Journal of Engineering Education Research
Vol. 15, No. 4, pp. 41~47, July, 2012

Applying Decision Tree Algorithms for Analyzing HS-VOSTS Questionnaire Results

Dae-Ki Kang
Division of Computer & Information Engineering, Dongseo University

ABSTRACT

Data mining and knowledge discovery techniques have shown to be effective in finding hidden underlying rules inside large database in an automated fashion. On the other hand, analyzing, assessing, and applying students' survey data are very important in science and engineering education because of various reasons such as quality improvement, engineering design process, innovative education, etc. Among those surveys, analyzing the students' views on science-technology-society can be helpful to engineering education. Because, although most researches on the philosophy of science have shown that science is one of the most difficult concepts to define precisely, it is still important to have an eye on science, pseudo-science, and scientific misconducts. In this paper, we report the experimental results of applying decision tree induction algorithms for analyzing the questionnaire results of high school students' views on science-technology-society (HS-VOSTS). Empirical results on various settings of decision tree induction on HS-VOSTS results from one South Korean university students indicate that decision tree induction algorithms can be successfully and effectively applied to automated knowledge discovery from students' survey data.

Keywords: HS-VOSTS, Decision tree induction algorithms, Science-Technology-Society, Questionnaire analysis

I. Introduction

Data mining and knowledge discovery techniques have shown to be effective in finding hidden underlying rules inside large database in an automated fashion [1]. Several data mining algorithms such as decision tree induction algorithms [2,3], artificial neural network learning algorithms [4,5], Bayesian networks [6,7] and support vector machines [8,9] have been successfully applied to various real world applications including health informatics [10], bio-informatics [11], security informatics, etc.

On the other hand, analyzing, assessing, and applying students' survey data are very important in science and engineering education because of various reasons such as quality improvement [12], engineering design process [13], innovative education, etc. Among those surveys, analyzing the students' views on science-technology-society can be helpful to engineering education [14]. Because, although researches on the philosophy of science have shown that science is one of the most difficult concepts to define [15-18], it is still important to have an eye on science, pseudo-science, and scientific misconducts.

However, there have been few literatures on the application of data mining algorithms for the analysis of survey and questionnaire results for the problem of engineering education. Therefore, it would be interesting and significant contributions if we systematically explore and analyze students' survey and questionnaire data using data mining algorithms. Fig. 1 summarizes our idea explained above and shows the flow of our research contribution.

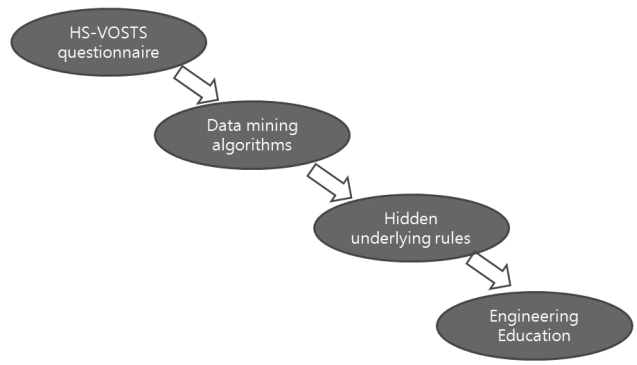

Fig. 1 Flow of research contribution

Received 4 October, 2011; Revised 12 June, 2012
Accepted 17 July, 2012
† Corresponding Author: dkkang@dongseo.ac.kr

Following these background, in this paper, we report the experimental results of applying decision tree induction algorithms for analyzing the questionnaire results of high school students' views on science-technology-society (HS-VOSTS) [19].

Empirical results of decision tree induction on HS-VOSTS from one South Korean university students are somehow mixed, but indicate that decision tree induction algorithms can be successfully and effectively applied to automated knowledge discovery from students' survey data.

II. Decision Tree Induction

Data mining and knowledge discovery techniques have shown to be effective in finding hidden underlying rules inside large database.

Several data mining algorithms such as decision tree induction algorithms, artificial neural network learning algorithms, Bayesian networks and support vector machines have been successfully applied to various real world applications including health informatics, bio-informatics, security informatics, etc.

The following list shows popular learning algorithms for analyzing databases.

- Decision tree induction
  - ID3 [2]
  - C4.5 [3,22]
  - C5.0 [3,23]
  - CART [3,24]
  - CHAID [3,25]
- Artificial neural network
  - Hopfield [4]
  - Kononen map [4]
  - Backpropagation [4]
- Bayesian networks
  - Naive Bayes [7]
  - K2 [7]
  - Independent Component Analysis [7]
- Support vector machines
  - Quadratic programming [8]
  - Sequential Minimal Optimization [9]

```
DecisionTreeInduction (given data D)
Begin
   Choose the best test T among the tests generated from D
      Partition D into D1, D2, D3, and Dn according to T
      If the partition is statistically significant enough (i.e.
      pre-pruning)
         Generate a child node with the test T
         For each Di of {D1,D2,D3,…,Dn} do
         DecisionTreeInduction(Di)
      Else
         Generate a leaf node with the best label L that
         represents the data D
End
```

Fig. 2 Traditional Decision Tree Induction Algorithm

There have been diverse researches on induction of decision trees from data [2,3,20,21]. That is, the decision trees are automatically generated from computer programs. (If they are manually generated by human experts, it is not quite interesting for data mining/machine learning experts.) The trees help users to make a decision by intuitively indicating a path from the root to one of the leaves as an ordered set of tests for the decision. Usually each test represents a specific value of an attribute in the data.

Traditional decision tree induction algorithms work as in Fig. 1. Note that we sometimes apply post-pruning to a full generated decision tree after the induction.

There have been many actual decision tree induction algorithms studied and widely used. Those widely used algorithms include C4.5 [22], C5.0 [23], Classification and Regression Tree (CART) [24], and CHi-squared Automatic Interaction Detector (CHAID) [25].

There have been diverse researches on induction of decision trees from data. The trees help users to make a decision by intuitively indicating a path from the root to one of the leaves as an ordered set of tests for the decision. Usually each test represents a specific value of an attribute in the data.

Popular decision tree learning algorithms include the following:

- ID3 uses information gain for splitting criteria.
- C4.5 uses gain ratio for splitting criteria.
- C5.0 is a commercial algorithm of which the splitting criteria is unknown.

- Classification and Regression Tree (CART) uses minimum description length (MDL) as a splitting criteria.
- CHi-squared Automatic Interaction Detector (CHAID) uses chi-squre measure as a splitting criteria.

III. High School Students' Views On Science-Technology-Society (HS-VOSTS)

Before we discuss HS-VOSTS and our approach, it is worth while to introduce "two cultures". There is a big gap between the researchers/scholars from humanities and natural sciences[29]. For example, Hardy questioned on the definition of "intellectual people" [29], and Feynman also counter-argued to an artist on the relation of scientific knowledge and artistic sense in "The pleasure of finding things out"[30]. In South Korea, students are forced to choose their major between "humanity" and "natural science" when they are in high schools. Now, many scholars in Korea insist that even graduate students need to have knowledge of both humanities and natural sciences.

There have been researches on the definition of so-called "science". Carnap [31] and Hempell [32] in Vienna circle tried to define science as an inductive process from repeated observations. Finally, Popper stated that scientific knowledge evolves itself through conjectures and refutations [33]. And many philosophers such as Feyerabend [34], Lakatos [35], Kuhn [36] continued researching on science.

In Asia, science is translated into '科學'. It is interesting that science was once referred as '格物', '格治', which were taken from Great Learning, one of classic textbooks on Confucianism.

Although researches on the philosophy of science have agreed that science is one of the most difficult concepts to define, it is still important for students to have an eye on science, pseudo-science, and scientific misconducts. Those pseudo-science includes water cures, dousing, iron pegs in Korean mountains.

In Science-Technology-Society (STS) field, researchers are interested in the science and technology conducted in society. [37]

HS-VOSTS [19] is an useful instrument for monitoring students' beliefs and viewpoints on STS topics. It has 23 multiple-choice items on four categories. When Lim et al. [19] have developed HS-VOSTS, they firstly have constructed categorical scheme based on many instruments for evaluating students' understanding of STS, literature review, and STS learning goal. Then, they have developed multiple-choice items through four steps.

In the first step, they formed some pairs of statement on each subordinate category. Next, 772 students responded the student statement questionnaires which were based on the pairs of statement. In the second step, they analyze the response written by the students to common viewpoints and constructed the first multiple-choice items. In the third step, they implemented the semistructured interview with 28 high school students and the constructed second multiple-choice items. In the fourth step, they developed the final version of the instrument through the analysis of the students' response on the second multiple-choice items.

IV. Experiments

It would be interesting and significant contributions if we systematically explore and analyze students' survey and questionnaire data using data mining algorithms.

We report the experimental result of applying a decision tree induction algorithm for analyzing the questionnaire results of high school students' views on science-technology-society (HS-VOSTS)

We distributed HS-VOST questionnaires to 190 students in one South Korean university.

The task of the decision tree learning algorithm (C4.5) is to detect where a student majors in humanities, natural sciences, or media art & athlete.

If the algorithm performs well, that indicates that there might be a fundamental gap on understanding STS between the students from humanities, natural sciences, or media art & athlete.

If the algorithm performs poorly, that indicates that, in spite of separate training, there might be no big gap on understanding STS between the students from humanities, natural sciences, or media art & athlete.

Empirical results of decision tree induction on HS-VOSTS

from one South Korean university students are somehow mixed, shown in the Figs 3,4,5,6 and Table 1,2,3,4. However, these results indicate that decision tree induction algorithms can be successfully and effectively applied to automated knowledge discovery from students' survey data. Ten-fold stratified cross-validation of data has been used for the evaluation in the experiments.

Fig. 3 shows the generated decision tree using C4.5 algorithm for the task of classifying a student's major (humanities, natural sciences, or media art & athlete). The number of nodes in the tree is 51 including 26 leaves.

The confusion matrix for the 10-fold cross validation scheme is in the Table 1. The accuracy is 47.59%.

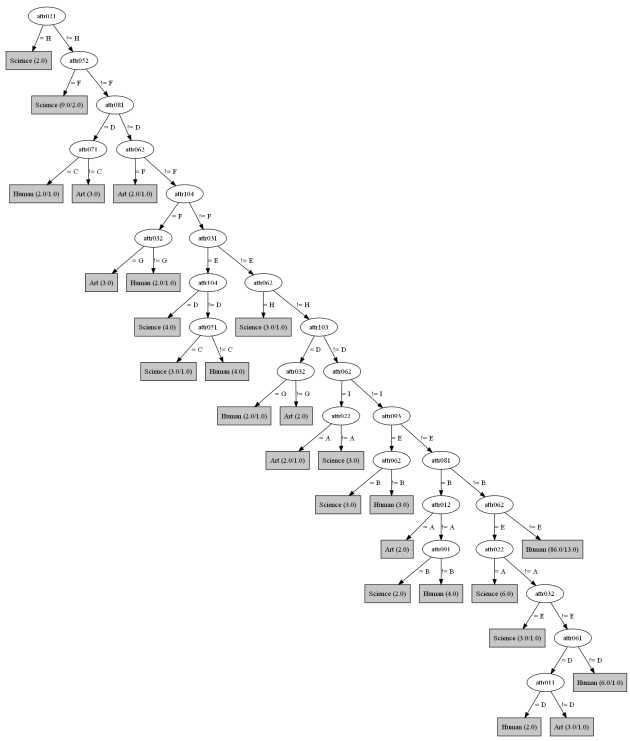

Fig. 3 Results of decision tree on HS-VOSTS to classify humanities, natural sciences, and media art & athlete

Table 1 Confusion matrix from the classification of humanities, natural sciences, and media art & athlete

Actual \ Predicted	Art	Human	Science
Art	3	16	4
Human	12	64	23
Science	6	26	12

Fig. 4 shows the generated decision tree using C4.5 algorithm for the task of classifying a student's major (media art & athlete or others). The number of nodes in the tree is 5 including 3 leaves.

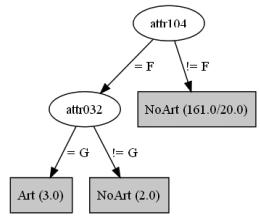

Fig. 4 Results of decision tree on HS-VOSTS to classify media art & athlete and other majors

Table 2 Confusion matrix from the classification of media art & athlete and others

Actual \ Predicted	Art	Else
Art	2	21
Else	7	136

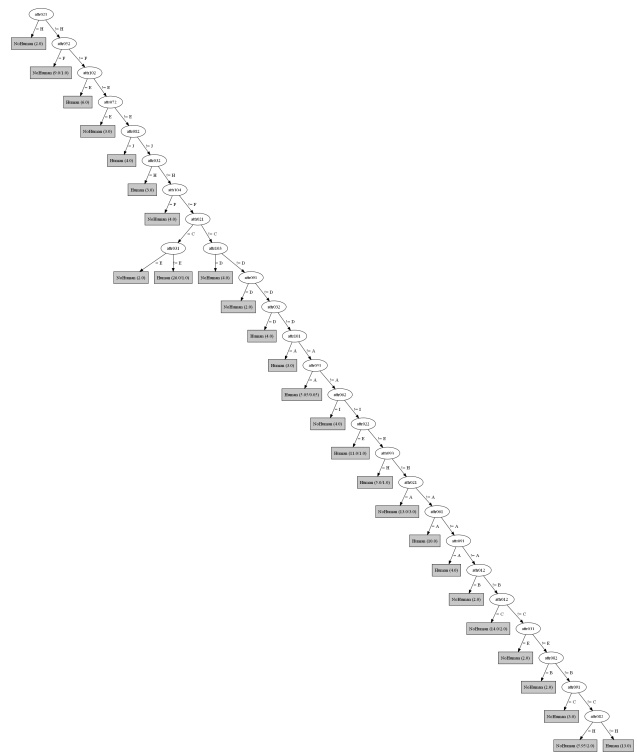

Fig. 5 Results of decision tree on HS-VOSTS to classify humanities & athlete and other majors

Table 3 Confusion matrix from the classification of humanities and others

Actual \ Predicted	Humanities	Else
Humanities	56	43
Else	37	30

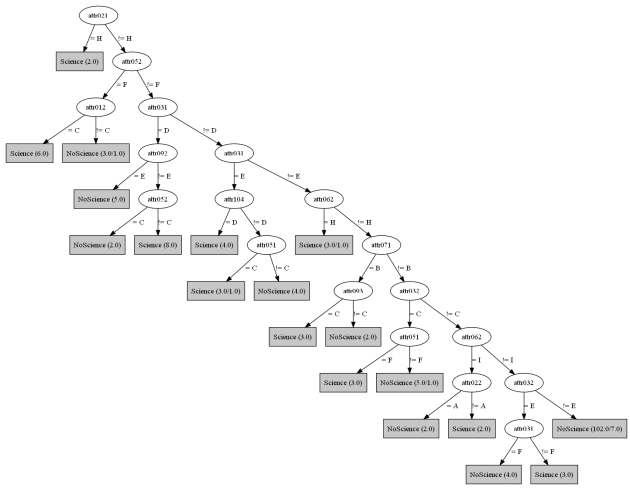

Fig. 6 Results of decision tree on HS-VOSTS to classify science and other majors

Table 4 Confusion matrix from the classification of science and others

Actual \ Predicted	Science	Else
Science	16	28
Else	34	88

The confusion matrix for the 10-fold cross validation scheme is in the Table 2. The accuracy is 83.13%, which is the best among the different experimental setups, while the size of the generated tree is smallest.

Fig. 5 shows the generated decision tree using C4.5 algorithm for the task of classifying a student's major (humanities or others). The number of nodes in the tree is 53 including 27 leaves.

We have applied binary split for the decision tree generation, but note that the generated tree for this experimental setting is highly skewed. The confusion matrix for the 10-fold cross validation scheme is in the Table 3. The accuracy is 51.81%.

Fig. 6 shows the generated decision tree using C4.5 algorithm for the task of classifying a student's major (media art & athlete or others). The number of nodes in the tree is 5 including 3 leaves.

The confusion matrix for the 10-fold cross validation scheme is in the Table 4. The accuracy is 62.65%.

V. Related Work

There have been many investigations about the ethical views and opinions of students on science and engineering.

Lee [26,27] has summarized the survey results on the information and communication ethics of college students. He has applied statistical techniques for the analysis of the survey, however has not applied machine learning techniques.

Kim [28] has integrated three decision tree algorithms (C5.0, CART, and CHAID) for the study on factors of education's outcome. In our work, we applied C4.5 decision tree algorithm to university students' survey results of HS-VOSTS.

VI. Conclusion and Future Work

We report the experimental results of applying decision tree induction algorithms for analyzing the questionnaire results of high school students' views on science-technology -society (HS-VOSTS).

Empirical results of decision tree induction on HS-VOSTS from one South Korean university students are somehow mixed, shown in the Figs 3,4,5,6 and Table 1,2,3,4. However, these results indicate that decision tree induction algorithms can be successfully and effectively applied to automated knowledge discovery from students' survey data.

This research was supported by 2011 Dongseo University research grants and Dongseo University's Ubiquitous Appliance Regional Innovation Center research grants from Ministry of Knowledge Economy of the Korean government. (No. B0008352).

References

1. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2nd ed., (2006).
2. Ross Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, (1993).
3. Ross Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research, 4*: 77-90, (1996).
4. Christopher M. Bishop Neural Networks for Pattern Recognition, Oxford: Oxford University Press, (1995).
5. Richard O. Duda, Peter E. Hart and David G. Stork, Pattern classification (2nd edition), Wiley, (2001).
6. Judea Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, (2000).
7. Judea Pearl, "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning," Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA: 329–334, (1985).
8. Corinna Cortes and Vladimir Vapnik, "Support-Vector Networks," Machine Learning, 20, (1995).12. Sipser, M (1996). Introduction to the Theory of Computation, PWS Pub. Co.; 1 edition.
9. Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik, "A training algorithm for optimal margin classifiers," The 5th Annual ACM Workshop on COLT: 144–152, Pittsburgh, PA, (1992).
10. Barry Robson, O. K. Baek, and Sean Ekins, The engines of Hippocrates: From the Dawn of Medicine to Medical and Pharmaceutical Informatics. Hoboken, NJ: John Wiley & Sons, (2009).
11. Srinivas Aluru, Handbook of Computational Molecular Biology. Chapman & Hall/Crc, (2006).
12. J. S. Lyons, and A. M. Bayoumi, "CQI processes, results, and program improvements for engineering design," *IEEE Transactions on Education, 43*(2): 174-181, (2000).
13. A. Ertas, and J. Jones, The Engineering Design Process. 2nd ed. New York, N.Y., John Wiley & Sons, Inc., (1996).
14. Yong kil Lee and Kyung hee Kang, "Analyzing opinions which university students from engineering and social science department have about science-technology-society literacy", *Journal of Engineering Education Research, 13*, (4): 43-50, (2010).
15. Larry Laudan, "The Demise of the Demarcation Problem". In Adolf Grünbaum, Robert Sonné Cohen, Larry Laudan. Physics, Philosophy, and Psychoanalysis: Essays in Honor of Adolf Grünbaum. Springer, 1983.
16. Karl Popper, The logic of scientific discovery, New York: Basic Books, 1959.
17. Thomas. S. Kuhn, The Structure of Scientific Revolutions, 2nd ed., Chicago: Univ. of Chicago Press, 1970.
18. Paul Feyerabend, Against Method: Outline of an Anarchistic Theory of Knowledge, 1975.
19. Jai-Hang Lim, Soon-Min Kang, Young-Tae Kong, Byung-Soon Choi, and Jeong-Hee Nam, "The Development of an Instrument to Assess High School Students' Views on Science-Technology-Society," *Journal of the Korean Association for Science Education, 24*(6): 1143-1157, (2004).
20. D.-K. Kang and K. Sohn, "Learning decision trees with taxonomy of propositionalized attributes," *Pattern Recognition, 42*(1): 84-92, Jan. 2009, Elsevier B.V.
21. D.-K. Kang and M.-J. Kim, "Propositionalized Attribute Taxonomies from Data for Data-Driven Construction of Concise Classifiers," *Expert Systems With Applications, 38*(10): 12739-12746, September 2011, Elsevier B.V.
22. J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
23. Is See5/C5.0 Better Than C4.5?, Rulequest Research 2009. Online: http://www.rulequest.com/see5-comparison.html.
24. L. Breiman, Classification and Regression Trees, Chapman & Hall, 1984.
25. G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics 29*(2): 119–127, 1980.
26. Y. J. Lee, "Factors in Information and Communication Ethics Behavior of College Students Majoring in Information and Communication Engineering," *Journal of Engineering Education Research, 13*(3): 68-77, (2010).
27. Y. J. Lee, "A Study on the Information and Communication Ethics from the Survey of College Students," *Journal of Engineering Education Research, 13*(3): 96-103, (2010).
28. W. S. Kim, "A Study on Factors of Education's Outcome using Decision Trees," *Journal of Engineering Education Research, 13*(4): 51-59, (2010).
29. C. P. Snow, The Two Cultures, London: Cambridge University Press, 1959.
30. R. P. Feynman, The Pleasure of Finding Things Out: The Best Short Works of Richard P. Feynman, Basic Books; 1 edition, 2000.
31. R. Carnap, The Continuum of Inductive Methods. University of Chicago Press, 1952.
32. C. Hempel, Philosophy of Natural Science, 1966.

33. K. Popper, Conjectures and Refutations: The Growth of Scientific Knowledge, 1963.
34. P. Feyerabend, Against Method: Outline of an Anarchistic Theory of Knowledge, 1975.
35. I. Lakatos, The Methodology of Scientific Research Programmes: Philosophical Papers Volume 1. Cambridge: Cambridge University Press, 1978.
36. T. Kuhn, The Structure of Scientific Revolutions. Chicago: University of Chicago Press, 1962.
37. W. Bauchspies, J. Croissant, and S. Restivo, Science, Technology, and Society: A Sociological Approach, 2005.

Dae-Ki Kang

is an assistant professor at Dongseo University in South Korea. He was a senior member of engineering staff at the attached Institute of Electronics & Telecommunications Research Institute in South Korea.

He earned a Ph.D. in computer science from Iowa State University in 2006. Prior to joining Iowa State, he worked at two Bay-area startup companies and at Electronics and Telecommunication Research Institute in South Korea. He received a science master degree in computer science at Sogang University in 1994 and a bachelor of engineering (BE) degree in computer science and engineering at Hanyang University in 1992.

His research interests include Cloud Computing, Big Data, Multi-Relational Learning, Social Networking Services and Complex System Network, Statistical Graphical Model, Ontology Learning, Datamining based Intrusion Detection, Context Sensitive Mobile Advertising, Malicious Bots Detection, Whole Body Interaction, Web application firewall, Web Mining and Computer Vision.

Phone: +82-51-320-1724

E-mail: dkkang@dongseo.ac.kr