ArcGIS Bayesian Classification Tool Addin

User's Guide

CSER http://www.gpem.uq.edu.au/cser/
University of Queensland
Queensland, Australia

Getting Started

Background

The Bayesian Classification Tool applies probabilistic inference to each feature in a GIS layer to generate new information. A Bayesian network is a graphical model that encodes probabilistic relationships between variables of interest. There are many good texts on Bayesian networks (see References section), so we will not explain their concepts here. We assume the user is familiar with Bayesian networks and their construction; and our focus is on the linkage between Bayesian networks - as a powerful means of knowledge representation and inferencing – and GIS. A feature in a GIS layer contains descriptive attributes for entities in the world. Likewise a Bayesian network is a model of the world that represents relationships between variables; these variables may match up with the attribute fields in a GIS layer. Hence the benefits of combing Bayesian networks and GIS are apparent; one provides a model of the world and other provides data about the world.

The main use of this tool is to enter feature data into a Bayesian network for *evidence* variables, and allow the Bayesian network to infer the possible states of *classified* variables (called beliefs in Bayesian terminology) which then update designated feature attributes. We will see in a later section on Concepts that the attributes have different roles within practical classification problems.

Use

The Bayesian classification tool may is used to do classification with a Bayesian network model. Technically this is called belief updating or probabilistic inference where beliefs (posterior probabilities) are found for classification attributes to reflect the evidence attributes entered for a feature. The prior probabilities for variables are captured as joint probability distributions for nodes in a Bayesian network.

The two modes of use for the tool are:

i) Classification where the user has developed a Bayesian network and apply this to data in a GIS
ii) Scenario analysis where the user modifies the prior probabilities by changing the condition probabilities for certain variables in the network.

We note that a GIS is also useful for exploring data as part of the development stages in a Bayesian network model. For instance, statistical mapping of data provide valuable insights for discovering states and discretization of continuous variables.

To use the Bayesian Classification Tool it is assumed you have a valid Bayesian network. However it is useful to have guidelines on ways to develop the network for common assessment problems. The Concepts and Quick Start sections give useful insights and templates for developing Bayesian network solutions.

Software / Prerequisites

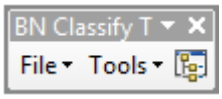
BN Classification requires:

- ESRI ArcMap version 10.0 or later from ESRI www.esri.com
- Netica version 3.1 or later from Norsys Software Corp. www.norsys.com. You can download a free copy of Netica, but a license to run Netica is built into this application. However it is advisable to purchase Netica to build unlimited sized models.

Installation

To install BN Classification:

1. Download AddIn and demo data for ArcGIS from: www.gpem.uq.edu.au/cser-tools
2. Double click on the file 'BNClassify.esriAddIn' in the Windows Explorer. This will run the ESRI Add-In installation utility - in the dialog click *Install Add-In*.
3. Start ArcMap. If the menubar (see below) is not visible then click on *Customize* in ArcGIS menu, and select sub-menu for *Toolbars*; click *BN Classify Toolbar* to enable.

   Instuctions for using software is given in later section *Menus and Windows*, but would advise new users following subsequent section for *Quick Start* tutorial.

4. Optional. You can manage AddIns from the *Customize* menu select *Add-In Manager*. You should see BNClassify listed. 4. The BN Classification toolbar should appear.

Concepts

Bayesian Network (BN) Model Structure

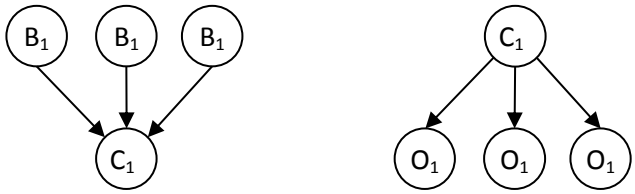
The BN Classification tool follows a common structure for diagnostic networks; namely we classify one or more variables based upon information given in other variables. We assume a BN model has been developed and the objective is to apply this to a database to infer certain attributes. The variables in the BN model and GIS can be categorized into types by their roles and dependencies. For instance, we identify classification variables that compute findings to write back to the GIS.

BN Classification adopts a general model structure that distinguishes four types of variables. This structure is explained in more detail by Kjaerulff and Madsen (2008).

1. *Classification variables* are nodes whose values we want to know, or in other words we want to compute the posterior probability for. We assume a BN model exists and it expresses what we know or believe about the information in our GIS, and the classification variables are the output or explanations inferred from the model.

2. *Information variables* provide information in the model and GIS relevant to solving the classification. If the classification is the output, then information variables can be considered as inputs. We distinguish two types of information variables:
    a. *Background variables* define explanatory conditions that have a *causal* influence on the classification. They are referred to as common-effects (Korb and Nicolson, 2004) as the classification is seen as correlated to certain background conditions. In a landscape setting, background information may be the land use, the type of vegetation, slope, etc.

    b. *Observation variables* indentify information that is observed as a consequence of a particular classification. They are referred to as common-causes (Korb and Nicolson, 2004) as the classification is correlated with observations. In a landscape setting, observation information may be infiltration rates, erosion occurrence, etc.



a) Common-effects b) Common-cause

3. *Mediating variable* defines internal conditions to moderate the combined causes of nodes. This may reflect logical or weighted interactions between causal nodes which properly represent independence between the effect nodes. In a landscape setting you may be classifying threats related to flooding and drought, and want to enforce they do not occur together. The mediating variables are considered unobservable and are defined only in terms of a conditional probability table.

A generic overall structure of problems in a BN is shown below. We do not claim is has universal applicability, but many problems may be conceptualized in terms of this structure.

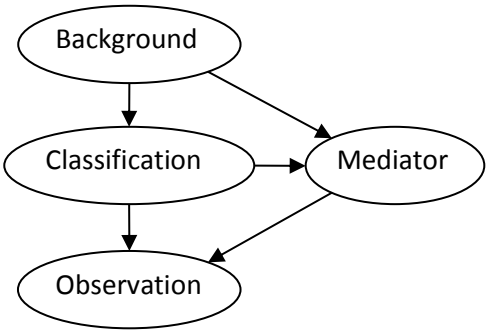


Figure. Generic causal structure for types of variables.

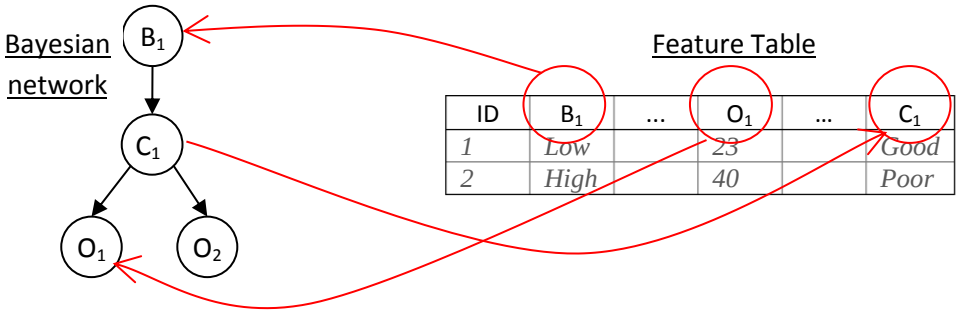
Bayesian Network (BN) Model-GIS Interface

We use the causal structure to constrain the way nodes in a BN match up with attributes in a GIS. For instance, classification nodes must be discrete. The table below summarizes constraints on nodes. Scalar types define if nodes have discrete states and/or associated real values. Attributable defines if nodes have associated attributes in the GIS and their read/write mode.

Table 1. Data type and attribute access for classification categories.

	Background	Classification	Mediator	Observation
Scalar type:	*Discrete or Continuous*	*Discrete*	*Discrete*	*Discrete or Continuous*
Attribution:	*Read*	*Write*	-	*Read*

The model-GIS interface allows subsets of nodes in a BN to be associated with feature attributes in a layer. Each feature or row in a layer is an instance whose attribute values can be: i) input as findings or evidence in a BN, or ii) output for beliefs from a classification. Not all nodes in a BN need to be associated with an attribute in a GIS; this is shown schematically below.



Attributes for each feature in a layer are entered as findings in a BN and the inferred classification (namely the marginal posterior probability distribution) is output back to the respective feature. This is repeated for each feature, i.e. row in GIS layer.

The variables in a BN - that we deal with - use a multinominal representation for a probability distribution. Or in other words, the probability distribution is discretized into sub-ranges with an associated label. For example, a variable may have *states* {low,high} with state sub-ranges [low=25%,high=75%], these are called *chance* variables. All variables use probability distributions internally, but sometimes we want variables to have an absolute state, these are called deterministic variables. This signifies the variable has to be completely in one state or another, i.e. [low=100%,high=0%] or [low=0%,high=100%] for states low and high respectively.

Internally all variables in a BN are probability distributions, but they are related to real world quantities by a name (i.e. the named states) or a value. A BN variable is discrete if it has a finite number of possible values, and it is quantified by a name or number. For example, a discrete variable may have a finding quantified by a state of 'low' or a number value of 1. Alternatively a variable may be *continuous* if it has values within a continuous range, these are quantified by a number. For example, a continuous variable may have a finding quantified by a value of 1.5 which is in a range 0..5 associated with a state. Knowing the type of variable - discrete or continuous – is important for relating BN variables to GIS attributes.

Variables and GIS Attributes

The variables in a BN are matched to attributes in a GIS layer. They may be given in the GIS as: i) a precise quantity (e.g. a point estimate), ii) an imprecise quantity (e.g. a distribution). There are two ways to give imprecise values that depend on the variable type. A discrete variable may have a certainty attribute from 0% to 100%. For example, an attribute value of 'high' with 75% certainty is an imprecise estimate with a distribution of [low=25%,high=75%]. For discrete variables, the uncertainty (100%-certainty attribute) is uniformly given to the other possible states. A continuous variable may have a standard deviation attribute value, that is we assume it has a normal distruibution. For example, an attribute value of 1.5 with a standard deviation of 5 given for a continuous variable with state 'low' in the range 0..5 and state 'high' in range 5..10 will have an imprecise estimate with a distribution of [65%,35%]. Quantities are represented in the GIS with one or more attributes; these options are shown below.

Table 2. Types of BN nodes and the way they are attributed in a GIS. For example assume a BN variable A has states ['low','high']; as a discrete type this may be quantified as $A_{low}=1,A_{high}=2$, or as a continuous type may be quantified by ranges $A_{low}=0..5,A_{high}=5..10$.

Quantity	BN Type	BN Kind	GIS Attributes	Example in GIS	Distribution in BN
Precise	Discrete	Deterministic	*<value‡>*	*A= 'low' or A=1*	[low=100%,high=0%]
		Chance	*<value†>*	*A= 'high' or A=2*	[low=0%,high=100%]
	Continuous	Deterministic	*<value‡>*	*A=1.5*	[low=100%,high=0%]
		Chance	*<value†>*	*A=6*	[low=0%,high=100%]
	Discrete	Chance	*<value‡, certainty†>*	*A='low',80*	[low=80%,high=20%]

Imprecise	Continuous	Chance	*<mean†, stddev†>*	*A=1.5,5*	[low=65%,high=35%]
Probability	Discrete or Continuous	Chance	*<value†, valve†>*	*A=25,75*	[low=25%,high=75%]

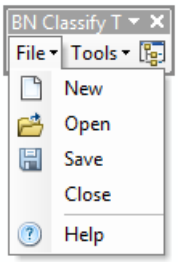
‡ is *string or number data type* † *is number data type*

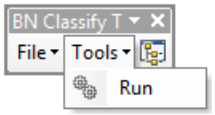
Menus and Windows

BN Classification Toolbar

The Toolbar has menus for file management of data and other support functions. All data for a BN classification may be stored in a file, and re-opened and changed at a later date. The file has an XML-encoded ASCII format and is best stored in the same folder as the Netica BN files.

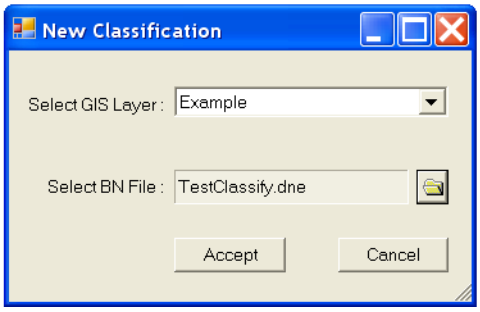
To start a new classification you need to identify the GIS layer and the Netcia BN file. A window will appear with the classification categories, you use this to identify associations between the BN nodes and GIS attributes.





Toolbar menu

New	*Start new classification (see form below)*
Open	*Open stored classification*
Close	*Close current classification*
Save	*Save current classification*
Save As	*Save current classification as*
Help	*Show this help*
Run	*Run classification*
Window	*Toggle display of classification window*



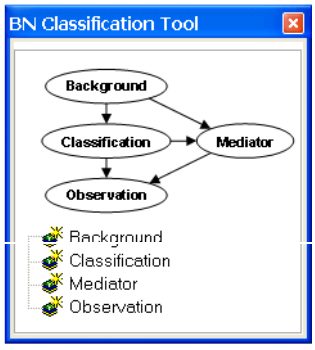
New Classification Form

Select a GIS layer

Open a Netica BN file

BN Classification Window

The Window lists the BN classification categories. A popup menu appears when you right click on each category. Generally you will add, view properties, view the conditional property table, and delete BN nodes under each category. See Table 1.

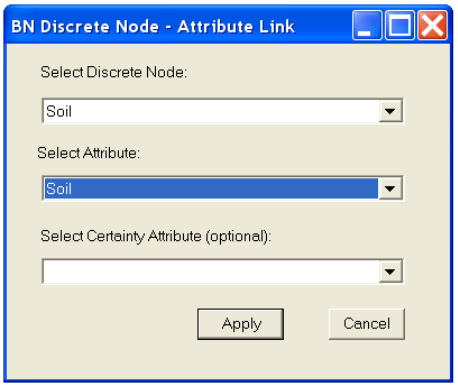


Classification window

Right click for popup menu options

Add nodes from the BN model under the categories for Background, Classification, Mediator, and Observation. Different forms are used to add nodes that are discrete, continuous or a probabilities. See table 2 in section on Variables and GIS Attributes.

Associate a (discrete) BN node with GIS attribute(s)

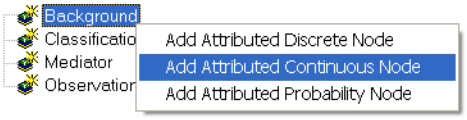


Select node in BN file

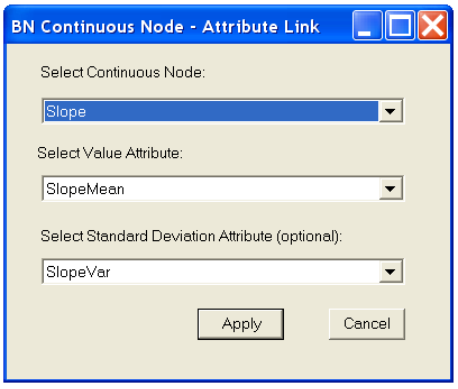
Select attribute in GIS layer for node

Optionally select an attribute in GIS layer with a certainty probability (0%-100%)

(Only Chance nodes can have a certainty)



Associate a (continuous) BN node with GIS attribute(s)

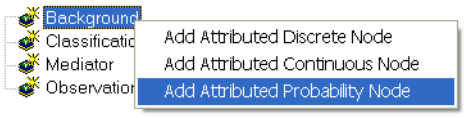


Select node in BN file

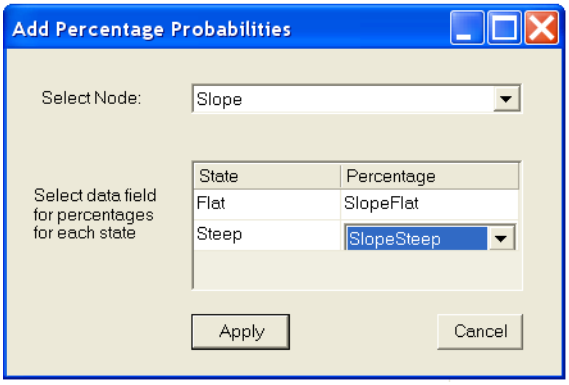
Select attribute in GIS layer for node

Optionally select an attribute in GIS layer with standard deviation

(Only Chance nodes can have a standard deviation)

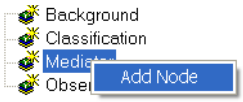


Associate a (probabilistic) BN Node with GIS attributes

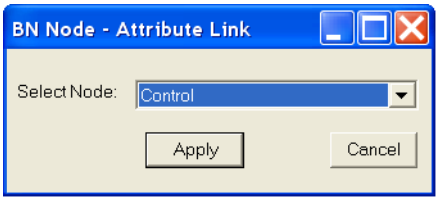


Select node in BN file

Select attributes in GIS layer from drop down list for each respective state (probability sub-range)



Associate a (table) BN Node



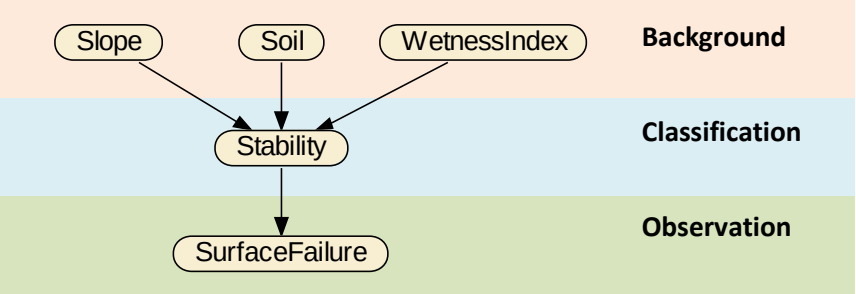
Select node in BN file

Quick Start

A simple application is explained to demonstrate how to setup and run a BN classification.

Assessing slope stability is a frequent engineering problem. We use simple rules of thumb to develop a BN model. We assume that slope stability has influential factors for: i) terrain slope, ii) soil type, and iii) soil wetness. We can also visit a site and observe the slope conditions for surface failures; there could be evidence of soil creep or embankment slips.

The figure below shows relevant variables and their causal dependency for this problem. Background variables for influential factors include: slope, soil and wetness index. An observation variable for an indicator of slope stability is surface failure. The classification variable we wish to compute is stability.



The GIS layer has attributes corresponding to nodes. One or more attributes are required for a node. The table below shows how BN variables are matched to GIS attributes.

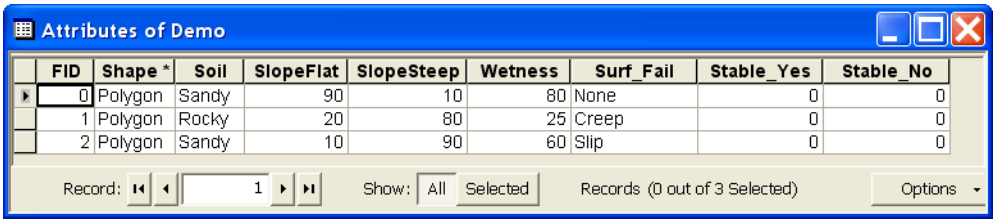
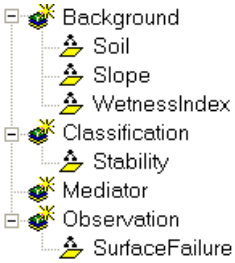


Table 3. Setup to match BN nodes with GIS attributes

Node	Model	BN Node Type	BN Attribute	GIS Attribute	Comments
Slope	*Background*	*Probability*	*Flat* *Steep*	*SlopeFlat* *SlopeSteep*	*Probability % for each state stored as number 0-100*
Soil	*Background*	*Discrete*	*Soil*	*Soil*	*Label for soil state stored as string*
Wetness Index	*Background*	*Continuous*	*Wetness (cm)*		*Sub-ranges for low:0-50, high: 50-100*
Stability	*Classification*	*Probability*	*Stable* *Unstable*	*Stable_Yes* *Stable_No*	*Probability % stored for each state as number 0-100*
Surface Failure	*Observation*	*Discrete*	*SurfaceFailure*	*Surf_Fail*	*Label for fail state stored as string*

Add each attribute to the appropriate classification category. Save the classification to a file. Your BN classification window should similar to the figure below.



From the toolbar menu Run the classification. A message appears when calculations are complete. Attributes associated with a classification component are updated in the GIS.

A Netica application opens in the background when you run a classification. Inspect the results and compare to the BN model.

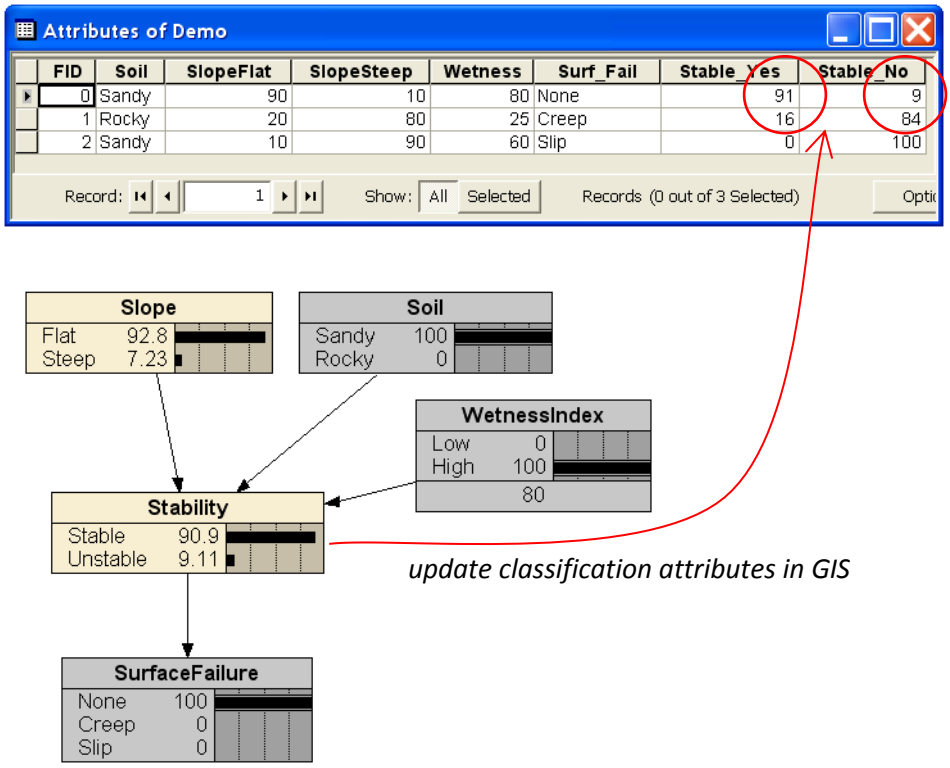


Figure. Findings are entered in the BN, and the classifications beliefs update attributes in the GIS.

References

Kjaerulff, Uffe B. and Madsen, Anders L. (2008) Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis. Springer, p.318.

Korb, Kevin B. and Nicholson, Ann E. (2004) Bayesian Artificial Intelligence Bayesian Artificial Intelligence. Chapmann & Hall, p.364.