International Journal of Innovative
Computing, Information and Control
Volume **12**, Number **5**, October **2016**

 ISSN 1349-4198
pp. **1679–1690**

USING SVM TO COMBINE BAYESIAN NETWORKS FOR EDUCATIONAL TEST DATA CLASSIFICATION

Yen-Chun Tseng, Chih-Wei Yang and Bor-Chen Kuo

Graduate Institute of Educational Information and Measurement
National Taichung University of Education
No. 140, Minsheng Rd., West Dist., Taichung City 40306, Taiwan
yjtzeng@mail2000.com.tw; {yangcw; kbc}@mail.ntcu.edu.tw

Received April 2016; revised August 2016

Abstract. *The goal of this paper is trying to develop fusion methods for combining multiple Bayesian networks for modeling students' learning bugs and skills. Seven methods, maximum, minimum, mean, product, majority vote, sub-structure and support vecter machine, are proposed and evaluated based on educational assessment data. There are five test data, and each has five different Bayesian networks designed by experts. Experiment results show that the proposed fusion methods, maximum, minumum, mean, majority vote, sub-structure and SVM can improve the classification rates but only sub-structure and SVM methods can increase classification rates stably across all datasets. Using SVM to combine the different experts judgments is a more proper method which outperforms other methods.*

Keywords: Bayesian network, Fusion, Support vector machine, Combining information, Diagnostic testing

1. **Introduction.** Knowing student's state of learning, for instance, lack of concepts, proficiency level of skills, or presence of bugs, is benefit to teachers and remedies instructors planning instructions. Especially bugs are hard to diagnose with their uncertainty and instability, meaning that the bugs may arise inconsistently between items even in a single test.

Ketterlin-Geller and Yovanoff [1] compare three types of diagnostic assessment approaches, cognitive diagnostic assessment, skills analysis assessment, and error analysis assessment. Skills analysis assessment involves aggregating student's responses data to determine skills mastery for personalized profiles and is usually used to identify students who may be at risk for failure in the domain. Error analysis is the process of reviewing student's item responses to identify a pattern of misunderstanding and can be used to provide information for designing remedial instructional program [1,2]. Ketterlin-Geller and Yovanoff take some examples to describe only using one of the skills or error analysis assessment is not enough, by integrating multiple information and the principles of cognitive psychology with response analysis. Diagnostic assessments can be created to provide detailed information into sustained errors that disturb student thinking [1].

One of the most popular diagnostic assessment methods of modeling uncertainty is Bayesian networks [3-5] which consider unstable event occurrences using a probabilistic approach. It is a powerful tool to diagnose, explain, and model student's cognitive skills and has been applied widely in educational assessment, including mixed-number subtraction [3], physics problem solving [6,7], proportional reasoning [8], and diagnostic student's learning bugs and sub-skills [9,10]. Bayesian network is a probabilistic graphical model that can represent the relationships between variables such as concepts and bugs

by conditional probabilities. For example, given learning bugs, the network can be used to compute the probabilities of the lack of corresponding sub-skills.

Nevertheless, the probabilities of bugs and skills occurrences are estimated according to the architectures which are constructed by the domain experts. Different experts may construct different Bayesian networks based on their own domain knowledge, judgment and teaching experiences. These differences may let some nodes be estimated better in some specific networks. Building a good Bayesian network that means all nodes estimations are better than other networks is very difficult and also time-consuming.

Many studies [11,12] show that combining multiple information or fusing the outputs of different classifiers may improve classification of complicated datasets. Accordingly, the goal of this paper is trying to develop fusion methods for combining multiple Bayesian networks and compare to the classification results of the proposed methods.

2. **Bayesian Network.** Bayesian networks are graphical models for probabilistic relationships among a set of domain variables. These graphical structures are used to illustrate knowledge about comprising uncertainty.

Bayesian networks not only enable efficient uncertainty reasoning with hundreds of variables, but also help humans understand the modeled domain better. It has been applied to expert systems in many fields. A Bayesian network is composed by a directed acyclic graph and a corresponding set of conditional probability distributions. The graph consists of a set of nodes and directed arcs, where the nodes represent variables, and the arcs signify direct dependence between the connected nodes. In addition to the graphical structure, $\mathbf{P} = \{p(x_1\,|\pi_1), \ldots, p(x_n\,|\pi_n)\,\}$ is a set of conditional probability distributions (CPDs) associated with each node in the network, where π_i is the set of parents of node X_i in D. According to conditional independence property in Bayesian network, the joint probability distribution of all variables can be simplified and reduce complexity of inference desired probabilistic information.

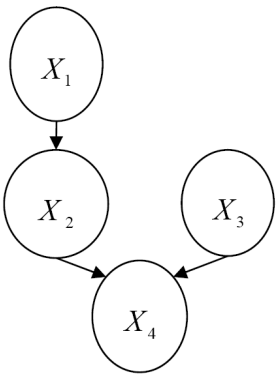

FIGURE 1. An example of a Bayesian network

Figure 1 is an example of a Bayesian network, X_1 is parent node for X_2, and it can be represented by conditional probability $P(X_2|X_1)$. X_2 and X_3 are parent nodes for X_4, so it can be represented by $P(X_4|X_2, X_3)$. And the joint probability can be shown as:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3)P(X_4|X_2, X_3) \tag{1}$$

If X_4 is observed, then Bayesian network can compute the posterior of X_1 by using Bayesian estimation theory. Thus, complex relationships between lots of variables can be decomposed to smaller subsets multiplication of variables.

Mislevy et al. [13] described three key points of a Bayesian network framework that can be used for proceeding probability-based inference in cognitive diagnosis. The first

one is building Bayesian networks for a student model; the second is constructing tasks that may let students demonstrate their performance in targeted knowledge and skill; the third is creating Bayesian networks for evidence models that describe how to extract the evidence from the task, and indicator to parent student-model variables. Different experts may have their own view in student model that means different connection relationship between variables, so this study tries to use multiple student models to diagnose students' skills and bugs.

3. **Multiple Classifiers Fusion Methods for Bayesian Networks.** In [14], six fusion methods were applied to combining Bayesian networks. Five of them, maximum, minimum, mean, product, and majority vote methods, were adopted from Duin's multiple classifiers fusion methods [15]. A brief description of those six methods is in the following.

Once a set of posterior probabilities $\{p_{ij}(x),\ i = 1, \ldots, m;\ j = 1, \ldots, c\}$ for m classifiers and c classes is computed for test object x, they have to be combined into a new set $q_j(x)$ that can be used for the final classification. In this study, it only has two classes, master and non-master of the skills, and then the new confidence $q(x)$ for class j is now computed by

$$q(x) = q_j(x) = rule\left(p_{1j}(x), p_{2j}(x), \ldots, p_{mj}(x)\right) \tag{2}$$

There are two sets of rules: fixed fusions and trained fusions [15]. The following fixed fusions are used for rules: maximum, minimum, mean, product, and majority vote [11,16,17]. The maximum rule selects the network producing the highest posterior probability estimates. In contrast, the minimum rule selects the network having the lowest probabilities. Mean rule averages the posterior probability thereby reducing estimation errors.

According to the Bayesian theory, given networks measurements $P(x|BN_i)$ as the posterior $p_i(x) = p_{i1}(x)$, $i = 1, \ldots, m$, the node, x, should be assigned to 1 to provide the interpretation of a posteriori probability by comparing with a threshold ε, in which 0.5 is used in this study.

$$assign \quad x \to 1 \quad if\ q(x) = rule\{P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)\} \geq \varepsilon \tag{3}$$

3.1. **Maximum method.**

$$\begin{aligned} &assign \quad x \to 1 \quad if \\ &Max\left\{P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)\right\} \\ &= \arg\max_i \left\{P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)\right\} \geq \varepsilon \end{aligned} \tag{4}$$

3.2. **Minimum method.**

$$\begin{aligned} &assign \quad x \to 1 \quad if \\ &Min\left\{P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)\right\} \\ &= \arg\min_i \left\{P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)\right\} \geq \varepsilon \end{aligned} \tag{5}$$

3.3. **Mean method.**

$$\begin{aligned} &assign \quad x \to 1 \quad if \\ &Mean\left\{P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)\right\} = \frac{1}{m}\left(\sum_{i=1}^{m} P(x|BN_i)\right) \geq \varepsilon \end{aligned} \tag{6}$$

3.4. **Product method.**

$$\begin{aligned} &assign \quad x \to 1 \quad if \\ &Prod\left\{P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)\right\} = \frac{\prod_{i=1}^{m} P(x|BN_i)}{\prod_{i=1}^{m} P(x|BN_i) + \prod_{i=1}^{m} (1 - P(x|BN_i))} \geq \varepsilon \end{aligned} \tag{7}$$

3.5. **Majority vote method.**

$$\begin{aligned} &assign \quad x \to 1 \quad if \\ &\Delta_i = \begin{cases} 1 & P(x|BN_i) \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}, \ i = 1, 2, \ldots, m \\ &Vote\{P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)\} = \frac{1}{m}\sum_{i=1}^{m} \Delta_i \geq 0.5 \end{aligned} \tag{8}$$

3.6. **Sub-structure fusion method.** As the first step of majority vote method, assign posterior probability into master or non-master, and then $\psi(\Delta_i)$ will equal 1, if the ith Bayesian network has the highest correct classification on node x. The sub-structure fusion method selects the best Bayesian network diagnosis as the output for each node.

$$\begin{aligned} &assign \quad x \to 1 \quad if \\ &\Delta_i = \begin{cases} 1 & P(x|BN_i) \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}, \ i = 1, 2, \ldots, m \\ &M^{sub} = [\psi(\Delta_1), \psi(\Delta_2), \ldots, \psi(\Delta_m)]^T, \ \psi(\Delta_i) \in \{0, 1\} \\ &Sub\{P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)\} \\ &= [P(x|BN_1), P(x|BN_2), \ldots, P(x|BN_m)] * M^{sub} = P(x|BN_i) \geq \varepsilon \end{aligned} \tag{9}$$

In the experiment of [14], three Bayesian networks were generated based on one educational test data. The performances of single Bayesian network, and six fusion methods were evaluated. The experimental result of [14] showed that maximum, mean, majority vote, and sub-structure fusion methods outperformed single Bayesian network and sub-structure fusion method has the best performance. This finding may be very limited and restricted because of using only one data.

Moreover, sub-structure fusion method tended to select the best sub-structure among different Bayesian networks, and then combined all the best sub-structure to form a new Bayesian network, which is very similar to the feature selection method in machine learning. Basically, the performance of sub-structure fusion method cannot exceed the union of all Bayesian networks.

In this study, a new fusion method using support vector machine will be proposed and the classification accuracy of this proposed fusion method is expected to exceed that of the union of all Bayesian networks. In addition, more datasets and Bayesian networks are used in experiments for obtaining a more robust result.

4. **Multiple Bayesian Networks Fusion by Support Vector Machine.** Support vector machine (SVM) [18,19], a success learning algorithm commonly used for classification and regression issues, is motivated by designing a linear discriminate function with the consideration of the margins. The following is a brief introduction of SVM [20]. Given a training dataset of labelled pairs (x_i, y_i), where $x_i \in \Re^n$, $y_i \in \{+1, -1\}$, and $i = 1, 2, \ldots, N$, the goal of SVM is to find the separating hyperplane $w^T\varphi(x)$ that maximizes the margin b, and it requires the solution of the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}w^T w + C\sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & y_i\left(w^T\varphi(x_i) + b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \\ & y_i\left(w^T\varphi(x_i) + b\right) \geq 1 - \xi_i \end{aligned} \tag{10}$$

where C and ξ are penalty parameter and slack variables, respectively, for the soft-margin SVM. In this study, the posterior probability of bug, skill, and indicator nodes in every single Bayesian network will be treated as the input x_i of SVM and related expert's

classification of each bug, skill, and indicator nodes will be treated as the label y_i. By this way, a training dataset can be generated to train an SVM for classifying bug, skill, or indicator. For this fusion method, we used a set of off-the-shelf classifiers taken from Matlab toolbox PRTools [21].

5. **Experiment Design.** For evaluating the performance of the proposed SVM fusion method, six educational assessment datasets about math are collected from the 6th and 7th grade students in Taiwan. The research developed computerized diagnostic assessment by using multiple Bayesian Network [22-27] and the units of assessment include four fundamental operations of fraction, cylinders and pyramid, girth of fan-shapes, algebra, circular area, and linear equation with two variables.

Table 1 represents the abstract of six datasets including sample sizes, number of items, bugs and skills. Every test was taken for forty-five minutes. Each dataset comprises five Bayesian networks which were constructed by domain experts individually within the same skills and bugs. Figure 2 to Figure 6 show different Bayesian networks of test 1. There are four layers designed in proposed Bayesian networks, including competence indicator, sub-skill, bug, and item. The first layer is defined to targeted domain, and links to the second layer, corresponded sub-skills. This structure implies mastery of competence indicator is evaluated by mastery of sub-skills. The third layer is defined to correspond bugs with sub-skills. And the fourth layer is tasks which are designed to detect students' profiles.

Table 1. Summary of six tests

	Sample size	Number of items	Number of bugs	Number of skills
Test 1	140	20	13	10
Test 2	289	32	30	20
Test 3	260	29	14	13
Test 4	233	35	20	21
Test 5	256	39	32	25
Test 6	302	26	20	17

Table 2. Correct classification rate index

Experts' judgment \ BN diagnostic	Master(1)	Non-master(0)
Correct(1)	f_{11}	f_{10}
Incorrect(0)	f_{01}	f_{00}

Take "fundamental operations of fraction" unit as an example. Competence indicator can be fraction addition and subtraction, and one of the sub-skills is fraction addition. The most constantly occurring bug is "numerator plus numerator, and denominator plus denominator", so the corresponding task can be designed as "1/2 + 3/8". Bayesian network will diagnose students' learning status according to their response of tasks. (More error types detail can see [1], p.6.)

Evaluation index. To evaluate Bayesian network, which means how consistency of a diagnostic of networks with that from experts, a 5 folds cross-validation method is used. Four fifths of samples are chosen to be training data which are used to train the Bayesian network and the others of samples are retained as the validation data to test the model.

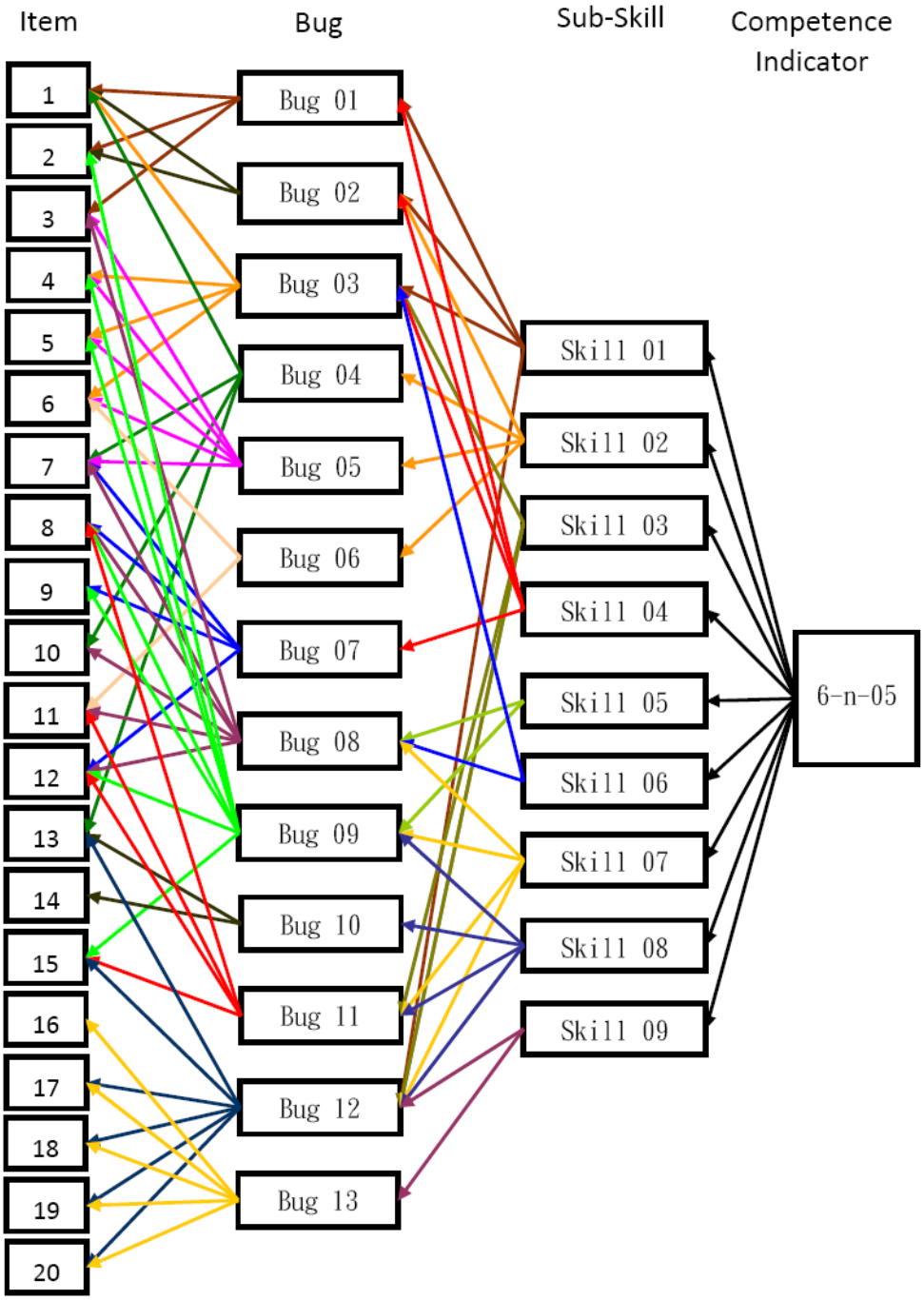

FIGURE 2. The first Bayesian networks of test 1

This cross-validation process is executed five times, with each of the one fifth subsamples selected as the validation data. An average of five folds results is computed as the final result that means the evaluation index, correct classification rate. The correct classification rates shown are the percentage agreement between the models diagnostic and experts' judgment [9,10,28]. The computing formula is as follows:

$$\text{The correct classification rate} = \frac{f_{11} + f_{00}}{N}$$

N: the number of testing samples.

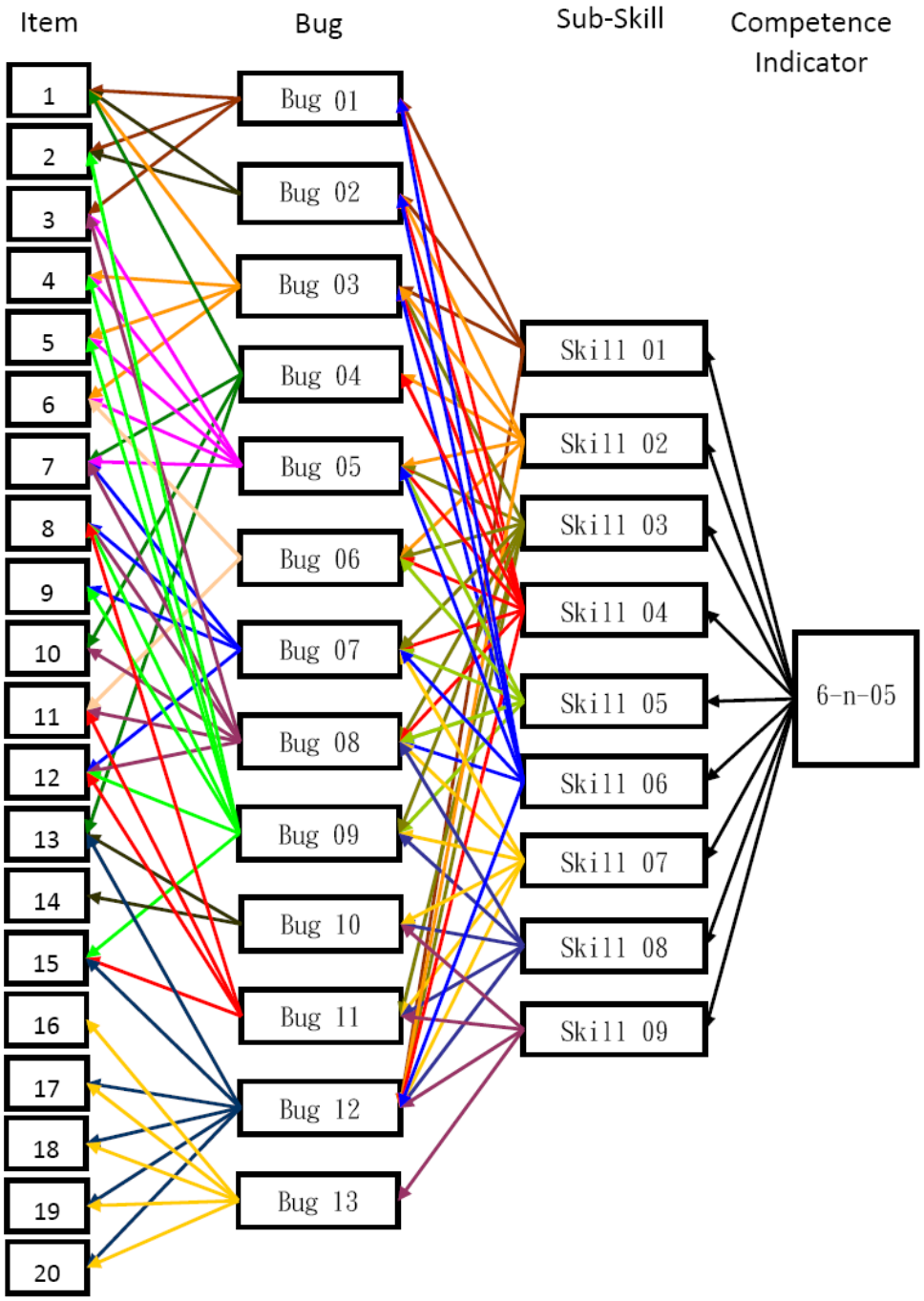

FIGURE 3. The second Bayesian networks of test 1

6. **Results.** There are several skills and bugs in each Bayesian network, so the mean of all skills and bugs results is used to represent the performance of networks. Table 3 represents the correct classification rates of each test cross network and the cells of character shading in every row show the highest classification rates in each test. In test 1, the classification rates are greater than 90% across the five networks and the difference between networks is smaller than 2%. The similar results can also be found in other tests. These results imply experts may have some better opinions when modeling the Bayesian networks, but classification rate is difficult to substantially exceed others'.

For more detail, the highest classification rate does not imply every node is higher than that in other networks. For example, the Bayesian network 3 classification rate is the

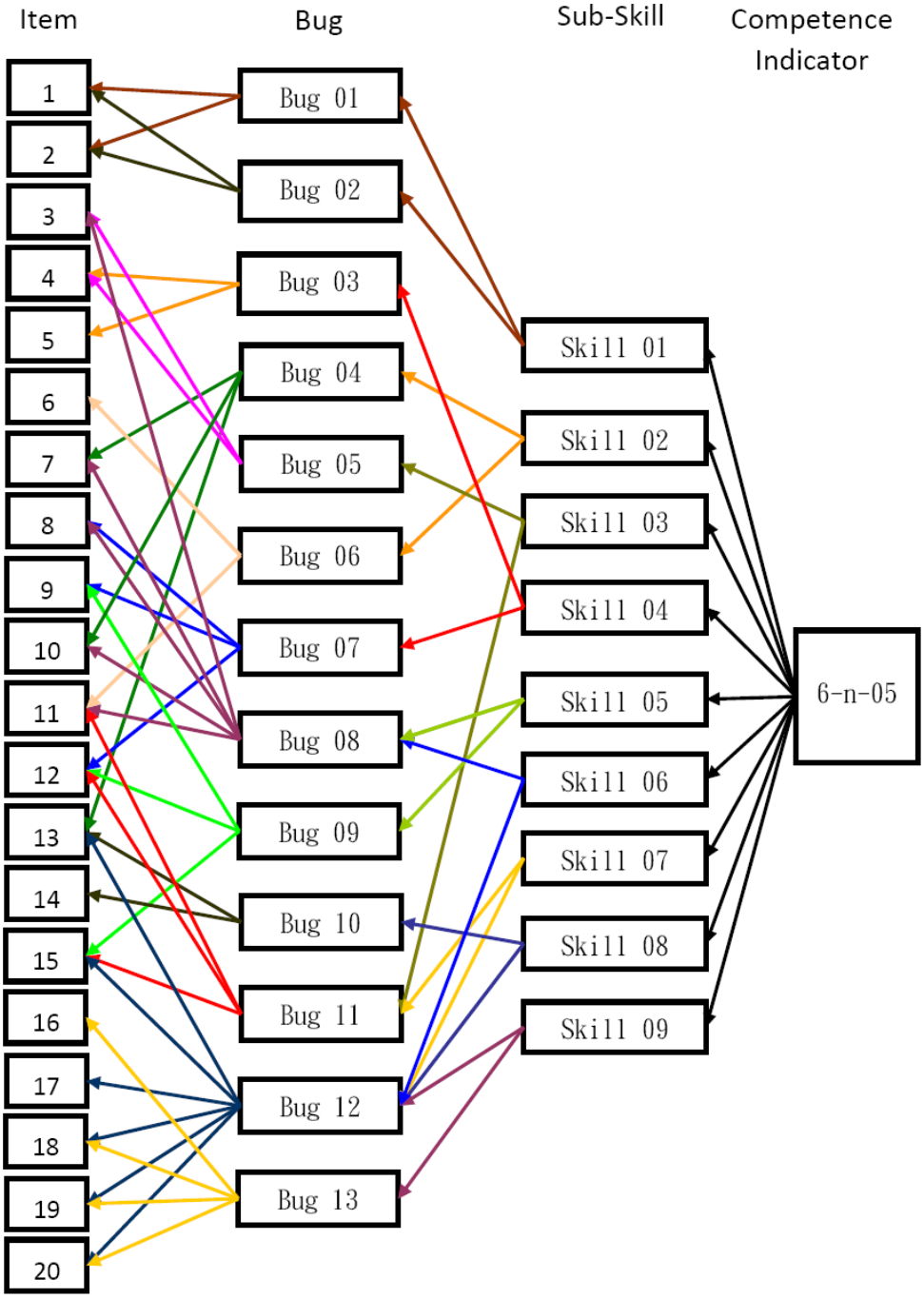

FIGURE 4. The third Bayesian networks of test 1

TABLE 3. Correct classification rates of single Bayesian networks

	BN1	BN2	BN3	BN4	BN5
Test 1	91.30%	90.68%	91.77%	91.27%	91.02%
Test 2	95.53%	95.48%	95.40%	95.88%	95.86%
Test 3	88.35%	88.12%	88.63%	87.75%	89.06%
Test 4	92.94%	93.00%	93.21%	92.97%	92.91%
Test 5	91.42%	91.60%	91.59%	91.41%	91.16%
Test 6	92.79%	92.19%	92.06%	92.75%	92.60%

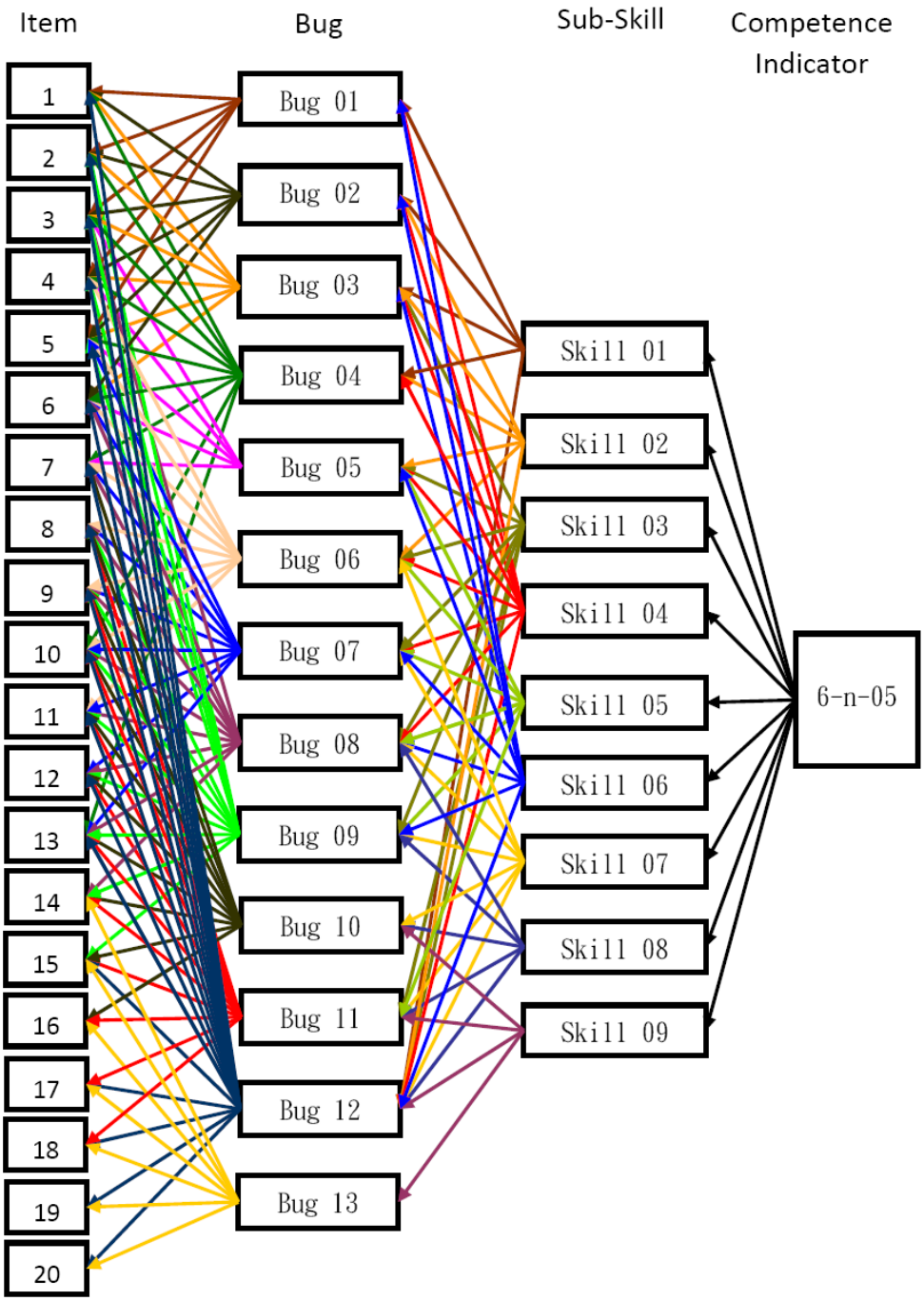

FIGURE 5. The fourth Bayesian networks of test 1

highest result in test 1, but some nodes like skills in Bayesian network 1 are better than Bayesian network 3. So that is why we need to combine different Bayesian networks' classifications.

Table 4 represents the fusion method results of combining five networks classifications and the cells of character shading mean the correct classification rates higher than single Bayesian network result. Compared with the highest classification rates in each test, the results of maximum and minimum fusion methods are improved in tests 4 and 5. The results of vote method are increased in tests 3, 4, and 5. The results of mean method are increased in tests 2, 3, 4, and 5. Only sub-structure and SVM methods results are higher

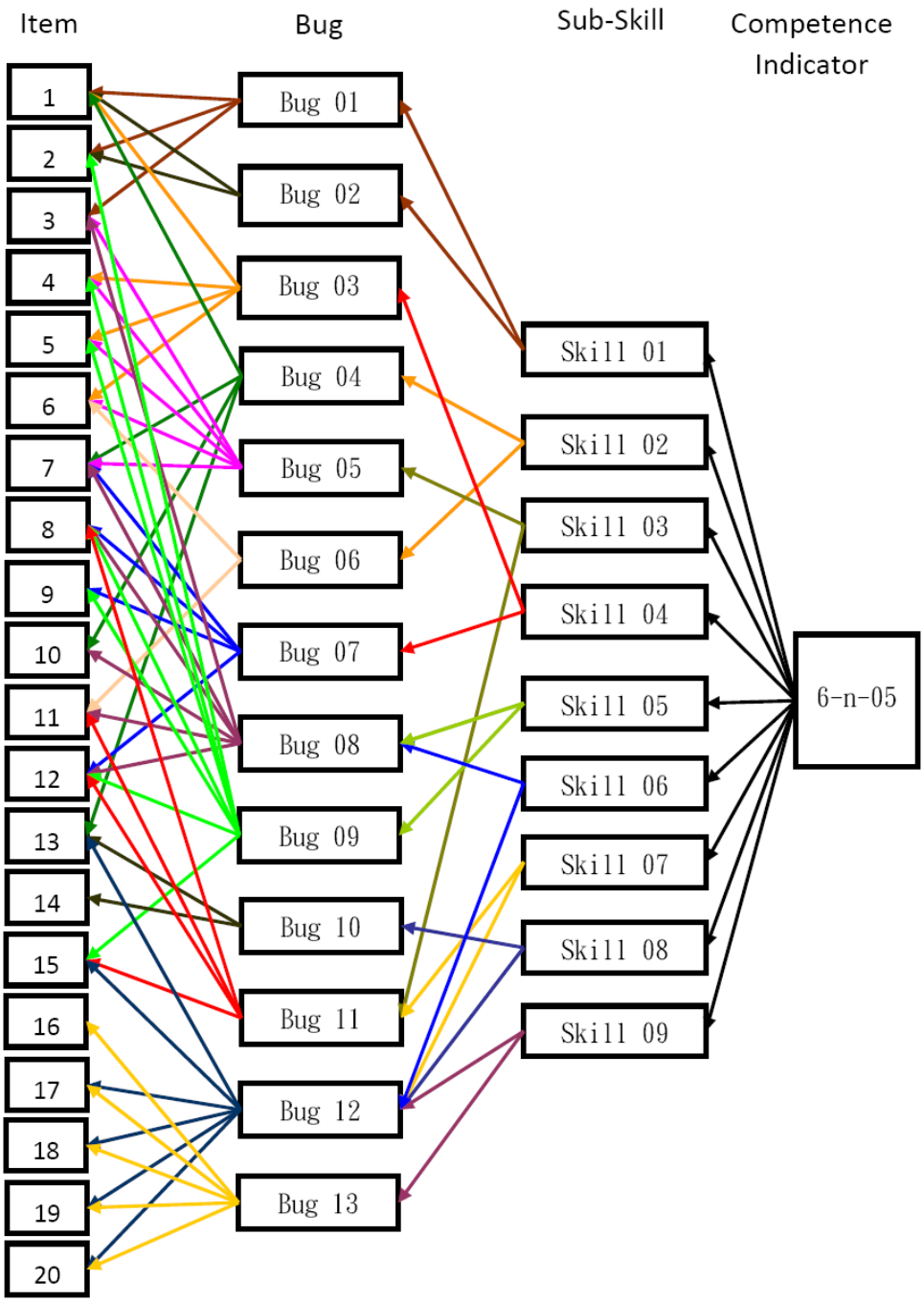

FIGURE 6. The fifth Bayesian networks of test 1

than before in all tests. Moreover, SVM method's results outperform sub-structure in every test. Conversely, the results of product method are decreased in all tests.

7. **Discussion.** Bayesian networks are powerful and useful tools of diagnosing assessment dataset of bugs and skills. Clearly defining the relations between bugs and skills, and simultaneously diagnosing bugs and skills can improve the diagnostic results [9].

Combining multiple Bayesian networks information can increase the diagnosis consistency [14]. The results of this study are conducted from multiple datasets and Bayesian networks. Seven fusion methods were evaluated on six educational test data. Some fusion methods can increase the classification rates but are not stable across different tests. Only

TABLE 4. Classification rates of seven fusion methods

	Max	Min	Product	Vote	Mean	Sub-structure	SVM
Test 1	90.68%	91.49%	88.08%	91.55%	91.71%	91.96%	92.76%
Test 2	95.02%	95.87%	95.36%	95.71%	96.01%	96.09%	97.45%
Test 3	87.78%	89.00%	80.07%	89.36%	89.39%	90.11%	90.63%
Test 4	93.25%	93.24%	90.19%	93.27%	93.31%	93.34%	93.93%
Test 5	91.61%	91.73%	81.02%	91.70%	91.73%	91.82%	92.43%
Test 6	92.44%	92.08%	87.68%	92.69%	92.68%	92.82%	92.98%

sub-structure fusion and SVM methods can promote classification rates stably, and SVM method outperforms sub-structure method in all tests. The results show the robustness of SVM [29] holds in combining educational test data classification. Thus, the methodology proposed here may provide a suggestion to deal with different experts' opinions.

Acknowledgment. This research is partially supported by the Ministry of Science and Technology under the grant number NSC 102-2511-S-142-008-MY3.

REFERENCES

[1] L. R. Ketterlin-Geller and P. Yovanoff, Diagnostic assessments in mathematics to support instructional decision making, *Practical Assessment, Research & Evaluation*, vol.14, no.16, pp.1-11, 2009.
[2] U. K. De and D. Sengupta, Error analysis in mathematics in relation to secondary school students, *Indian Journal of Educational Research*, vol.3, pp.105-125, 2014.
[3] R. J. Mislevy, *Probability-Based Inference in Cognitive Diagnosis*, ETS Research Report Series, pp.1-31, 1994.
[4] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, Morgan Kaufmann, CA, 1988.
[5] M. J. Culbertson, Bayesian networks in educational assessment: The state of the field, *Applied Psychological Measurement*, pp.1-19, 2015.
[6] J. D. Martin and K. VanLehn, A Bayesian approach to cognitive assessment, *Cognitively Diagnostic Assessment*, pp.141-165, 1995.
[7] J. D. Martin and K. VanLehn, Evaluation of an assessment system based on Bayesian student modeling, *International Journal of Artificial Intelligence in Education*, vol.8, no.2, pp.179-221, 1998.
[8] A. Beland and R. J. Mislevy, Probability-based inference in a domain of proportional reasoning tasks, *Journal of Educational Measurement*, vol.33, pp.3-27, 1996.
[9] J. Lee and J. E. Corter, Diagnosis of subtraction bugs using Bayesian networks, *Applied Psychological Measurement*, vol.35, no.1, pp.27-47, 2011.
[10] S.-C. Shih and B.-C. Kuo, Using Bayesian networks for modeling students' learning bugs and sub-skills, *Lecture Notes in Artificial Intelligence*, vol.3681, pp.69-74, 2005.
[11] J. Kittler, M. Hatef, R. P. Duin and J. Matas, On combining classifiers, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.20, no.3, pp.226-239, 1998.
[12] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley, Chichester, 2004.
[13] R. J. Mislevy, R. G. Almond, D. Yan and L. S. Steinberg, Bayes nets in educational assessment: Where do the numbers come from? *Proc. of the 15th Conference on Uncertainty in Artificial Intelligence*, pp.437-446, 1999.
[14] B.-C. Kuo, T.-Y. Hsieh and Y.-Y. Chang, Combining multiple Bayesian networks for modeling students' learning bugs and skills, *The 7th International Conference on Intelligent Technologies*, Taipei, Taiwan, 2006.
[15] R. P. Duin and D. M. Tax, Experiments with classifier combining rules, *Multiple Classifier Systems*, pp.16-29, 2000.
[16] L. Xu, A. Krzyżak and C. Y. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, *IEEE Trans. Systems, Man and Cybernetics*, vol.22, no.3, pp.418-435, 1992.
[17] R. E. Schapire, The strenght of weak learnability, *Machine Learning*, vol.5, pp.197-227, 1990.

[18] B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers, *Proc. of the 5th Annual Workshop on Computational Learning Theory*, pp.144-152, 1992.
[19] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, vol.20, no.3, pp.273-297, 1995.
[20] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.
[21] R. P. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. De Ridder, D. M. J. Tax and S. Verzakov, A Matlab toolbox for pattern recognition, *PRTools Version*, 2000.
[22] C.-P. Lin, *The Design of Individualized Test and Digitalized E-Learning Content by the Case of "Calculation of Fractions" in Elementary School Mathematics*, Master Thesis, Asia University, Taiwan, 2007.
[23] S.-J. Lin, *A Study of Developing Adaptive Diagnostic Test and Remedial Instruction System – Using "Prism, Pyramid, Cylinder and Cone" Unit as an Example*, Master Thesis, Asia University, Taiwan, 2007.
[24] L.-Y. Yeh, *On-Line Diagnostic Test and Adaptively Remedial Instruction of the Girth and Area of a Fan-Shaped Unit in Elementary School Curriculum, Based on Combining Multiple Bayesian Networks*, Master Thesis, Asia University, Taiwan, 2007.
[25] C.-M. Chang, *Integrating Advantages of Diverse Expert Bayesian Networks in Developing an Adaptive Learning System for the Sixth-Year Elementary School Algebra Class*, Master Thesis, Asia University, Taiwan, 2007.
[26] C.-J. Huang, *Diagnostic Test and Adaptively Remedial Instruction System Based on the Combination of Different Bayesian Networks – Using the "Circumference and Circle Area" as an Example*, Master Thesis, Asia University, Taiwan, 2007.
[27] M.-J. Hung, *Application of Bayesian Network – Use Two-Variable Linear Equation Unit in Junior High School Mathematics Course to Diagnostic Test Design and Adaptive Remedial Instruction*, Master Thesis, Asia University, Taiwan, 2007.
[28] J. Chen and J. de la Torre, A general cognitive diagnosis model for expert-defined polytomous attributes, *Applied Psychological Measurement*, vol.37, no.6, pp.419-437, 2013.
[29] H. Xu, C. Caramanis and S. Mannor, Robustness and regularization of support vector machines, *Journal of Machine Learning Research*, vol.10, pp.1485-1510, 2009.