•Biostatistics in psychiatry (26)•

Decision tree methods: applications for classification and prediction

Yan-yan SONG[1,2]*, Ying LU[2,3]

Summary: Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure. When the sample size is large enough, study data can be divided into training and validation datasets. Using the training dataset to build a decision tree model and a validation dataset to decide on the appropriate tree size needed to achieve the optimal final model. This paper introduces frequently used algorithms used to develop decision trees (including CART, C4.5, CHAID, and QUEST) and describes the SPSS and SAS programs that can be used to visualize tree structure.

Key words: decision tree; data mining; classification; prediction

[*Shanghai Arch Psychiatry*. 2015; **27**(2): 130-135. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.215044]

1. Introduction

Data mining is used to extract useful information from large datasets and to display it in easy-to-interpret visualizations. First introduced in 1960's, decision trees are one of the most effective methods for data mining; they have been widely used in several disciplines[1] because they are easy to be used, free of ambiguity, and robust even in the presence of missing values. Both discrete and continuous variables can be used either as target variables or independent variables. More recently, decision tree methodology has become popular in medical research. An example of the medical use of decision trees is in the diagnosis of a medical condition from the pattern of symptoms, in which the classes defined by the decision tree could either be different clinical subtypes or a condition, or patients with a condition who should receive different therapies.[2]

Common usages of decision tree models include the following:

- Variable selection. The number of variables that are routinely monitored in clinical settings has increased dramatically with the introduction of electronic data storage. Many of these variables are of marginal relevance and, thus, should probably not be included in data mining exercises. Like stepwise variable selection in regression analysis, decision tree methods can be used to select the most relevant input variables that should be used to form decision tree models, which can subsequently be used to formulate clinical hypotheses and inform subsequent research.
- Assessing the relative importance of variables. Once a set of relevant variables is identified, researchers may want to know which variables play major roles. Generally, variable importance is computed based on the reduction of model accuracy (or in the purities of nodes in the tree) when the variable is removed. In most circumstances the more records a variable have an effect on, the greater the importance of the variable.

[1] Department of Pharmacology and Biostatistics, Institute of Medical Sciences, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2] Division of Biostatistics, Department of Health Research and Policy, Stanford University, Stanford, CA, USA
[3] Veterans Affairs Cooperative Studies Program Palo Alto Coordinating Center, the VA Palo Alto Health Care System, Palo Alto, CA, USA
*correspondence: yanyansong@sjtu.edu.cn

A full-text Chinese translation of this article will be available at http://dx.doi.org/10.11919/j.issn.1002-0829.215044 on June 6, 2015.

- Handling of missing values. A common – but incorrect – method of handling missing data is to exclude cases with missing values; this is both inefficient and runs the risk of introducing bias in the analysis. Decision tree analysis can deal with missing data in two ways: it can either classify missing values as a separate category that can be analyzed with the other categories or use a built decision tree model which set the variable with lots of missing value as a target variable to make prediction and replace these missing ones with the predicted value.
- Prediction. This is one of the most important usages of decision tree models. Using the tree model derived from historical data, it's easy to predict the result for future records.
- Data manipulation. Too many categories of one categorical variable or heavily skewed continuous data are common in medical research. In these circumstances, decision tree models can help in deciding how to best collapse categorical variables into a more manageable number of categories or how to subdivide heavily skewed variables into ranges.

2. Basic concepts

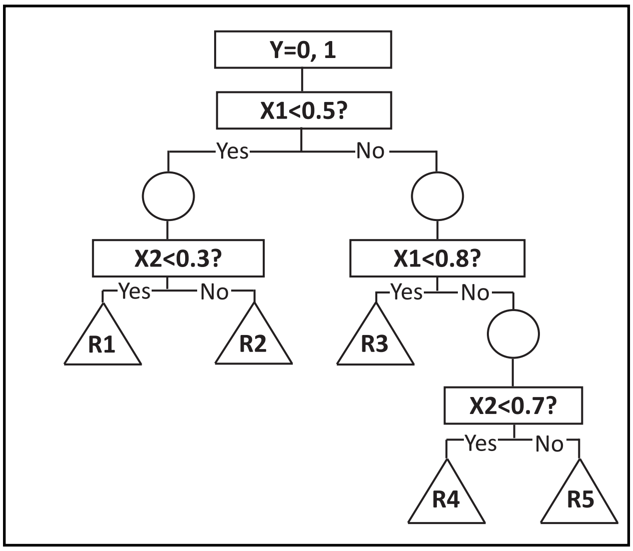
Figure 1 illustrates a simple decision tree model that includes a single binary target variable Y (0 or 1) and two continuous variables, x1 and x2, that range from 0 to 1. The main components of a decision tree model are nodes and branches and the most important steps in building a model are splitting, stopping, and pruning.

Nodes. There are three types of nodes. (a) A root node, also called a decision node, represents a choice that will result in the subdivision of all records into two or more mutually exclusive subsets. (b) Internal nodes, also called chance nodes, represent one of the possible choices available at that point in the tree structure; the top edge of the node is connected to its parent node and the bottom edge is connected to its child nodes or leaf nodes. (c) Leaf nodes, also called end nodes, represent the final result of a combination of decisions or events.

Branches. Branches represent chance outcomes or occurrences that emanate from root nodes and internal nodes. A decision tree model is formed using a hierarchy of branches. Each path from the root node through internal nodes to a leaf node represents a classification decision rule. These decision tree pathways can also be represented as 'if-then' rules. For example, "if condition 1 and condition 2 and condition … and condition k occur, then outcome j occurs."

Splitting. Only input variables related to the target variable are used to split parent nodes into purer child nodes of the target variable. Both discrete input variables and continuous input variables (which are collapsed into two or more categories) can be used.

Figure 1. Sample decision tree based on binary target variable Y



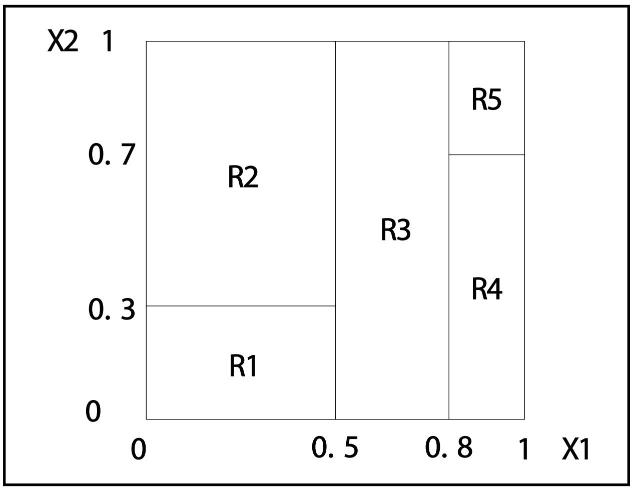
When building the model one must first identify the most important input variables, and then split records at the root node and at subsequent internal nodes into two or more categories or 'bins' based on the status of these variables. Characteristics that are related to the degree of 'purity' of the resultant child nodes (i.e., the proportion with the target condition) are used to choose between different potential input variables; these characteristics include entropy, Gini index, classification error, information gain, gain ratio, and twoing criteria.[3] This splitting procedure continues until pre-determined homogeneity or stopping criteria are met. In most cases, not all potential input variables will be used to build the decision tree model and in some cases a specific input variable may be used multiple times at different levels of the decision tree.

Stopping. Complexity and robustness are competing characteristics of models that need to be simultaneously considered whenever building a statistical model. The more complex a model is, the less reliable it will be when used to predict future records. An extreme situation is to build a very complex decision tree model that spreads wide enough to make the records in each leaf node 100% pure (i.e., all records have the target outcome). Such a decision tree would be overly fitted to the existing observations and have few records in each leaf, so it could not reliably predict future cases and, thus, would have poor generalizability (i.e., lack robustness). To prevent this from happening, stopping rules must be applied when building a decision tree to prevent the model from becoming overly complex. Common parameters used in stopping rules include: (a) the minimum number of records in a leaf; (b) the minimum number of records in a node prior to splitting; and (c) the depth (i.e., number of steps) of any leaf from the root node. Stopping parameters must be selected based on the goal of the analysis and the characteristics of the dataset being used. As a rule-of-thumb, Berry and Linoff[4] recommend avoiding overfitting and underfitting by setting the target proportion of records in a leaf node to be between 0.25 and 1.00% of the full training data set.

Pruning. In some situations, stopping rules do not work well. An alternative way to build a decision tree model is to grow a large tree first, and then prune it to optimal size by removing nodes that provide less additional information.[5] A common method of selecting the best possible sub-tree from several candidates is to consider the proportion of records with error prediction (i.e., the proportion in which the predicted occurrence of the target is incorrect). Other methods of selecting the best alternative is to use a validation dataset (i.e., dividing the sample in two and testing the model developed on the training dataset on the validation dataset), or, for small samples, cross-validation (i.e., dividing the sample in 10 groups or 'folds', and testing the model developed from 9 folds on the 10th fold, repeated for all ten combinations, and averaging the rates or erroneous predictions). There are two types of pruning, pre-pruning (forward pruning) and post-pruning (backward pruning). Pre-pruning uses Chi-square tests[6] or multiple-comparison adjustment methods to prevent the generation of non-significant branches. Post-pruning is used after generating a full decision tree to remove branches in a manner that improves the accuracy of the overall classification when applied to the validation dataset.

Decision trees can also be illustrated as segmented space, as shown in Figure 2. The sample space is subdivided into mutually exclusive (and collectively exhaustive) segments, where each segment corresponds to a leaf node (that is, the final outcome of the serial decision rules). Each record is allocated to a single segment (leaf node). Decision tree analysis aims to identify the best model for subdividing all records into different segments.

Figure 2. Decision tree illustrated using sample space view



3. Available algorithms and software packages for building decision tree models

Several statistical algorithms for building decision trees are available, including CART (Classification and Regression Trees),[7] C4.5,[8] CHAID (Chi-Squared Automatic Interaction Detection), [9] and QUEST (Quick, Unbiased, Efficient, Statistical Tree).[10] Table 1 provides a brief comparison of the four most widely used decision tree methods.[11,12]

Decision trees based on these algorithms can be constructed using data mining software that is included in widely available statistical software packages. For example, there is one decision tree dialogue box in SAS Enterprise Miner[13] which incorporates all four algorithms; the dialogue box requires the user to specify several parameters of the desired model.

The IBM SPSS Modeler[14] software package is more user-friendly; it includes four separate dialog boxes, one for each of four algorithms (it uses C5.0,[15] an upgraded

Table 1. Comparison of different decision tree algorithms

Methods	CART	C4.5	CHAID	QUEST
Measure used to select input variable	Gini index; Twoing criteria	Entropy info-gain	Chi-square	Chi-square for categorical variables; J-way ANOVA for continuous/ordinal variables
Pruning	Pre-pruning using a single-pass algorithm	Pre-pruning using a single-pass algorithm	Pre-pruning using Chi-square test for independence	Post-pruning
Dependent variable	Categorical/ Continuous	Categorical/ Continuous	Categorical	Categorical
Input variables	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous
Split at each node	Binary; Split on linear combinations	Multiple	Multiple	Binary; Split on linear combinations

version of C4.5). Based on the desired method of selecting input variables, the user goes to the dialog box for the corresponding algorithm (i.e., using the following steps: Analyze menu ==> Classify ==> Tree==>select algorithm of choice). For example, the SPSS syntax associated with the CART algorithm dialog box[16] would be as follows:

```
tree y [n] by x1 [s] x2 [c] x3 [o]
/tree display=topdown nodes=statistics branchstatistics=yes nodedefs=yes scale=auto
/depcategories usevalues=[valid]
/print modelsummary classification risk
/method type=crt maxsurrogates=auto prune=none
/growthlimit maxdepth=auto minparentsize=100 minchildsize=50
/validation type=none output=bothsamples
/crt impurity=gini minimprovement=0.0001
/costs equal
/priors fromdata adjust=no.
```

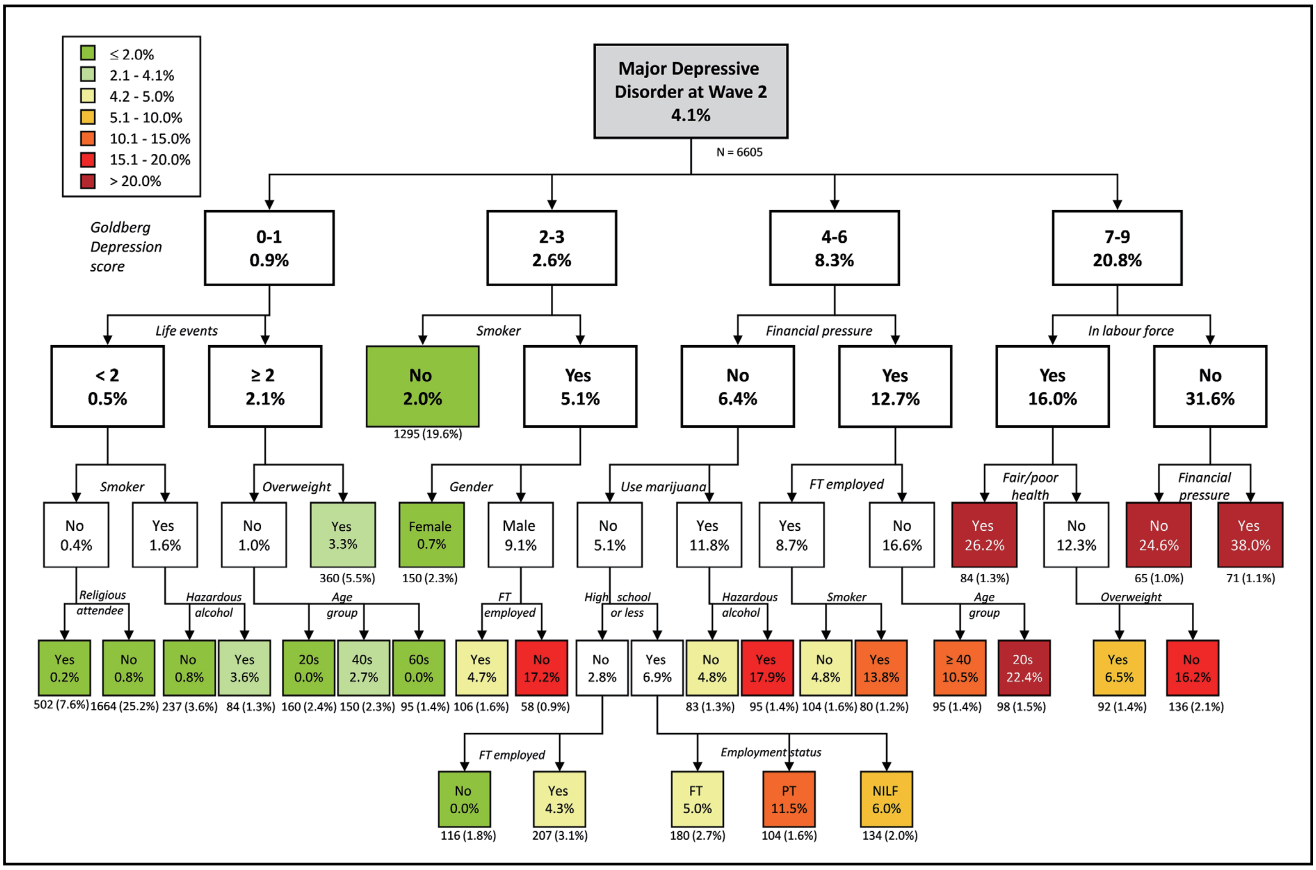
NOTE: [n], [s], [c], [o] indicate the variables are nominal, scale, categorical and ordinal.

4. Example

We use the analysis of risk factors related to major depressive disorder (MDD) in a four-year cohort study[17] to illustrate the building of a decision tree model. The goal of the analysis was to identify the most important risk factors from a pool of 17 potential risk factors, including gender, age, smoking, hypertension, education, employment, life events, and so forth. The decision tree model generated from the dataset is shown in Figure 3.

All individuals were divided into 28 subgroups from root node to leaf nodes through different branches. The risk of having depressive disorder varied from 0 to 38%. For example, only 2% of the non-smokers at baseline had MDD four years later, but 17.2% of the male smokers, who had a score of 2 or 3 on the Goldberg depression scale and who did not have a fulltime job at baseline had MDD at the 4-year follow-up evaluation. By using this type of decision tree model, researchers can identify the combinations of factors that constitute the highest (or lowest) risk for a condition of interest.

Figure 3. Decision tree predicting the risk of major depressive disorder based on findings from a four-year cohort study (reprinted with permission from Batterham et al.[17])



5. Discussion

The decision tree method is a powerful statistical tool for classification, prediction, interpretation, and data manipulation that has several potential applications in medical research. Using decision tree models to describe research findings has the following advantages:

- Simplifies complex relationships between input variables and target variables by dividing original input variables into significant subgroups.
- Easy to understand and interpret.
- Non-parametric approach without distributional assumptions.
- Easy to handle missing values without needing to resort to imputation.
- Easy to handle heavy skewed data without needing to resort to data transformation.
- Robust to outliers.

As with all analytic methods, there are also limitations of the decision tree method that users must be aware of. The main disadvantage is that it can be subject to overfitting and underfitting, particularly when using a small data set. This problem can limit the generalizability and robustness of the resultant models. Another potential problem is that strong correlation between different potential input variables may result in the selection of variables that improve the model statistics but are not causally related to the outcome of interest. Thus, one must be cautious when interpreting decision tree models and when using the results of these models to develop causal hypotheses.

Lohand and Strobl[18,19] provided a comprehensive review of the statistical literature of classification tree methods that may be useful for readers who want to learn more about the statistical theories behind the decision tree method. There are several further applications of decision tree models that have not been considered in this brief overview. We have described decision tree models that use binary or continuous target variables; several authors have developed other decision tree methods to be employed when the endpoint is the prediction of survival.[20-27] Our discussion was limited to cases in which the selection of input variables was based on statistical properties, but in the real world selection of input variables may be based on the relative cost of collecting the variables or on the clinical meaningfulness of the variables; Jin and colleagues[28,29] introduced an alternative classification tree method that allows for the selection of input variables based on a combination of preference (e.g., based on cost) and non-inferiority to the statistically optimal split. Another extension of the decision tree method is to develop a decision tree that identifies subgroups of patients who should have different diagnostic tests or treatment strategies to achieve optimal medical outcomes.[30]

Conflict of interest

The authors report no conflict of interest related to this manuscript.

Funding

The author received no funding for the preparation of this report.

用于分类与预测的决策树分析

Song YY, Lu Y

概述：决策树是一种常用的数据挖掘方法，用于多变量分析时建立分类系统或制定预测结果变量的算法。此方法将一个数据群分割成分枝状节段，构造出包括根节点、内部节点和叶节点的倒置形树状模型。该算法运用非参数方法，不需要套用任何复杂的参数模型就能有效地处理大型复杂的数据库。当样本足够大时，可将研究数据分为训练数据集和验证数据集。使用训练数据集构建决策树模型，使用验证数据集来决定树的适合大小，以获得最优模型。本文介绍了构建决策树的常用算法（包括 CART，C4.5，CHAID 和 QUEST），并描述了 SPSS 和 SAS 软件中将树结构可视化的程序。

关键词：决策树；数据挖掘；分类；预测

本文全文中文版从 2015 年 06 月 06 日起在
http://dx.doi.org/10.11919/j.issn.1002-0829.215044 可供免费阅览下载

References

1. Hastie TJ, Tibshirani, RJ, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Second Edition. Springer; 2009. ISBN 978-0-387-84857-0
2. Fallon B, Ma J, Allan K, Pillhofer M, Trocmé N, Jud A. Opportunities for prevention and intervention with young children: lessons from the Canadian incidence study of reported child abuse and neglect. *Child Adolesc Psychiatry Ment Health.* 2013; **7**: 4
3. Patel N, Upadhyay S. Study of various decision tree pruning methods with their empirical comparison in WEKA. *Int J Comp Appl*; **60** (12): 20-25
4. Berry MJA, Linoff G. *Mastering Data Mining: The Art and Science of Customer Relationship Management.* New York: John Wiley & Sons, Inc.; 1999
5. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer; 2001. pp: 269-272

6. Zibran MF. *CHI-Squared Test of Independence*. Department of Computer Science, University of Calgary, Alberta, Canada; 2012

7. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont, California: Wadsworth, Inc.; 1984

8. Quinlan RJ. *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufmann Publishers, Inc.; 1993

9. Kass, GV. An exploratory technique for investigating large quantities of categorical data. *Appl Stat.* 1980; **29**: 119-127

10. Loh W, Shih Y. Split selection methods for classification trees. *Statistica Sinica.* 1997; **7**: 815-840

11. Bhukya DP, Ramachandram S. Decision tree induction-an approach for data classification using AVL–Tree. *Int J Comp d Electrical Engineering.* 2010; **2**(4): 660-665. doi: http://dx.doi.org/10.7763/IJCEE.2010.V2.208

12. Lin N, Noe D, He X. Tree-based methods and their applications. In: PhamH, Ed. *Springer Handbook of Engineering Statistics*. London: Springer-Verlag; 2006. pp. 551-570

13. SAS Institute Inc. *SAS Enterprise Miner12.1 Reference Help*, Second Edition. USA: SAS Institute Inc; 2011

14. IBM Corporation. *IBM SPSS Modeler 17 Modeling Nodes*. USA: IBM Corporation; 2015

15. Is See5/C5.0 Better Than C4.5? [Internet]. Australia: Rulequest Research; c1997-2008 [updated 2008 Oct; cited 2015 April]. Available from: http://rulequest.com/see5-comparison.html

16. IBM Corporation. *IBM SPSS Statistics 23 Command Syntax Reference*. USA: IBM Corporation; 2015

17. Batterham PJ, Christensen H, Mackinnon AJ. Modifiable risk factors predicting major depressive disorder at four-year follow-up: a decision tree approach. *BMC Psychiatry*. 2009; **9**: 75. doi: http://dx.doi.org/10.1186/1471-244X-9-75

18. Loh WY. Fifty years of classification and regression trees. *Int Stat Rev*.2014; **82**(3): 329-348 .doi: http://dx.doi.org/10.1111/insr.12016

19. Strobl C. Discussions. *Int Stat Rev.* 2014; **82**(3): 349-352. doi: http://dx.doi.org/10.1111/insr.12059

20. Segal M. Regression trees for censored data. *Biometrics.* 1988; **44**: 35-47

21. Segal M, Bloch D. A comparison of estimated proportional hazards models and regression trees. *Stat Med.* 1989; **8**: 539-550.

22. Segal M. Features of tree-structured survival analysis. *Epidemiol*. 1997; **8**: 344-346

23. Therneau T, Grambsch P, Fleming T. Martingale based residuals for survival models. *Biometrika*. 1990; **77**: 147-160

24. LeBlanc M, Crowley J. Relative risk trees for censored survival data. *Biometrics.* 1992; **48**: 411-425

25. Keles S, Segal M. Residual-based tree-structured survival analysis. *Stat Med.* 2002; **21**: 313-326

26. Zhang HP. *Splitting criteria in survival trees. 10th International Workshop on Statistical Modeling*. Innsbruck (Austria): Springer- Verlag; 1995. p.305–314

27. Jin H, Lu Y, Stone K, Black DM. Alternative tree structured survival analysis based on variance of survival time. *Med Decis Making.* 2004; **24**(6): 670-680. doi: http://dx.doi.org/10.1177/0272989X10377117

28. Jin H, Lu Y, Harris ST, Black DM, Stone K, Hochberg MC, Genant HK. Classification algorithm for hip fracture prediction based on recursive partitioning methods. *Med Decis Making*. 2004; **24**(4): 386-398. doi: http://dx.doi.org/10.1177/0272989X04267009

29. Jin H, Lu Y. A procedure for determining whether a simple combination of diagnostic tests may be non-inferior to the theoretical optimum combination. *Med Decis Making.* 2008; **28**(6): 909-916. doi: http://dx.doi.org/10.1177/0272989X08318462

30. Li C, Gluer CC, Eastell R, Felsenberg D, Reid DM, Rox DM, Lu Y. Tree-structured subgroup analysis of receiver operating characteristic curves for diagnostic tests. *Acad Radiol.*2012; **19**(12): 1529-1536. doi: http://dx.doi.org/10.1016/j.acra.2012.09.007

(received, 2015-04-01; accepted, 2015-04-09)

Dr. Yan-yan Song is a Lecturer in the Department of Biostatics at the Shanghai Jiao Tong University School of Medicine who is currently a visiting scholar in the Division of Biostatistics, Department of Health Research and Policy, Stanford University School of Medicine. She is responsible for the data management and statistical analysis platform of the Translational Medicine Collaborative Innovation Center of the Shanghai Jiao Tong University. She is a fellow in the China Association of Biostatistics and a member on the Ethics Committee for Ruijin Hospital, which is Affiliated with the Shanghai Jiao Tong University. She has experience in the statistical analysis of clinical trials, diagnostic studies, and epidemiological surveys, and has used decision tree analyses to search for the biomarkers of early depression.