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 Critical Analysis

This paper aims to provide a critical overview of the given paper to highlight the positive and negative aspects,along with bringing forth errors observed in the method or during experimentation.Technical suggestions and enhancements have been offered along with directions of future research in this domain.

1.1 Positive Aspects/Pros

The highlight of the paper was the clear and concise explanation of various possible classifiers as well as the basic functioning and underlying mathematical/graphical principles. This made way for the reader to understand the approach taken and the reason for choosing each classifier.Furthermore almost all algorithms for sequential feature selections were mentioned and discussed and various combinations of these approaches were taken to highlight the performance of various algorithms to provide a clearer picture for the primary problem statement chosen for this research paper.
The paper had various references mentioned over the text so the reader could be directed to each segment or reference paper and clearly understand the underlying process and methods in continuity to the paper.
The paper includes two experiments,the first experiment laying a foundation for deciding the best performing Bayesian Network Classifier and how it compares with respective k-NN Classifiers. The second experiment further demonstrates the performance and strength of the Selective Unrestricted Bayesian Network Classifier in comparison to the k-NN methods.This two-fold approach is quite effective as it consolidates the idea of the project and reinforces the same thereby highlighting the final verdict to the proposed problem statement.

1.2 Negative Aspects/Cons

The paper fails to throw enough light on the Search and Score learning technique on which the Selective Unrestricted Bayesian Network Classifier is based upon. Hill Climbing search

and floating search algorithms could have been explained in more depth.Only the more efficient process is mentioned however the principals and algorithms behind these techniques could have been enumerated.
Various Sequential Feature Selection Algorithms were employed during experiments however more explanation could have been provided as to why certain techniques fare poorly in comparison to other techniques instead of listing only optimal subset sizes for classification.
Although the paper was published in 2004 keeping in mind the current computational capabilities,future insight could have been shown by giving less weight-age in criticising k-NN algorithms as computationally expensive and demanding algorithms. Accuracy could have been made the primary determining factor with space and computational efficiency being relevant factors only as the ability to alleviate such problems would be easy with future advancements.

1.3 Technical Errors

This paper primarily employs known algorithms with different feature selection methods which do not show any technical shortcomings in the method.However in showcasing the outcomes of the first experiment an experimental technical error is seen in this paper."Table 1: Comparison of classification approaches" is a table provided in the paper that shows the various results as the outcome of the first experiment with the bold values representing the best results.According to the table the average percentage of 5 cross validation data sets is highest for SFFS-1-NN-C followed by SFFS-3-NN-C.
Hence this table clearly highlights that the SFFS-1-NN-C classifier outperforms the CFS-SUN classifier in the cross validation sets and scores the same accuracy result in the Hold-out(H) data set.As a result the most optimal classifier for this experiment must be the SFFS-1-NN-C classifier however it is not accredited with this due to the reason that the k-NN decision rule searches through a labeled reference set for the nearest neighbors which might be time consuming in case of a large number of samples as well as a large amount of memory is required.Thus Bayesian Network Classifiers such as CFS-SUN is said to outperform selective k NN methods in terms of memory requirements and computational demands during classification. However there must be mention that SFFS-1-NN-C classifier and other similar methods provide the best cross validation accuracy and are more optimised for smaller number of samples and even for larger set of samples provided computational and memory requirements are present.

1.4 Technical Suggestions

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" 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 such data sets the CFS-SUN classifiers should consider and take in all the parameters(probabilities) in case of a total small number of probabilities in the given data sets,that is there should not be parameter selection to

only select the best performing parameters but that all parameters must be used which would lead to an exponential increase in the number of nodes in the Bayesian Network thereby increasing the accuracy of estimation.

1.5 Directions For Future Research

In the research paper computational and memory limitations were mentioned at several stages which should be relaxed going forth. The computational limits of modern machines will with the future grow at an exponential rate hence k-NN methods which were deemed computationally sub optimal to the CFS-SUN method,should be revisited. Further more emphasis in the research must be given to high accuracy rather than computational or memory limitations and various factors for Unrestricted Selective Bayesian Network Classifiers can be worked and improved upon such as increasing the feature attributes by taking in probability values which leads to an exponential increase in nodes in the Bayesian Network.This would lead to better estimation accuracy despite using more memory.