2012 International Conference on Image, Vision and Computing (ICIVC 2012)
IPCSIT vol. 50 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V50.63

Improving K2 Algorithm by Single Link Clustering For Bayesian Network Structural Learning

Saman Poursiah [1 +] and Alireza Sadeghi Hesar [2]

[1] Department of Computer Engineering, Quchan Branch, Islamic Azad University, Quchan, Iran
[2] Young Researchers Club, Mashhad Branch, Islamic Azad University, Mashhad, Iran

Abstract. A Bayesian network (BN) is an appropriate tool to work with the uncertainty that is typical of real-life applications. Basically, A BN provides an effective graphical language for factoring joint probability distributions. Two important methods of learning bayesian are parametric learning and structural learning. Finding bayesian network structure is a NP-hard problem. In this paper we introduced structural learning in bayesian network and key learning algorithms, like Hill Climbing and K2 and briefly. As a second step, We presented a new structural learning method using composite of k2 search algorithm and single link clustering method. Finally we compared proposed method with other structural learning methods based on accuracy and construction time on meteorological databases. Our experimental results show that the proposed method can find the good structure with the less construction time and also improves the classification accuracy.

Keywords: bayesian network, learning algorithms, structure learning, K2 algorithm, single link clustering.

1. Introduction

A Bayesian network (BN) is a probabilistic graphical model that relies on a structured dependency among random variables to represent a joint probability distribution in a compact and efficient manner. It is composed by a directed acyclic graph (DAG) where nodes are associated to random variables and conditional probability distributions are defined for variables given their parents in the graph. The automated creation of a Bayesian network can be separated into two tasks: 1) structure learning, which consists of creating the structure of the Bayesian network from the collected data, and 2) parameter learning, which consists of calculating the numerical parameters for a given structure. Because of huge number of possible structures, finding the best structure is a NP-hard problem. Therefore, heuristic search procedures have been tried. In this paper, a new structure learning method of the BNs using composite of k2 search algorithm and single link clustering method is proposed. The K2 algorithm is greedy algorithm, which obtains the best structure through a iterative process among all possible arrangements. In order to compensate for this, the algorithm has to iterate through many structures to ensure that the best scoring one is found. We will introduce, a new structural learning method using composite of k2 search algorithm and single link clustering method. The rest of the paper is organized as follows. In this paper, In Section 2, we review the main concepts related to Bayesian networks. In Section 3, we discuss structural learning in Bayesian networks. Next, sub-Section 3.1 and 3.2 is described the K2 search algorithm and single link clustering method. In section 5 we will illustrate the implementation results. Finally, Section 6 summarizes the main conclusions.

2. Bayesian Networks

Bayesian networks which represent the joint probability distributions for a set of domain variables are proved to be useful as a method of reasoning under uncertainty. A Bayesian network is a structure that shows

[+] Corresponding author.
E-mail address: samanpoursiah@gmail.com.

the conditional dependencies between domain variables and may also be used to illustrate graphically the probabilistic causal relationships among domain variables. A Bayesian network consists of a directed acyclic graph and probability tables. The nodes of the network represent the domain variables and an arc between two nodes (parent and child) indicates the existence of a causal relationship or dependency among these two nodes. Associated with each node there exist a conditional probability table (CPT). If the node has no parents, its probability table contains the prior probabilities else the conditional probabilities between the node and its parents. Although the domain variables can be continuous, they are discretized most of the time for simplicity and efficiency. Besides representing the dependencies between domain variables, a Bayesian network is used to inference the probability of a variable given the observations of other variables. A simple bayesian network shown in figure 1.

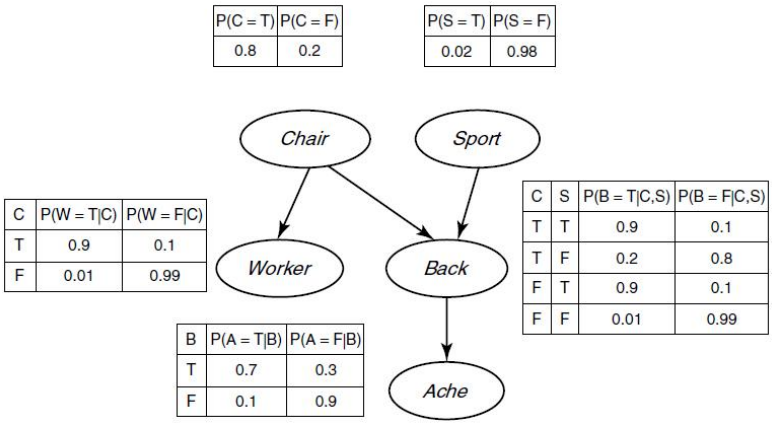


Fig. 1: A Bayesian Network (DAG with CPTs for each node)

3. Bayesian Structural Learning

The two major tasks in learning a BN are: learning the graphical structure, and then learning the parameters (CPTs entries) for that structure. We focused on structure learning in this paper. The structure learning methods can be separated in two ways including constraint-based and score-based methods. Constraint-based algorithms learn the network structure by analyzing the probabilistic relations entailed by the Markov property of Bayesian networks with conditional independence tests and then constructing a graph. But score- based algorithms assign a score to each candidate structure and try to maximize it with some heuristic search algorithm. On other hand score based algorithms are simply applications of various general purpose heuristic search algorithms, such as hill climbing, k2 search, simulated annealing and etc.

3.1. Hill climbing

The idea of a hill climbing search algorithm is to generate a model in a step by step fashion by making the maximum possible improvement in an objective quality function at each step. Initialize with a network structure, possibly random, evaluate the change in the score for all arc changes on this network and choose the one that has the maximum change. Continue this process until no more arc changes increase the score. This algorithm generally sticks into local maxima.

3.2. K2 algorithm

K2 algorithm (see figure .2) use a greedy technique and is initialized with an ordering of nodes such that the parents of a node are listed above the node .Starting with an empty structure and tracing the node list in order, one adds a parent to the node that increases the score. The number of parents to be added to a single variable may be limited to a predetermined constant for fast inference. K2 finds structures quickly if given a reasonable ordering. The scoring function of K2 algorithm for i nodes is in equation 2. The final score of network will obtains by multiplying the individual score of nodes.

$$f(\mathrm{i}, \pi_{\mathrm{i}}) = \prod_{j=1}^{|\Phi_i|} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} \alpha_{ijk}! \qquad (2)$$

i is the current node, r_i is the number of states, π_i is the parent of, $|\varphi_i|$ is the number of values within the CPT of, α_{ijk} is The number of cases in the dataset in which has its k_{th} value and have their j_{th} value in CPT and N_{ij} is sum of α_{ijk} for each state of i.

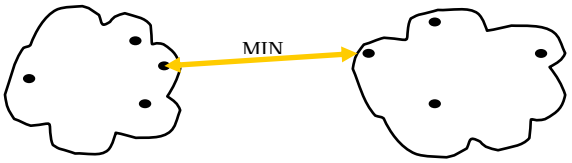
K2 Algorithm

```
1. procedure K2;
2. {Input: A set of n nodes, an ordering on the nodes,
an upper bound u on the
3. number of parents a node may have,
and a database D containing m cases.}
4. {Output: For each node, a printout of the parents of the node.}
5. for i:= 1 to n do
6. πi  := φ;
7. Pold := f(i, πi);
8. OKToProceed := true;
9. While OKToProceed and | πi | < u do
10. let z be the node in Pred(xi) -  πi  that maximizes f(i, πi  ∩{z});
11. Pnew:= f(i, πi  ∩{z});
12. if  Pnew  > Pold  then
13. Pold  := Pnew;
14. πi  := πi  ∩ {z};
15. else OKToProceed := false;
16. end {while};
17. write("Node: ", xi, " Parent of xi: ",πi);
18. end {for};
19. end k2;
```

Fig. 2: The pseudo code for the K2 algorithm

3.3. Modified K2 with single link clustering

In this section an improvement on K2 is proposed. We propose that within the context of problems that the nodes have calculable coordinate such as meteorology, the number of potential parents for the K2 algorithm could be reduced based on the geographical location of the nodes or stations. This algorithm is known as the SLK2 (Single Link K2) algorithm. In this algorithm, only the nearest nodes are chosen to be potential parents for a node. By reducing the number of parents based on their location, they were able to show a decrease in the construction time of the network. Now we apply single link clustering method to find nearest nodes. Single link algorithm is an example of agglomerative hierarchical clustering method. We recall that is a bottom-up strategy: compare each point with each point. Each object is placed in a separate cluster, and at each step we merge the closest pair of clusters, until certain termination conditions are satisfied. This requires defining a notion of cluster proximity. For the single link, the proximity of two clusters is defined as the minimum of the distance between any two points in the two clusters. We consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, we consider the similarity between one cluster and another cluster to be equal to the greatest similarity from any member of one cluster to any member of the other cluster (see Figure .3).



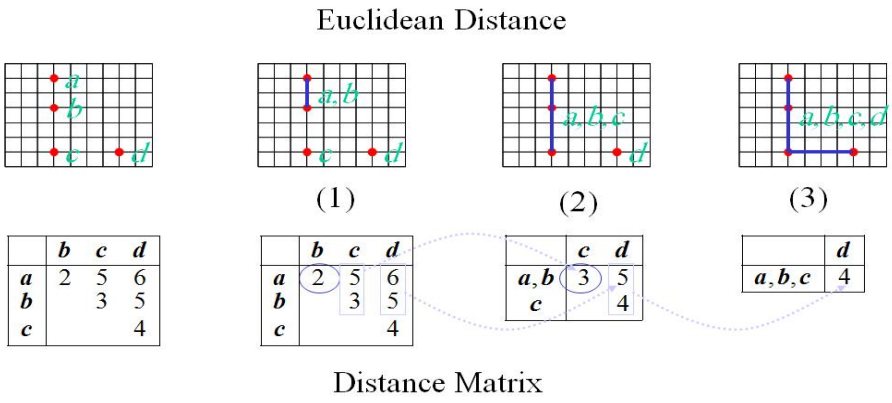


Fig. 3: Clustering method of Single Link Algorithm

4. Experimental Results

As first test, we consider the Kermanshah province in Islamic Republic of Iran as the geographical area of interest, and use monthly data (rainfall) from a 15 weather stations network (see Figure . 4) provided by the Iran weather service. The data covers the period from 1985 to 2010 and is representative of the local climatology and synoptic of this area. The variables are considered continuous for some applications, but are also quantized for other applications. Rainfall is quantized into four different states, 0=dry, 1=light-rain, 2=moderate-rain, 4=heavy-rain. According to the thresholds 0, 2, 10 and 20 mm/day, respectively (for instance, the event "heavy rain" corresponds to rainfall > 20 mm). The SLK2 algorithm output for 15 weather stations in area of study is shown in figure 5.

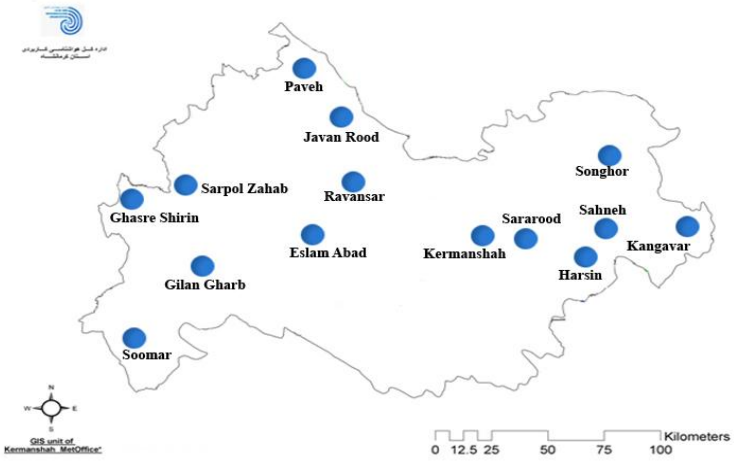


Fig. 4: Indicating 15 Stations In The Kermanshah Province

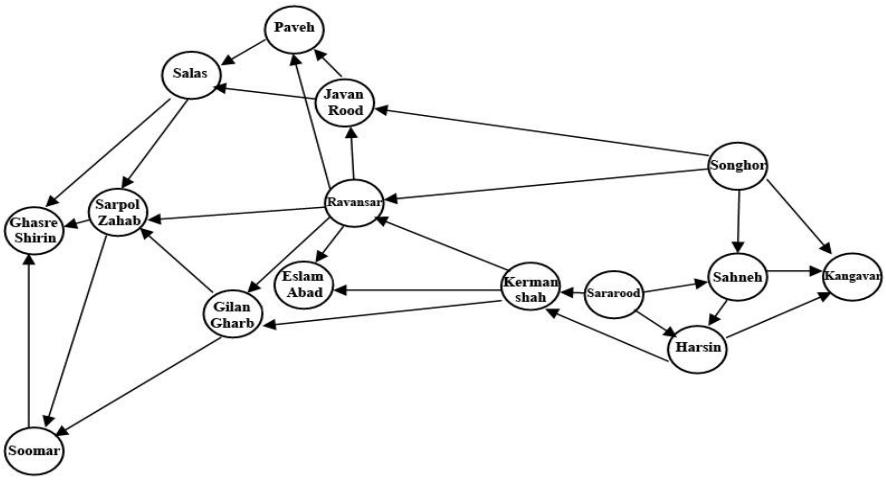


Fig. 5: SLK2 Algorithm Output For 15 Weather Stations In Area Of Study

As second test, we run out experiments on the 3 databases listed in table 1.

Table. 1: Experimental datasets

		Number of Instances	Number of Attributes	Number of Cluster
Rainfall	Database 1	9000	4	3
Ave. Temperature	Database 2	4500	5	3
Wind Speed	Database 3	1500	5	3

The rainfall dataset (Database 1), contains 9000 examples. There are four attributes: dry, light-rain, moderate-rain, heavy-rain. (all measured in millimeter). The Average of temperature data set (Database 2), contains 4500 examples. There are five attributes: very cold, cold, moderate, warm, hot (all measured in centigrade). The wind speed dataset (Database 3), contains 1500 examples. There are five attributes: no wind, low, moderate, high, very high (all measured in centigrade). The all databases have 3 cluster including ACM stations, synoptic stations and climatology stations. Thus the train and test procedures are executed a total of 10 times and the total error rate is the average of errors in all conducted tests. We observed classification performance for each algorithm. The accuracy of each classifier is based on the percentage of successful predictions on the test sets of each dataset. In Figure 2 and Table 2, the accuracies of the four learning procedures discussed in this paper are summarized. Computations were Computations were done on a computer with a 2.6 GHz Core 2 CPU, and 2 GB of RAM memory.

Table. 2: Clustering Performance

	Hill Climbing	K2	SLK2
Database 1	67.65	84.54	90.08
Database 2	74.92	90.34	94.57
Database 3	71.66	88.33	94.29

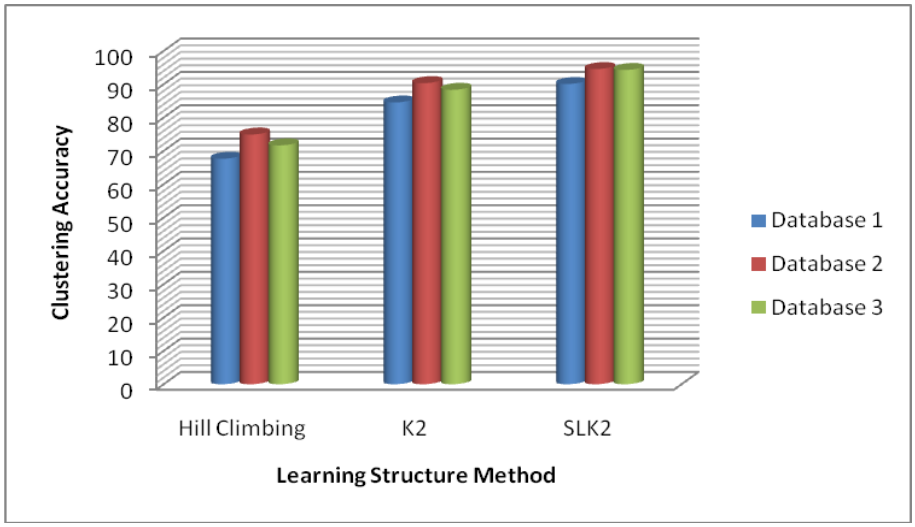


Fig. 6: Clustering Accuracy on databases

Table 3, presents the total time for structural learning versus search algorithm when BN classifier is used for classification of datasets.

Table. 3: Construction times

	Hill Climbing	K2	SLK2
Database 1	1.78	1.02	1.03
Database 2	0.65	0.11	0.11
Database 3	0.16	0.14	0.15

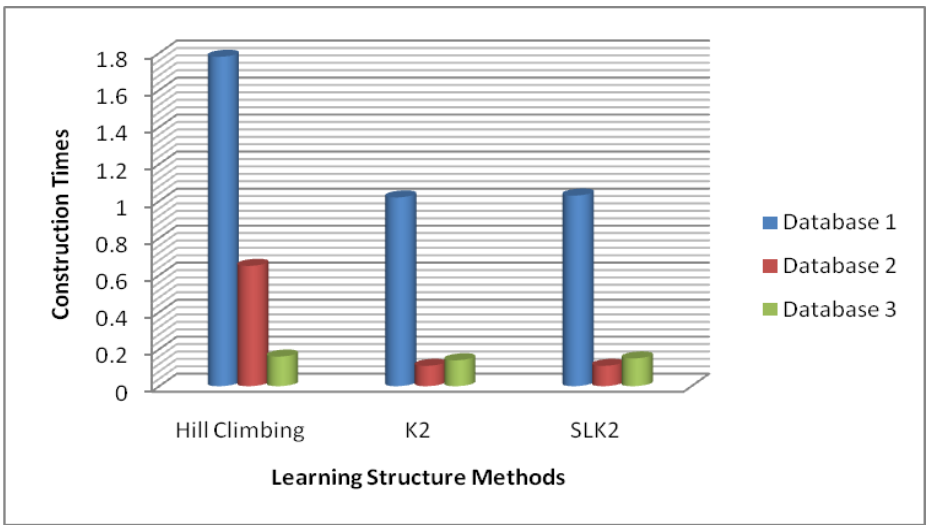


Fig. 7: Clustering Accuracy on databases

5. Conclusion

In this paper we introduced structural learning in bayesian network and described key learning algorithms such as Hill Climbing and K2 briefly. As a second step, We presented a new structural learning method using composite of k2 search algorithm and single link clustering method. Finally we compared proposed method with other structural learning methods on meteorological databases. As a result, it was shown that from the point of clustering accuracy, the SLK2 search is the best algorithm and K2 is less time consuming compared to other algorithms.

6. References

[1] Cofino Antonio, Cano Rafael , Sordo Carmen & Jose M. Gutierrez, "Bayesian Networks for probabilistic Weather Prediction", Proceedings of the 15th European Conference on Artificial Intelligence, IOS Press ,2000, 695-700.

[2] Cano Rafael , Sordo Carmen, M. Guti érrez Jos é, "Applications of Bayesian Networks in Meteorology Advances in Bayesian Networks", in Advances in Bayesian Networks, G ámez et al. eds., Springer, 2004, 309-327.

[3] A Kent Michael, Le Hanh , Tadross Mark, Potgieter Anet , Weather Forecasting With Bayesian Network, Ph.d, Department of Computer Science University of Cape Town , Cape Town, South Africa, 2008.

[4] Lee Byoungkoo , Joseph Jacob ," Learning a probabalistic model of rainfall graphical models" Quarterly Journal of the Royal Meteorological Society, 1998, 90–96.

[5] Korb Kevin, Nicholson Ann, Bayesian Artificial Intelligence, 75-83, Chapman & Hall/CRC Press LLC, 2004.

[6] Khanteymoori A.R, Homayounpour M.M, Menhaj M.B, "A Bayesian Network Based Approach For Data Classification Using Structural Learning",Communications in Computer and Information Science, Springer, 25-32, springer, ,2009.

[7] Lamma Evelina , Riguzzi Fabrizio, Storari Sergio Stor ," Improving the K2 Algorithm Using Association Rule Parameters", Elsevier B.V , 1-11 , Dipartimento di Ingegneria, Universit`a di Ferrara, Via Saragat 1, 41100 Ferrara, Italy , 2005.

[8] Heckerman, D., A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), Learning in Graphical Models, 301-354. MIT Press, 1999.

[9] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 2, 131–163, 1997.

[10] E. Gyftodimos & P. Flach, Hierarchical Bayesian networks: an approach to classification and learning for structured data, Methods and Applications of Artificial Intelligence: Third Hellenic Conference on AI, SETN 2004, Samos, Greece, 2004.