# Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis

Miho Ohsaki[1], Shinya Kitaguchi[2], Kazuya Okamoto[2], Hideto Yokoi[3], and Takahira Yamaguchi[4]

[1] Department of Information Systems Design, Doshisha Univerisity
1-3, Tataramiyakodani, Kyotanabe-shi, Kyoto 610-0321, Japan
mohsaki@mail.doshisha.ac.jp
[2] Department of Computer Science, Shizuoka University
3-5-1, Johoku, Hamamatsu-shi, Shizuoka 432-8011, Japan
{cs8037,cs9026}@cs.inf.shizuoka.ac.jp
[3] Department of Administration Engineering, Keio University
3-14-1, Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa 223-8522, Japan
yamaguti@ae.keio.ac.jp
[4] Division of Medical Informatics, Chiba University Hospital
1-8-1, Inohana, Chuo-ku, Chiba-shi, Chiba 260-0856, Japan
yokoih@telemed.ho.chiba-u.ac.jp

**Abstract.** This research empirically investigates the performance of conventional rule interestingness measures and discusses their practicality for supporting KDD through human-system interaction in medical domain. We compared the evaluation results by a medical expert and those by selected measures for the rules discovered from a dataset on hepatitis. Recall, Jaccard, Kappa, CST, $\chi^2$-M, and Peculiarity demonstrated the highest performance, and many measures showed a complementary trend under our experimental conditions. These results indicate that some measures can predict really interesting rules at a certain level and that their combinational use will be useful.

## 1 Introduction

Rule interestingness is one of active fields in Knowledge Discovery in Databases (KDD), and there have been many studies that formulated interestingness measures and evaluated rules with them instead of humans. However, many of them were individually proposed and not fully evaluated from the viewpoint of theoretical and practical validity. Although some latest studies made a survey on conventional interestingness measures and tried to categorize and analyze them theoretically [1–3], little attention has been given to their practical validity – whether they can contribute to find out really interesting rules.

Therefore, this research aims to (1) systematically grasp the conventional interestingness measures, (2) compare them with real human interest through an experiment, and (3) discuss their performance to estimate real human interest and their utilization to support human-system interaction-based KDD. The

J.-F. Boulicaut et al. (Eds.): PKDD 2004, LNAI 3202, pp. 362–373, 2004.

experiment required the actual rules from a real dataset and their evaluation results by a human expert. We determined to set the domain of this research as medical data mining and to use the outcome of our previous research on hepatitis [4] because medical data mining is scientifically and socially important and especially needs human-system interaction support for enhancing rule quality.

In this paper, Section 2 introduces conventional interestingness measures and selects dozens of measures suitable to our purpose. Section 3 shows the experiment that evaluated the rules on hepatitis with the measures and compared the evaluation results by them with those by a medical expert. It also discusses their performance to estimate real human interest, practicality to support KDD based on human-system interaction, and advanced utilization by combining them. Section 4 concludes the paper and comments on the future work.

## 2 Conventional Rule Interestingness Measures

The results of our and other researchers' surveys [1–3, 5] show that interestingness measures can be categorized with the several factors in Table 1. The subject to evaluate rules, a computer or human user, is the most important categorization factor. Interestingness measures by a computer and human user are called objective and subjective ones, respectively. There are more than forty objective measures at least. They estimate how a rule is mathematically meaningful based on the distribution structure of the instances related to the rule. They are mainly used to remove meaningless rules rather than to discover really interesting ones for a human user, since they do not include domain knowledge [6–17]. In contrast, there are only a dozen of subjective measures. They estimate how a rule fits with a belief, a bias, or a rule template formulated beforehand by a human user. Although they are useful to discover really interesting rules to some extent due to their built-in domain knowledge, they depend on the precondition that a human user can clearly formulate his/her own interest and do not discover absolutely unexpected knowledge. Few subjective measures adaptively learn real human interest through human-system interaction.

The conventional interestingness measures, not only objective but also subjective, do not directly reflect the interest that a human user really has. To avoid the confusion of real human interest, objective measure, and subjective measure, we clearly differentiate them. **Objective Measure**: The feature such as the

**Table 1.** The factors to categorize interestingness measures.

| Factors | Meaning | Sub-factors |
|---|---|---|
| Subject | Who evaluates? | Computer / Human user |
| Object | What is evaluated? | Association rule / Classification rule |
| Unit | By how many objects? | A rule / A set of rules |
| Criterion | Based on what criterion? | Absolute criterion / Relative criterion |
| Theory | Based on what theory? | Number of instances / Probability / Statistics / Information / Distance of rules or attributes / Complexity of a rule |

correctness, uniqueness, and strength of a rule, calculated by the mathematical analysis. It does not include human evaluation criteria. **Subjective Measure**: The similarity or difference between the information on interestingness given beforehand by a human user and those obtained from a rule. Although it includes human evaluation criteria in its initial state, the calculation of similarity or difference is mainly based on the mathematical analysis. **Real Human Interest**: The interest which a human user really feels for a rule in his/her mind. It is formed by the synthesis of cognition, domain knowledge, individual experiences, and the influences of the rules that he/she evaluated before.

This research specifically focuses on objective measures and investigates the relation between them and real human interest. We then explain the details of objective measures here. They can be categorized into some groups with the criterion and theory for evaluation. Although the criterion is absolute or relative as shown in Table 1, the majority of present objective measures are based on an absolute criterion. There are several kinds of criterion based on the following factors: Correctness – How many instances the antecedent and/or consequent of a rule support, or how strong their dependence is [6, 7, 13, 16], Generality – How similar the trend of a rule is to that of all data [11] or the other rules, Uniqueness – How different the trend of a rule is from that of all data [10, 14, 17] or the other rules [11, 13], and Information Richness – How much information a rule possesses [8]. These factors naturally prescribe the theory for evaluation and the interestingness calculation method based on the theory. The theory includes the number of instances [6], probability [12, 14], statistics [13, 16], information [7, 16], the distance of rules or attributes [10, 11, 17], and the complexity of a rule [8] (See Table 1). We selected the objective measures in Table 2 as many and various as possible for the experiment in Section 3. Note that many of them do not have the reference numbers of their original papers but those of survey papers in Table 2 to avoid too many literatures. We call GOI with the dependency coefficient value at the double of the generality one GOI-D, and vice versa for GOI-G and adopt the default value, 0.5, for the constant $\alpha$ of Peculiarity.

Now, we explain the motivation of this research in detail. Objective measures are useful to automatically remove obviously meaningless rules. However, some factors of evaluation criterion have contradiction to each other such as generality and uniqueness and may not match with or contradict to real human interest. In a sense, it may be proper not to investigate the relation between objective measures and real human interest, since their evaluation criterion does not include the knowledge on rule semantics and are obviously not the same of real human interest. However, our idea is that they may be useful to support KDD through human-system interaction if they possess a certain level of performance to detect really interesting rules. In addition, they may offer a human user unexpected new viewpoints. Although the validity of objective measures has been theoretically proven and/or experimentally discussed using some benchmark data [1–3], very few attempts have been made to investigate their comparative performance and the relation between them and real human interest for a real application [5]. Our investigation will be novel in this light.

**Table 2.** The objective measures of rule interestingness used in this research. **N:** Number of instances included in the antecedent and/or consequent of a rule. **P:** Probability of the antecedent and/or consequent of a rule. **S:** Statistical variable based on P. **I:** Information of the antecedent and/or consequent of a rule. **D:** Distance of a rule from the others based on rule attributes.

| Measure Name (**Abbreviation**) [Reference Number of Literature] | **Theory** |
|---|---|
| *Mathematical Definition* | |
| **Coverage** [5] | **P** |
| $P(A)$, $P(A)$: Probability of antecedent. | |
| **Prevalence** [5] | **P** |
| $P(C)$, $P(C)$: Probability of consequent. | |
| **Precision** [3, 5] | **P** |
| $P(C\|A)$, $P(C\|A)$: Conditional probability of consequent for antecedent. | |
| **Recall** [5] | **P** |
| $P(A\|C)$, $P(A\|C)$: Conditional probability of antecedent for consequent. | |
| **Support** [1, 3, 5] | **P** |
| $P(C\|A) * P(A)$ | |
| **Specificity** [5] | **P** |
| $P(\neg C\|\neg A)$, $\neg X$: Negation of $X$. | |
| **Accuracy** [5] | **P** |
| $P(C\|A) * P(A) + P(\neg C\|\neg A) * P(\neg A)$ | |
| **Lift** [5] | **P** |
| $P(C\|A)/P(C)$ | |
| **Leverage** [5] | **P** |
| $P(C\|A) - P(A) * P(C)$ | |
| Added Value (**AV**) [3] | **P** |
| $P(C\|A) - P(C)$ | |
| Relative Risk (**RR**) [1] | **P** |
| $P(C\|A)/P(C\|\neg A)$ | |
| **Jaccard** [3] | **P** |
| $P(A \cap C)/\{P(A) + P(C) - P(A \cap C)\}$<br>$P(A \cap C)$: Probability of antecedent and consequent. | |
| Certainty Factor (**CF**) [3] | **P** |
| $\{P(C\|A) - P(C)\}/\{1 - P(C)\}$ | |
| Odds Ratio (**OR**) [3] | **P** |
| $\{P(A \cap C) * P(\neg A \cap \neg C)\}/\{P(A \cap \neg C) * P(\neg A \cap C)\}$ | |
| **Yule's Q** [3] | **P** |
| $(OR - 1)/(OR + 1)$ | |
| **Yule's Y** [3] | **P** |
| $(\sqrt{OR} - 1)/(\sqrt{OR} + 1)$ | |
| **Kappa** [3] | **P** |
| $\frac{P(A\cap C)+P(\neg A\cap\neg C)-P(A)*P(C)-P(\neg A)*P(\neg C)}{1-P(A)*P(C)-P(\neg A)*P(\neg C)}$ | |
| Klosgen's Interestingness (**KI**) [1, 3] | **P** |
| $\sqrt{P(A \cap C)} * \{P(C\|A) - P(C)\}$ | |
| Brin's Interest (**BI**) [3] | **P** |
| $P(A \cap C)/\{P(A) * P(C)\}$ | |
| Brin's Conviction (**BC**) [3] | **P** |
| $\{P(A) * P(\neg C)\}/P(A\|\neg C)\}$ | |

| | |
|---|---|
| Gray and Orlowska's Interestingness weighting Dependency (**GOI-D**) [1, 5, 12] | **P** |
| $((\frac{P(C\|A)}{P(A)*P(C)})^k - 1) * ((P(A) * P(C))^m, \quad k, m$: Coefficients of dependency and generality. | |
| GOI weighting Generality (**GOI-G**) [1, 5, 12] | **P** |
| Definition is the same of GOI-D. | |
| Collective Strength (**CST**) [3] | **P** |
| $\frac{P(A\cap C)+P(\neg C\|\neg A)}{P(A)*P(C)+P(\neg A)*P(\neg C)} * \frac{1-P(A)*P(C)-P(\neg A)*P(\neg C)}{1-P(A\cap C)-P(\neg C\|\neg A)}$ | |
| **Credibility** [5, 9] | **P, N** |
| $\beta_i * P(C) * \|P(R_i\|C) - P(R_i)\| * T(R_i), \quad \beta_i$: Coefficient of normalization. | |
| $P(R_i)$: Probability of the rule $R_i$. $T(R_i)$: Number of instances in $R_i$. | |
| Laplace Correction (**LC**) [3] | **N** |
| $\{N(A \cap C) + 1\}/\{N(A) + 2\}, \quad N(X)$: Number of instances in $X$. | |
| $\chi^2$ Measure ($\chi^2$**-M**) [5, 13] | **S** |
| $\sum_{event} \frac{(T_{event} - O_{event})^2}{T_{event}}, \quad event$: $A \rightarrow C$, $A \rightarrow \neg C$, $\neg A \rightarrow C$, $\neg A \rightarrow \neg C$ | |
| $T_{event}$: Theoretical number of instances in $event$, $O_{event}$: Observed one. | |
| Gini Index (**Gini**) [3] | **S** |
| $P(A) * \{P(C\|A)^2 + P(\neg C\|A)^2\} + P(\neg A) * \{P(C\|\neg A)^2 + P(\neg C\|\neg A)^2\}$ $-P(C)^2 - P(\neg C)^2$ | |
| Goodman and Kruskal's Interestingess (**GKI**) [3] | **S** |
| $\frac{\sum_i max_j P(A_i \cap C_j) + \sum_j max_i P(A_i \cap C_j) - max_i P(A_i) - max_j P(C_j)}{2 - max_i P(A_i) - max_j P(C_j)}$ | |
| Normalized Mutual Information (**NMI**) [3] | **I** |
| $\sum_i \sum_j P(A_i \cap C_j) * log_2 \frac{P(A_i \cap C_j)}{P(A_i) * P(C_j)} / \{-\sum_i P(A_i) * log_2 P(A_i)\}$ | |
| J-Measure (**J-M**) [1, 3, 5, 7] | **I** |
| $P(C) * (KLD(C\|A; C) + KLD(\neg C\|\neg A; \neg C), \quad KLD$: Kullback-Leibler Distance | |
| Yao and Liu's Interestingness 1 based on one-way support (**YLI1**) [1] | **I** |
| $P(C\|A) * log_2 \frac{P(A \cap C)}{P(A) * P(C)}$ | |
| Yao and Liu's Interestingness 2 based on two-way support (**YLI2**) [1] | **I** |
| $P(A \cap C) * log_2 \frac{P(A \cap C)}{P(A) * P(C)}$ | |
| Yao and Liu's Interestingness 3, the sum of possible YLI2 variations (**YLI3**) [1] | **I** |
| $P(A \cap C) * log_2 \frac{P(A \cap C)}{P(A) * P(C)} + P(A \cap \overline{C}) * log_2 \frac{P(A \cap \overline{C})}{P(A) * P(\overline{C})} +$ $P(\overline{A} \cap C) * log_2 \frac{P(\overline{A} \cap C)}{P(\overline{A}) * P(C)} + P(\overline{A} \cap \overline{C}) * log_2 \frac{P(\overline{A} \cap \overline{C})}{P(\overline{A}) * P(\overline{C})}$ | |
| K-Measure (**K-M**) [5] | **I** |
| $KLD(C\|A; C) + KLD(\neg C\|\neg A; \neg C) - KLD(C\|A; \neg C) + KLD(\neg C\|\neg A; C)$ | |
| $\phi$ Coefficient ($\phi$) [3] | **N** |
| $\{P(A \cap C) - P(A) * P(C)\}/\sqrt{P(A) * P(C) * P(\neg A) * P(\neg C)}$ | |
| Piatetsky-Shapiro's Interestingness (**PSI**) [3, 5, 6] | **N** |
| $N(A \cap C) - \frac{N(A) * N(C)}{N(U)}, \quad N(U)$: Number of instances in universe. | |
| $N_C$: That in consequent. $N_U$: That in rule. | |
| Cosine Similarity (**CSI**) [3] | **N** |
| $P(A \cap C)/\sqrt{(P(A) * P(C))}$ | |
| Gago and Bento's Interestingness (**GBI**) [11] | **D** |
| $\sum_{j=1}^{N_R} D(R_i, R_j)/N_R, \quad R_i$ : i-th rule. $N_R$: Number of rules. | |
| $D(R_i, R_j)$: Distance based on attribute overlap degree between i-th and j-th rules. | |
| **Peculiarity** [17] | **D** |
| $\sum_{i=1}^{N_a} \sum_{k=1}^{N_i} \|x_{ij} - x_{ik}\|^\alpha / N_a, \quad x_{ij}$: j-th value of i-th attribute. | |
| $N_a$: Number of attributes. $N_i$: Number of values of i-th attribute. $\alpha$: Constant. | |

## 3 Evaluation Experiment of Objective Measures

### 3.1 Experimental Conditions

The experiment examined the performance of objective measures to estimate real human interest by comparing the evaluation by them and a human user. Concretely speaking, the selected objective measures and a medical expert evaluated the same medical rules, and their evaluation values were qualitatively and quantitatively compared. We used the objective measures in Table 2 and the rules and their evaluation results in our previous research [4].

Here, we note the outline of our previous research. We tried to discover new medical knowledge from a clinical dataset on hepatitis. The KDD process was designed to twice repeat a set of the rule generation by our mining system and the rule evaluation by a medical expert for polishing up the obtained rules. Our mining system was based on the typical framework of time-series data mining, a combination of the pattern extraction by clustering and the classification by a decision tree. It generated prognosis-prediction rules and visualized them as graphs. The medical expert conducted the following evaluation tasks: After each mining, he gave each rule the comment on its medical interpretation and one of the rule quality labels, which were Especially-Interesting (**EI**), Interesting (**I**), Not-Understandable (**NU**), and Not-Interesting (**NI**). **EI** means that the rule was a key to generate or confirm a hypothesis.

A few rules in the first mining inspired the medical expert to make a hypothesis, a seed of new medical knowledge: Contradict to medical common sense, GPT, which is an important medical test result to grasp hepatitis symptom, may change with three years cycle (See the left side in Fig. 1). A few rules in the second mining supported him to confirm the hypothesis and enhanced its reliability (See the right side in Fig. 1). As a consequence, we obtained a set of rules and their evaluation results by the medical expert in the first mining and that in the second mining. Three and nine rules received **EI** and **I** in the first mining, respectively. Similarly, two and six rules did in the second mining.

In our current research, the evaluation procedure by the objective measures was designed as follows: For each objective measure, the same rules as in our previous research were evaluated by the objective measure, sorted in the descending

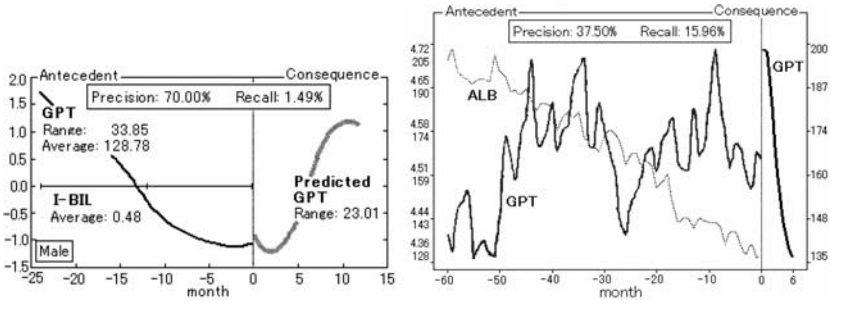


**Fig. 1.** The examples of highly valued rules in first (left) and second mining (right).

order of evaluation values, and assigned the rule quality labels. The rules from the top to the $m$-th were assigned **EI**, where $m$ was the number of **EI** rules in the evaluation by the medical expert. Next, the rules from the $(m+1)$-th to the $(m+n)$-th were assigned **I**, where $n$ was the number of **I** rules in the evaluation by the medical expert. The assignment of **NU** and **NI** followed the same procedure. We dared not to do evaluation value thresholding for the labeling. The first reason was that it is quite difficult to find the optimal thresholds for the all combinations of labels and objective measures. The second reason was that although our labeling procedure may not be precise, it can set the conditions of objective measures at least equal through simple processing. The last reason was that our number-based labeling is more realistic than threshold-based labeling. The number of rules labeled with **EI** or **I** by a human user inevitably stabilizes at around a dozen in a practical situation, since the number of evaluation by him/her has a severe limitation caused by his/her fatigue.

### 3.2 Results and Discussion

Fig. 2 and 3 show the experimental results in the first and second mining, respectively. We analyzed the relation between the evaluation results by the medical expert and the objective measures qualitatively and quantitatively. As the qualitative analysis, we visualized their degree of agreement to easily grasp its trend. We colored the rules with perfect agreement white, probabilistic agreement gray, and disagreement black. A few objective measures output same evaluation values for too many rules. For example, although eight rules were especially interesting (**EI**) or interesting (**I**) for the medical expert in second mining, the objective measure OR estimated 14 rules as **EI** or **I** ones (See Fig. 3). In that case, we colored such rules gray. The pattern of white (possibly also gray) and black cells for an objective measure describes how its evaluation matched with those by the medical expert. The more the number of white cells in the left-hand side, the better its performance to estimate real human interest.

For the quantitative analysis, we defined four comprehensive criteria to evaluate the performance of an objective measure. #1: Performance on **I** (the number of rules labeled with **I** by the objective measure over that by the medical expert. Note that **I** includes **EI**). #2: Performance on **EI** (the number of rules labeled with **EI** by the objective measure over that by the medical expert). #3: Number-based performance on all evaluation (the number of rules with the same evaluation results by the objective measure and the medical expert over that of all rules). #4: Correlation-based performance on all evaluation (the correlation coefficient between the evaluation results by the objective measure and those by the medical expert). The values of these criteria are shown in the right side of Fig. 2 and 3. The symbol '+' besides a value means that the value is greater than that in case rules are randomly selected as **EI** or **I**. Therefore, an objective measure with '+' has higher performance than random selection does at least. To know the total performance, we defined the weighted average of the four criteria as a meta criterion; we assigned 0.4, 0.1, 0.4, and 0.1 to #1, #2, #3, and

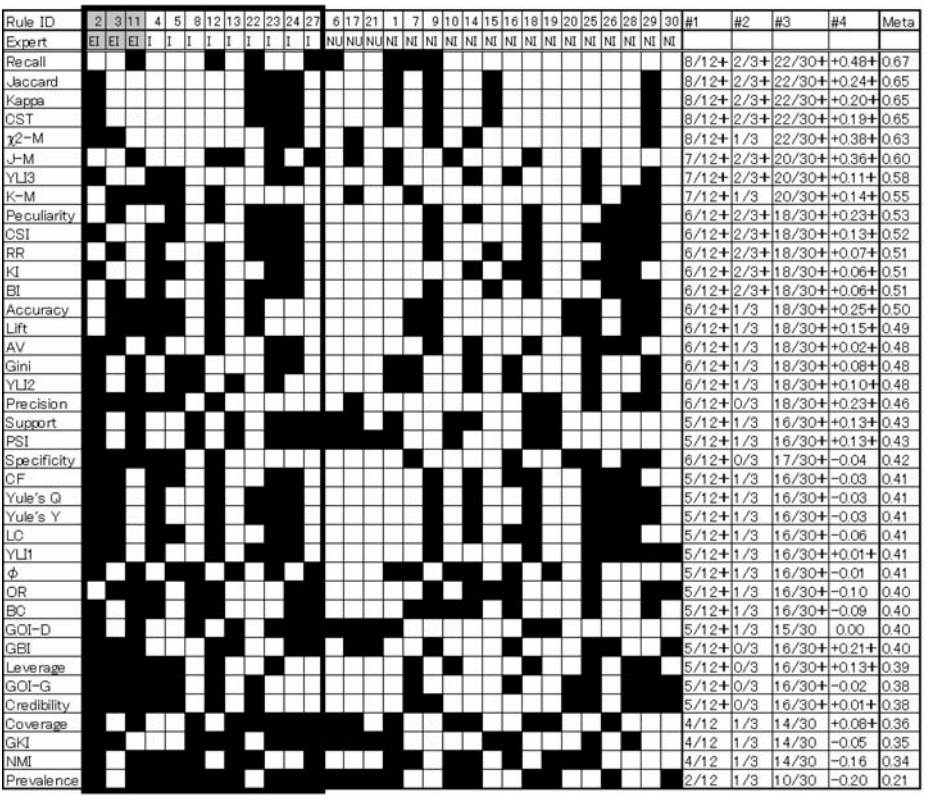


**Fig. 2.** The evaluation results by a medical expert and objective measures for the rules in first mining. Each column represents a rule, and each row represents the set of evaluation results by an objective measure. The rules are sorted in the descending order of the evaluation values given by the medical expert. The objective measures are sorted in the descending order of the meta criterion values. A square in the left-hand side surrounds the rules labeled with **EI** or **I** by the medical expert. White, gray, and black cells mean that the evaluation by an objective measure was perfectly, was probabilistically, and was not the same by the medical expert, respectively. The five columns in the right side show the performance on the four comprehensive criteria and the meta one. '+' means the value is greater than that of random selection.

#4, respectively, according to their importance. The objective measures were sorted in the descending order of the values of meta criterion.

The results in the first mining in Fig. 2 show that Recall demonstrated the highest performance, Jaccard, Kappa, and CST did the second highest, and $\chi^2$-M did the third highest. Prevalence demonstrated the lowest performance, NMI did the second lowest, and GKI did the third lowest. The results in the second mining in Fig. 3 show that Credibility demonstrated the highest performance, Peculiarity did the second highest, and Accuracy, RR, and BI did the third highest. Prevalence demonstrated the lowest performance, Specificity did the

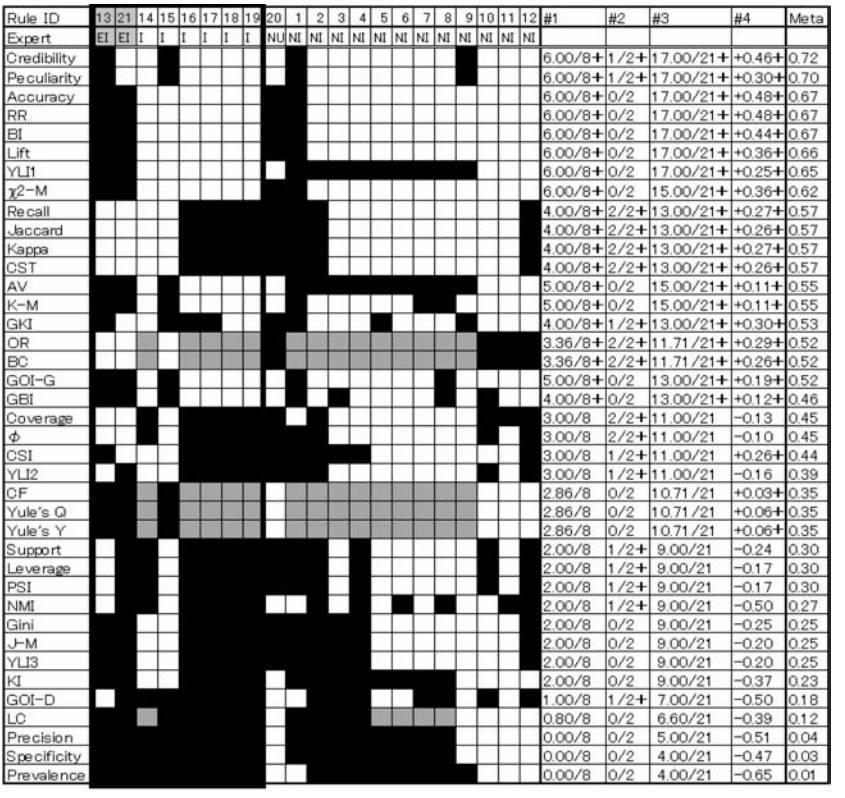


**Fig. 3.** The evaluation results by a medical expert and objective measures for the rules in second mining. See the caption of Fig. 2 for the details.

second lowest, and Precision did the third lowest. We summarized these objective measures in Table 3. As a whole, the following objective measures maintained their high performance through the first and second mining: Recall, Jaccard, Kappa, CST, $\chi^2$-M, and Peculiarity. NMI and Prevalence maintained their low performance. Only Credibility changed its performance dramatically, and the other objective measures slightly changed their middle performance.

More than expected, some objective measures – Recall, Jaccard, Kappa, CST, $\chi^2$-M, and Peculiarity – showed constantly high performance. They had comparatively many white cells and '+' for all comprehensive criteria. In addition, the mosaic-like patterns of white and black cells in Fig. 2 and 3 showed that the objective measures had almost complementary relationship for each other. The results and the medical expert's comments on them imply that his interest consisted of not only the medical semantics but also the statistical characteristics of rules. The combinational use of objective measures will be useful to reductively analyze such human interest and to recommend interesting rule candidates from various viewpoints through human-system interaction in medical KDD. One method to obtain the combination of objective measures is to formu-

**Table 3.** The summary of the objective measures with the highest or the lowest performance in the first and second mining. ($N$) means the rank in the other mining.

Top 3

| Ranking | First Mining (Second Mining) | Second Mining (First Mining) |
|---|---|---|
| 1 | Recall(9) | Credibility(34) |
| 2 | Jaccard(9), Kappa(9), CST(9) | Peculiarity(9) |
| 3 | $\chi^2$-M(8) | Accuracy(14), RR(11), BI(11) |

Last 3

| Ranking | First Mining (Second Mining) | Second Mining (First Mining) |
|---|---|---|
| 37 | GKI(15) | Precision(19) |
| 38 | NMI(30) | Specificity(22) |
| 39 | Prevalence(39) | Prevalence(39) |

late a function consisting of the summation of weighted outputs from objective measures. Another method is to learn a decision tree using these outputs as attributes and the evaluation result by the medical expert as a class. We can conclude that although the experimental results are not enough to be generalized, they gave us two important implications: some objective measures will work at a certain level in spite of no consideration of domain semantics, and the combinational use of objective measures will help medical KDD.

Our future work will be directed to two issues including some sub-issues as shown in Table 4. Here, we describe their outlines. On Issue (i) the investigation/analysis of objective measures and real human interest, Sub-issue (i)-1 and (i)-2 are needed to generalize the current experimental results and to grasp the theoretical possibility and limitation of objective measures, respectively. Sub-issue (i)-3 is needed to establish the method to predict real human interest. We have already finished an experiment on Sub-issue (i)-1 and (i)-3, and will show their results soon. Sub-issue (i)-2 is now under discussion. The outcome of those empirical and theoretical researches will contribute to solving Issue (ii). Issue (ii) the utilization of objective measures for KDD support based on human-system interaction, assumes that the smooth interaction between a human user and a mining system is a key to obtain really interesting rules for the human user. Our previous research in Section 3.1 [4] and others' researches using the same dataset of ours [18] led us to this assumption. We think that smooth human-system interaction stimulates the hypothesis generation and confirmation of a human user, and actually it did in our previous research. Sub-issue (ii)-1 is needed to support such a thinking process in the post-processing phase of data mining. Our current idea is to develop a post-processing user interface in which a human user can select one among various objective measures and see the rules sorted with its evaluation values. We expect that the user interface will enhance the thinking from unexpected new viewpoints. Sub-issue (ii)-2 is the extension of Sub-issue (ii)-1; It comprehensively focuses on the spiral sequence of mining algorithm organization and post-processing. As the one of Sub-issue (ii)-2 solutions, now we are implementing an evaluation module, which uses the predicted real

**Table 4.** The outlines of issues in our future work.

| |
|---|
| Issue (i) Investigation/analysis of objective measures and real human interest. |
| Sub-issue (i)-1 Experiments with different datasets and medical experts. |
| Sub-issue (i)-2 Mathematical analysis of objective measures. |
| Sub-issue (i)-3 Reductive analysis of real human interest using the combination of objective measures. |
| Issue (ii) Utilization of objective measures for KDD support based on human-system interaction. |
| Sub-issue (ii)-1 Development of a post-processing user interface. |
| Sub-issue (ii)-2 Development of a comprehensive KDD environment. |

human interest with objective measures in Sub-issue (i)-3, into a constructive meta-learning system called CAMLET [19].

## 4 Conclusions and Future Work

This paper discussed how objective measures can contribute to detect interesting rules for a medical expert through an experiment using the rules on hepatitis. Recall, Jaccard, Kappa, CST, $\chi^2$-M, and Peculiarity demonstrated good performance, and the objective measures used here had complementary relationship for each other. It was indicated that their combination will be useful to support human-system interaction. Our near-future work is to obtain the generic trend of objective measures in medical KDD. As an empirical approach, we have already finished another experiment with a clinical dataset on meningoencephalitis and a different medical expert and are comparing the experimental results on hepatitis and meningoencephalitis. As a theoretical approach, we are conducting the mathematical analysis of objective measure features. We will utilize these outcomes for supporting medical KDD based on system-human interaction.

## References

1. Yao, Y. Y. Zhong, N.: An Analysis of Quantitative Measures Associated with Rules. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-1999 (1999) 479–488
2. Hilderman, R. J., Hamilton, H. J.: Knowledge Discovery and Measure of Interest. Kluwer Academic Publishers (2001)
3. Tan, P. N., Kumar V., Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns. Proceedings of International Conference on Knowledge Discovery and Data Mining KDD-2002 (2002) 32–41
4. Ohsaki, M., Sato, Y., Yokoi, H., Yamaguchi, T.: A Rule Discovery Support System for Sequential Medical Data, – In the Case Study of a Chronic Hepatitis Dataset –. Proceedings of International Workshop on Active Mining AM-2002 in IEEE International Conference on Data Mining ICDM-2002 (2002) 97–102
5. Ohsaki, M., Sato, Y., Yokoi, H., Yamaguchi, T.: Investigation of Rule Interestingness in Medical Data Mining. Lecture Notes in Computer Science, Springer-Verlag (2004) will appear.

6. Piatetsky-Shapiro, G.: Discovery, Analysis and Presentation of Strong Rules. in Piatetsky-Shapiro, G., Frawley, W. J. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press (1991) 229–248
7. Smyth, P., Goodman, R. M.: Rule Induction using Information Theory. in Piatetsky-Shapiro, G., Frawley, W. J. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press (1991) 159–176
8. Hamilton, H. J., Fudger, D. F.: Estimating DBLearn's Potential for Knowledge Discovery in Databases. Computational Intelligence, 11, 2 (1995) 280–296
9. Hamilton, H. J., Shan, N., Ziarko, W.: Machine Learning of Credible Classifications. Proceedings of Australian Conference on Artificial Intelligence AI-1997 (1997) 330–339
10. Dong, G., Li, J.: Interestingness of Discovered Association Rules in Terms of Neighborhood-Based Unexpectedness. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-1998 (1998) 72–86
11. Gago, P., Bento, C.: A Metric for Selection of the Most Promising Rules. Proceedings of European Conference on the Principles of Data Mining and Knowledge Discovery PKDD-1998 (1998) 19–27
12. Gray, B., Orlowska, M. E.: CCAIIA: Clustering Categorical Attributes into Interesting Association Rules. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-1998 (1998) 132–143
13. Morimoto, Y., Fukuda, T., Matsuzawa, H., Tokuyama, T., Yoda, K.: Algorithms for Mining Association Rules for Binary Segmentations of Huge Categorical Databases. Proceedings of International Conference on Very Large Databases VLDB-1998 (1998) 380–391
14. Freitas, A. A.: On Rule Interestingness Measures. Knowledge-Based Systems, 12, 5–6 (1999) 309–315
15. Liu, H., Lu, H., Feng, L., Hussain, F.: Efficient Search of Reliable Exceptions. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-1999 (1999) 194–203
16. Jaroszewicz, S., Simovici, D. A.: A General Measure of Rule Interestingness. Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery PKDD-2001 (2001) 253–265
17. Zhong, N., Yao, Y. Y., Ohshima, M.: Peculiarity Oriented Multi-Database Mining. IEEE Transaction on Knowledge and Data Engineering, 15, 4 (2003) 952–960
18. Motoda, H. (eds.): Active Mining, IOS Press, Amsterdam, Holland (2002).
19. Abe, H. and Yamaguchi T.: Constructive Meta-Learning with Machine Learning Method Repository, IEA/AIE2004, LNAI3029 (2004) pp.502–511.