One Bayesian Network Construction Algorithm Based On Dimensionality Reduction

Shuo Quan[a], Pengfei Sun[b], Guoshi Wu[c],Jie Hu[d]

Beijing University of Posts and Telecommunication, Beijing, China

[a]quanshuo_bupt@126.com, [b]sunpengfei@bupt.edu.cn, [c]guoshiwu@bupt.edu.cn, [d]hujiesse@gmail.com

Keywords: feature clustering, feature mapping, Bayesian network, dimensionality reduction

Abstract. Bayesian network is a common probabilistic graphical model. It often can achieve good result in expression of uncertainty for knowledge and regular as well as in data classification. Since Bayesian network construction is a relatively complex problem, we propose a Bayesian network construction dimensionality reduction algorithm (BNDR). This algorithm maps a set ofassociated features to an abstract feature node by feature clustering and mapping. It can ensure no loss in accuracy and improve the time efficiency. For more complex industrial scenario, using the BNDR, you can get better practical efficiency.

1. Introduction

Bayesian network is a common probabilistic graphical model, which indicates uncertain relationship between features by directed graph. It is widely used in the field of classification, prediction, reason analysis and so on. Since Bayesian network is proposed, it has been widely used.The theoretical research of Bayesian network and practical application has become one of the hot focuses.

Construction of Bayesian network is a NP problem.For a graph which has$N$ nodes,there will be 3^{N^2} kinds of patterning methods(for any two nodesC_N^2, there are three cases,which include boundless and two directional edge). The time complexity of the algorithm will be very high while the number of nodes becomes more and more.To solve this problem,many experts raised some methods to improve the efficiency by simplifying the structure of Bayesian network,such as Naive Bayes Classifier[1], Tree Augmented Naive Bayesian Classifier(TAN)[2], Dimensionality reduction method based on Principal Components Analysis(PCA)[3], etc. However, the methods listed above make Bayesian network have limitations on expressing the causal relationship of the nodes: The Naive Bayes Classifier assumes that non-classification nodes are independent to each other. The classification node is known as parent node, and all non-classificationnodes only link with the classification node,so we can't distinguish which of non-classification nodes has closer relationship with the classification node; TAN, which has a loose independent constraint and allows each non-classification node has not only a classification node, but also at most one non-classification node as parent nodes. As a result, each node has at most two parent nodes, and we can't find allits parent nodes; Dimensionality reduction method based on PCA chooses the top M features which have larger eigenvalues to build a network. The disadvantages of this method are that for the selected features, it may have information redundancy, and for the missed features,it may lose some important information.For instance,Bayesian network analyzes the reason of traffic accidents, considersthe weather conditions(Fog, snow, rain, sunny) and visibility division([0, 50),[50,100), [100,200),[200, +∞)). If the weather is Fog, the visibility division is likely to be [0,50), andobviously the visibility is directly dependent on weather conditions. In this paper, we define features which has such dependency like weather conditions and visibility division as *associated features*.At the same time, a singlefeature with smaller eigenvalue may be not as important as features with larger eigenvalue for traffic accidents, but the combined contribution of these features to accident may be relatively large.So PCA deletes features with smaller eigenvalue that may be lose important information.

Based on the above analysis, this paper uses a feature fusion method called BNDR, which aggregates similar features into abstract features, and then using abstract features to build Bayesian network. The structure of the usual Bayesian network is shown in Fig. 1(a),*C* represents the classification feature, *V* represents the non-classification features, and it adds edges between non-classification features (*V*) and classification feature (*C*). The structure of the BNDR Bayesian network is shown in Fig. 1(b), where *C*represents the classification feature, *V* represents the non-classification features,*A*represents the abstract features, and it adds edges between abstract features (*A*) and classification feature (*C*).The advantages of this method are as follows：For the complex network, because of feature clustering, it is easier to analyze which abstract features impact the classification feature. For example, when analyzing the reason of traffic accidents, there may be hundreds of features, and each pair of similar non-classification feature nodes will have an edge, such as weather conditions and visibility division, etc. These edges may influence the overall structure of the Bayesian network. That we can’t focus on the important features.However, if we use abstract features to describe car information, weather conditions, road conditions, driver information, etc. Then we can use such abstract features to buildnetwork.It may be easier to analyze the reason of the traffic accidents.By mapping a set of associated features to an abstract feature, BNDR can ensure no loss in accuracy and improve the time efficiency.

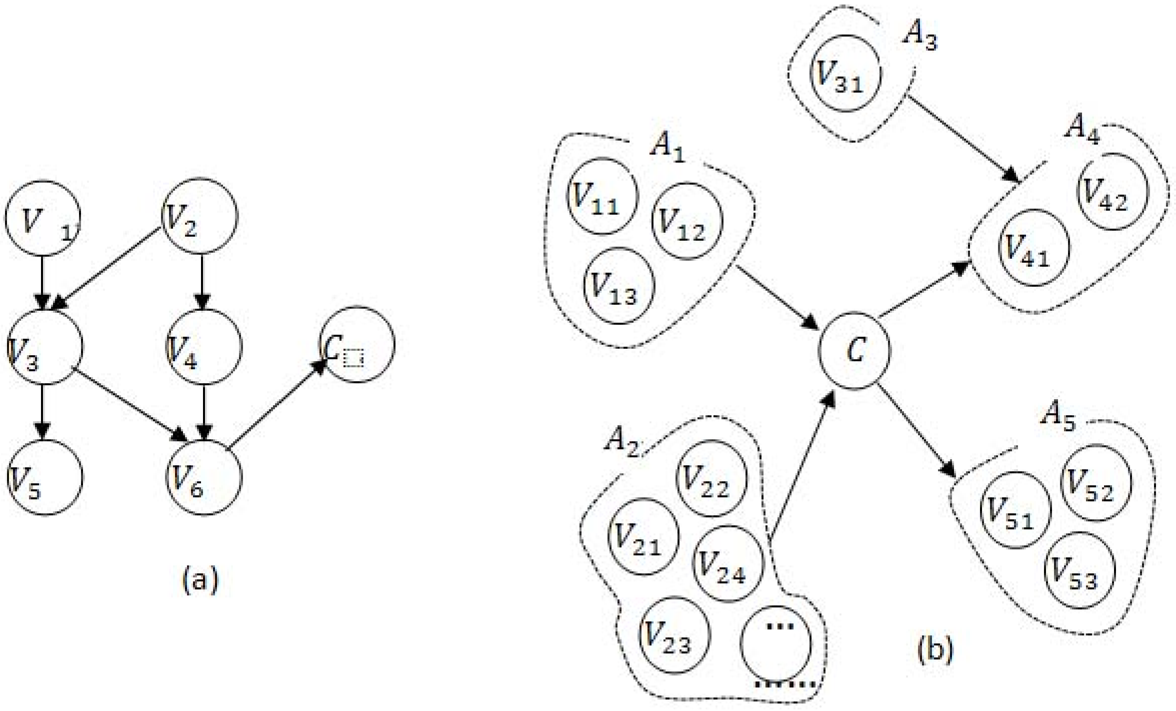


Fig.1 (a) represents the structure chart of usual Bayesian Network, *C* represents the classification feature, *V* represents the non-classification features (b) represents the structure chart of BNDR, *C* represents the classification feature, *V* represents the non-classification features,*A* represents the abstract features

2. The overall algorithm introduced

Flowchart of the overall algorithm shown in Fig. 2:

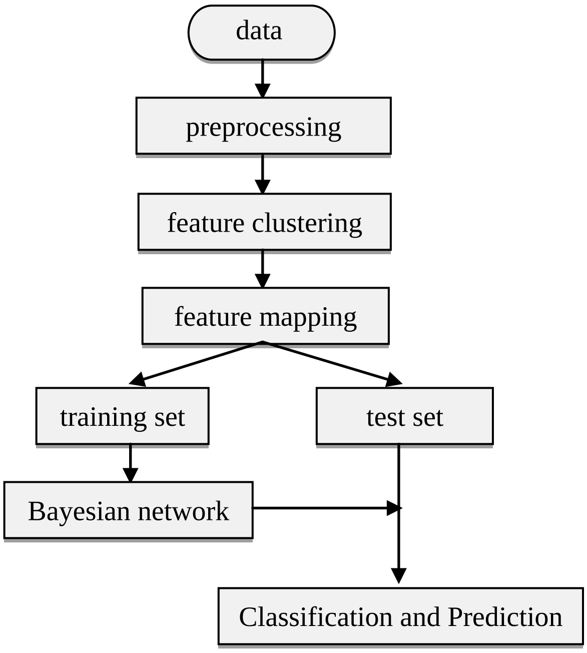


Fig. 2 Flowchart of the overall algorithm

From Fig. 2, flowchart will be following steps:

1) BNDR algorithm uses search scoring method to build Bayesian network.Itrequires statistical probability distribution anddiscrete variables. So the first step, we should do data preprocessing, discretizing the continuous variables.
2) The second step puts the similar features together by feature clustering. In this step, we use density clustering method to put the associated original featuresinto a category,and declare it as an abstract feature.
3) In order to achieve the purpose of reducing dimensions, for each category got from the second step, we use self-learning neural network algorithm to complete the mapping from similar features to an abstract feature.All original features are sub feature of abstract features.
4) Merge abstract features and the classification feature as new data set. Divide the new data sets into training and testing sets.
5) For the training set, we use MMHC(max-min hill-climbing) algorithm[4] to build Bayesian network.
6) For the test set, we use Bayesian network model trained by thestep fiveand probabilistic reasoning based on Markov border to predict the classification.

3. Algorithm steps described in detail

The focus of this paper is aggregatingthe entire feature set and mapping similar features to the corresponding abstract feature. For these two parts, we will give a detailed elaboration. And for the Bayesian network construction algorithm MMHC, we will only give a brief description.

LetD_{N*M} represents the original data set, the$N$ represents the number of instances, M represents the number of original features,V represents the set of features,and C represents the classification feature.

1) Feature clustering

Feature clustering puts similar features together. It can use following formula expression.

$$V = \{V_1, V_2, V_3 \dots\} \xRightarrow{feature\ clustering} A = \{A_1\{V_{11}, V_{12} \dots\}, A_2\{V_{21}, V_{22} \dots\} \dots\} \qquad (1)$$

In the formula above, V represents the non-classification feature set, and A represents the abstract feature set, $V_i$represents original non-classification feature, after feature clustering,V_i becomeV_{ij}, A_i represents an abstract feature. In this paper, density clustering is used to cluster the features. The basic idea is that for an unclassified node, if the number of its neighborhood nodes is greater than the threshold value, treating it as a new category. Then finds all nodes belonging to this category and affixes the category labels to them. Repeat the process until all feature nodes are classified.

In the process of clustering, we need to choose the appropriate distance metric formula to make similarity measure for features. Mutual Information [5]has a great advantage on correlation determination then we choose it as appropriate distance metric formula.Mutual Information is an information metric in information theory. It can be seen as the amount of information that one random variable contains another. In other words, a random variable will reduce its uncertaintyif it knows another random variable. Formulas are as follows:

$$I(X\ ;Y\ ) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\quad \log\frac{p(x,y)}{p(x)*p(y)} (2)$$

Assume that joint distribution of random variable$X$ and random variable Y is $p(x,y)$, marginal distribution is $p(x)$ and $p(y)$, the Mutual Informationis the relative entropy of joint distribution $p(x,y)$ and product of $p(x)$* $p(y)$.

2) Feature mapping

After feature clustering, we get the abstract feature set$A = \{A_1\{V_{11}, V_{12} \dots\},\ A_2\{V_{21}, V_{22} \dots\} \dots\}$. Next, we need to map the value domain of similar features to the value domain of the abstract feature. It is described by the following formula.

$$\{V_{i1}, V_{i2} \dots\} \xRightarrow{feature\ mapping} A_i\ (3)$$

In the formula above, $i$represents thei-th category. Here, we achieve feature mapping by Back Propagation(BP) Neural Network.

BP Neural Network is come upwith a team of scientists led by Rumelhart and McCelland. It is a multi-layer feedforward network trained by an error back propagation algorithm. And it is also one of the most widely used neural network model. It can learn and store lots of input - output mode mapping relationship, without to reveal the mathematical equations, another advantage is that it has a better description for linear and non-linear relationship [6].

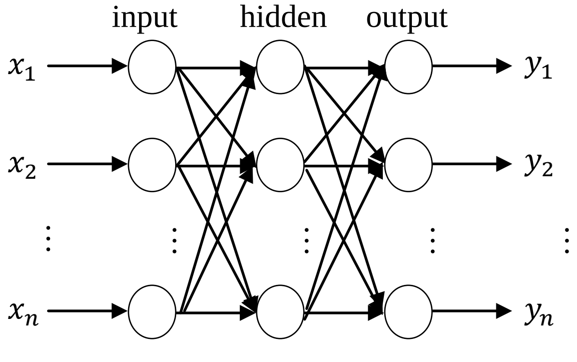
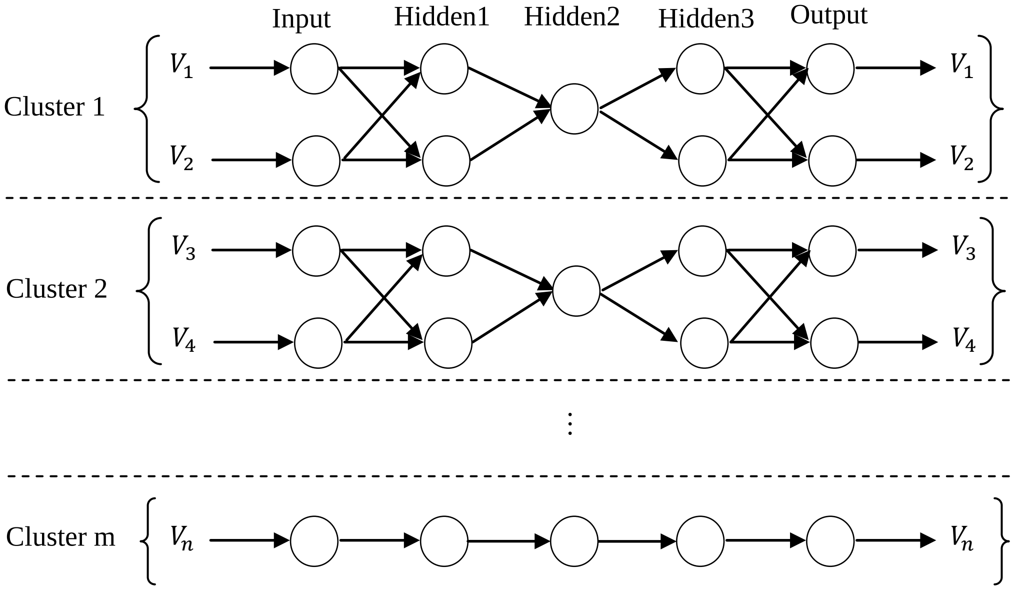


Fig. 3 structure chart of BP Neural Network

As shown in Fig. 3, BP Neural Network contains an input layer, one or more hidden layer and an output layer. Among them, the number of neurons in the input layer is determined by the number of non-classification features, the number of neurons in the output layer is determined by the number of classification features, the number of hidden layers and neurons in each layer is specified by experts.

BP neural network is a supervised learning. In order to correct layers' weights, it needs to know the output layer's results in advance. In this paper, the neural network intends to solve the problem of mapping similar features to an abstract feature. It do not know the mapping results in advance. So it is an unsupervised learning.

Fig. 4 structure chart of Self-learning Neural Network



Therefore, we make some changes to BP Neural Network: the number of neurons and neurons' values between the input and output layer should be same. Set the hidden layer as three layers. The first hiddenlayer completes the mapping of input layer to abstract node. The second hidden layer represents abstract node, so the output value of this hidden layer is the mapping result that we want to get from the Self-learning Neural Network.And the third hidden layer completes the mapping of abstract node to output layer. Combined with feature clustering, the network structure is shown in Fig. 4. The algorithm 1 gives the pseudo-code of Self-learning Neural Network in each category.

Algorithm 1BuildNeuralNetwork

```
(1)initialize network weight $W_{ij}$ and $\theta_j$
(2) while($Err$>threshold && m < iteration){
(3)    foreach instance {
(4)   foreach cell j in three hidden-layer and output-layer {
(5)$I_j = \sum_i W_{ij} * O_i + \theta_j$
(6)$O_j = 1/(1 + e^{-I_j})$
(7)        }
(8)        foreach cellj in output-layer{
(9)$Err_j = O_j * (1 - O_j) * (T_j - O_j)$
(10)}
(11)        foreach cell j in three hidden-layer{
(12)$Err_j = O_j * (1 - O_j) * \sum_k (Err_k * W_{jk})$
(13)        }
(14)foreach $W_{ij}$ and $\theta_j$ in network{
(15)$W_{ij} = W_{ij} + \delta * Err_j * O_j$
(16)$\theta_j = \theta_j + \delta * Err_j$
(17)     } } }
```

Step(1) initializes the network weight and error; Step (2) sets the end condition for the loop, if the error is less than the threshold or reach the number of cycles,the program stops; Step (3) completes a iteration for each instance; Step (4)~(7) is forward propagation; Step (8)~(13) is reverse propagation; Step ⒁~⒃ updates the weight and error. Among them, step（5）calculates the net input value of each hidden and output layer, according to the output value and weight of its previous layer; Step（6）uses Sigmoid function to map the net input value to the interval [0,1], as the output value of this layer; Step（9）calculates the error of the output layer; Step ⑿ calculates the error of the hidden layer; Step ⒂ updates the weight of the entire network; Step ⒄ updates the error of the entire network.

3) Bayesian network construction and probabilistic reasoning

There are two main methods to build Bayesian network: (a)one method is based on dependencies measured by conditional independence test.The representative algorithm is TPDA [7]; (b) the other method is based on search scoring. It needs a heuristic search algorithm and scoring function. The representative algorithm is hill-climbing [8]. Now popular MMHC algorithm combines the two methods mentioned above and has some advantages. So in this paper, we use MMHC algorithm to build the Bayesian network.Firstly, it reduces the search space complexity by dependencies method. Then it uses search scoring method to find the best network structure. Compared with other algorithms, MMHC has great advantages in accuracy,time complexity and network construction.

After Bayesian network set up, we should do probabilistic reasoning to predict the value of classification feature. That is, for the classification feature V_c to be predicted, we should find evidence features V_o associated withV_c. Then get the most likely value ofV_c. The probabilistic reasoning adopted here based on the information flow and Markov border. Specific meaning are as follows: if a node’s parent nodes, child nodes, and child nodes’ parent nodes are known, it is conditional independent of other nodes in the diagram. Formulas are as follows:

$$\max_{h_i} \left\{ P\left(h_i | \pi(h_i)\right) * \prod_{ch \in child(h_i)} P\left(ch | h_i \bigcup \pi(ch)\right) \right\} \tag{4}$$

Wherein, h_i is the i-th value of classification feature V_c, $\pi(h_i)$ represents the parent nodes of h_i, $ch \in child(h_i)$ represents the child nodes of h_i, $P\left(h_i | \pi(h_i)\right)$ means the conditional probability of h_i and its parent nodes, $P(ch | h_i \cup \pi(ch))$means the conditional probability of ch and its parent nodes.

According to the result of reasoning, each candidate value of classification feature will be assigned a relative probability value. Then choose the maximum probability of the candidate value as the predicted value of classification feature.

4. The results

Here, we use BNDR algorithm, MMHC algorithm and Naive Bayes Algorithm to do the experiment. And the programming language is java language. To ensure the authenticity of the experiment, the test data are real data sets selected from UCI(University of California Irvine) Machine Learning Repository. In order not to introduce errors, we delete instances which contains the missing value. For each data set, we choose training and test data sets randomly, and the ratio of training and test data sets is 1 to 3. To prevent data over-fitting, for the same data set, we divide it into different training and test data sets, and do three times experiment to get averages as results.

Table 1 the results

Data set	features	instances	BNDR			MMHC		Naive Bayes	
			clusters	time	accuracy	time	accuracy	Time	accuracy
Nursery	9	12960	6	139s	92.8%	504s	94.3%	1.97s	90.5%
mushroom	23	5643	13	209s	99.2%	1966s	99.8%	<1s	97.2%
Connect-4	43	6937	25	62m	74.5%	563m	76.8%	2.64s	72.3%

The results of three experiments are shown in table 1. Here, we take mushroom experiment as example analyzing the results. Mushroom data set has 23 features, including 22 non-classification features and a classification feature. The non-classification features describe the characteristics of mushroom, and the classification feature represents if it’s a poisonous mushroom. The feature clustering result is shown in Fig. 5:

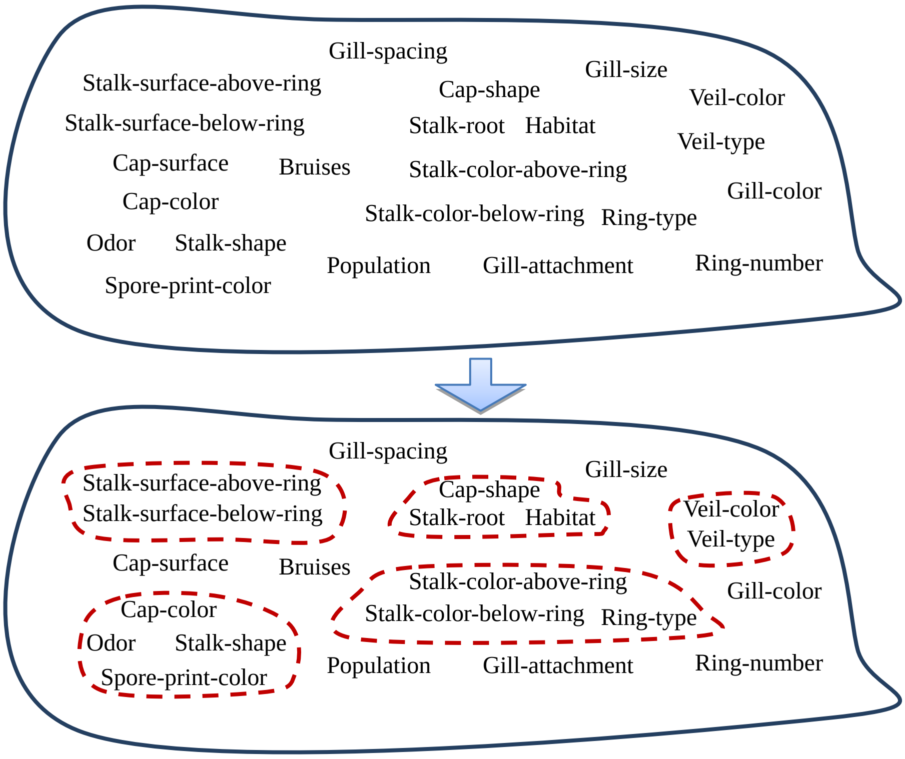


Fig. 5 the feature clustering result of mushroom data set

We can see that associated features are put together. In terms of the algorithm accuracy, the BNDR algorithm is 99.2%, the MMHC algorithm is 99.8%, and the Naïve Bayes algorithm is 97.2%. The precision ofBNDR algorithm is only a little less than MMHC algorithm. But it is more accurate than Naïve Bayes algorithm. In terms of the algorithm time complexity, BNDR algorithm only used 209 seconds, and it is almost the tenth MMHC algorithm.

According to the results, we can see that the time performance of BNDR algorithm has been greatly improved, in the case of a slight loss of precision.

5. Summary

Combined feature clustering, feature mapping and MMHC algorithm,we propose the BNDR algorithm. It gathers similar features into clusters, maps each cluster to an abstract node, and uses these abstract nodes to build Bayesian network. Then it can reduce information redundancy and improve time efficiency. If the number of nodes become more and more,it is easier to analyze which abstract features impact the classification feature. The algorithm model was verified by some UCI datasets. It achieved good accuracy and better time performance, and confirmed the feasibility of BNDR.

However, this algorithm also has some limitations. It’s more suitable for complex structure which contains information redundancy between features. Meanwhile, the method is mainly used to analyze the causal relationship between abstract features and classification feature. Next work, it’s recommended to build sub Bayesian network forassociated features. And add this sub Bayesian network to the whole Bayesian network. Then the network can describe the relationship between original features. At the same time, optimizing the feature clustering is another the target.

Acknowledgments

This research was supported by the Science and Technology Projects of MOT China (NO.

2014364222100).This work was supported by the fundamental research funds for the central university.

Reference

[1] LiangxiaoJianga,Zhihua Cai, Naive Bayes text classifiers: a locally weighted learning approach, Journal of experimental and theoretical artificial intelligence, 2013, 25(2)

[2] Liangxiao Jiang, Zhihua Cai, Improving Tree augmented Naive Bayes for class probability estimation, Knowledge-based systems,2012, 26

[3] Jose L. Godoy, Jorge R. Vega, Jacinto L. Marchetti, Relationships between PCA and PLS-regression, Chemometrics and Intelligent Laboratory Systems, 2014, 130

[4] Ioannis Tsamardinos, Laura E. Brown, Constantin F. Aliferis, The max-min hill-climbing Bayesian network structure learning algorithm, Machine learning,2006, 65(1)

[5] Vinh, N.X., Bailey, J. Comments on supervised feature selection by clustering using conditional mutual information-based distances, Pattern Recognition: The Journal of the Pattern Recognition Society,2013, 46(4)

[6] Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J, Learning representations by back-propagating errors, Nature 323 (6088): 533–536

[7] Y. Tang, K. Cooper, C. Cangussu, “Bayesian Belief Network Structure Learning Algorithms”, Technical Report, University of Texas at Dallas, UTDCS-25-09, 2009

[8] Russell, Stuart J.; Norvig, Peter (2003), Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, pp. 111–114

[9] J. Wang "A scalable data science workflow approach for big data Bayesian network learning", *Proc. Int. Symp. Big Data Comput.*

[10] J. Cheng, R. Greiner, J. Kelly, D. A, Bell, W. Liu, “Learning Bayesian networks from data: An information-theory based approach”, Artificial Intelligence, Vol.137, pp. 43-90, 2002