Review of Rule Quality Measurement: Metrics and Rule Evaluation Models

Munirah Muslim[1,3]*, E. Winarko[2]

[1]* Computer Science Department, Gadjah Mada University of Yogyakarta, Indonesia
[2] Computer Science Department, Gadjah Mada University of Yogyakarta, Indonesia
[3] Informatics Engineering Department, Muhammadiyah University of Ponorogo, Indonesia

e-mail: munirah.mt@gmail.com, ewinarko@mail.ugm.ac.id

*Corresponding Author: munirah.mt@gmail.com

Available online at: http://www.ijcert.org

Abstract: - A rule-based system is a system based on the set of rules used to make inference knowledge. The system gathers knowledge into the representation of knowledge in the form of a rule. However, the knowledge in the form of the rule is inductive, meaning that the algorithm can construct the rule by studying a limited number of cases and then the induced rule of a limited number of cases and then generalize it to the general reality from time to time. This, of course, has the degree of inaccuracy in expressing reality into knowledge, or an experienced expert builds it but it is not absolute that the knowledge it possesses is 100% accurate or always consistently true from one time-space location to another time-space location. Therefore, the need for a formula that can measure the quality of the resulting rule and assess the consistency of the rule. In this study, we did a review of the ideas of people trying to measure knowledge built inductively by either the algorithm or the experts. These measurements are based on several parameters defined by them according to the underlying assumptions. This review seeks to partially present how ideas to measure the rule as knowledge representation from a varied viewpoint and how people construct evaluation models to assess the resulting regulations either from the experts or human experts as well as those resulting from the induction rule algorithm much developed.

Keywords: Review, Quality, Measurement, Metric, Rule, Evaluation, Model.

----

1. Introduction

A rule is a form of knowledge representation packed into a knowledge-based system. Knowledge representation is a factor that plays a significant role in the artificial intelligence especially rule-based system (rule-based system). The key to rule-based system success is solving problems with existing knowledge covered in the knowledge base. This knowledge base contains the knowledge gained from transforming human expertise into computer-compliant forms and computer systems. Troubleshooting in a domain requires a collection of knowledge of objects or particles or term affiliated with the domain. On a rule-based system, each rule is organized as follows:

IF A, B and C THEN D

Example rule above consist of antecedent part (condition) that is A, B, C and D is the consequence (conclusion). The facts of the constituent or consequent composer can be either a single fact or a combination of several effects at once connected by logic connectors such as AND and OR. As an attempt to translate and model human knowledge into machine-processable form, it takes a form that is flexible enough to represent knowledge (rules) as well as simple enough to be implemented into a programming language.

The use of the IF-THEN model is considered adequate by associating facts with premises. Thus, in general, the term premise is used to represent facts in its representation.

Evaluation on the rule is done to see or observe whether or not the changes in the condition of the rule to the state of data or facts that occur, along with the development of life and changing lifestyle. Suppose the diagnosis of the case of X disease, the rule that was taken in a few years ago to be used to diagnose an X disease only 100 rules, could be changed if paired with the data that occurs today.

In some studies, many modeling and use of several metrics to evaluate the rules induced by the data using mathematical calculation approach both statistically and non-statistically. In this study, we will discuss some calculation models to measure and evaluate the rules based on the results of previous research.

In this paper also provides an overview of the measurement and method of evaluation of the rule commonly used as a search heuristics in learning rules. The behavior of various heuristics in the variation of research is analyzed by visualizing their dynamics in the scope of space based on the way research is conducted. Several variants of the research results related to the measurement model and the evaluation model of the widely used rule will be reviewed.

2. Review Method

In conducting this study, the methods undertaken are the searching of the literature and references related to the study discussed. Some references are mapped into the picture as follows (Figure 1):
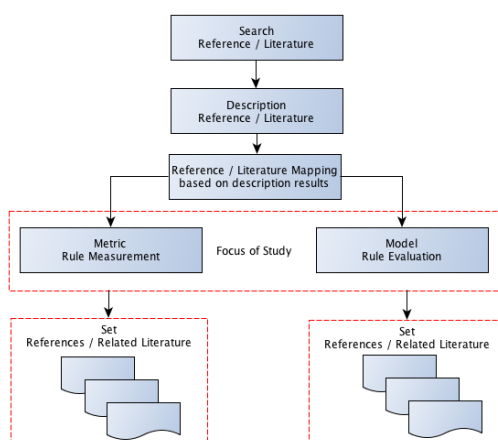

Figure 1. Mapping of review method

The focus of the study in this paper is the metric of rule measurement and model of rule evaluation based on fig.1.

There is 32 references/literature obtained based on the focus to be studied, namely:
a) Focus on the review of measurement metrics rule; will review how a or a set of metrics is obtained, from the basic philosophy to the form of the equation.
b) Focus on study evaluation model of the rule; will review how to build a model for evaluation of the rule.

The following sessions will explain the results of each paper search.

3. Rule Quality Measurement Metrics

The quality of the rule is one thing that many do research. The definition of quality determination also varies according to the viewpoint examined. Many metrics with various approaches have been found to determine the quality measure of the rule. Some of these approaches, among others, approach the theory of probability statistics, as well as from information theory.

A. Information Theory Approach

Viewed from the information theory approach, many ideas of methods and metrics are made, starting from Smith & Goodman's idea (1992) by constructing J-Measure metrics and algorithms that support those metrics called ITRule [1]. This research proposes a way to measure the rule using the information content brought by a rule. It sees that each rule is essentially visible as an expression of the information flow channel. Suppose the rule if X then Y, this rule can be seen as a transmission of information from transmitter X to receiver Y, which then generalizes a metric that measures the average entropy received on the Y side rather than using an entropy difference that occurs before, and the information is sent. The metrics proposed in this paper are called J-measure which is a generalization of the J-measure (average entropy of information occurring at the receiving end). Using the J-measure metric, this paper builds an algorithm named ITRULE that uses the metric to perform rule selection from any randomly generated rule, then selects it as a way to generate rules from the data. The built metric is called the J-measure metric, which measures the information content of a rule, expressed by the following formula:

$$(X:Y = y) = p(y).j(X:Y = y), where:$$
$$(JC:Y = y) = \sum_x p(x|y).\log\frac{p(x|y)}{p(x)} \qquad (1)$$

This metric (1) is then widely used by researchers to evaluate the rule by classifying the rules, such as searching

for rules against interesting rules, said Interestingness Rule based on the character of rule and evaluation of the rule by classifying and ranking rules based on the hypothesis or the discovery of the rule using artificial intelligent systems [2],[3]. The method of the proposed metrics is then reviewed into a survey-based study of research results. In a research survey conducted by JF., Roddick, aims to explore the latest issues that exist on the topic of research-based data mining, either from the side of the paradigm or perspective or the side of the method used [4]. While in the research survey conducted by F., Provost attempted to provide information on what methods and algorithms are already available regarding improving the results of induction rule in data mining techniques [5].

B. The Probability & Bayes Statistical Probability Approach

From the results of the survey, some researchers conducted experiments for development, including the invention of methods for automated learning using a learning machine using genetic algorithms (GA) and Bayesian Network (BN) to search for probability of searching a randomly populated population and can be run when the knowledge domain is not available, called the FSS-EBNA method (Feature Subset Selection by Estimation of Bayesian Network Algorithm)[6]. Furthermore, a new method of classification is made using probability trees (PETs = Probability Estimation Trees)[7]. The idea of this method basically has the same features as the classification tree in general (e.g. the ability to understand, accuracy and efficiency in high dimensions and large data sets), is given little development by using the simple common smoothing method as well as the uniform Laplace correction so can substantially increase ranking results and more efficiently because it can improve the accuracy. Then the results of the experiments were made comparative studies by Sulzman et al. in 2009 to obtain a good class probability estimate to study predictable rules with a more accurate level [8].

Meanwhile, A., Freitas and P., Flach et al. did the same in doing their research on measurement rules. Freitas emphasizes the measurement of several factors that influence the evaluation of the level of Interestingness Rule found by data mining algorithms. Interestingness Rule is a degree or measure of confidence in each rule measured by the factors that influence it. The novelty of this measurement method is the addition of a new criterion called *Surprisingness Attribute* as one of the factors influencing the interest of the found rule. This study proposes a new metric for measuring the rule. This new metric is a combination of several previous metrics built on metrics that measure the degree of rule interestingness, then in this study proposes a metric that measures the rule of surprisingness. Metric

groups that measure rule interestingness are metrics that measure disjunct size, imbalance of class distributions, attribute interestingness, misclassification costs, the asymmetry of classification rules, coverage, completeness and confidence [9]. The proposed metrics are:

$$(|A\&B|) - \left(\frac{|A||B|}{N}\right) * AttUsef * MisclasCost, \; where:$$
$$MisclasCost = 1/\sum_{j=1}^{k} Prob(j)Cost(i,j)$$
$$(2)$$

From the results of a survey written by K., McGarry can be seen below (Figure 2):
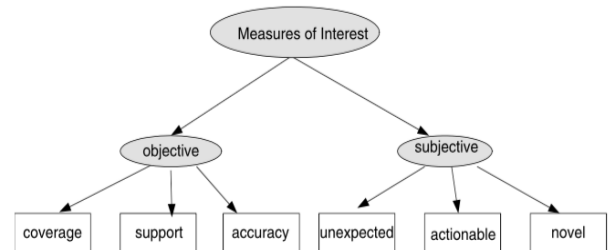

Figure 2. Taxonomic measurement of interestingness rule

There are two main groups of interestingness rule measurement models, namely the arrangement of objective and subjective groups as in Fig.2 above shows. Objective criteria such as coverage or content rules, completeness or support of the rule and also the accuracy of the rules that are considered attractive. As for subjective criteria in the form of unexpected patterns of rules, the ability of the rule and novelty of the rule in building designs [10].

One of the implementations of the interestingness rule measurements for objective group arrangements is found in the experimental experiments conducted by X..Huynh, F.,Guillet, J.,Blanchard, and P.,Kuntz presenting a new approach implemented by the new tool, ARQAT, to make comparisons. This approach is based on a correlation graph analysis that presents an objective grouping of objective interestingness rules using the association rule process. This graphical clustering approach is used to compare and discuss the behavior of thirty-six striking steps on two high-correlated and low correlated prototypical and contrast datasets [11].

While the study conducted by B., Vaillant, et al. is one example of the implementation of the interestingness rule measurement for subjective group arrangements, in which a statistically approach is introduced and used to link the behavior of existing criteria by applying a minimum threshold of confidence to datasets prepared in this experiment [12].

The research related to the measurement of this interestingness rule was done by Jon Hill et al. by building a set of rules that are classified according to the consequence of each rule measured by the criterion of confidence level by conducting a proof test against thirteen metric measurements of interestingness rule available, there are metrics of confidence, satisfaction, ohsaki's conviction, added value, interest/lift/strength, brin's conviction, certainty factor/loevinger, mutual information, interestingness, sebag-schonauer, ganascia index, odd multiplier, and counter-example rate. From the results of the measurement, experiments can be further simplified by the data mining process to make it easier to find the rule that has the optimal value for each class size [13].

While P., Flach, et al. do the same thing only more commonly, it means for all rules, not just for the purpose of measuring the interestingness rule. This research proposes a measurement framework that unifies predictive (category or classification) rule induction and descriptive rule induction (just constructing combinations of attributes conjunctionally on the antecedents section and possibly conjunction combinations consequently without categorizing objectives). The resulting framework is a set of metrics built on a contingent table. The contingent table is obtained from frequency count to dataset. Then, based on the frequency count in the contingency table, the probabilities constructing the metrics are calculated [14].

Philosophically, any rule is modeled as:

$$\text{Head} \neg \text{Body}$$
$$H \leftarrow B$$

Furthermore, this study produces a framework that contains a collection of metrics generated to measure the rule, there are:

$$R\ Acc(H \leftarrow B) = p(H|B) - p(H)$$
$$R\ NegRel(H \leftarrow B) = p(\overline{H}|\overline{B}) - p(\overline{H})$$
$$R\ Sens(H \leftarrow B) = p(B|H) - p(B)$$
$$R\ Spec(H \leftarrow B) = p(\overline{B}|\overline{H}) - p(\overline{B})$$
$$(3)$$

From the above mentioned of metrics framework, further research by F., Johannes et al. by analyzing and evaluating some of the measurement rule metrics to find the optimal point in performing the best rule measurements by not requiring the large effort. There are several metric frameworks that are analyzed in different angles, resulting in a measurement balance with little difference from the weighting side of the predefined rule classification [15].

In a slightly different viewpoint, this study was conducted at 1999 by C., David and renewed at 2016 by G., Salvatore et al. which explores all measures that measure a premise confirmation of a hypothesis or conclusion in a rule

[16]. Beginning the formulation of confirmation of rules from C., David as follows:

$$S(E,H) = \frac{pr\left(\frac{H}{E}\right) - pr(H)}{pr(\sim E)} \tag{4}$$

From the above metrics, the novelty is present in 4 measurement perspectives that measure confirmation in a rule, i.e., Bayesian perspective, strong Bayesian perspective, likelihood, strong likelihood. The four formulations of the confirmation are as follows:

(i)   *Bayesian Confirmation* $(\Pr(H|E) > \Pr(H))$,
(ii) *Strong Bayesian Confirmation* $(\Pr(H|E) > \Pr(H|\neg E))$,
(iii) *Likelihoodist Confirmation* $(\Pr(E|H) > \Pr(E))$,
(iv) *Strong Likelihoodist Confirmation* $(\Pr(E|H) > \Pr(E|\neg H))$.

These four confirmations are then shown to have logical equivalence using the ad-bc term, where a, d, b, and c are each probability values in the contingency table. The researcher then proposed a new measurement condition which was a generalization of the four confirmation perspectives. These four confirmations are each measurement method for measuring the monotony of rules in 4 different perspectives. Together with several measurements made in previous research, researchers explored the monotony and symmetry properties of the metrics [17].

Further research is done by M.,Michalak, et al, which also builds measurement metrics against rule quality induced by algorithms [18]. The measurement metrics selected and used in this experiment are as follows:

$$g = \frac{p}{p + n + 2}$$
$$wLap = \frac{(p+1)(P+N)}{(p+n+2)P}$$
$$LS = \frac{pN}{nP}$$
$$Rss = \frac{p}{P} - \frac{n}{N}$$
$$MS = \frac{p}{n+P}$$
$$C1 = Coleman.\frac{2 + Cohen}{3}$$
$$C2 = Coleman.\,0.5(1 + \frac{P}{P})$$
$$Corr = \frac{pN - Pn}{\sqrt{PN(p+n)(P-p+N-n)}}$$
$$s = \frac{p}{p+n} - \frac{P-p}{P-p+N-n}$$

$$Cohen = \frac{(P + N)\left(\frac{p}{p+n}\right) - P}{\frac{P + N}{2} \frac{p + n + P}{p + n} - P}$$

$$Coleman = \frac{(P+N)\frac{p}{p+n} - P}{N} \quad\quad (5)$$

This study focuses on studying the effect size on conflict resolution on the rule during the classification process. Research conducted by F., Johannes et al. and P., Flach et al. focuses on determining quality measures in probability approaches. There are 30 rules analyzed concerning efficiency, while for checking the effectiveness of each rule, the size used in the same induction algorithm used in each induction stage and during the classification conflict resolution. Then the classification capability of the classifier of rule obtained will be checked. Classification skills are analyzed for overall classification accuracy. This measurement is suitable for verifying the classifier working on unbalanced data. In 2004, P., Flach, et al. researched the implementation of measurement metrics that had been built into a decision support system model [19].

At 2012, the metric of the rule's measurement of consistency and certainty in the classification of some rules has also been reviewed by DMW., Powers. The measurement model is built on a statistical approach where renewal of the F-Measure is carried out in the normalization stage with the arithmetic mean of bias and prevalence, thereby obtaining the mean value of the recall and precision. Thus, his contribution results in a relationship expressed as follows: AUC (Area Under the Curve) = Consistency + Certainty [20].

There is also the acquisition of a new framework-based measurement method conducted by P. Salgado. In the research, a new method is produced to model the structure of the rule set obtained into the HPS model to manage the information that occurs in a rule, which aims to reduce the number of rules that have small match value with other rules [21].

C. Summary of Rule Quality Measurement Metrics Review

From various previous reviews, it can be formulated some results of reviews related to metric variants for the measurement of rules that have resulted from several studies, a list of variants of the results of metrics reviews can be seen in table 1 as follows:

Table 1. Rule Quality Measurement Metrics

Reference	Metrics Definition	Metrics Aim
[1]	*J-Measure*	Measure the average
		entropy of information in a rule using IT-Rule (the rule selection algorithm)
[16]	*Measuring Confirmation*	Measuring the potential probability of conclusions from the rule using the Bayes method.
[9]	*Measuring Interestingness Rule*	Measure the interestingness rule on several factors that influence using the probability approach.
[14]	*Measuring Rule Evaluation*	Measure the performance of rule using Bayes probability approach
[21]	*SLIM (Separation of Linguistic Information Methodology)*	Modeling the rule structure into the HPS structure (Hierarchy Priority Structure) to organize the information that occurs in the rule so it can be possible to reduce the number of rules that are judged to have a low relevance value to other rules.
[7]	*PETs (Probability Estimation Trees)*	Measure the ranking of ranking results against classified induction rules using updated decision tree concepts.
[13]	*Interestingness Measures for Fixed Consequent Rules*	Measure the consistent performance of a predefined rule of some interestingness rule.
[20]	*ROC-ConCert*	Measure the performance of the rule against consistency and certainty in the classification of some rules using the method of F-Measure and ROC (Relative Operating Characteristic Curve).

4. Models Of Rule Evaluation

In recent years, large amounts of data are stored in information systems in the fields of science, social science, and business domains. People have been able to gain valuable knowledge because of the development of information technology. Also, data mining techniques combine different types of technologies such as database

technology, statistical methods, and machine learning methods. Then, data mining has been renowned for utilizing data stored on database systems. In particular, the IF-Then rule, generated by the induction rule algorithm, is considered one of the most useful and readable data mining outputs. However, for large datasets with hundreds of attributes including noise, the process often gets a lot of thousands of rules. From such great rules, it is difficult for human experts or experts to find valuable knowledge that can be obtained in the tens, hundreds or even thousands of rules generated by the induction rule algorithm. To support the selection of the rule, many attempts are made using objective rules based on rule evaluation metrics such as recall, precision, and other interesting measurements as discussed in previous sessions.

A. Constructing Rule Evaluation Models

In constructing an evaluation model, it is usually based on an assumption that the model will produce results according to the expected assumption. Call it the model of rule evaluation constructed by F., Coenen & P., Leng. Model evaluation of the rule that is built in the selection techniques used to find the best number of rules in the set of classification rules that are formed [22]. The datamining technique used is CSA (Support, Confidence, Size of Antecedent), while for selection techniques use measurement metrics Weighted Relative Accuracy (metrics formed by P., Flach, et al. and Laplace Accuracy metrics (metrics formed by I., Inza, et al).

Also, the evaluation model constructed by Y., Yao & B., Zhou is focused on linking two types of evaluation models called micro and macro evaluations. Micro evaluation is based on a single rule that can be measured by common empirical measures while the macro evaluation is based on a collection of interdependent rules in its set, here different resolutions can be applied [23]. This research applies many of the metrics formulated by P.,Flach, et al and A., Freitas to perform calculations in the evaluation model built. Similarly, on research conducted by DMW., Powers [24].

The next evaluation model was published in 2007 by H., Abe, et al. This research proposes an evaluation model based on post-processing mining of the rule [25]. In this post-processing concept, mined rules are evaluated using metrics which they refer to as objective indices, i.e., a number of metric formulas that are made objectively rather than subject judgments. Then the results of this evaluation build a meta-data containing the rule-rule of mining and metric measurement results. Each rule then gets a direct assessment (without metrics) by the expert on purely subjective opinion. Above this meta-data, an algorithm is constructed that performs meta-learning that does attribute the value of objective indices and then predicts the expert's subjective values. This algorithm then becomes a machine that learns to evaluate rules as the experts evaluate the rule directly. The evaluation model is divided into two stages, the training stage and the prediction stage. It is expected that the presence of this algorithm provides the rule evaluation speed and decrease the cost of evaluation rule by the expert significantly. Then this evaluation model was also applied to follow-up research also by H., Abe, et al in 2008 [26],[27], [28].

The construction of an evaluation model built by H., Abe et al. were developed in several study research by A., Gruca, A.,&M., Sikora. This study research focuses on a selection of rules by using the UTA method to perform rule selection. UTA method uses Q-Uta measure which is a multi-criterion of interestingness. Then, based on the multi-criterion assessment along with the expert preference assessment in order of importance of the selected rule, a meta-dataset is created, i.e. a dataset that describes the rules to be selected in Q-Uta measure and expert assessment or preference. This method is described first with the dataset of gene ontology, then from this dataset is used RuleGO method to generate all possible rules as a representation of genes. Then each rule is presented to experts to be selected and sorted according to expert preference. From each selection and expert sorting, measurements were made using Q-Uta measure which resulted in a multi-criteria assessment of each rule. This multi-criteria assessment illustrates the selection and sorting of expert preferences. Then from this Quota measurement that represents the expert preference, the UTA method estimates and performs sorting or filtering that is consistent with the expert's preference [29],[30].

The evaluation model constructed by A., Gruca, A.,&M., Sikora is implemented by U., Stanczyk in 2016. The results of research focus on the need to provide a decision from the selection of rules to the pruning rule process in order to find the optimal solution when in the domain conditions knowledge is inadequate and limit the learning data and or regulate the induction of rules that only attract the rules of the most interesting course (interestingness rule) [31].

Also, there is a construction of the model evaluation model conducted in 2014 by KK., Sethi, et al by proposing a way to select the learning algorithm for a given dataset [32]. In the formulation of the problem, if a given dataset then the best learning algorithm that can be used to perform data mining (extraction rule) on it. The way proposed in this study is to first for each dataset built metadata from the dataset. The study defines some meta-data attributes that can be used to express the characteristics of a dataset. For example, a dataset is represented by metadata attributes such as the number of attributes, number of instances, number of categorization classes, number of categorizing symbols, number of continuous attributes, an indication of data availability, rough indication of the dimensionality of the problem, and so on. Then from the values of these metadata attributes, they are used as inputs for a meta-learning algorithm that maps metadata values to appropriate learning algorithms. Meta-learning learns to map a dataset to the best learning algorithm by comparing the model of a learning algorithm with a model compiled by another algorithm. The model in question here is the set of rules generated by a learning algorithm, a set of rules representing knowledge that can be derived from a dataset. In general, this meta-learning evaluates the model generated by a learning algorithm compared to the meta data dataset being extracted rule. The idea of this study can be stated as follows:

Dataset → *Metadata* → *Meta-learning* → *appropriate algorithm.*

B. Summary of Rule Evaluation Models Review

Several studies have been conducted to produce a good model in evaluating the rules in different perspectives and needs of each researcher; the model variation can be seen in table 2 below:

Table 2. Rule Evaluation Model

Refe-France	Models Definition	Models Aim
[25]	*Evaluating for Selection Learning Algorithm Rule*	The construction of an evaluation model to evaluate the results of the rule assessment based on the metrics and the subjective judgment of the expert.
[23]	*Micro and Macro Evaluation of Classification Rules*	Construction of the evaluation model by specifying the classification of rules on the micro and macro set based on each indicator inside.
[29]	*The Multicriteria Rule*	Construction of evaluation model for selection of interestingness rule on several
	Interestingness Measure by The UTA Method	criteria according to measurement requirement.
[32]	*KD Resemblance Score Algorithm To Evaluate Consistency*	Construction of evaluation models to select suitable and consistent learning algorithms from the provided dataset.

5. Conclusion and Future Scope

From the various reviews of the results of research both related to the results of the metric of measuring the quality of the rule as well as related to the construction of the model of evaluation of the rule, there are many ways with the use of other methods that can be done to do both things. Comparative study is needed in constructing which evaluation model to use and which metrics to use, it all depends on what fact or condition we will do. Acquiring knowledge from the data is the work of experts in the field, but every expert is a human who has limitations related to his basic nature as a human being. The most common phenomenon is that if the acquisition of knowledge submitted to an expert is a bottleneck phenomenon, which an expert can produce only a small amount of knowledge over a limited time interval and a matter of accuracy that sometimes accompanies the extraction of an expert's rule.

But for machines, by using appropriate datamining algorithms, a machine can perform fast rule acquisition, careful evaluation in a very short time and in a very large amount of knowledge for a limited time. Smart machines in the future are expected to construct their science, in a way that is not trapped in the phenomenon of the bottleneck. Therefore, the selection of methods and models should be adjusted to the conditions, because not all models or methods can be appropriate to use, this will impact on the results of experiments that will be obtained later.

The results of this review-based study can also be used as a reference to the need of how to measure a large-scale data in a particular data mining technique, which of course is implemented into rules-based [33].
.

6. References

[1] P. Smyth and R. M. Goodman, "A n Information Theoretic Approach to Rule Induction from Databases," vol. 4, no. 4, pp. 301–316, 1992.

[2] R. Shinghal, "Evaluating the Interestingness of Characteristic Rules," pp. 263–266, 1996.

[3] S. Dreiseitl, M. Oil, C. Baumgartner, and S. Viterbo, "An evaluation of heuristics for rule ranking," Artif. Intell. Med., vol. 50, no. 3, pp. 175–180, 2010.

[4] J. F. Roddick and M. Spiliopoulou, "A survey of temporal knowledge discovery paradigms and methods," IEEE Trans. Knowl. Data Eng., vol. 14, no. 4, pp. 750–767, 2002.

[5] F. Provost, "A Survey of Methods for Scaling Up Inductive Algorithms," vol. 169, pp. 131–169, 1999.

[6] I. Inza, P. Larrañaga, R. Etxeberria, and B. Sierra, "Feature Subset Selection by Bayesian network-based optimization," vol. 123, pp. 157–184, 2000.

[7] F. Provost, "Tree Induction for Probability-Based Ranking," vol. 5, pp. 199–215, 2003.

[8] J. Sulzmann and F. Johannes, "An Empirical Comparison of Probability Estimation Techniques for Probabilistic Rules," no. 2003, pp. 317–331, 2009.

[9] A. A. Freitas, "On rule interestingness measures," Knowledge-Based Syst., vol. 12, no. March, pp. 309–315, 1999.

[10] K. E. N. Mcgarry, "A survey of interestingness measures for knowledge discovery," pp. 39–61, 2005.

[11] X. Huynh, F. Guillet, J. Blanchard, and P. Kuntz, "A Graph-based Clustering Approach to Evaluate Interestingness Measures : A Tool and a Comparative Study," vol. 50, pp. 25–50, 2007.

[12] B. Vaillant, S. Lallich, and P. Lenca, "On the behavior of the generalizations of the intensity of implication : A data-driven comparative study," vol. 447, pp. 421–447, 2008.

[13] J. Hills, L. M. Davis, and A. Bagnall, "Interestingness Measures for Fixed Consequent Rules," pp. 68–75, 2012.

[14] P. Flach, N. Lavrac, and B. Zupan, "Rule Evaluation Measures: A Unifying View," Proc. 9th Int. Work. Inductive Log. Program., pp. 174–185, 1999.

[15] F. Johannes and P. A. Flach, "An Analysis of Rule Evaluation Metrics," 2003.

[16] D. Christensen, "David Christensen - Measuring Confirmation.pdf." pp. 437–461, 1999.

[17] S. Greco, R. Słowi, and I. Szcz, "Measures of rule interestingness in various perspectives of confirmation," vol. 347, pp. 216–235, 2016.

[18] M. Michalak, M. Sikora, and Ł. Wróbel, "Rule Quality Measures Settings in a Sequential Covering Rule Induction Algorithm - an Empirical Approach," vol. 5, pp. 109–118, 2015.

[19] P. F. Nada Lavrac, Bojan Cestnik, Dragan Gamberger, "Decision Support Through Subgroup Discovery : Three Case Studies and the Lessons Learned," no. 1994, pp. 115–143, 2004.

[20] D. M. W. Powers, "ROC-ConCert," pp. 12–15, 2012.

[21] P. Salgado, "Relevance as a new measure of relative importance: of sets of rules," no. 3, pp. 3770–3777, 2000.

[22] [22] F. Coenen and P. Leng, "An Evaluation of Approaches to Classification Rule Selection," IEEE Int. Conf. Data Min., pp. 2–5, 2004.

[23] Y. Yao and B. Zhou, "Micro and Macro Evaluation of Classification Rules," Proc. Seventh IEEE Int. Conf. Cogn. Informatics, ICCI 2008, Stanford Univ. California, USA, 2008.

[24] D. M. W. Powers, "Evaluation: From Precision, Recall And F-Measure To Roc, Informedness, Markedness & Correlation," vol. 2, no. 1, pp. 37–63, 2011.

[25] H. Abe, S. Tsumoto, M. Ohsaki, and T. Yamaguchi, "Evaluating Learning Algorithms to Construct Rule Evaluation Models Based on Objective Rule Evaluation Indices," 2007.

[26] H. Abe, S. Tsumoto, M. Ohsaki, and T. Yamaguchi, "Evaluating Learning Algorithms to Support Human Rule Evaluation with Predicting Interestingness Based on Objective Rule Evaluation Indices," vol. 282, no. 2008, pp. 269–282, 2008.

[27] H. Abe and S. Tsumoto, "Comparing Accuracies of Rule Evaluation Models to Determine Human Criteria on Evaluated Rule Sets," pp. 1–7, 2008.

[28] H. Abe and S. Tsumoto, "Rule Evaluation Model as Behavioral Modeling," pp. 8–15, 2009.

[29] A. Gruca and M. Sikora, "Rule-based functional description of genes – Estimation of the multicriteria rule interestingness measure by the UTA method," Integr. Med. Res., vol. 33, no. 4, pp. 222–234, 2013.

[30] A. Gruca and M. Sikora, "Data- and expert-driven rule induction and filtering framework for functional interpretation and description of gene sets," pp. 1–14, 2017.

[31] U. Stanczyk, "Weighting and Pruning of Decision Rules," pp. 106–114, 2016.

[32] K. K. Sethi, D. K. Mishra, and B. Mishra, "Novel Algorithm to Measure Consistency between Extracted Models from Big Dataset and Predicting Applicability of Rule Extraction," IEEE Trans. Knowl. Data Eng., 2014.

[33] H. Mutluri and P. Sujatha, "Challenges in Big Data using Data Mining Techniques," Int. J. Comput. Eng. Res. Trends, vol. 2, no. 12, pp. 924–930, 2015.

AUTHORS PROFILE

Munirah Muslim is a lecturer and researcher at the Muhammadiyah University of Ponorogo since 2008. Received the master's degree in Information Technology from Departement of Electrical Engineering, Gadjah Mada University, Indonesia in 2007.

Her research interest in datamining, expert system also artificial intelligent. Now, she is a doctoral student in Departement of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University, Indonesia.

Edi Winarko is a lecturer and researcher at Departement of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University. He received his bachelor degree from Statistics in Gadjah Mada University, Indonesia.

Then his master degree from School of Computing and Information Science, Queen's University, Canada and his doctoral degree from Computer Science School of Informatics and Engineering Flinders University, Australia. His current research interest in Machine Learning, Text Mining, and Spatio-Temporal Mining, also Information Retrieval.