Discovering and Visualizing Attribute Associations Using Bayesian Networks and Their Use in KDD

Gou Masuda[1], Rei Yano[1], Norihiro Sakamoto[2], and Kazuo Ushijima[1]

[1] Graduate School of Information Science and Electrical Engineering, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan
{masuda,yano,ushijima}@csce.kyushu-u.ac.jp
[2] Department of Medical Informatics, Kyushu University Hospital, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan
nori@med.kyushu-u.ac.jp

Abstract. In this paper we describe a way to discover attribute associations and a way to present them to users using Bayesian networks. We describe a three-dimensional visualization to present them effectively to users. Furthermore we discuss two applications of attribute associations to the KDD process. One application involves using them to support feature selection. The result of our experiment shows that feature selection using visualized attribute associations works well in 17 data sets out of the 24 that were used. The other application uses them to support the selection of data mining methods. We discuss the possibility of using attribute associations to help in deciding if a given data set is suited to learning decision trees. We found 3 types of structural characteristics in Bayesian networks obtained from the data. The characteristics have strong relevance to the results of learning decision trees.

1 Introduction

Remarkable progress in data collecting and storing technologies has been generating a large number of huge databases such as astronomy databases and human genome databases. Knowledge Discovery in Databases (KDD) [3] aims automatically to analyze such huge databases and extract useful and interesting knowledge. A number of heuristic methods and strategies have been proposed for improving efficiency and accuracy in KDD. In general there is no single best method or strategy for all knowledge discovery tasks. Users therefore have to select an appropriate method for their specific task. However there are no clear theoretical metrics for selecting an appropriate method under a given circumstance. Consequently users have to apply a range of methods to their own data and repeatedly compare results to determine which provides the best fit. The KDD process thus has an iterative and interactive nature. In this situation, it is essential to visualize and present to users as much information on data as possible.

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 61–70, 1999.

The purpose of this study is to discover attribute associations and to present them to users in the KDD process. An attribute association is one kind of information implicit among data and it possesses at least two features. One is the degree of relevance between a pair of attributes the data have. The other is the structure that exists between them. Discovering such attribute associations and presenting them to users make it possible to conduct data mining effectively. We propose a way using Bayesian networks [4], which are one of the graphical representations of knowledge that employ directed acyclic graph.

Further consideration is given to the applicability of attribute associations to two different steps of the KDD process. One application we describe involves utilization of attribute associations to support feature selection [5,6]. The other application we discuss is the possibility of using attribute associations to support the selection of data mining methods. In this study we use them for deciding whether a given data set is suited to learning decision trees [9].

The remainder of this paper is organized as follows. Section 2 reviews the Bayesian networks and describes a way of discovering and visualizing attribute associations using Bayesian networks. Section 3 presents an application using attribute associations to support feature selection in the KDD process. Section 4 argues the possibility of using attribute associations to support discrimination of data suitable for learning decision trees. Section 5 concludes this paper.

2 Discovering and Visualizing Attribute Associations

2.1 Attribute Associations

To begin, we describe the data format that we deal with in this study and define several terms. A case, a tuple of data, is expressed in terms of a fixed collection of attributes. Each attribute has either discrete or numeric values. A case has also a predefined category of a target concept. A data set is a set of cases for an event. Attribute associations of data are information on the degree of relevance between a pair of attributes and on the structure existing between them. "Degree of relevance between attributes" is a numeric value which represent the strength of relevance such as covariance, correlation coefficient and mutual information. "Structure existing between attributes" is an indication of which pair of attributes have relevance. Since data we deal with contain a target concept, we need to consider associations among not only attributes but also attributes and a target concept. We simply deal with the target concept as an attribute of the data because a target concept can be regarded as a discrete attribute.

2.2 Discovering Attribute Associations Using Learning Bayesian Networks

The first question to be discussed here is how we obtain attribute associations that exist implicitly in data. We propose a method of discovering attribute associations via learning Bayesian networks. A Bayesian network is a directed

acyclic graph with a conditional probability distribution for each node. Each node represents an attribute in data. Arcs between nodes represent probabilistic dependencies among the attributes. A set of conditional probability distributions defines these dependencies.

The task of discovering attribute associations from data is equivalent to learning Bayesian networks from the data. The problem of learning a Bayesian network can be informally stated as: given a training set of data, find a network that best matches the data [2]. We used the following algorithm for learning Bayesian networks, which is based on a greedy search strategy.

1. Let a network $N(V, A)$ where $V = \{$ all the nodes corresponding to the attributes of a data set $\}$, $A = \{\}$. Let L and I be empty lists which are used in this algorithm. For each arc $(v_i, v_j) \in V$, compute a score for the arc. For all the arcs, sort them by score and put them into list L in decreasing order.
2. Select arcs from L created in step 1 and put them into list I which is used as input for the construction of a network.
3. Create 3 candidates N_1, N_2 and N_3 adding an arc a_i from the head of I to the current network N. N_1 is a network to which a_i is added in some direction. N_2 is a network to which a_i is added in the opposite direction to N_1. N_3 is a network to which no edge is added. Remove a_i from I.
4. Compute scores for the 3 candidates created in step 3. Select the network that gives the best score.
5. In a case in which N_1 or N_2 is selected in step 4, add a_i to A and go back to step 3. If I becomes empty, return N.

In step 1, we use the mutual information of each pair of nodes as the score for the arc. The mutual information of two nodes X_i, X_j is defined as

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \tag{1}$$

where x_i is a possible attribute value of the attribute correspondent to the node X_i, $P(x_i)$ is the probability that is calculated as a ratio of the number of cases which have x_i to the total number of cases in a data set, and $P(x_i, x_j)$ is the probability that is calculated as a ratio of the number of cases which have x_i and x_j. We apply the drafting [1] to select arcs in step2. It selects $n - 1$ arcs from the head of L, where n is the number of nodes in N. This prevents there being an excess of edges in a network. We adopt the Bayesian Dirichlet (BD) metric [4] as a scoring metric for a Bayesian network in step 3. It calculates the relative posterior probability of a network structure given a data set. Let D be the data set, B_S^h be the hypothesis that a data set D is generated by network structure B_S and ξ be given background knowledge. The BD Metrics is calculated as

$$p(D, B_S^h|\xi) = p(B_S^h|\xi) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N_{ij}^{'})}{\Gamma(N_{ij}^{'} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk}^{'} + N_{ijk})}{\Gamma(N_{ijk}^{'})} \tag{2}$$

where r_i is the number of possible attribute values of the i-th node X_i, q_i is the number of state of $\prod_i$, N_{ijk} is the number of cases in D in which $X_i = k$ and

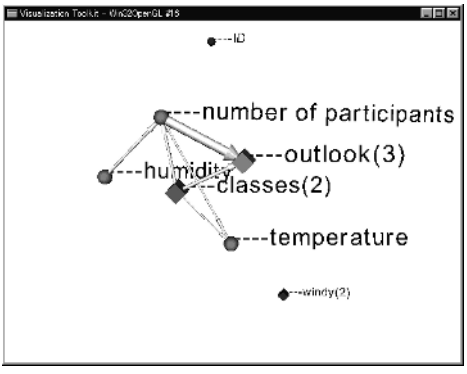

Fig. 1. Visualization of attribute associations

$\prod_i = j$, N'_{ijk} is the Dirichlet exponents, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ and $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$. $\prod_i$ denotes a parent node set of the X_i such that X_i and $\{X_1, \ldots, X_{i-1}\}$ are conditionally independent. Γ is the Gamma function which satisfies $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$. As the Dirichlet exponents we use the equation $N'_{ijk} = \frac{2}{r_i q_i}$. Because this learning algorithm is not able to deal with numeric attributes, a discretization is required beforehand. We adopt the gain criterion [9] to find a threshold value that divides numeric attribute values into two discrete ones.

2.3 Visualization of Attribute Associations

We propose a three-dimensional visualization of attribute associations obtained from data. We describe how we visualize them in this section.

1. **Degree of relevance between a pair of attributes:** We represent the degree of relevance by the size of a node, the color of a node and the thickness of an arc between nodes. We arrange a node indicating the target concept in the center. Attributes relevant to the target concept are arranged on the inner circumference surrounding it. Attributes irrelevant to the target concept are arranged on the outer circumference. When an attribute is a constituent of the network including the target concept it is regarded as relevant to the target concept. On the other hand an attribute is regarded as irrelevant to the target concept when it is not contained in the network. As regards the color and size of nodes, we give a different color and size to each node according to its kind. A node indicating the target concept is red. Attributes relevant to the target concept are purple while attributes irrelevant to the target concept are blue and are smaller than those relevant to the target concept. Thickness of an arc indicates the strength of mutual information the arc has.
2. **Structure exists between a pair of attributes:** We represent an arc in a Bayesian network by an arrow and the structure of cause and effect by

the direction of the arrow from cause to effect. Attributes are topologically sorted by their causal direction and are arranged on the circumference. This enables users intuitively to understand the causal direction among attributes.

3. **Extra basic information:** We represent a type of attribute by the shape of the node indicating the attribute. A discrete attribute is expressed by a square, while a numerical attribute is expressed by a sphere. An attribute name is labeled on each node. The number of possible discrete attribute values is also labeled on discrete attributes. Attributes which have a large number of attribute values (the default is 5) are colored in yellow in order that users can easily distinguish them.

We implemented a tool for discovering and visualizing attribute associations from data. It visualizes them in a three-dimensional view and provides manipulations such as rotation and zoom for users. These manipulations allow users closely to examine an area of interest. Figure 1 is an example of visualized attribute associations from a data set for a golf tournament. Each case in the data set indicates the target concept, whether a golf tournament takes place or not, under a set of conditions such as outlook, humidity and temperature. In this example, the target concept classes is arranged in the center of the figure. The attributes relevant to the target concept are arranged on the inner circumference (outlook, humidity, temperature, number of participants). These attributes are topologically sorted clockwise by their causal direction. Attributes irrelevant to the target concept are arranged on the outer circumference (ID, windy). Users can find that the attributes outlook and humidity have relevance to whether or not a golf tournament takes place. Furthermore users can see that the attribute outlook has strong relevance to the attribute number of participants.

3 Using Attribute Associations to Support Feature Selection

3.1 Feature Selection Using Attribute Associations

Feature selection [5,6] eliminates irrelevant and/or redundant attributes in a data set in order to obtain simple and interpretable patterns and to decrease the size of search space in data mining. We propose an interactive feature selection using visualized attribute associations. Our visualization shows which attributes have relevance to the target concept and the strength of the relevance. It enables users interactively to select attributes by looking at the visualized attribute associations for their data set. We believe that it is important for users to be able easily to reflect their intention in the KDD process.

The following are examples of policies for feature selection which users can lay down.

- Rule 1: Eliminate all the attributes arranged on the outer circumference.
- Rule 2: Eliminate the discrete attributes which have a large number of possible attribute values.

Table 1. Results of experiment on feature selection

Data set	Tree size			Predicted error rate(%)			Types of
	No FS	FS	Rate	No FS	FS	Rate	Network
australian	**53**	**37**	**0.70**	14.0	14.5	1.04	DENSE_EX
balance-scale(*)	41	41	1.00	34.1	34.1	1.00	STAR
breast-cancer-wisconsin	**7**	**3**	**0.43**	8.6	9.0	1.05	DENSE
breast-cancer	**23**	**6**	**0.26**	**29.3**	**28.6**	**0.98**	SPARSE
crx	**46**	**31**	**0.67**	14.3	14.3	1.00	DENSE_EX
german	118	125	1.06	**25.9**	**25.0**	**0.96**	SPARSE
glass	**59**	**43**	**0.73**	21.1	25.9	1.23	—
hayes-roth	**34**	**25**	**0.74**	32.2	34.1	1.06	STAR
heart	**45**	**28**	**0.62**	19.3	20.0	1.04	—
hepatitis	15	17	1.13	17.1	17.3	1.01	—
iris(**)	7	7	1.00	6.3	6.3	1.00	DENSE
labor(**)	7	7	1.00	17.4	17.4	1.00	—
liver	**81**	**79**	**0.98**	23.1	27.1	1.17	SPARSE
lymphography	**33**	**15**	**0.45**	**25.5**	**24.9**	**0.98**	DENSE_EX
monk1	18	29	1.61	**28.6**	**26.5**	**0.93**	STAR
pima-diabet	25	27	1.08	22.5	22.5	1.00	—
post-operative(*)	1	1	1.00	32.9	32.9	1.00	SPARSE
primary-tumor	**61**	**59**	**0.97**	56.6	56.6	1.00	STAR
segmentation	25	25	1.00	10.6	10.7	1.01	DENSE
tic-tac-toe(*)	139	139	1.00	18.4	18.4	1.00	SPARSE
vehicle(*)	183	183	1.00	15.4	15.4	1.00	SPARSE
voting(**)	7	7	1.00	6.9	6.9	1.00	DENSE
wine(**)	9	9	1.00	5.1	5.1	1.00	DENSE
zoo	**21**	**13**	**0.62**	**15.8**	**12.2**	**0.77**	DENSE
Average			0.877			1.009	—

Boldface denotes some improvement. * denotes that no attribute was eliminated in feature selection. ** denotes that several attributes were eliminated but there was no change in the result.

Attributes arranged on the outer circumference do not have relevance to the target concept. It is therefore a reasonable policy to eliminate such attributes (Rule 1). Moreover an attribute which has a large number of possible discrete attribute values tends to affect the size of patterns obtained from the data. Eliminating such attributes would also be a reasonable policy (Rule 2).

3.2 Experimental Results

We carried out an experiment in order to confirm the effectiveness of our proposed feature selection on the 24 data sets stored in the UCI Machine Learning Repository [10]. The two rules we described above were used as a policy of feature selection. We used a decision tree learning system [7,8] developed in our research group as a data mining method. It is based on C4.5 [9]. However flexibility

and extensibility of the system are emphasized in order easily to modify parts of the system using object-oriented technology. We adopted tree size, namely the number of nodes and the predicted error rate described in [9] as criteria for evaluation of results. In general it is to be desired that the decision tree has both a small size and a low predicted error rate. For comparison, we also analyzed the same data sets without feature selection.

Table 1 shows the results. "FS" represents the results with feature selection, while "No FS" represents the results without feature selection. The results show that tree size obtained with feature selection is on average 0.88 times as large as that obtained without feature selection. However the predicted error rate did not worsen greatly with feature selection compared to without feature selection.

To discuss these results in some detail, differences in either tree size or predicted error rate were found in 16 data sets out of 24. Of these, both tree size and predicted error rate were improved in 3 data sets. Tree size alone was improved in 8 data sets. Predicted error rate alone was improved in 2 data sets. Both got worse in 2 data sets. For the remaining 8 data sets, no attribute could be eliminated in 4 data sets (mark * in Table 1), while the eliminated attributes were not used originally in the other 4 data sets (mark ** in Table 1). However computation time was improved in these 4 data sets. These results indicate that feature selection using visualized attribute associations works well.

4 Using Attribute Associations to Support the Selection of Data Mining Methods

As [3] states, the ability to suggest to users the most appropriate data mining method is an important requirement for KDD tools. We describe our attempt to discriminate if a given data set is suitable for learning decision trees by catching the characteristic of the data via visualized attribute associations. We investigated 24 data sets and analyzed the results of learning decision trees used in the previous section. We found 3 types of structural characteristic in Bayesian networks obtained from the data. Moreover we found that these characteristics have a strong relevance to the analysis results of learning decision trees. We call these 3 characteristics of data DENSE, SPARSE and STAR according to the topology of Bayesian networks we obtained from the data. We present the 3 types as follows.

DENSE: Bayesian networks in Fig. 2 have thick arcs between attributes arranged on the inner circumference. They also have thick arcs between the target concept and attributes arranged on the inner circumference. We classify the data that produce such Bayesian networks as the DENSE type. We found 6 data sets belonged to the DENSE type in the 24 data sets we used in the previous section ("DENSE" type in Table 1). It shows that decision trees derived from the data tend to be small and to have a low predicted error rate. We believe the reason to be that attributes which have strong relevance to the target concept tend to be used as nodes of the decision trees.

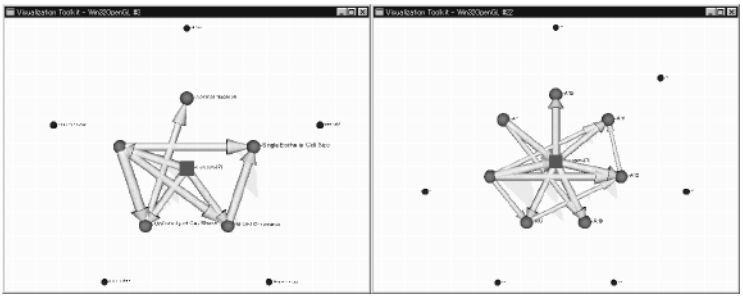

Fig. 2. Examples of the DENSE type: *breast-cancer-wisconsin*(left) and *wine*(right)

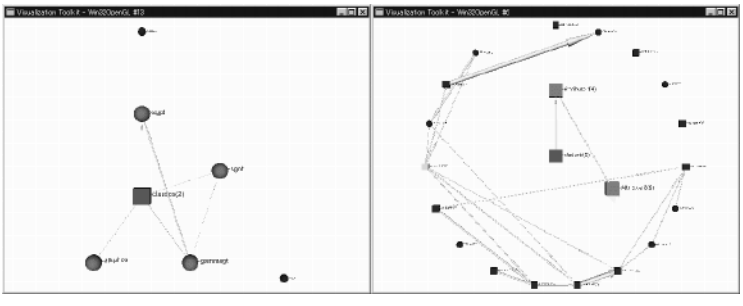

Fig. 3. Examples of the SPARSE type: *liver*(left) and *german*(right)

Some exceptions were observed. Tree size tends to be large when the target concept has strong relevance to discrete attributes which have a large number of possible attribute values. For example, the data set *lymphography* includes 2 discrete attributes which have 8 possible attribute values. By eliminating these 2 attributes using feature selection, we were able to obtain a smaller decision tree while keeping its predicted error rate almost the same. We found similar results in the data sets *crx* and *australian* ("DENSE_EX" type in Table 1).

SPARSE: In the Bayesian networks shown in Fig. 3, the arcs between attributes arranged on the inner circumference are very narrow and the networks look quite sparse. In the Bayesian network obtained from the data set *german* (Fig. 3 right), in particular, almost all attributes are arranged on the outer circumference. We classify the data from which such Bayesian networks are obtained as the SPARSE type. We found 6 data sets belonged to the SPARSE type out of the 24 data sets we used ("SPARSE" type in Table 1). Decision trees obtained from such data tend to be large and to have high predicted error rates. We believe the reason is that few attributes have strong relevance to the target concept, and it is hence difficult to classify data with such attributes.

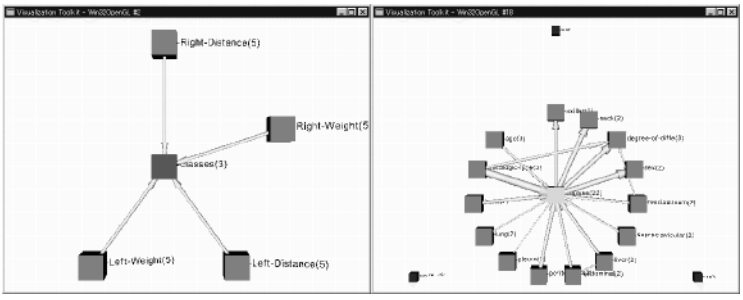

Fig. 4. Examples of the STAR type: *balance-scale*(left) and *primary-tumor*(right)

STAR: In the Bayesian networks shown in Fig. 4, each attribute is arranged like a star surrounding the target concept in the center. In other words, each attribute arranged on the inner circumference is relevant to the target concept, while attributes themselves are irrelevant to each other. We classify the data from which such Bayesian networks are obtained as the STAR type. We found 4 data sets belonged to the STAR type out of 24 data sets ("STAR" type in Table 1). It was found that decision trees derived from the STAR type data tend to be large and to have high predicted error rates. We believe the main reason for this is that such data have the rule which is a conjunction of all the attributes relevant to the target concept. Decision trees therefore tend to be large and to overfit the training data, which worsen their predicted error rate.

Up to this point we have presented 3 types of structural characteristics which are found in the Bayesian networks we obtained from the data. As a result we were able to set up the following criteria for learning decision trees.

- In cases in which a data set belongs to the DENSE type, learning decision trees is suited for analyzing the data as a data mining method.
- In cases in which a data set belongs to the SPARSE or STAR types, learning decision trees is not suited for analyzing the data.

5 Conclusion

In this paper we have described a way of discovering and visualizing attribute associations using Bayesian networks. In addition we have described two applications using visualized attribute associations to the KDD process. As regards feature selection with attribute associations, we ascertained that our proposed method worked well in 17 data sets out of the 24 tested, all of which come from the UCI Machine Learning Repository. In our approach users can directly reflect their intentions in feature selection. This advantage is very important for KDD tools because of the interactive nature of the KDD process. As regards supporting the selection of data mining methods, we found 3 types of structural characteristic in Bayesian networks which have strong relevance to the results

of learning decision trees. Based on these characteristics we set up criteria for discrimination of data suitable for learning decision trees. Users can estimate whether a given data set should be analyzed by learning decision trees according to these criteria. However they are based on a visual characteristic. There may exist other essential characteristics which affect the correlation between the types of Bayesian networks and the result of learning decision trees. Further analysis on this will be necessary as part of our future work.

We have not considered conditional probabilities of Bayesian networks in our Bayesian networks learning algorithm. Using the conditional probabilities enables users to obtain more information on their data. Introducing conditional probabilities into our learning algorithm is a future task.

Acknowledgments

We would like to thank Nick May for his advice on the presentation of English. We would also like to thank the maintainers and contributors to the UCI repository of machine learning databases.

References

1. Cheng, J., Bell, DA and Liu, W. "Learning belief networks from data: an information theory based approach," Proceedings of the Sixth ACM International Conference on Information and Knowledge Management, 1997.
2. Friedman, N., Geiger, D. and Goldszmidt, M. Bayesian Network Classifiers, *Machine Learning* 29, pp. 131-163, 1997.
3. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, H. *Advances in Knowledge discovery and data mining*, AAAI/MIT Press, 1996.
4. Heckerman, D., Geiger, D. and Chickering, D. "Learning Bayesian networks: The Combination of Knowledge and Statistical Data," Technical Report MSR-TR-94-09, Microsoft Research, 1994.
5. John, G., Kohave, R. and Pfleger, K. "Irrelevant Features and the Subset Selection Problem," In *Machine Learning: Proceedings of the Eleventh International Conference*, pp. 121-129, Morgan Kaufmann Publisher, 1994.
6. Liu, H., Setiono, R. "Feature Selection and Classification – A Probabilistic Wrapper Approach,", Proc. 9th Int. Conf. on IEA/AIE, pp. 419-424, 1996.
7. Masuda, G. Sakamoto, N. and Ushijima, K. "A Practical Object-Oriented Concept Learning System in Clinical Medicine," Proc. 9th Int. Conf. on IEA/AIE, pp. 449-454, 1996.
8. Masuda, G., Sakamoto, N. and Ushijima, K. "Applying Design Patterns to Decision Tree Learning System," Proc. of ACM SIGSOFT Sixth International Symposium on the Foundations of Software Engineering, pp.111-120, 1998.
9. Quinlan, J. R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
10. Blake, C., Keogh, E. and Merz, C. J. UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science, 1998.