Shahid Beheshti University
Mathematical Sciences Faculty

Title: the summery of

# Multi – dimensional Bayesian Network Classifier

Lynda C. Van der Gaag. &   Peter R. de Waal
Department of information and computing science

Utrecht University , Nederland

**Course:**
**Probabilistic Graphical Models**

**اسمعیل مفاخری**

July 2019

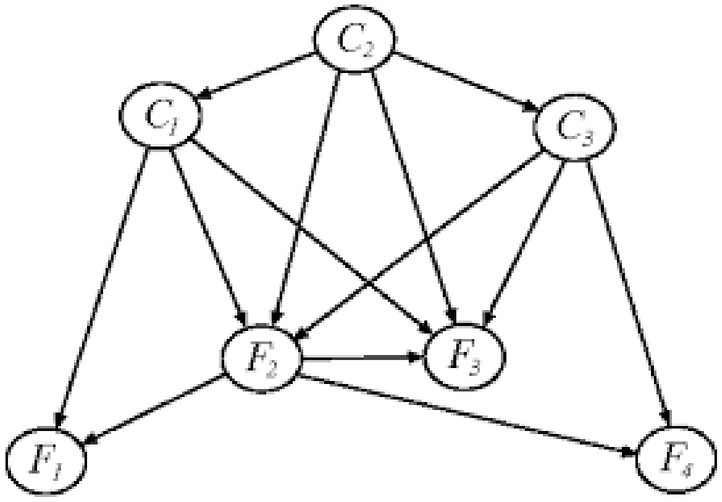
# Definition

- Bayesian Network Classifiers for solving classification problems where an instance described by a number of features has to be classified in one of several classes
- A Multi- dimensional BN Classifier include one or more class variables and one or more features variables
- It models the relationship between the variables by acyclic directed graph over the class and over the features and connect tow set by a bi-partite directed graph

# Graph

An Example multi dimensional BN classifier

with class variables $C_i$ and feature variables $F_j$



# Definition Bayesian Networks

- we consider BN over a finite set $V=\{ X_1,\dots , X_k\}$ k>=1
- $X_i$ random variables with distinct values in finite set $Val(X_i)$
- BN is a pair

$$B = < G , \theta >$$

G : acyclic directed graph that vertices are random variables **V** & **θ** parameter

- B define a joint probability distribution over V according to

$$P(X_1 ,\dots\dots , X_k) = \Pi\theta_{xi|\Pi xi}$$

- The set V of BN partitioned into tow set

$$V_f = \{F_1, \dots . F_n\} \quad \& \quad V_c=\{C\}$$

of features and class variables

# The problem

- The problem of BN classifier from a dataset $D = \{u_1, ..... u_N\}$ N>=1 is to fined one that the best matches the available data

  often the log likelihood is used

- The log-likelihood of B given data set D define as :

$$\mathbf{LL\ (B|D)} = \sum_{i=1}^{N} \mathbf{Log\ (PB(ui))}$$

This problem is solved in polynomial time

Example: NB and TAN

# Multi- dimensional Classifier definition and notations

- a Multi- dimensional Bayesian Network Classifier is a BN of graph $G = \langle V , A \rangle$
- $V$ : Random Variable and A set of arcs

$$V_C = \{C_1, \ldots , C_n\}$$ of class Variable

$$V_F = \{F_1, \ldots , F_m\}$$ of Feature Variable

- $A$ partitioned to three part :

$$A_C , A_F , A_{CF}$$

With following properties:

# Multi- dimensional Classifier notations

- For each $F_i \in V_F$ there is a $C_j \in V_C$ with $(C_j , F_i) \in A_{CF}$

and for each $C_i \in V_C$ there is an $F_j \in V_F$ with $(C_i , F_j) \in A_{CF}$

- The subgraph of G that is induced by Vc equals

$$G_C = < V_C , A_C >$$

- The subgraph of G that is induced by $V_F$ equals

$$G_F = < V_F , A_F >$$

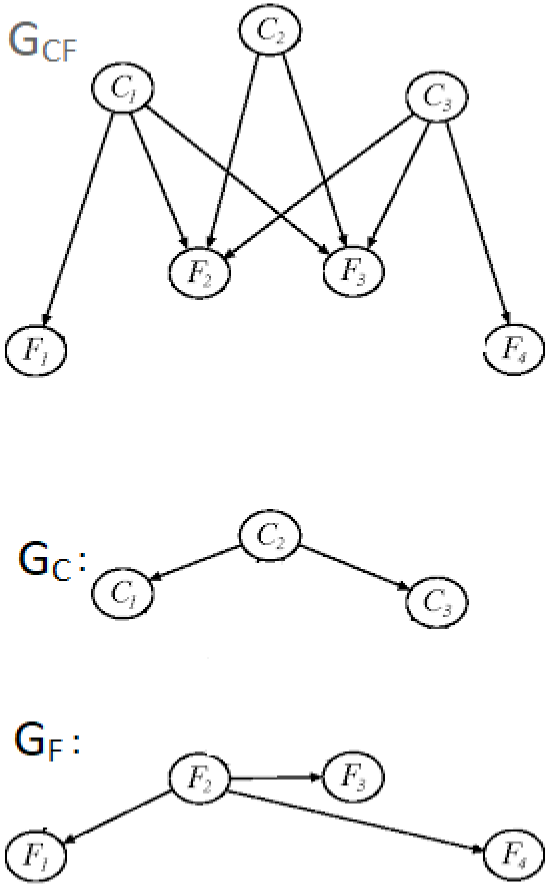
# Notations of Multi-dimensional BN

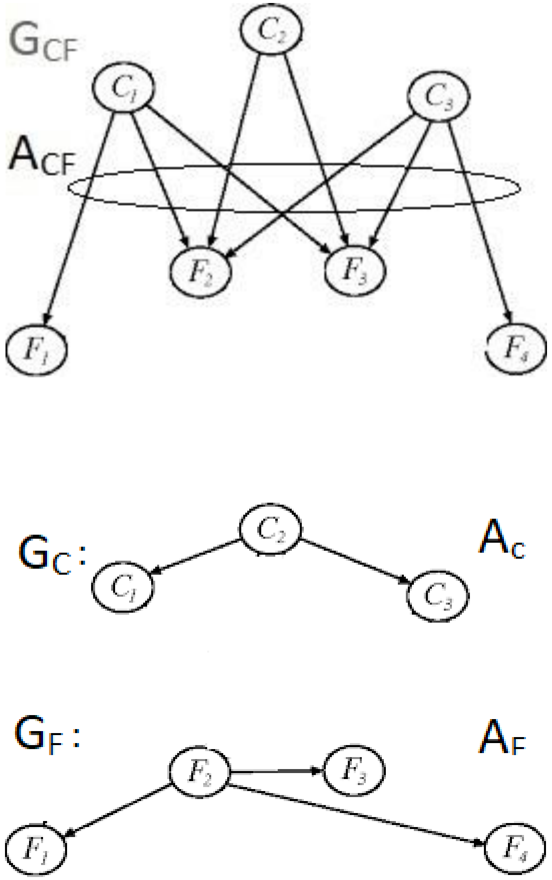
- $G_C$ is called **classifier subgraph**
- $G_F$ is called **feature true subgraph**
- $G_{CF}$ is called **feature selection subgraph** that is a bi-partied graph that relates the features to the classes.

  $G_{CF} = < V , A_{CF} >$
- $A_{CF}$ set of arcs called feature selection arc set
- $\Pi_C X$ denote as **class parent** of X in G
- $\Pi_F X$ denote as **feature parent** of X in G

## Multi- dimensional Bayesian Network Classifier

# Multi- dimensional Classifier subgraphs



# Multi- dimensional Classifier examples

- Nave Bayes as a special case of multi dimensional classifier with both class subgraph $G_C$ and feature subgraph $G_F$ have empty arc set
- This subfamily of bi-partite classifier includes the one dimensional Nave Bayes
- Another type of multi-dim in which $G_C$ and $G_F$ are directed trees Or fully tree-augmented Multi- dimensional classifiers that this paper focus on

# Complexity

- finding highest posterior probability for a Multi- dimensional Bayesian Network Classifier in general is

**NP-hard**

**But yet can be solved in**

**Polynomial time**

**for networks of bounded treewidth**

**(Bodlaender 2002)**

Multi- dimensional Bayesian Network Classifier

# Learning problem of Fully tree-augmented

- We defined $A_C$ , $A_F$ , $A_{CF}$ before, now we define another subset $\underline{A}_{CF}$ of $V_C \times V_F$ such that <V , $\underline{A}_{CF}$> is a feature selection subgraph
- A fully tree augmented is admissible for $\underline{A}_{CF}$ if we have

$$\underline{A}_{CF} = A_{CF}$$

The set of all admissible $A_{CF}$ classifier for $\underline{A}_{CF}$ is denote as $\beta_{CF}$

# Learning problem of Fully tree-augmented

- The learning problem now is to fine the set of admissible classifier that best fit the available data
- How well a model describe the data we use its log-likelihood given the data
- Formally the learning problem for tree augmented multi-dimensional with a fixed feature selection arc set $\underline{A}_{CF}$ is to fined a classifier B in $\boldsymbol{\beta}_{\underline{A}_{CF}}$ that maximize :

$$\mathbf{LL\ (B\,|\,D)}$$

# Solving the learning problem

- Consider B with class variables $V_c$ and feature $V_F$ that is admissible for feature selection $A_{CF}$ the log-likelihood of B given a data set D can be written as:

(Freidman 1977)

- LL(B|D) =

= -N.$\sum_{i=1}^{n} HPD( Ci|\Pi Ci)$ + N. $\sum_{j=1}^{m} HPD( Fj|\Pi Fj)$

=N.$\sum_{i=1}^{n} IPD( Ci, \Pi Ci)$- N.$\sum_{i=1}^{n} HPD( Ci) +$
N. $\sum_{j=1} IPD( Fj, \Pi Fj)$ - N.$\sum_{i=1} HPD( Ci)$

- $P_D$ is the empirical distribution from D
- $H_P(X)$ = -$\sum_{i=1}^{n} P( X) \, logP(X)$ is Entropy of X
- $H_P(X|Y)$ = -$\sum P( x, y) \, logP(x|y)$ is conditional Entropy of X given Y
- $I_P(X|Y)$ = $\sum P( x, y) \log(P(x|y)/P(x)P(Y)$ is mutual information of X and Y

# Solving the learning problem

- Hp (Ci) and Hp (Fj) depend on empirical distribution not on graphical structure of classifier this implies that can maximize the log likelihood given the data sum of its tow mutual information terms
- Finally : a classifier that solve the learning problem for fully tree augmented multi dimensional with the fixed feature selection $\underline{A}_{CF}$ is a classifier from $\beta_{ACF}$ that Maximizes :
- $\sum_{i=1}^{n} IPD(Ci, \Pi Ci) + \sum_{j=1} IPD(Fj, \Pi Fj | \Pi cFj)$
- The learning can be decomposed into tow separate optimization problem which can be solve in polynomial time

# Solving the learning problem

- Class variables have only class parents depend on the Ac class subgraph only this term depend on feature selection arc set ACF

Fixed on AF

**mutual-information class variables** maximized by using algorithm chow – Liu (1968)

1- construct a full undirected G over $V_c$

2- assign $I_{PD}(C_i,C_j)$ to each arc $c_i$-$c_j$ $i \neq j$

3- build a maximum weighted spanning tree (Kruskal algorithm 1956)

4-transfer undirected tree to direct one by arbitrary variable for its root and setting all arc direction from the root outward

# Solving the learning problem

- **mutual-information Feature variables** maximize by finding maximum likelihood directed spanning tree over feature by following

1- make complete directed graph over $V_F$

2- assign weight $I_{PD}(F_i , F_j \mid \Pi_C F_j)$ to each arc from $F_i$ to $F_j$ $i \neq j$

3-build maximum-weighted directed spanning tree by (chow & Liu 1968 or Edmonds algorithm1967)

# Solving the learning problem

Note that:

- Maximize mutual information feature need to directed spanning tree but for class variables we compute undirected one
- IPD(Ci,Cj) = IPD(Cj,Ci) for **class variables**
- IPD(Fi , Fj | ΠCFj) ≠ IPD(Fj , Fi | ΠCFi) for **feature variables**
- This algorithm can be formulated classifiers which $A_c$ or $A_F$ is empty ( NB or TAN)
- Complexity of weights for undirected tree O($\boldsymbol{n}^{\mathbf{2}}\mathbf{N}$) and make tree itself is O($\boldsymbol{n}^{\mathbf{2}}\mathbf{logn}$)
- Complexity weights for directed tree O($\boldsymbol{m}^{\mathbf{2}}\mathbf{N}$) and make the tree itself is O($\boldsymbol{m}^{\mathbf{3}}$)

# Feature subset selection

- **Feature subset selection** is finding a minimum subset of feature such that selective classifier constructed has highest accuracy
- This problem in general is **NP-Hard** (Tsamardinos 2003)
- Wrapper approach (kohavi and john 1997) for feature selection (forward selection) or (backward elimination)

1- choose the empety feature selection subgraph

2-generate all possible feature selection obtained by adding an arc from class to variables

3- compute the accuracy of the best classifier

4- select the best subgraph that is feature selection

5- if accuracy is higher than current subgraph go to 2 if not **stop** and choose the best classifier for current subgraph

# Experimental Result

- Three Data set: from the oesophageal cancer with 42 variables of 25 arc feature variables 100, 200 , 400 samples
- Construct fully NB and fully tree augmented multi dimensional
- And using forward selection wrapper approach
- Using 10-fold cross- validation the result summarized in table
- attention to accuracy and number of parameter that were estimate for classifiers

| Classifier type | Accuracy 200 samples | Accuracy 400 samples | # parameter 200 samples | # parameter 400 samples |
|---|---|---|---|---|
| Compound nave | 0.420 | 0.550 | 661 | 732 |
| Multi- dim nave | 0.555 | 0.605 | 179 | 276 |
| Compound TAN | 0.305 | 0.505 | 3060 | 4604 |
| Multi- dim FTAN | 0.475 | 0.585 | 1092 | 386 |

# Conclusion

- We introduce a new family of BN classifier include one or more class and multiple feature that need not be modeled as being dependent upon every class variable.

- We formulated learning problem for this family and present a solution algorithm that is polynomial in the number of involved.

- Our experimental result illustrate the benefit of multi-dimensionality of our BN network classifier.

- In the future we perform more experimentation study of our learning algorithm for other data sets and other approaches for feature subset selection.

# References


- H.L. Bodlaender , F. van der Eijkhof and L.C. van der Gaag. (2002) complexity of MAP problem
- V.K. Chow and C.N. Liu. (1968) approximating discrete probability distribution
- Y.J Chu and T.H Liu (1965) shortest arborescence of directed graph
- T.M. Cover and JA Thomas (1991) Elementary of Information Theory
- J. Edmonds (1967) optimum branching's.
- N. Friedman D. Geiger(1997) BN classifier Machin Learning
- E.J. Keogh and M.J. Pazzani (1999) Learning Augmented Bayesian Classifier
- R. Kohavi and GH John (1997) Wrappers for Feature subset selection Artificial Intelligence
- J.B. Kruskal (1956) spanning subtree of graph
- J. Park (2002) MAP complexity result
- I. Tsamardinos and C.F. Aliferis (2003) Towards principled feature selection
- L.C van der Gaag S. Renooij (2002) Probabilities for a probabilistic network a case study of oesophageal cancer Artificial Intelligence in medicine.