Multiple Clustering

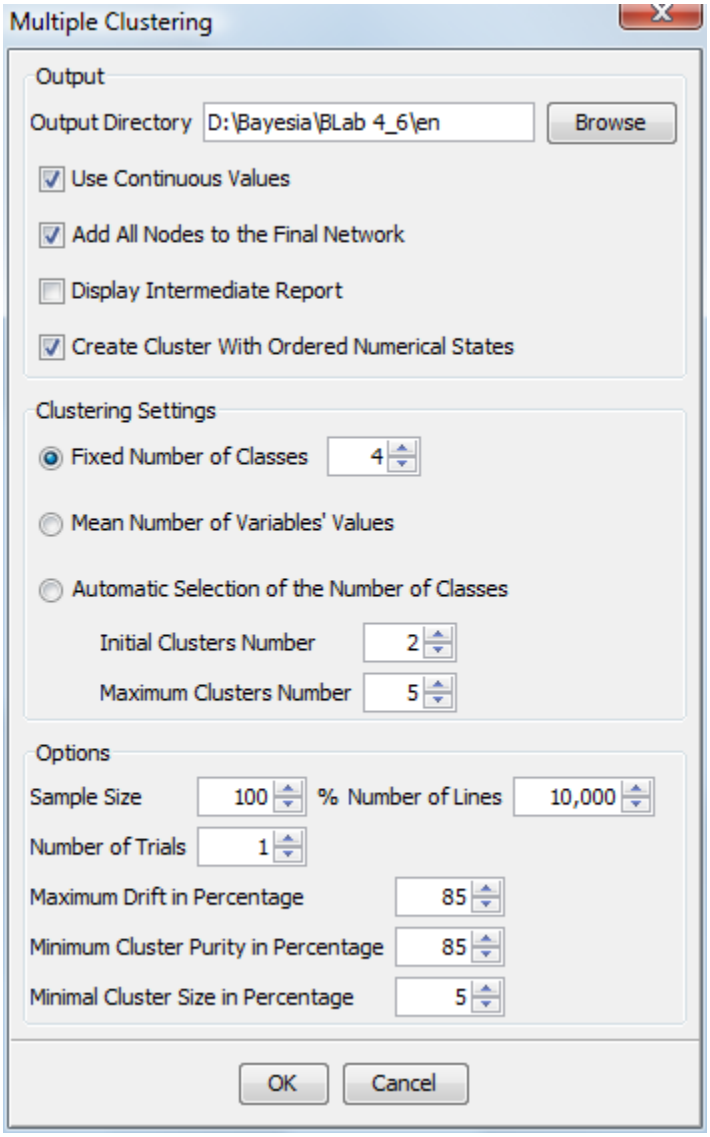
This tool allows carrying out **data clustering** on various subsets of variables of the same Bayesian network. The subsets are defined by the **classes** that ar named **[Factor_i]**. The variables of each class **[Factor_i]** are then used to induce a new variable (the latent variable, i.e. without any corresponding data in the file) that summarizes them.

Even if it is possible to manually define these classes, they can be created automatically by the **variable clustering** tool.

A Bayesian network is created for each of these classes. This network contains the variables that belong to the class and the synthetic variable Cluster na med **[Factor_i]**. This variable is added to the corres- ponding class **[Factor_i]**. It is set as **not observable**. If the initial network contains fixed arcs, they wil l be added to each new Bayesian network if they are included in it, i.e. the extremities of each arc are contained in the network. At the end of the last data c lustering, a final Bayesian network is created. It contains all the Clusters and comes with an internal database where all the values of these latent variables have been inferred with respect to their corresponding network. It is also possible to add to that final Bayesian network the initial variables (the manifest variables).

You must note that the sets containing only one node are not taken in account for data clustering.

The following dialog box allows entering the different parameters for multiple clustering. It naturally reuses most of the parameters used for **data clustering**.



In the Output area, the wizard allows selecting the directory where the various generated networks will be saved (a network by class **[Factor_i]** and the final network with all the latent variables). The interme- diate and final databases are saved with the generated networks. The continuous values can be saved with the databases. This wizard also allows us to add or not all the nodes of the initial network to the final one. It is also possible to display or not the interm ediate report generated for each intermediate network. However, the intermediate reports will always be saved in the target directory.

It is possible to create, for each network, a cluster node with ordered numerical states. These values are the mean of the score of each connected node for each state of the cluster node. If two of these values are strictly identical, an epsilon is added to one of them to obtain two different values.

As in **data clustering**, the number of values of the latent variables can be a priori fixed or found by a random walk. It can also be defined as being equal to the average number of values of the variables belonging to **[Factor_i]**. The remainder is strictly identical to **data clustering**.

At the end of each clustering, an algorithm allows finding automatically if one of the **[Factor_i]** node's states is a **filtered state** or not. If so, this state is marked as filtered.

At the end of each clustering, an automatic analysis of the obtained Bayesian network is carried out and a **target analysis report** is generated. This report is identical to the one generated by **data clustering**. It can be displayed if the convenient option has been selected. However, it is always saved in the targ et directory.

At the end of the last clustering, a synthetic report is generated. At the beginning of the report, a summary indicates the number of factors found, and the mi nimum, average and maximum number of clusters, mean purity and contingency table fit.

This report describes, for each latent variable, the mean purity, the contingency table fit and the deviance. Those indices are described in the **Correlation with Target Node's report**. This new measure can be used to qualify the clustering result, to measure how well the Joint Probability Distribution is represe nted through each **[Factor_i]** variable.

It also describes the distribution of its values on the learning set, and the list of the nodes sorted according to the quantity of information brought to its know ledge (cf. **Target analysis report**). The final network is automatically opened and the final database is associated with it. If the initial database contains a t est set, it is also transferred and the missing value imputation is performed on the new **[Factor_i]** variables. At least, the final database is saved in the target directory.