



A Novel Approach of Multiple Submodel Integration Based on Decision Forest Construction

Limin Wang

College of Computer Science and Technology, JiLin University

Changchun 130012, China

Tel: 86-431-8517 2081 E-mail: wanglim@jlu.edu.cn

Xiaolin Li (Corresponding author)

School of Business, Nanjing University

Nanjing 210093, China

Tel: 86-431-8517 0836 E-mail: lixl_126@126.com

Yuting Mao

Information dissemination Academy of Engineering, ChangChun University of Technology

Changchun 130012, China

Tel: 86-431- 8571 6001 E-mail: valeriazuo@126.com

Abstract

An analytical general solution is derived for reasoning uncertain knowledge by multiple sub-model integration. By choosing decision rule for each specific instance, a decision forest rather than a tree will be constructed, thus all relatively independent attribute sets can be determined automatically without any human intervention. Necessary discretization for mixed-mode subset will be processed based on post-discretization strategy to minimize information loss.

Keywords: Multiple submodel integration, Decision forest, Post-discretization strategy, Conditional independence assumption

The volume of data for discovery of decision rules and recognition of patterns is growing at an exponential rate, both in the number of attributes (features) and objects (instances). One way to reduce computational complexity of knowledge discovery is dimensionality reduction, which includes projection pursuit, factor analysis, and principal components analysis. In artificial intelligence, decomposition methodology is a major tactic both for ensuring the transparent end-product and for avoiding the combinatorial explosion. And for this, the conditional independence assumption has been widely used, e.g. in Bayesian network structure learning. However, despite its popularity, the independence assumption always supposes that all attributes are discrete and continuous ones have to be discretized before learning even at the cost of information loss. Wang et al. reported that an unsteady special solution, which supposed that continuous feature subset and discrete subset are independent. This paper presents an analytical general solution to further handle mixed-mode subset based on decision forest construction to divide original feature space into several parts automatically.

Suppose instance space D with mixed-mode data has two types of attribute sets. The first k attributes are continuous and others are discrete. After pre-discretization the conditional independence assumption can be expressed as:

$$P\left(\bigwedge_{i=1}^k x_i \leq X_i \leq x_i + \Delta_i, \bigwedge_{j=k+1}^n x_j | c\right) = \prod_{i=1}^k P(x_i \leq X_i \leq x_i + \Delta_i | c) \prod_{j=k+1}^n P(x_j | c) \quad (1)$$

Where lower-case letters denote specific values taken by corresponding attributes (for instance, x_i represents the event that $X_i = x_i$). And Δ_i is arbitrary interval of the values of attribute X_i , $P(\cdot)$ refers to the probability. The following result can be formulated based on bayes theorem and differential theorem:

$$P(c|x_1, \dots, x_n) = \prod_{i=1}^k p(x_i | c) \prod_{j=k+1}^n P(x_j | c) P(c) / \alpha \quad (2)$$

Where $\alpha = p(x_1, \Lambda, x_k | x_{k+1}, \Lambda, x_n) P(x_{k+1}, \Lambda, x_n)$ is constant irrelevant to class value. $p(\cdot)$ refers to the probability

density function. Maximum likelihood estimation is chosen to estimate probability and joint probability, and Kernel-based density estimation is chosen to estimate conditional probability density function:

$$\begin{cases} \hat{p}(x_i|c) = \frac{1}{mh_i} \sum_{k=1}^l K\left(\frac{x_i - x_{ik}}{h_i}\right) \\ \hat{P}(c) = \frac{\text{Count}(C=c)}{N} \\ \hat{P}(x_j|c) = \frac{\text{Count}(C=c, X_j=x_j)}{\text{Count}(C=c)} \end{cases} \quad (3)$$

where $x_{ik} (k=1, \dots, l)$ is the corresponding value of attribute X_i when $C = c$, $K(\cdot)$ is a given kernel function $K(t) = (2\pi)^{-1/2} e^{-t^2/2}$. And h_i is the corresponding kernel width, m is the number of training instances when $C = c$.

Let $\{T_1, \dots, T_p\}$ denotes a decomposition of the attribute set A into p mutually independent subsets, each containing discrete and pre-discretized attributes or continuous input attributes. The aim of classification is to decide and choose the class that maximizes the posteriori probability, an analytical general solution based on conditional independence assumption can be obtained as follows:

$$v_{MAP}(C) = \arg \max_{c \in C} \prod_{i=1}^p P(c|t_i)/P(c)^{p-1} \quad (4)$$

Where t_i and $P(c|t_i)$ denote any reasonable combinations of attribute values in subset T_i and the classification accuracy of submodel constructed by t_i , respectively. t_i and $P(c|t_i)$ can be determined flexibly during the learning procedure of decision forest, which is composed of a set of decision trees. Attributes in the same tree should be dependent, while independent classification rules should be in different trees as the independent assumption suggests.

The original information entropy of class attribute C for instance space D is:

$$H(C|D) = - \sum_{c \in C} P(c) \log P(c) \quad (5)$$

The information entropy of C for subspace D' which satisfied $X_i = x_i$ is:

$$Gini = H(C|D) - H(C|D') \quad (6)$$

The Gini index defined above just consider the information that each attribute value rather than specific attribute gave to class label. Since it is applicable to both continuous and discrete attributes, the information loss caused by pre-discretization can be effectively avoided. The first part of Eq.(6), which describes the information entropy of class label itself, is the same to all attribute values. Thus the second part of Eq.(6) should be considered only during test selection procedure, that is to maximize the conditional information entropy. The construction procedure of decision forest can be described as follows:

Input: Training set D with n predictive attributes and N instances.

Output: Decision forest composed of n decision trees at most.

1. As to any given instance $\{x_1, \dots, x_n\}$, sort all attribute values according to the Gini index defined in Eq.(6) and select the one x_i which maximize the Gini index as the root node.
2. As to continuous values, the discretized interval Δ_i and the scope of the next subspace are determined to minimize information loss.
3. Search for the next attribute value as the branch node in the instance subspace which satisfies $X_i = x_i$ if X_i is discrete or $x_i \leq X_i < x_i + \Delta_i$ if X_i is continuous, until the class label is the same or height of the decision tree is n . Then a leaf node is generated. Each path from root node to leaf node is corresponding to a classification rule, and the pre-condition is the combination of all attribute values in this path.
4. Apply the learning procedure described above recursively, after N iterations each instance can be assigned a classification rule.
5. Combine those rules which have the same root node, then subtrees or relatively independent classification rules can be determined automatically without any human intervention.
6. Prune rule sets in each subtree repeatedly until this will not help to improve classification accuracy.
7. Eliminate those rules that will result in high misclassification rate. Then decision forest with more powerful expressive ability to uncertain knowledge is constructed.

The continuous attributes in mixed-mode subset have to be discretized in step 3. According to post-discretization strategy [8], the boundary of continuous attribute X_i can be decided based on information gain:

$$Gain(X_i, B; S) \geq \frac{\log_2(N-1)}{N} + \frac{\Delta(X_i, B; S)}{N}$$

Where S is sorted sequence of the attribute values, N is the number of instances in set S , $\Delta(X_i, B; S) = \log_2(3^k - 2) - [k \times Ent(S) - k_1 \times Ent(S_1) - k_2 \times Ent(S_2)]$, and $\{S_1; S_2\}$ are any given adjacent partitions. k_i is the number of class labels represented in set S_i . In order to evaluate the performance of submodel integration of decision forest, we conducted an empirical study on 12 data sets from the UCI machine learning repository to compare it with C4.5 release 8. Each data set consists of a set of classified instances described in terms of continuous or discrete attributes. Since the essence of submodel integration can be considered as partial leave-one-out validation, we also applied it to C4.5 release 8. Figure 1 summarizes the experimental results and from it, the superior generalization accuracy of submodel integration can be clearly seen.

References

- Duntelman G.H. (1989). *Principal Components Analysis*. Sage Publications, 221-254.
- Friedman J. H., & Tukey J.W. (1974). A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*. 23(9), pp. 881-889.
- Friedman J.H. (1997). On bias, variance, 0/1 loss and the curse of dimensionality. *Data Mining and Knowledge Discovery*. 1(1): 55-77.
- Kim J.O. (1978). *Factor Analysis: Statistical Methods and Practical Issues*. Sage Publications, 81-110.
- Kononenko I. (1991). Seminaive Bayesian classifier. In: Proceedings of the 6th European Working Session on Learning. New York, 206-219.
- Langley P., & Sage S. (1994). Oblivious decision trees and abstract cases. Working Notes of the AAAI-94 Workshop on Case-Based Reasoning. Seattle, WA: AAAI Press, 113-117.
- Michie D. (1995). Problem decomposition and the learning of skills. In: *Proceedings of the 8th European Conference on Machine Learning*. London, UK, 17-31.
- Silverman B. W. (1986). Density Estimation for Statistics and Data Analysis. *Monographs on Statistics and Applied Probability*. London: Chapman and Hall, 1-40.
- Wang L. M., & Li, X.L. (2006). Combining decision tree and Naive Bayes for classification. *International Journal of Knowledge-Based Systems*, 19(7): 511-515.
- Wang L. M., & Yuan S.M. (2004). Induction of hybrid decision tree based on post-discretization strategy. *Progress in Natural Science*, 14(6): 541-545.

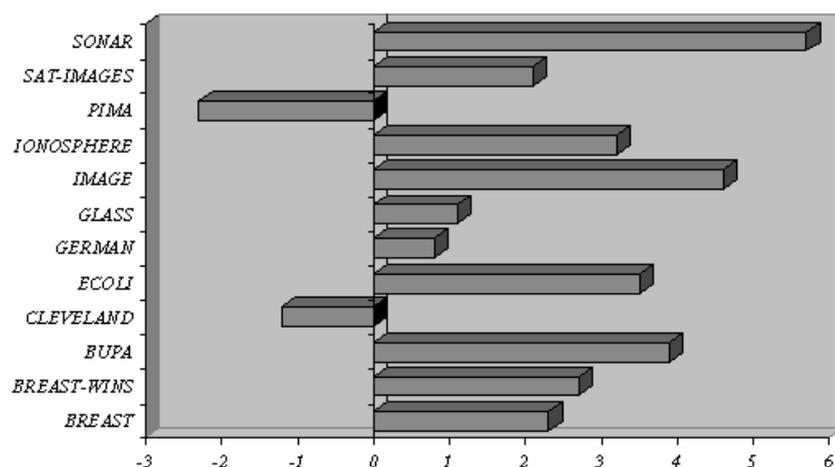


Figure 1. Comparison of classification performance % (Decision Forest-C4.5)