequential Feature Selection" and justify how various technical enhancements would lead to better and desired performance and results.

1.1 Parameter Selection for Bayesian Network Classifiers

The primary aim behind the paper is to showcase the strength and utility of Bayesian Network Classifiers.Bayesian Network Classifiers such as CFS-SUN are compared to k-NN methods in the two experiments depicted in the project.The first experiment showcases CFS-SUN as the the best performing Bayesian Network Classifier,while the second experiment further demonstrates it's efficiency in comparison to other k-NN classifiers.
A technical suggestion can be given for the second experiment in this paper.The experiment is performed on 8 distinct data sets using CFS-SUN and k-NN classifiers."Table 3: Comparison of classification approaches"of the experiment showcases the results of the classification of the distinct data sets by the various classifiers.The CFS-SUN performs the best in most data sets however it is sub optimal in data sets where the number of parameters used are significantly smaller. This can be correlated to the fact that in these data sets the number of samples for learning parameters of the Bayesian network is also reduced and small,has sub optimal results for CFS-SUN classifiers.
Hence instead of selecting only a few parameters for data sets with less attribute parameters(probabilities) the CFS-SUN classifiers should consider and take in all the available attribute parameters(probabilities) to carry on all experiments and not just the second experiment in particular.To put it in another way there should not be parameter selection to select the best performing parameters using any methods but instead in such a circumstance all parameters must be used which would lead to an exponential increase in the number of attribute features.As it is known that CFS-SUN is an Unrestricted Bayesian Network Classifier every attribute is a node and can be interconnected to as many nodes as required. As a result the size of total connections and the number of nodes increase exponentially thereby training the

Bayesian Network Classifier and increasing the performance and efficiency of the estimation even if it may lead to the classifier requiring more memory and computational requirements.

1.2 Hyperparameter Tuning

A Parameter is a value that is used to define a part of the algorithm or learning process. Hyperparameter is a parameter which controls the learning process.Hyperparameter tuning or optimisation is the choosing or selection of an optimal set of hyper parameters for an algorithm or learning process which gives us the desired results and accuracy prediction. This is the reason why hyper parameter optimisation must be a necessary step in all learning processes.
K-NN classifiers is used throughout the experiment in order to compare efficiency of Bayesian Network Classifiers(CFS-SUN).Primary hyper parameters for k-NN algorithm include nearest neighbour,leaf node size and distance metrics.Out of which nearest neighbour is predecided and the other two hyper parameters can be optimised to find the best performing variables for each data set with respect to the training model.Hence optimised performance can be seen for k-NN classifiers.
Bayesian Network Classifiers are the primary algorithms or learning processes used in the experiment and can be optimised as well with the help of hyper parameter tuning.Parameter space can be estimated and the number of fold iterations to be done can be decided along with learning rate for the process.Bayes Optimisation search is invaluable in giving back the optimal features to be used in improving the Bayesian Network Classifiers' accuracy.

1.3 Avoiding Overfitting

Overfitting is a major issue that affects many machine learning algorithms and may hamper the outcome of the given experiments.It is caused due to the over training of the model with excess features and data points,a technical improvement would be to restrict the number of features selected,this could be done with an iterative approach to determine how many features would be the optimal for each algorithm to train upon to give the best accuracy result.As a result the model will not be constrained to many parameters and be flexible enough to assign classifications even in case of slightly abnormal data points which could not be done in an over fitted or over trained model.
Data Processing is a step in any machine learning application where in the data to be experimented on is transformed into a state or factor that is more sought after or beneficial in execution of the entire code.In both the experiments depicted in this experiment the data sets are specified,however there is no mention of any data preprocessing that is done before or during the experiment.The first step is to always ensure standardised data.Standardised data is essentially scaling all the values of features so that all values are between 0 and 1 in value.This helps in having a more smooth distribution and ensures less classification problems when comparing features.Another data processing technique is to extract outliers or noisy data values.These data points do not have coherent value and have feature values much different from other data points.Such data could either be manually classified or their could be a separate algorithm for such outliers. This would help train the model in a more uniform way without the interference of outliers which could hamper the efficiency during classification.