AN EVALUATION OF VALIDATION CRITERIA ON INTELLIGENT SYSTEM VALIDATION PROCESS

Dyah Rahayu and Siti Rochimah

Department of Informatics Engineering, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Jalan Teknik Kimia,Gedung Teknik Informatika Kampus ITS Sukolilo, Surabaya, 60111, Indonesia

E-mail: dyah.s.rahayu11@mhs.if.its.ac.id

Abstract

In the software development cycle, validation is the important stage which is held in final stage especially in intelligent system. Validation obtains the validity, credibility and trustworthy of the system. It is needed to ensure that the intelligent system has same manner as human experts. Whilst with the importance of validation stage, determining the validation criteria is also important. This paper presents the evaluation of validation criteria which is commonly used in intelligent system validation process. The evaluation is carried out by reviewing the literature of intelligent system validation process. The result shows that the validation criteria have its own characteristic so it requires for understanding the validation criteria characteristics, purposes of validation and also the intelligent system itself to hold validation process.

Keywords: *intelligent system, validation, validation criteria*

Abstrak

Pada siklus pengembangan perangkat lunak, validasi adalah tahap penting yang diadakan ditahap akhir terutama dalam bidang sistem cerdas. Validasi dilakukan untuk memperoleh validitas, kredibilitas, dan kepercayaan terhadap sistem. Hal ini diperlukan untuk memastikan bahwa sistem cerdas memiliki cara yang sama seperti para ahli. Sementara itu dengan pentingnya tahap validasi, penentuan kriteria validasi juga menjadi penting. Makalah ini menyajikan evaluasi kriteria validasi yang umum digunakan dalam proses validasi sistem cerdas. Evaluasi dilakukan dengan melakukan *review* literatur dari proses validasi sistem cerdas. Hasil penelitian menunjukkan bahwa kriteria validasi memiliki karakteristik tersendiri sehingga untuk melaksanakan proses validasi diperlukan pemahaman terhadap karakteristik kriteria validasi, tujuan validasi dan juga sistem cerdas itu sendiri.

Kata Kunci: *sistem cerdas, validasi, kriteria validasi*

1. Introduction

The final phase of software development cycle is to test its quality and to obtain the credibility of the system. The process is well known as verification and validation (V&V). Verification refers to “how to build the system right” and validation is about “how to build the right system” [1].

For intelligent system, verification is a process to obtain the correctness of the implementation of intelligent system, the correctness of the representation of input parameter and also the correctness of the logical structure being built [1]. Validation of intelligent system is utilized to ensure the validity of the system is in a reasonable level so that the human may use the output of the system as a recommendation of decision making process. The validation process is usually done by comparing the system results with the expert knowledge.

Many research have proposed verification and validation method because its necessity in software development cycle. Although many verification methods have been proposed, it is difficult to comparing them directly because of its special method that is based on each software structure. The main objective of validation method is to ensure that the system can be used in real world. The other important thing is that the software output should satisfy expert demand. So, the validation method should consist of comparison between the system and the expected performance.

Based on Mosquera-Rey and Moret-Bonillo[1], there are two kinds of validation, i.e. result oriented validation and usage oriented validation. The most well-known one is the results-oriented validation which utilizes statistical approach to measure the performance of system compared with the expert-knowledge [1][2][3]. The result oriented validation is carried out by measuring some statistical criterion between system output and expert knowledge. Some statistical criterion in the process is called as validation criteria. There are four types of comparison based on the presence of expert, namely validation against a single expert, group of expert, consensus of expert and the standard knowledge [1]. Basically, the first step to obtain the value of validation expert is creating table contingency. The next step is to calculate the suitable validation criteria and make the interpretation of the validation criteria result.

Because of many validation criteria that has been proposed and commonly used, it seems to be difficult to determine the validation criteria in accordance with the characteristic of the system. This problem occurred because each system has each output with its own characteristic.

Therefore, this paper presents an evaluation of validation criteria on intelligent system validation process based on the system output characteristic. The main purpose is to identify the best validation criteria to be used in measuring statistical criterion of result oriented validation of each system based on their output characteristic. If the validation criterion which is used to validate the system is suitable, so the result of validation may be able to represent the ability of system in true way. The study was done by researching and reviewing many validation of intelligent system and finding their own characteristic and also their limitations.

2. Methodology

Result oriented validation compares the system performance with the expected performance demand based on standard reference. The aim of the process is to find out whether the system is feasible to use. When there is none of standard reference, the comparison is conducted on system performance and the expert knowledge.

The different process is made on different amount of standard reference (standard reference itself or expert comment), here are: Process of validation using single expert is done by measuring validation criteria between a single human expert and system output. Whether the result of this process trustworthy or not depends on the credibility of the human expert. It is such a risk to do the validation in this way because the outcome is highly dependent on the consistency of a single expert. The measurement process is called as pair measure.

Validation uses group of expert is commonly used because of its advantages. The outcome from the process does not depend on a single expert, so it is more credible. The other one is the possibility to measure more validation criteria than the other type of validation have. The process of measurement using group of expert is called as group measurement.

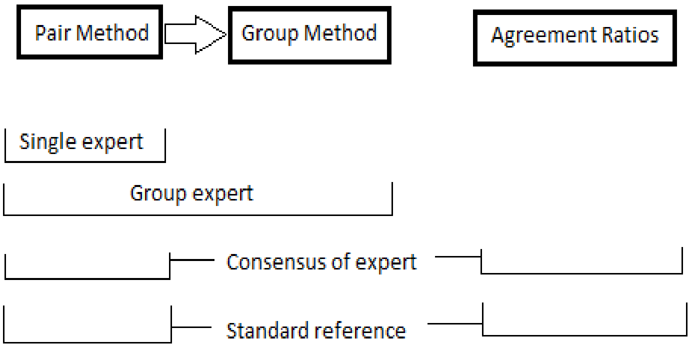


Fig 1. The method used based on presence of expert.

The process to carry out Validation uses consensus of expert is as same process as validation uses single expert which uses pair method, but as the output of system is compared with the consensus of expert, it is more reliable. A consensus of expert is created within the agreement of group experts. Therefore the output is objective and does not depend on single expert.

The standard reference is relevant with the consensus of expert. It is a well-known and publicized material of consensus of expert. To validate using standard reference utilized pair method and the special one is agreement ratios.

Which one the method of measurement done according the type of validation based on presence of expert is explained in figure 1. The pair method is used for validating using single expert, consensus of expert, standard reference and also group expert. The group expert uses pair method to obtain each pair measurement between system output and each expert, but the result is difficult to interpret because each single expert has its value of validation criteria. So, the pair method outcome is used as the input for group method. The agreement ratios is used for standard reference as well as for consensus of expert because its relevancy before.

The pair method are done by 1) creating validation database which records expert comment and system output 2) constructing the contingency table of expert and system and 3) measuring some validation criteria from contingency table. The validation criteria for pair method explained in next section.

For group method, the early step is as same as the step of pair method. The remaining ones are 1) summarized the contingency table of each pair method, 2) construct the summarized table and 3) measure some validation criteria from the summarized table. The validation criteria explained in next section.

The agreement ratios are used for validation when the standard reference or consensus of expert are exist. The process for agreement ratios is as pair method process, make the validation database and then construct the contingency table. The validation criteria of agreement ratios are explained in next section.

Validation criteria are employed to each test case to identify their correct result. The criteria are usually interpreted by some statistical measurement which is belonging to quantitative evaluation methods. Based on [3], statistical measures are separated into three classes: pair measurement, group measurement and agreement ratios, so the validation criteria will be explained in such way too

Pair measurement is done by constructed the contingency table of all possible pair of expert and system and then calculated the agreement measure and association measure.

Agreement measurements use an index which corresponds to probability value of same interpretation between an expert with another expert. There are four popular measurements i.e agreement index, within-one agreement index, kappa coefficient and weighted kappa.

The agreement index I is probability value of same interpretation between two experts of all events as seen in equation (1), where N is the number of all events and n_{ij} is the number of same interpretation of each category. Although the measurement is simple to be implemented, the limitation of this measurement is there is not any consideration of disagreement number.

$$I = \frac{\sum_{i=j}^{k} n_{ij}}{N} = \sum_{i=j}^{k} p_{ij} \qquad (1)$$

The difference of the within-one agreement index with agreement index is the within-one agreement index adopts the agreement of interpretation which have single differentiate category. The formulation of the within-one agreement index is shown in equation (2).

$$WI = \frac{\sum^{k}{}_{\substack{i=j \\ i=j\pm 1}} n_{ij}}{N} = \sum^{k}{}_{\substack{i=j \\ i=j\pm 1}} p_{ij} \qquad (2)$$

The third of agreement measurement is kappa measurement which was proposed by Cohen [4]. Kappa is calculated by equation (3) where p_o is probability of agreement observed, pc is probability agreement expected which is obtained by summing product of marginal probability of agreement.

$$k = \frac{p_o - p_c}{1 - p_c} \qquad (3)$$

$$p_c = \sum_{i=j}^{k} p_i \cdot p_j \qquad (4)$$

The weighted kappa was proposed to overcome the disadvantage of kappa measurement that the kappa is taking much consideration of disagreement. The weighted kappa [4] is shown in equation (5).

$$k_w = 1a \frac{\sum_{i=1}^{k} \sum_{j=1} v_{ij} P_{ij}}{\sum_{i=1}^{k} \sum_{j=1} v_{ij} P_i P_j} \qquad (5)$$

Association measurements compute the degree of linier association between system and human expert. The measurements that belong to the category are Kendall's tau and Spearman's rho. Kendall's tau is utilized by calculating Equation (6), where C is the number of concordant observation, D is the number of discordant observations and N is the total number of events. Concordant observation occurs when $(x_i-x_j)(y_i-y_j)>0$ and discordant appears when $(x_i-x_j)(y_i-y_j)<0$ where (x_i,y_i) and (x_j,y_j) are pairs of observations.

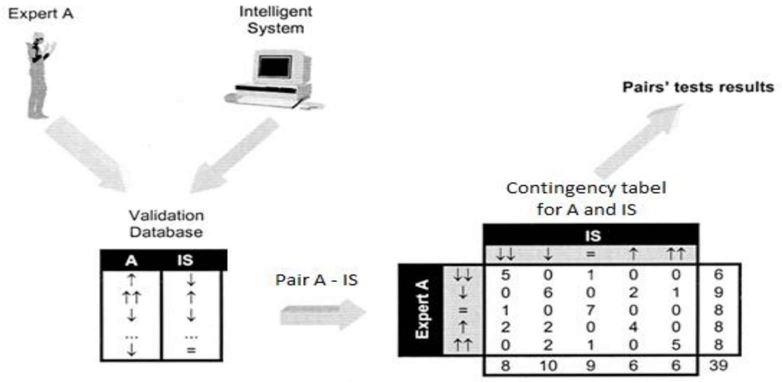
To overcome the problem of tied pairs when $(x_i-x_j)(y_i-y_j)=0$, Kendall's tau-b was proposed as seen in Equation (7). The modified part is the denominator where U represents the tied pairs in x and V is the tied pairs in y.

Another association measure which is more popular is Spearman's rho as seen in equation (8). R and S are ranks that is obtained by converting the pair of values(x,y) to be pair of ranks (R,S).

$$\tau = \frac{C-D}{n(n-1)/2} \tag{6}$$

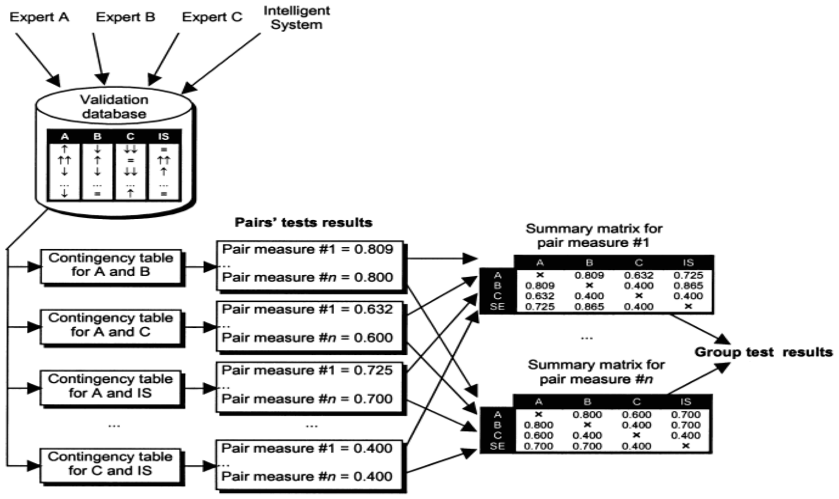
$$\tau_k = \frac{C-D}{\sqrt{\left[\frac{n(n-1)}{2}-U\right]\left[\frac{n(n-1)}{2}-V\right]}} \tag{7}$$

$$r_s = \frac{\sum_{i=1}^{n}(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i-\bar{R})^2\sum_{i=1}^{n}(S_i-\bar{S})^2}} \tag{8}$$



	↓↓	↓	=	↑	↑↑	
↓↓	5	0	1	0	0	6
↓	0	6	0	2	1	9
=	1	0	7	0	0	8
↑	2	2	0	4	0	8
↑↑	0	2	1	0	5	8
	8	10	9	6	6	39

Fig 2. Process of pair method.



Summary matrix for pair measure #1

	A	B	C	IS
A	×	0.809	0.632	0.725
B	0.809	×	0.400	0.865
C	0.632	0.400	×	0.400
SE	0.725	0.865	0.400	×

Summary matrix for pair measure #*n*

	A	B	C	IS
A	×	0.800	0.600	0.700
B	0.800	×	0.400	0.700
C	0.600	0.400	×	0.400
SE	0.700	0.700	0.400	×

Fig 3. Process of group method [5].

If the number of expert is large, the group measurement is needed. It is because of difficulties to take true interpretation of many result of pairs measurement between each expert. The result of pair measurement is used as input for group measurement. Some of group measurements are the William's index, cluster analysis, multidimensional scaling (MDS) and dispersion and bias measurements. Pair and group measurement is employed when the standard reference does not exist, therefore the result of the system and the expert interpretation should be compared.

Williams measurement is done to determine whether an isolated expert agrees with the expert in a concrete reference group where the expert group agree among themselves [5]. The Williams measurement is assessed by Equation (9) where Po represents the agreement between an isolated expert and a group of expert, Pn represents the agreement in the internal group, n is the number of reference expert and P(a,b) is pair measurements that interprets agreement in the internal group, n is the number of reference expert and P(a,b) is pair measurements that interprets the agreement of expert a and b.

$$I_w = \frac{P_0}{P_n} \qquad (9)$$

To test the system, the isolated expert is substituted by the system. If I_w value is greater than 1, it represents that the system has a high credibility because it has large agreement with the expert. The Williams measurement needs high value of agreement within the reference group to obtain the proper result of validation system.

There are 2 types of cluster analysis, hierarchical and non-hierarchical. The hierarchical analysis uses matrix of distance of each members. Distance can be calculated using agreement index or agreement percentage between experts. The output of hierarchical cluster analysis is hierarchical tree or dendogram that joins up the different experts depending on their similarity of their interpretations. Nonhierarchical analysis builds a division to minimize the sum of square distance between each point to their centroid.

There is no numerical value as the output that it may analyzed directly by its value. Although the analyzes is quite more complex than the other method, the cluster analysis has the following advantages: it is fast in the quite number of expert, it provides an overall view of the agreement between expert and it may cluster the expert to some group based n their similarity interpretations.

Multidimensional Scaling (MDS) is a data analysis method that represents the similarity between different experts by geometric space. The MDS is form based on eigen-values and eigen-vectors of a distance matrix.

The MDS is built by following steps: (a).Convert initial pairs matrix into dissimilarities matrix (D). (b).Build a semi-defined positive matrix A based on dissimilarities matrix (D). (c).Gain the co-ordinates for each element from eigen-values and eigen-vectors of semi-defined positive matrix (A)

The dispersion method is used to measure the dispersion of particular expert result and the rest of expert result. The dispersion is computed using Equation (10) where nc represents the number of cases, ne is the number of expert, and Dij is the ranking order made by expert i to the case j.

$$Disp = \frac{1}{n_c}\sum_{j=1}^{n_c}\left(\sqrt{\frac{1}{n_c-1}\sum_{i=1}^{n_c}\left(D_{kj} - D_{ij}\right)^2}\right) \qquad (10)$$

Bias is used to compare the magnitude of the result of particular expert and the rest of expert. Bias is gained by calculating equation (11).

$$Bias = \frac{1}{n_c}\sum_{j=1}^{n_c}\left(\frac{1}{n_c-1}\sum_{i=1}^{n_c}\left(D_{kj} - D_{ij}\right)\right) \qquad (11)$$

When the standard reference exists, we use agreement ratios measurement and Jaccard's coefficient. Agreement ratios measure the agreement between an intelligent system and a standard reference. The standard reference may be obtained from consensus between experts or actual solution which has known. The agreement ratios are calculated by constructing the contingency table between standard reference and expert result then achieve the similarity measure such as agreement index and Jaccard's coefficient as seen in equation (12).

$$Jc = \frac{TruePositive\ +TrueNegative}{Numberofcases} \qquad (12)$$

Others validation criteria are specificity, sensitivity, predictability which used in [2] and Youden's index [6]. The formulas of the five measurements are explained in equation (13), (14) and (15).

$$Sensitivity = \frac{\sum Truepositive}{\sum Truepositive\ + \sum Truenegative} \qquad (13)$$

$$Specificity = \frac{\sum Truenegative}{\sum Truenegative\ + \sum Falsepositive} \qquad (14)$$

$$PPV = \frac{\sum Truepositive}{\sum Truepositive\ + \sum Falsepostive} \qquad (15)$$

$$NPV = \frac{\sum Truenegative}{\sum Truenegative\ + \sum Falsenegative} \qquad (16)$$

$$Youden's\ index = sensitivity + specifity — 1 \qquad (17)$$

TABLE I
EVALUATION OF VALIDATIONCRITERIA

Validation Criteria	Ordinal/ Nominal	Number of category	Number of expert	Consensus of expert/ Standard reference
Agreement index	Nominal	>1	1	v
Within-one agreement index	Both	>1	1	-
Kappa	Both	>1	1	-
Weighted-kappa	Both	>1	1	-
Kendall's tau	Ordinal	>1	1	-
Spearman's rho	Ordinal	>1	1	-
William's index	Nominal	>1	>1	-
Cluster analysis	Both	>1	>1	-
MDS	Both	>1	>1	-
Dispersion & bias measurements	Ordinal	>1	>1	-
Jaccard's coefficient	Both	2	-	v
Sensitivity	Ordinal	2	1	v
Specificity	Ordinal	2	1	v
Predictability (PPV and NPV)	Ordinal	2	1	v
Youden's index	Ordinal	2	1	v

TABLE II
THE ADVANTAGES AND DISADVANTAGES OF VALIDATIONCRITERIA

Validation criteria	Advantages	Disadvantages
Agreement index	Simple to interpret	Not considering the disagreement
Within-one agreement index	Permit to analyzed tendency of system	Not considering the disagreement,
Kappa	Considering the disagreement	Treat the disagreement in same way
Weighted-kappa	Weighting the disagreement	Difficult to assign weight for nominal scale
Kendall's tau	Simple to interpret	Not considering the tied-pair
Spearman's rho	Considering the tied-pair	Difficult to interpret
William's index	Simple to interpret	Should ensure the expert in same decision
Cluster analysis	Give overall view of similarities on expert, permits the division of expert	Not considering the reverse process, give an error similarity when the number of group is large
MDS	Give exact similarities on expert	Give an error when number of expert is few
Dispersion & bias measurements	Not based on pair test	-
Jaccard's coefficient	Simple to interpret	Should not be used for more than 2 categories
Sensitivity		
Specificity		
Predictability (PPV and NPV)		
Youden's index		

3. Results and Analysis

We constructed table I that describes the validation criteria relates to the characteristic of the system output. The advantages and disadvantage of them is described briefly in table II. Agreement index, within-one agreement index, Kendall's tau and Spearman's rho measure only ordinal output. It is done because it only considers agreement without taking account how much the disagreement as well as the other number, nominal scale.

Others nominal scale validation criteria are Kappa, Weighted-kappa, William's index, Cluster analysis, and MDS. Although all criteria of validation can measure two or more categories, the best use of them is quite different, especially for Sensitivity, Specificity, Predictability, and Youden's index. It is not recommended to measure the criteria in more than two categories, absent and present. The four validation criteria are only considering the agreement and disagreement in binary value.

The number of experts which is explained in the table cannot be implemented directly. For example on Kappa criteria, the table describes the number of experts is only one expert. It does not mean that if there are 2 experts or more, the criteria may not be used. This Kappa is used by take the average of Kappa value of each expert.

The existing of consensus of experts or standard reference may be interpreted as one expert but the consensus of experts or standards reference is more trustable than one expert. So, it allows for implementing the validation criteria which has one expert to the problem when the consensus of expert or standard reference exists but not allows in others direction. It is not recommended to apply the criteria which standard reference required to the validation of problem in one expert. The disadvantages column of table I explained the reason that must be considered before using the criteria of validation. The disadvantages occur according the case of validation in intelligent system. If the validation is held to know the percentage of agreement so it does not matter to use agreement index and etc. When the validation is done to obtain how much the credibility and un-credibility of the system so the agreement index may not be used and be better to use weighted-kappa.

4. Conclusion

The explanation of table I is a short introduction of characteristic of the validation criteria that may be used as the guidance to determine the criteria of validation. The best validation is held by involving more than 1 experts or using consensus of experts or standard reference. The type of intelligent system output is ordinal scale or nominal scale or both of them. It should be considered well so the validation will not make wrong conclusion. It is recommended to use the validation criteria which taking account on the disagreement. As the un-trustable and un-credibility of the system is one point important.

All of the validation criteria above can be modified so it may be used to validate the system. How to modify the criteria is adjusted according the system and the parameter to be gained from the validation. Based on table 1 and table 2, user may carry on the right validation using the best validation criteria. So, the value of validation criteria is able to represent the credibility and range of acceptable of the system.

In this paper, we have presented the evaluation of validation criteria which used in intelligent system commonly. We have evaluated the characteristic by paper review and research. The results showed us that it is not an easy task to determine the best validation criteria to use in validation directly because there is none of the validation criteria which is suitable in all characteristic of intelligent system as well as in the purposes of validation.

But, there are some treat to validate the result of our discussion. First, the sources of literature that are used to evaluate the techniques are mainly from the published research papers, especially from the international journals and or the conference proceedings. The literatures usually contain brief information which is some other information probably were disappeared related to the long version one. Therefore the justifications of review are made from the concise information. Justifications are performed without any formal methodology. We use our comprehension from reviewing the papers and concluding the result based on our understanding. However, the initial result presented in the evaluation can be very useful to perform further and deeper evaluation of the subject for future improvement, and also to welcome any open discussions.

References

[1] P. Jaccard,"Étude comparative de la distribution floraledansune portion des Alpes et des Jura," *Bulletin de la SociétéVaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901

[2] K. Becker, B. Thull, H.Kasmacher-Leidinger, J. Stemmer, G. Rau, G. Kalff, &H. Zimmermann "*Design and validation of an intelligent patient monitoring and alarm based on a fuzzy logic process model*," Artificial Intelligence in Medicine. 1997

[3] Aguilar R. M., V. Munoz, M. Noda, A. Bruno,&L. Moreno, "*Verification and validation of an intelligent tutorial system*," Expert System with Applications. 2008

[4] J. Cohen, "*A coefficient of agreement for nominal scales*," Educational and Psychological Measurements. 1960

[5] E. Mosqueira-Rey, & V. Moret-Bonillo, "*Intelligent interpretation of validation data*," Expert System with Applications. 2002

[6] O'Keefe, Robert, O. Balci, &Smith, Eric P. "*Validation of expert system performance*," IEEE Expert, 2(4) 81-89