Artificial Intelligence in Medicine (2007) 41, 177—196

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Evaluation of rule interestingness measures in medical knowledge discovery in databases

Miho Ohsaki [a,*], Hidenao Abe [b], Shusaku Tsumoto [b], Hideto Yokoi [c], Takahira Yamaguchi [d]

[a] *Faculty of Engineering, Doshisha University, 1-3 Tataramiyakodani, Kyotanabe-shi, Kyoto 610-0321, Japan*
[b] *Department of Medical Informatics, Shimane University, 89-1 Enya-cho, Izumo-shi, Shimane 693-8501, Japan*
[c] *Department of Medical Informatics, Kagawa University Hospital, 1750-1 Ikenobe, Miki-cho, Kita-gun, Kagawa 761-0793, Japan*
[d] *Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kangawa 223-8522, Japan*

Received 18 October 2006; received in revised form 24 July 2007; accepted 24 July 2007

KEYWORDS
Data mining;
Knowledge discovery in databases;
Interestingness;
Postprocessing;
Clinical data

1991 MSC
68T30;
68U35

PACS
07.05.Kf;
07.05.Mh

Summary

Objective: We discuss the usefulness of rule interestingness measures for medical KDD through experiments using clinical datasets, and, based on the outcomes of these experiments, also consider how to utilize these measures in postprocessing.
Methods and materials: We first conducted an experiment to compare the evaluation results derived from a total of 40 various interestingness measures with those supplied by a medical expert for rules discovered in a clinical dataset on meningitis. We calculated and compared the performance of each interestingness measure to estimate a medical expert's interest using f-measure and correlation coefficient. We then conducted a similar experiment for hepatitis.
Results and conclusion: The comprehensive results of experiments on meningitis and hepatitis indicate that the interestingness measures, accuracy, chi-square measure for one quadrant, relative risk, uncovered negative, and peculiarity, have a stable, reasonable performance in estimating real human interest in the medical domain. The results also indicate that the performance of interestingness measures is influenced by the certainty of a hypothesis made by the medical expert, and that the combinational use of interestingness measures will contribute to support medical experts to generate and confirm their hypotheses through human—system interaction.

* Corresponding author. Tel.: +81 774 65 6468; fax: +81 774 65 6468.
E-mail address: mohsaki@mail.doshisha.ac.jp (M. Ohsaki).

0933-3657/$ — see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2007.07.005

1. Introduction

Rule interestingness is an active area of study in the fields of data mining (DM) and knowledge discovery in databases (KDD). Many studies have been conducted concerning the formalization of rule interestingness, and concerning human substitutive evaluation of rules using formalized interestingness measures. However, most interestingness measures introduced to date have been proposed independently, and comparisons and examinations of their usefulness from both practical and theoretical viewpoints are incomplete. Although several attempts have recently made to theoretically analyze various types of interestingness measures [1—5], the practicability of these measures has not been taken into account. It is therefore necessary to examine whether interestingness measures are useful in discovering interesting rules for users in real application domains [6—8].

Based on the above considerations, the present research aims to empirically examine the usefulness of interestingness measures, focusing on objective measures, which is one type of interesting measure (the other type of interesting measure is subjective measures). The reason for focusing on objective measures is explained in detail in Section 2. We should select a concrete domain for empirically examining the usefulness of interestingness measures. Our research concentrates on the clinical domain, which is both a scientifically and socially important domain, and discusses the usefulness of interestingness measures for medical KDD in this domain. In experiments, medical experts and interestingness measures evaluate the rules obtained from clinical data, and we estimate the performance of the interestingness measures to estimate real human interest according to the agreement between the evaluation results by the medical experts and those by the interestingness measures. Furthermore, on the basis of the experimental findings, we discuss how to utilize the interestingness measures to support users in medical KDD.

If, in the interest of rigor, there is an excessive strictness over experimental conditions of human-related experiments or the procedures to accurately carry out these experiments, unrealistic results may result. In particular, when KDD is the aim of an experiment, an expert in the particular domain will compare the rules output from the mining system and conclude new knowledge while gradually changing his/her own interest and point of observation. Thus, an approach to make the psychological scale stable by the strict control of experimental conditions and procedures in order to examine the relationship between prescribed explanatory variables and the psychological scale, would not be expected to be suitable for the observation and analysis of real KDD processes. In addition, it requires considerably much human resource and time cost to discover valuable knowledge in a real application domain. Therefore, although the degree of scientific strictness of real case studies is not high, we determined to consider real case studies in the present research.

We adopt two case studies in which new medical knowledge was discovered by the repetition of several steps over the course of 1 or 2 years. The steps used here were mining system adjustment, rule evaluation by a medical expert, and hearing investigation of the opinions expressed by the medical expert. One case study obtained medically interesting rules from a clinical dataset on meningitis, with rule evaluation results and opinions obtained from a medical expert [9]. The other study applied a similar KDD process to hepatitis (with the medical expert involved differing to that of the first case study) [10]. The present experiments make use of the rules and the rule evaluation results by the medical experts in these previous case studies. The number of case studies and that of medical experts are small, and the case study conditions are less well controlled than desired from a scientific viewpoint. However, we think that it is meaningful from an engineering viewpoint to obtain the utilization knowledge of interestingness measures through real case studies.

The organization of this paper is as follows: Section 2 describes the survey and selection of interestingness measures. Section 3 describes Experiment I, which concerns meningitis, while Section 4 describes Experiment II, which concerns hepatitis. In both the Experiments I and II, we estimate and compare the performance of each interestingness measure to estimate the interest of a medical expert. Moreover, we discuss the thinking process of a medical expert in KDD and some insights into the possible utilization of interestingness measures for supporting this thinking process. Section 5 discusses the results of Experiments I and II from a comprehensive standpoint and the possible utilization of interestingness measures. Section 6 presents the conclusions of the present research together with some considerations of future work.

2. Conventional rule interestingness measures

There are various interestingness measures, most of which have been proposed independently, and some measures are known by different names. In response to, attempts have been made to systematically describe interestingness measures including, for example, exhaustive surveys and classification of interestingness measures, mathematical proof of the equivalence of certain interestingness measures, and analysis of their behaviors using simulation techniques [1—5,8,11—13].

We herein list every possible factor by which to classify interestingness measures, as follows: Subject (What performs the evaluation? An objective subject, such as a mathematical formula/algorithm, or a subjective subject, such as a human user?). Object (What is being evaluated? Association rules or classification rules?). Concept (What concept is being sought? Conciseness, generality, reliability, peculiarity, diversity, novelty, surprisingness, utility, or actionability [8]?). Criterion (What criterion is being employed? Data distribution structure or domain knowledge/expectation?). Unit (What unit is being employed? An individual rule or a set of rules?). Base theory (What theory is being employed? Instance number, probabilistic magnitude, statistics value, information amount, distance among rules and/or attributes, or rule complexity?). The most important classifiable factors, subject, concept, and criterion, are mutually independent and should be inclusively discussed.

In some studies [12,13], interestingness measures are roughly classified into two groups: objective measures and subjective measures. In another study [8], interestingness measures are more finely classified into three groups: objective, subjective, and semantic measures. From the definition of semantic measures in this study, in a broad sense, subjective measures include semantic measures.

Generally, objective measures have been proposed together with mining algorithms as their learning metrics. According to the results of our literature search, there are at least several dozen objective measures. Objective measures analyze the distribution structure of the instances that satisfy and/or accompany the rule (i.e., instances that satisfy only the antecedent, or that are contained in the entire dataset, for example). Objective measures then estimate whether the rule has mathematically significant features based on the result of analysis [1—5,8,11—13].

Objective measures are effective for mining rules satisfying previously required mathematical features and filtering potentially interesting rules. However, it is not easy to extract rules that would be really of interest to a human user by preliminarily elaborating objective measures because of the lack of domain knowledge [12]. Although it would be difficult to obtain a consensus as to which concept can be described by objective measures, an assignment of concepts to objective measures was presented by Geng and Hamilton [8] along with some convincing explanations. They reported that objective measures are intended to formalize conciseness, generality, reliability, peculiarity, or diversity. The characteristics of objective measures can be summarized as follows [12]: objective measures are data-driven, domain-independent, and consequently guarantee mathematically significant features of a rule such as prediction accuracy, but do not guarantee domain-dependently favorable features.

Subjective measures have been proposed frequently as rule organization or evaluation metrics in preprocessing, postprocessing, or interactive mining frameworks. There are at least a dozen subjective measures. Subjective measures estimate the degree of agreement between the rule and the template based on domain knowledge/expectation prescribed by the user [3,12,13,8]. Among subjective measures, there are partly user-inclusive measures with user-specified belief, such as unexpectedness [14—16], interestingness [17], and preference [18], user-inclusive measures with postprocessing user interfaces for rule selection with user-specified templates and statistical significance [19,20], and user-exclusive measures with logical formalization of unexpectedness [21—23]. User-exclusive measures are objective rather than subjective.

Subjective measures can, to a certain extent, effectively discover interesting rules, because subjective measures include domain knowledge/expectation. However, the performance of subjective measures depends on the skill of the user in formulating the domain knowledge and their expectation. If the user provides a template that strictly specifies the type and scope of attributes and class, reasonable rules for a user can be discovered. Surprising rules that clearly contradict domain knowledge/expectation can also be discovered. However, unexpected rules that the user has never imagined are difficult to discover. Geng and Hamilton [8] presented an assignment of concepts to subjective measures as follows: subjective measures are intended to estimate novelty, surprisingness, utility, or actionability. The characteristics of subjective measures can be summarized as follows [12]: subjective measures are user-driven and domain-dependent, and consequently guarantee domain-dependently favorable features of a rule, such as medical reasonableness, but do not guarantee mathematically significant features.

Neither objective nor subjective measures perfectly describe real human interest. Real human interest is the interest that a human feels when viewing a rule. It is formed by synthesizing his/her recognition characteristics, domain knowledge, individual experiences and expectations, and the influence of previously encountered rules. Real human interest may change gradually or even drastically through a KDD process due to the above-described synthesizing elements. Thus, objective and subjective measures describe some parts of real human interest.

The characteristics of objective and subjective measures indicate that there are two problem-solving approaches on interestingness measures to the discovery of really interesting knowledge. One approach is the specification of subjective measures; namely, refining the specific definition and description of subjective measures to a target domain based on the explicit interest of a domain expert. This approach will contribute to the discovery of solid knowledge. The other approach is determining the implicit interest of a domain expert; namely, clarifying the relationship between objective measures and the implicit interest of a domain expert (and, as a further extension, using objective measures to estimate the implicit interest based on the relationship). This approach will contribute to the discovery of unexpected knowledge. Studies having the former approach and those having the latter approach should be conducted thoroughly, and on the next stage the fusion of objective and subjective measures should be discussed.

At present, the former approach has been adopted in several studies on subjective measures, while studies considering the latter approach have only recently begun [6,7]. In the studies adopting the latter approach, a method to examine the relationship has not yet been established and/or the number of interestingness measures used in these studies has been limited.

Our research takes the latter approach and tries to examine the relationship and to estimate the usefulness of objective measures in real medical KDD. We now more finely address the motivation for this study, i.e., "why should the relationship between objective measures and real human interest be studied?" Objective measures illustrate those features of a rule, such as generality and reliability, which can be formalized by mathematical formulae. Since this approach does not take into account the semantics of a rule in a specific domain, it is difficult for such objective measures to always match real human interest, and so high degrees of matching performance cannot be expected. However, if objective measures can perform not so highly but well-enough in predicting real human interest, then the use of objective measures to support a user in the evaluation of rules to discover new knowledge may become possible. This was the motivation for the present research.

For the preparation of experiments in this research, we select general and various other objective measures, and uniform the definitions of these measures so as to evaluate a classification rule, which is in the form *IF A THEN C*. The antecedent, A, of the classification rule is a conjunction of attributes A_1 (possible attribute values are $a_{11}, a_{12}, \ldots$), A_2 (possible attribute values are $a_{21}, a_{22}, \ldots$),... For example, $A_1 = a_{14}$ and $A_2 = a_{21}$ and .... The consequent, C, is one of the classes of $c_1, c_2, \ldots$. We here give a description of the most commonly used objective measures, coverage, prevalence, precision, recall, support, and uncovered negative. These measures were originally based on a confusion matrix that divides a set of instances into true positive (TP), false positive (FP), true negative (TN), and false negative (FN), and are defined by the first term of Eqs. (1)–(6), respectively. By corresponding the combinations of the antecedents and consequents of a classification rule with the upper part of Table 1, we generate the new confusion matrix shown in the lower part of Table 1. With reference to this matrix, we transform the first terms of Eqs. (1)–(6) to obtain the last terms, which are expressed in terms of the antecedent A, its negation $\neg A$, the consequent C, and its negation $\neg C$. We can express many of the other objective measures as combinations of these basic objective measures.

$$\begin{aligned}\text{Coverage} &= \frac{N_{\mathrm{TP}} + N_{\mathrm{FP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{TN}} + N_{\mathrm{FN}}} \\ &= \frac{N_{A\wedge C} + N_{A\wedge\neg C}}{N_{A\wedge C} + N_{A\wedge\neg C} + N_{\neg A\wedge\neg C} + N_{\neg A\wedge C}} \\ &= \frac{N_A}{N} = P(A) \qquad (1)\end{aligned}$$

Table 1 The upper part represents a confusion matrix that divides a set of instances into TP, FP, TN, and FN. The lower part represents a confusion matrix for a classification rule of *IF A THEN C*, where A is the conjunction of attribute conditions such as $A_1 = a_{14}$ and $A_2 = a_{21}$ and ..., and C is one of the classes of $c_1, c_2, \ldots N_X$ is the number of instances that satisfy X

Fact	Prediction		
	Pos	Neg	Sum
Pos	N_{TP}	N_{FN}	$N_{\mathrm{TP}} + N_{\mathrm{FN}}$
Neg	N_{FP}	N_{TN}	$N_{\mathrm{FP}} + N_{\mathrm{TN}}$
Sum	$N_{\mathrm{TP}} + N_{\mathrm{FP}}$	$N_{\mathrm{FN}} + N_{\mathrm{TN}}$	N

Fact	Prediction		
	A	$\neg A$	Sum
C	$N_{A\wedge C}$	$N_{\neg A\wedge C}$	N_C
$\neg C$	$N_{A\wedge\neg C}$	$N_{\neg A\wedge\neg C}$	$N_{\neg C}$
Sum	N_A	$N_{\neg A}$	N

$$\mathrm{Prevalence} = \frac{N_{\mathrm{TP}} + N_{\mathrm{FN}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{TN}} + N_{\mathrm{FN}}} = \frac{N_{A \wedge C} + N_{\neg A \wedge C}}{N_{A \wedge C} + N_{A \wedge \neg C} + N_{\neg A \wedge \neg C} + N_{\neg A \wedge C}} = \frac{N_C}{N} = P(C) \quad (2)$$

$$\mathrm{Precision} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} = \frac{N_{A \wedge C}}{N_{A \wedge C} + N_{A \wedge \neg C}} = \frac{N_{A \wedge C}}{N_A} = P(C|A) \quad (3)$$

$$\mathrm{Recall} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} = \frac{N_{A \wedge C}}{N_{A \wedge C} + N_{\neg A \wedge C}} = \frac{N_{A \wedge C}}{N_C} = P(A|C) \quad (4)$$

$$\mathrm{Support} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{TN}} + N_{\mathrm{FN}}} = \frac{N_{A \wedge C}}{N_{A \wedge C} + N_{A \wedge \neg C} + N_{\neg A \wedge \neg C} + N_{\neg A \wedge C}} = \frac{N_{A \wedge C}}{N} = P(A \wedge C) \quad (5)$$

$$\text{Uncovered negative} = \frac{N_{\mathrm{TN}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{TN}} + N_{\mathrm{FN}}} = \frac{N_{\neg A \wedge \neg C}}{N_{A \wedge C} + N_{A \wedge \neg C} + N_{\neg A \wedge \neg C} + N_{\neg A \wedge C}} = \frac{N_{\neg A \wedge \neg C}}{N} = P(\neg A \wedge \neg C) \quad (6)$$

We now list the selected uniformed objective measures in our research: coverage (Coverage), prevalence (Prevalence), precision (Precision), recall (Recall), support (Support), uncovered negative (UncNeg), accuracy (Accuracy), specificity (Specificity), lift (Lift), added value (AV), leverage (Leverage), Klösgen's interestingness (KI), relative risk (RR), Brin's interest (BI), Brin's conviction (BC), certainty factor (CF), Jaccard coefficient (Jaccard), f-measure (F-M), odds ratio (OR), Yule's q (Yule's Q), Yule's y (Yule's Y), kappa (Kappa), Gray and Orlowska's interestingness (GOI), collective strength (CST), Gini index (Gini), credibility (Credibility), chi-square measure for one quadrant (χ^2-M1), chi-square measure for four quadrants (χ^2-M4), j-measure (J-M), k-measure (K-M), Yao and Liu's interestingness based on one-way support (YLI1), Yao and Liu's interestingness based on two-way support (YLI2), mutual information (MI), Yao and Zhong's interestingness (YZI), cosine similarity (CSI), Piatetsky—Shapiro's interestingness (PSI), Laplace correction (LC), phi coefficient (ϕ), Gago and Bento's interestingness (GBI), and peculiarity (Peculiarity). In Table 2, these measures are described in detail with their base theories, denominations, abbreviations, literature reference numbers, and definitions.

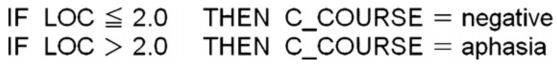
3. Experiment I: evaluation of interestingness measures in the research domain of meningitis

3.1. Experimental conditions

We compare the rule evaluation results obtained by objective measures with those obtained by a medical expert, and examine the performance of the objective measures in estimating the interest of the medical expert in the domain of meningitis research. More specifically, the medical expert evaluates the rules obtained from a clinical dataset on meningitis, with the objective measures also evaluating the same rules. We then examine the agreement degree of these two sets of evaluations. Here, the objective measures shown in Table 2 were used, while both the rules, and also the rule evaluation results of a medical expert, were taken from a medical KDD study on meningitis [9].

In a previous study [9], Hatazawa et al. discovered rules that describe meningitis prognosis on a clinical dataset [24]. They applied the mining system CAMLET [25], whose classification learning criterion was Accuracy to create text-based prognosis classification rules, and obtained evaluation results for the rules from a medical expert. They used two types of classes; binary classes and multivalued classes. An example of a binary class is, "the possible cause is either a virus or bacteria", and while an example of a multivalued class is "what type of bacteria can be the cause". Some rules were discovered that offer the prospect of new medical knowledge. Fig. 1 shows one example of the rules found to be interesting by the medical expert.

The rule evaluation process for the medical expert was as follows: The medical expert medically interpreted each rule, and gave a quality label to each rule. A quality label is either interesting (**I**), not-understandable (**NU**), or not-interesting (**NI**). As a result, the medical expert evaluated 23 rules out



IF LOC ≦ 2.0 THEN C_COURSE = negative
IF LOC > 2.0 THEN C_COURSE = aphasia

Figure 1 An example of rules on meningitis prognosis to which a medical expert showed interest. The upper part suggests that a language disorder may be avoided if the patient is admitted to a hospital within 2 days after losing consciousness. The lower part indicates the converse of contraposition.

Table 2 The objective measures used in the present research. For the cases in which the definition of an objective measure depends on the type of rule, we adopted the definition of the objective measure for classification rules. Also shown are the base theories, the denomination (abbreviation) and [literature reference number] for each objective measure

P	Coverage (Coverage) [1,37]	$P(A)$	P	Prevalence (Prevalence) [38]	$P(C)$
P	Precision (Precision) [1,37]	$P(C \mid A)$	P	Recall (Recall) [1,37]	$P(A \mid C)$
P	Support (Support) [1]	$P(A \wedge C)$	P	Uncovered negative (UncNeg) [5]	$P(\neg A \wedge \neg C)$
P	Accuracy (Accuracy) [1]	$P(A \wedge C) + P(\neg A \wedge \neg C)$	P	Specificity (Specificity) [1]	$P(\neg A \mid \neg C)$
P	Lift (Lift) [39]	$\frac{P(C \mid A)}{P(C)}$	P	Added value (AV) [4]	$P(C \mid A) - P(C)$
P	Leverage (Leverage) [40]	$P(C \mid A) - P(A) \times P(C)$	P	Klösgen's interestingness (KI) [41]	$\sqrt{P(A \wedge C)} \times \{P(C \mid A) - P(C)\}$
P	Relative risk (RR) [42]	$\frac{P(C \mid A)}{P(C \mid \neg A)}$	P	Brin's interest (BI) [43]	$\frac{P(A \wedge C)}{P(A) \times P(C)}$
P	Brin's conviction (BC) [43]	$\frac{P(A) \times P(\neg C)}{P(A \wedge \neg C)}$	P	Certainty factor (CF) [4]	$\frac{P(C \mid A) - P(C)}{1 - P(C)}$
P	Jaccard coefficient (Jaccard) [44,4]	$\frac{P(A \wedge C)}{P(A) + P(C) - P(A \wedge C)}$	P	F-measure (F-M) [28]	$\frac{2 \times P(C \mid A) \times P(A \mid C)}{P(C \mid A) + P(A \mid C)}$
P	Odds ratio (OR) [4]	$\frac{P(A \wedge C) \times P(\neg A \wedge \neg C)}{P(A \wedge \neg C) \times P(\neg A \wedge C)}$	P	Yule's q (Yule's Q) [4]	$\frac{OR - 1}{OR + 1}$
P	Yule's y (Yule's Y) [4]	$\frac{\sqrt{OR} - 1}{\sqrt{OR} + 1}$	P	Kappa (Kappa) [45,4]	$\frac{P(A \wedge C) + P(\neg A \wedge \neg C) - P(A) \times P(C) - P(\neg A) \times P(\neg C)}{1 - P(A) \times P(C) - P(\neg A) \times P(\neg C)}$
P	Gray and Orlowska's interestingness (GOI) [46]	$\left(\left(\frac{P(C \mid A)}{P(A) \times P(C)}\right)^k - 1\right) \times ((P(A) \times P(C))^m),\ k = m$ k, m: coefficients of independency and generality	P	Collective strength (CST) [47,4]	$\frac{P(A \wedge C) + P(\neg A \wedge \neg C)}{P(A) \times P(C) + P(\neg A) \times P(\neg C)} \times \frac{1 - P(A) \times P(C) - P(\neg C) \times P(\neg A)}{1 - P(A \wedge C) - P(\neg A \wedge \neg C)}$
P	Gini index (Gini) [48,4]	$P(A) \times \{P(C \mid A)^2 + P(\neg C \mid A)^2\} + P(\neg A) \times \{P(C \mid \neg A)^2 + P(\neg C \mid \neg A)^2\} - P(C)^2 - P(\neg C)^2$			
PN	Credibility (Credibility) [49]	$\beta_i \times P(C) \times \lvert P(R_i \mid C) - P(R_i) \rvert \times T(R_i)$ $\beta_i = \frac{1}{2 \times P(R_i) \times (1 - P(R_i))}$, $P(R_i)$: occurrence probability of rule R_i $T(R_i)$: threshold value of the number of instances contained in rule R_i	S	Chi-square measure for one quadrant (χ^2- M1) [50,51]	$\sum_{event} \frac{(T_{event} - O_{event})^2}{T_{event}}$, *event*: $A \rightarrow C$ T_{event}: the number of instances contained in *event* (theoretical value) O_{event}: the number of instances contained in *event* (observational value)
S	Chi-square measure for four quadrants (χ^2- M4) [50,51]	The definition is the same as χ^2-M1 But *event*: $A \rightarrow C, A \rightarrow \neg C, \neg A \rightarrow C, \neg A \rightarrow \neg C$			
I	J-measure (J- M) [52]	$P(A) \times (I_{\text{diff}}(C \mid A; C) + I_{\text{diff}}(\neg (C \mid A); \neg C))$ $I_{\text{diff}}(X; Y) = P(X) \times \log_2 \frac{P(X)}{P(Y)}$	I	K-measure (K-M) [6]	$I_{\text{diff}}(C \mid A; C) + I_{\text{diff}}(\neg (C \mid A); \neg C) -$ $I_{\text{diff}}(C \mid A; \neg C) - I_{\text{diff}}(\neg (C \mid A); C)$
I	Yao and Liu's interestingness based on one-way support (YLI1) [53,2]	$P(C \mid A) \times \log_2 \frac{P(A \wedge C)}{P(A) \times P(C)}$			

Table 2 (*Continued*)

I	Yao and Liu's interestingness based on two-way support (YLI2) [53,2]	$P(A \wedge C) \times \log_2 \frac{P(A \wedge C)}{P(A) \times P(C)}$
I	Mutual information (MI) [4]	$\log_2 \frac{P(A \wedge C)}{P(A) \times P(C)}$
I	Yao and Zhong's interestingness (YZI) [2]	$\sum_{X,Y} P(X \wedge Y) \times \log_2 \frac{P(X \wedge Y)}{P(X) \times P(Y)}$, X: A or $\neg A$, Y: C or $\neg C$
N	Cosine similarity (CSI) [4]	$\frac{N(A \wedge C)}{\sqrt{N(A) \times N(C)}}$
N	Piatetsky—Shapiro's interestingness (PSI) [40]	$N(A \wedge C) - \frac{N(A) \times N(C)}{N(U)}$
N	Laplace correction (LC) [4]	$\frac{N(A \wedge C)+1}{N(A)+2}$
N	Phi coefficient (ϕ) [4]	$\frac{N(A \wedge C) \times N(\neg A \wedge \neg C) - N(\neg A \wedge C) \times N(A \wedge \neg C)}{\sqrt{N(A) \times N(C)} \times \sqrt{N(\neg A) \times N(\neg C)}}$
D	Gago and Bento's interestingness (GBI) [54]	$\frac{\sum_{j=1}^{N_R} D(R_i, R_j)}{N_R}$ R_i : ith rule. N_R: the number of rules. $D(R_i, R_j)$: distance between the ith rule and the jth rule obtained based on the degree of duplications of their attributes
D	Peculiarity (Peculiarity) [55]	$\sum_{a_{ij} \in Rule} \frac{PF(a_{ij})}{N}$, $PF(a_{ij}) = \sum_k \lvert a_{ij} - a_{ik} \rvert^{\alpha}$ a_{ij}: the value at ith row at jth column in the data table. N: the number of attributes contained in the rule. α: constant (default value 0.5)

The nomenclature in the theory column is as follows: P, probabilistic magnitude; S, statistic value; I, information amount; N, number of instances; D, distance among rules or among attributes. The nomenclature in the definition row is as follows: U, the universal set of instances; A, the antecedent of a rule; C, the consequent of a rule; $A \wedge C$, the condition in which a rule comprises; $\neg X$, negation of X; $N(X)$, the number of instances that satisfy X; $P(X)$, the occurrence probability of X, namely the ratio of $N(X)/N(U)$; $P(X \wedge Y)$, the co-occurrence probability of $X \wedge Y$, namely the ratio of $N(X \wedge Y)/N(U)$; $P(Y|X)$, the conditional probability of Y after X, namely the ratio of $N(X \wedge Y)/N(X)$.

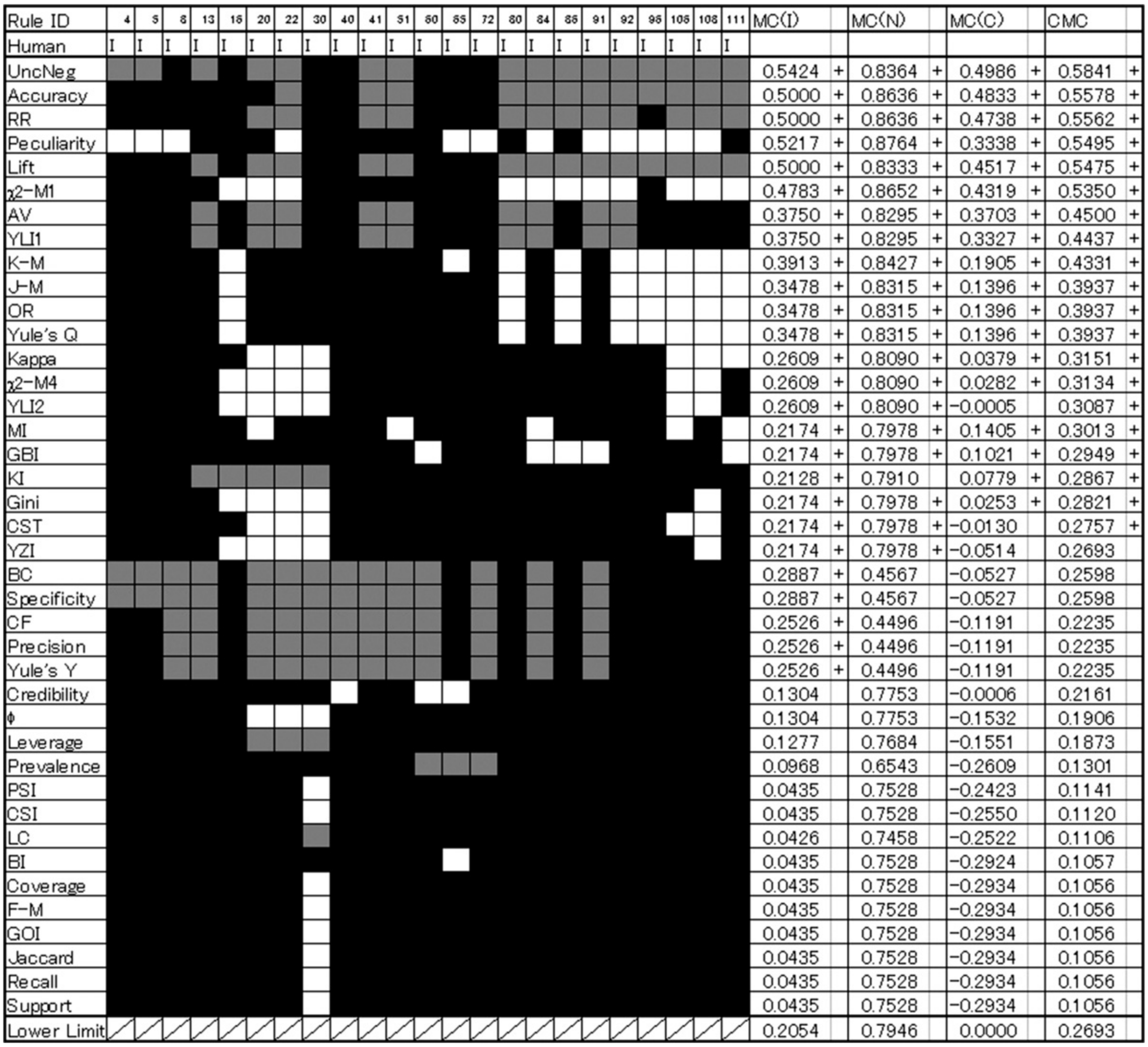
of a total of 112 binary class rules, and 25 rules out of a total of 132 multivalued class rules, as **I**. Here, we briefly note the expertise of the medical expert at the time when the rule evaluation was conducted [9]. The medical expert was a medical doctor and a professor at a national university hospital and had an approximately 10-year career in clinical internal medicine and medical informatics. Furthermore, the medical expert had been the chief evaluator of data-mined rules provided from all research groups in the MEXT Japan project "Discovery Science" [26] for 2 years.

Conversely, for the objective measures, we designed the rule evaluation process as follows: All objective measures evaluate the same rules, and we sort the rules into descending order of evaluation values for each objective measure. Next, according to the number of rules for each quality label allocated by the medical expert, we allocate **I**, **NU**, and **NI** to the rules sorted in descending order.

3.2. Results, analysis, and discussion

Fig. 2 shows the experimental results for binary classes, while Fig. 3 shows those for multivalued classes. Note that we omitted the data of **NI** and **NU** due to space limitations. We examine the extent to which rule evaluations made by each objective measure agree with those made by the medical expert, both in a qualitative and quantitative manner.

In the qualitative analysis, we visualized the degree of agreement to make it easier to understand the general trends. We used the following shading scheme: white indicates that the evaluations agree, gray indicates that the agreement appears to be probabilistic, i.e., to have partly occurred by chance, and black indicates that the evaluations do not agree. Here "probabilistic agreement" refers to cases in which an objective measure gave the same evaluation value to more than m rules where m is the number of rules to which the medical expert gave the same quality label. For example, 23 binary class rules appear interesting to the medical expert, but Specificity gave the same high value, and thus the same quality label, **I**, to 74 rules. Even if the 74 rules include really interesting rules, the really interesting rules will not always be ranked as the top 23 depending on the sorting method used. In Figs. 2 and 3, the greater the number of white cells, the better the estimation performance of the objective measures. The pattern of white, gray, and black



Rule ID	4	5	8	13	16	20	22	30	40	41	51	60	65	72	80	84	86	91	92	96	106	108	111	MC(I)		MC(N)		MC(C)		CMC	
Human	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I								
UncNeg	▨	▨	■	▨	■	▨	▨	■	■	▨	▨	■	■	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	0.5424	+	0.8364	+	0.4986	+	0.5841	+
Accuracy	■	■	■	■	■	■	▨	■	■	▨	▨	■	■	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	0.5000	+	0.8636	+	0.4833	+	0.5578	+
RR	■	■	■	■	■	▨	▨	■	■	▨	▨	■	■	■	▨	▨	▨	▨	▨	■	▨	▨	▨	0.5000	+	0.8636	+	0.4738	+	0.5562	+
Peculiarity	□	□	□	■	■	■	□	■	■	■	■	■	□	□	■	□	■	□	□	□	□	□	■	0.5217	+	0.8764	+	0.3338	+	0.5495	+
Lift	■	■	■	▨	■	▨	▨	■	■	▨	▨	■	■	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	0.5000	+	0.8333	+	0.4517	+	0.5475	+
χ2-M1	■	■	■	■	□	□	□	■	■	■	■	■	■	■	□	□	□	□	□	■	□	□	□	0.4783	+	0.8652	+	0.4319	+	0.5350	+
AV	■	■	■	▨	■	▨	▨	■	■	▨	▨	■	■	■	▨	▨	■	▨	▨	■	■	■	■	0.3750	+	0.8295	+	0.3703	+	0.4500	+
YLI1	■	■	■	▨	■	▨	▨	■	■	▨	▨	■	■	■	▨	▨	■	▨	▨	■	■	■	■	0.3750	+	0.8295	+	0.3327	+	0.4437	+
K-M	■	■	■	■	□	■	■	■	■	■	■	■	□	■	□	■	□	■	□	□	□	□	□	0.3913	+	0.8427	+	0.1905	+	0.4331	+
J-M	■	■	■	■	□	■	■	■	■	■	■	■	■	■	□	■	□	■	□	□	□	□	□	0.3478	+	0.8315	+	0.1396	+	0.3937	+
OR	■	■	■	■	□	■	■	■	■	■	■	■	■	■	□	■	□	■	□	□	□	□	□	0.3478	+	0.8315	+	0.1396	+	0.3937	+
Yule's Q	■	■	■	■	□	■	■	■	■	■	■	■	■	■	□	■	□	■	□	□	□	□	□	0.3478	+	0.8315	+	0.1396	+	0.3937	+
Kappa	■	■	■	■	■	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	0.2609	+	0.8090	+	0.0379	+	0.3151	+
χ2-M4	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	□	□	■	0.2609	+	0.8090	+	0.0282	+	0.3134	+
YLI2	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	□	□	■	0.2609	+	0.8090	+	−0.0005		0.3087	+
MI	■	■	■	■	■	□	■	■	■	■	□	■	■	■	■	□	■	■	■	■	□	■	□	0.2174	+	0.7978	+	0.1405	+	0.3013	+
GBI	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	□	□	□	■	■	■	■	□	0.2174	+	0.7978	+	0.1021	+	0.2949	+
KI	■	■	■	▨	▨	▨	▨	▨	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.2128	+	0.7910		0.0779	+	0.2867	+
Gini	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	0.2174	+	0.7978	+	0.0253	+	0.2821	+
CST	■	■	■	■	■	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	□	□	■	0.2174	+	0.7978	+	−0.0130		0.2757	+
YZI	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	0.2174	+	0.7978	+	−0.0514		0.2693	
BC	▨	▨	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	■	▨	■	▨	■	▨	■	■	■	■	■	0.2887	+	0.4567		−0.0527		0.2598	
Specificity	▨	▨	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	■	▨	■	▨	■	▨	■	■	■	■	■	0.2887	+	0.4567		−0.0527		0.2598	
CF	■	■	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	■	▨	■	▨	■	▨	■	■	■	■	■	0.2526	+	0.4496		−0.1191		0.2235	
Precision	■	■	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	■	▨	■	▨	■	▨	■	■	■	■	■	0.2526	+	0.4496		−0.1191		0.2235	
Yule's Y	■	■	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	■	▨	■	▨	■	▨	■	■	■	■	■	0.2526	+	0.4496		−0.1191		0.2235	
Credibility	■	■	■	■	■	■	■	■	□	■	■	□	□	■	■	■	■	■	■	■	■	■	■	0.1304		0.7753		−0.0006		0.2161	
φ	■	■	■	■	■	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.1304		0.7753		−0.1532		0.1906	
Leverage	■	■	■	■	■	▨	▨	▨	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.1277		0.7684		−0.1551		0.1873	
Prevalence	■	■	■	■	■	■	■	■	■	■	■	▨	▨	▨	■	■	■	■	■	■	■	■	■	0.0968		0.6543		−0.2609		0.1301	
PSI	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0435		0.7528		−0.2423		0.1141	
CSI	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0435		0.7528		−0.2550		0.1120	
LC	■	■	■	■	■	■	■	▨	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0426		0.7458		−0.2522		0.1106	
BI	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	0.0435		0.7528		−0.2924		0.1057	
Coverage	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0435		0.7528		−0.2934		0.1056	
F-M	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0435		0.7528		−0.2934		0.1056	
GOI	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0435		0.7528		−0.2934		0.1056	
Jaccard	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0435		0.7528		−0.2934		0.1056	
Recall	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0435		0.7528		−0.2934		0.1056	
Support	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0435		0.7528		−0.2934		0.1056	
Lower Limit	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	0.2054		0.7946		0.0000		0.2693	

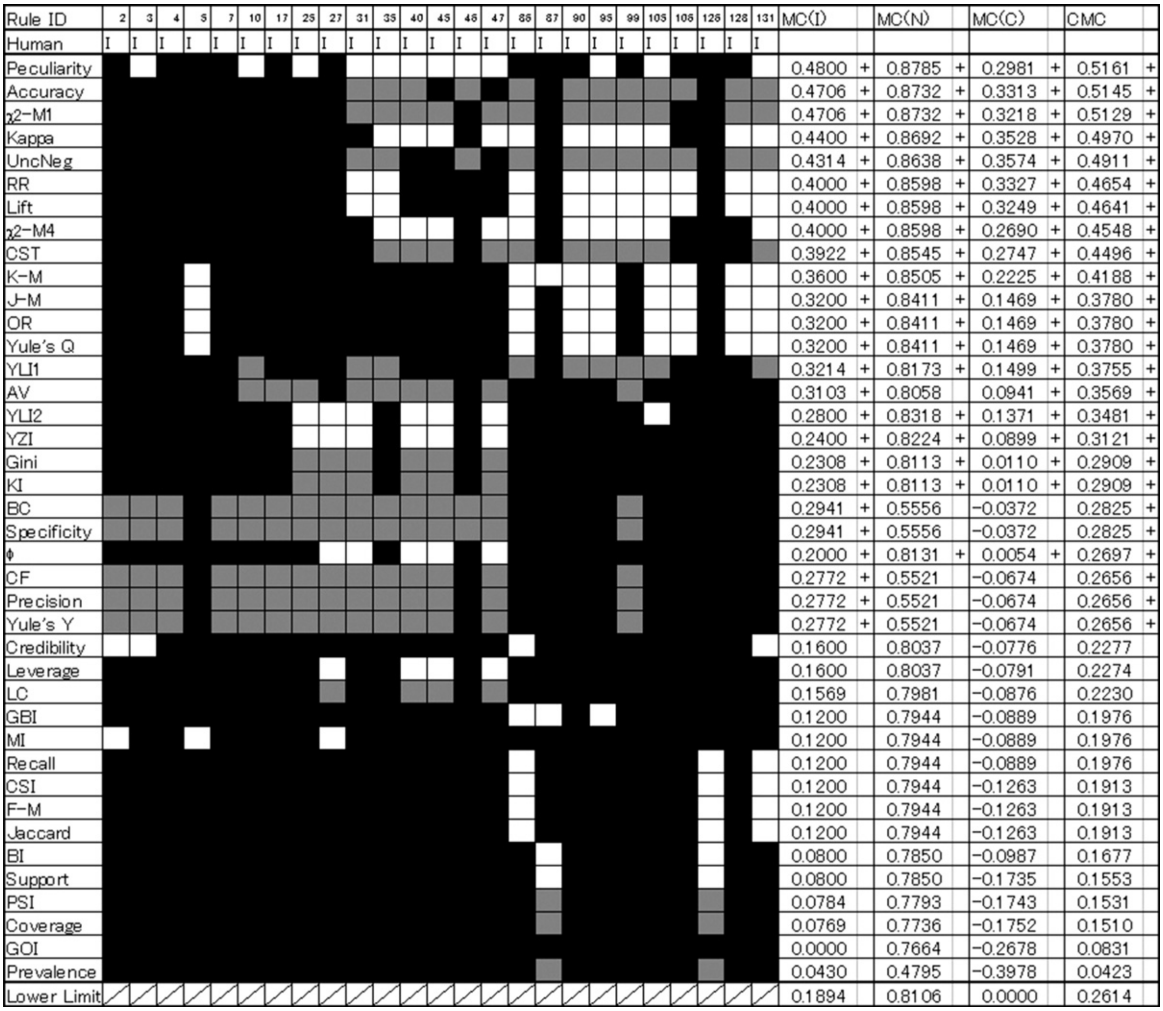
Figure 2 The results of Experiment I for binary classes. Each column represents a rule, while each row represents an objective measure.

cells in each row indicates the each measure's characteristics of estimating the interest of the medical expert.

For the quantitative analysis, we defined the criteria of evaluation, which we call meta-criteria, as follows: receiver-operating characteristics (ROC) curves and Precision—Recall (PR) curves are generally used to evaluate the performance of information retrieval and machine learning [27]. In the present research, we wish to ascertain how precisely and recallably objective measures predict all the rules that a medical expert would wish to see. Thus, we consider whether the area under the curve (AUC) of a PR curve satisfies the meta-criteria.

An objective measure can have a PR curve for each quality label; **I**, **NU**, and **NI**, and also every combination of these labels, such as **NU** ∨ **NI**. In order to draw a PR curve, it is necessary to sort the rules into descending order of evaluation values given by the objective measure, to determine the threshold value of the number of rules that provides the boundary between the focused quality label and the neighbor quality label, and to calculate the values of Precision and Recall parametrically changing the threshold value. We can then draw a curve by plotting the points of (Recall, Precision). Fig. 4 is an example of a real PR curve obtained in this way.

The interpolation method for the PR curves and the calculation method of AUC are still under discussion because Precision and Recall are not independent [27]. In addition, the present research is an attempt to make use of the estimation results obtained by the objective measures to support a user in discovering new knowledge. The number of rules should therefore be limited so as to limit the fatigue of the user. For example, if Precision and Recall take their maximum values on the PR curve of **I** when the threshold value of the number of rules is



Rule ID	2	3	4	5	7	10	17	25	27	31	35	40	45	46	47	86	87	90	95	99	105	106	126	128	131	MC(I)		MC(N)		MC(C)		CMC	
Human	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I								
Peculiarity	■	□	■	■	■	□	■	□	■	□	□	□	□	□	□	■	■	■	□	■	□	■	■	■	□	0.4800	+	0.8785	+	0.2981	+	0.5161	+
Accuracy	■	■	■	■	■	■	■	■	■	▨	▨	▨	■	▨	■	▨	■	▨	▨	▨	▨	▨	■	▨	▨	0.4706	+	0.8732	+	0.3313	+	0.5145	+
χ2−M1	■	■	■	■	■	■	■	■	■	▨	▨	▨	▨	■	▨	▨	■	▨	▨	▨	▨	■	■	▨	▨	0.4706	+	0.8732	+	0.3218	+	0.5129	+
Kappa	■	■	■	■	■	■	■	■	■	■	□	□	□	■	□	□	■	□	□	□	□	■	■	□	□	0.4400	+	0.8692	+	0.3528	+	0.4970	+
UncNeg	■	■	■	■	■	■	■	■	■	▨	▨	■	■	▨	■	▨	■	▨	▨	▨	▨	▨	■	▨	▨	0.4314	+	0.8638	+	0.3574	+	0.4911	+
RR	■	■	■	■	■	■	■	■	■	□	□	■	■	■	■	□	■	□	□	□	□	□	■	□	□	0.4000	+	0.8598	+	0.3327	+	0.4654	+
Lift	■	■	■	■	■	■	■	■	■	□	□	■	■	■	■	□	■	□	□	□	□	□	■	□	□	0.4000	+	0.8598	+	0.3249	+	0.4641	+
χ2−M4	■	■	■	■	■	■	■	■	■	■	□	□	□	■	□	□	■	□	□	□	□	■	■	■	□	0.4000	+	0.8598	+	0.2690	+	0.4548	+
CST	■	■	■	■	■	■	■	■	■	■	▨	▨	▨	■	▨	▨	■	▨	▨	▨	▨	■	■	■	▨	0.3922	+	0.8545	+	0.2747	+	0.4496	+
K−M	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	■	□	□	■	□	□	0.3600	+	0.8505	+	0.2225	+	0.4188	+
J−M	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	□	■	□	□	■	□	□	■	□	□	0.3200	+	0.8411	+	0.1469	+	0.3780	+
OR	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	□	■	□	□	■	□	□	■	□	□	0.3200	+	0.8411	+	0.1469	+	0.3780	+
Yule's Q	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	□	■	□	□	■	□	□	■	□	□	0.3200	+	0.8411	+	0.1469	+	0.3780	+
YLI1	■	■	■	■	■	▨	■	■	■	▨	▨	■	■	■	■	▨	■	▨	▨	▨	▨	■	■	■	▨	0.3214	+	0.8173	+	0.1499	+	0.3755	+
AV	■	■	■	■	■	▨	▨	▨	■	▨	▨	▨	▨	■	▨	■	■	■	■	▨	■	■	■	■	■	0.3103	+	0.8058		0.0941	+	0.3569	+
YLI2	■	■	■	■	■	■	■	□	□	□	■	□	□	■	□	■	■	■	■	■	□	■	■	■	■	0.2800	+	0.8318	+	0.1371	+	0.3481	+
YZI	■	■	■	■	■	■	■	□	□	□	■	□	□	■	□	■	■	■	■	■	■	■	■	■	■	0.2400	+	0.8224	+	0.0899	+	0.3121	+
Gini	■	■	■	■	■	■	■	▨	▨	▨	■	▨	▨	■	▨	■	■	■	■	■	■	■	■	■	■	0.2308	+	0.8113	+	0.0110	+	0.2909	+
KI	■	■	■	■	■	■	■	▨	▨	▨	■	▨	▨	■	▨	■	■	■	■	■	■	■	■	■	■	0.2308	+	0.8113	+	0.0110	+	0.2909	+
BC	▨	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	■	■	■	▨	■	■	■	■	■	0.2941	+	0.5556		−0.0372		0.2825	+
Specificity	▨	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	■	■	■	▨	■	■	■	■	■	0.2941	+	0.5556		−0.0372		0.2825	+
ϕ	■	■	■	■	■	■	■	■	□	□	■	□	□	■	□	■	■	■	■	■	■	■	■	■	■	0.2000	+	0.8131	+	0.0054	+	0.2697	+
CF	▨	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	▨	■	■	■	■	▨	■	■	■	■	■	0.2772	+	0.5521		−0.0674		0.2656	+
Precision	▨	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	▨	■	■	■	■	▨	■	■	■	■	■	0.2772	+	0.5521		−0.0674		0.2656	+
Yule's Y	▨	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	▨	■	■	■	■	▨	■	■	■	■	■	0.2772	+	0.5521		−0.0674		0.2656	+
Credibility	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	□	0.1600		0.8037		−0.0776		0.2277	
Leverage	■	■	■	■	■	■	■	■	□	■	■	□	□	■	□	■	■	■	■	■	■	■	■	■	■	0.1600		0.8037		−0.0791		0.2274	
LC	■	■	■	■	■	■	■	■	▨	■	■	▨	▨	■	▨	■	■	■	■	■	■	■	■	■	■	0.1569		0.7981		−0.0876		0.2230	
GBI	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	■	□	■	■	■	■	■	■	0.1200		0.7944		−0.0889		0.1976	
MI	□	■	■	□	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.1200		0.7944		−0.0889		0.1976	
Recall	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	□	■	□	0.1200		0.7944		−0.0889		0.1976	
CSI	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	□	■	□	0.1200		0.7944		−0.1263		0.1913	
F−M	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	□	■	□	0.1200		0.7944		−0.1263		0.1913	
Jaccard	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	□	■	□	0.1200		0.7944		−0.1263		0.1913	
BI	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	□	■	■	0.0800		0.7850		−0.0987		0.1677	
Support	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	□	■	■	0.0800		0.7850		−0.1735		0.1553	
PSI	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▨	■	■	■	■	■	▨	■	■	0.0784		0.7793		−0.1743		0.1531	
Coverage	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▨	■	■	■	■	■	▨	■	■	0.0769		0.7736		−0.1752		0.1510	
GOI	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0.0000		0.7664		−0.2678		0.0831	
Prevalence	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▨	■	■	■	■	■	▨	■	■	0.0430		0.4795		−0.3978		0.0423	
Lower Limit	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	0.1894		0.8106		0.0000		0.2614	

Figure 3 The results of Experiment I for multivalued classes. The configuration is the same as Fig. 2.

100, then objective measures will show 100 rules as interesting to the user. However, 100 rules will be too much for the user to evaluate, and this estimation cannot be said to be the best even though Precision and Recall are at their maximum values. It is difficult for the AUC to reflect such a restriction to the number of rules, since the AUC of a PR curve is given by the integration of all points of (Recall, Precision).

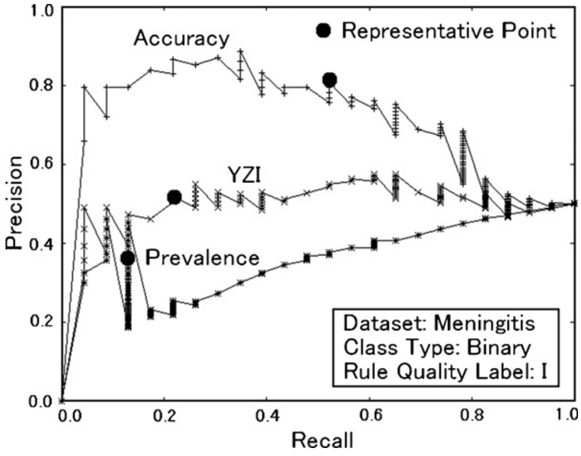


Figure 4 Examples of PR curves for the quality label I.

With respect to the above discussion, we do not adopt the AUC of a PR curve here. We regard the point (Recall, Precision) on a PR curve for a given quality label, whose threshold value is the number of rules with this quality label given by a medical expert, as a representative point that yields the best estimation performance. Furthermore, we adopt the harmonic mean of Precision and Recall, i.e., F-M [28], as the meta-criterion. For example, if a medical expert gave I to 20 rules, then we allocate I to the top 20 rules ranked according to the evaluation values obtained by an objective measure, and calculate F-M using the point (Recall, Precision) whose threshold value is 20.

We call this F-M the Precision—Recall-based (PR-based) meta-criteria and formalize it as follows. Eq. (7) describes the rule set R_{label}, which is the set of all rules given the quality label *label* by a medical expert, while Eq. (8) describes the rule set $\hat{R}_{label}$, which is the set of all rules given the quality label *label* by an objective measure. *R* is the set of all rules, *r* denotes one of these rules, $L_H(r)$ is the quality label given to rule *r* by the medical expert, and $L_M(r)$ is the quality label given to rule *r* by the objective measure. The rule set X_{label} expressed by Eq. (9) denotes those rules that have been given the quality label *label* by both the medical expert and the objective measure. Thus, the PR-based meta-criterion for quality label *label* is given by Eq. (10). Generally, it is more important to differentiate interesting rules from the other rules than to differentiate not-understandable rules from not-interesting rules. We therefore apply Eq. (10) to **I** and **NU** ∨ **NI** and name the application results MC(I) and MC(N), respectively.

$$R_{label} = \{r | r \in R \wedge L_H(r) = label\} \tag{7}$$

$$\hat{R}_{label} = \{r | r \in R \wedge L_M(r) = label\} \tag{8}$$

$$X_{label} = \{r | r \in R_{label} \wedge r \in \hat{R}_{label}\} \tag{9}$$

$$\mathrm{MC}(label) = \frac{(2(|X_{label}|/|R_{label}|)(|X_{label}|/|\hat{R}_{label}|))}{(|X_{label}|/|R_{label}|) + (|X_{label}|/|\hat{R}_{label}|)} \tag{10}$$

We add a correlation coefficient between quality labels given by a medical expert and those given by an objective measure in the meta-criteria because the order of quality labels for all rules should also be considered. Eq. (11) defines the correlation-based meta-criterion based on Spearman's rank correlation coefficient with the adjustment of ties [29]. The quantity $\mathrm{RANK}_H(r)$ represents the rank of rule *r* in the rule set R_{label}, which is the set of rules given quality label *label* by the medical expert. The quantity $\mathrm{RANK}_M(r)$ represents the rank of rule *r* in the rule set $\hat{R}_{label}$, which is the set of rules given quality label *label* by the objective measure. *S* and *K* are used to simplify Eq. (11). N_R is the number of all rules, T_H is the term of the adjustment of ties for R_{label}, and T_M is the corresponding term for $\hat{R}_{label}$.

$$\mathrm{MC}(C) = \frac{K - S}{\sqrt{K - 2T_H}\sqrt{K - 2T_M}} \tag{11}$$

where

$$S = \sum_{\mathbf{I},\mathbf{NU},\mathbf{NI}} (\mathrm{RANK}_H(r) - \mathrm{RANK}_M(r))^2 - T_H - T_M,$$

$$K = N_R(N_R^2 - 1)/6,$$

$$T_H = \sum N_{T_H}(N_{T_H}^2 - 1)/12,$$

and

$$T_M = \sum N_{T_M}(N_{T_M}^2 - 1)/12.$$

We introduce a comprehensive meta-criterion CMC defined by Eq. (12), which combines the PR-based meta-criteria MC(I) and MC(N), and the correlation-based meta-criterion MC(C), considering two extreme conditions, which are explained in a later paragraph. W_i^{equal} and W_i^{inequal} represent the weight for the *i*th meta-criterion under two extreme conditions, respectively. Eq. (12) averages the weighted sum of meta-criteria $\mathrm{CMC}^{\mathrm{equal}}$ and $\mathrm{CMC}^{\mathrm{inequal}}$ between the conditions. The purpose of this averaging process is to approximate the true values of weights and the true value of CMC that are unknowable but surely exist between two extreme conditions.

For the purposes of KDD, there are not a few cases in which it is more important to find even just one valuable rule in the real application domain than to precisely model the dataset with the full set of rules. In order to concentrate on extracting a few valuable rules, it is sometimes necessary to sacrifice some modeling accuracy. Therefore, the estimation performance for interesting rules are more important than those for not-understandable or not-interesting rules, or the rank of the quality labels for all the rules. The specific values of meta-criterion weights depend on the application domains and users. However, we should determine the weights on MC(I), MC(N), and MC(C) through an arithmetically reproducible procedure. We then consider two extreme conditions that logically offer the values of meta-criterion weights and define the comprehensive meta-criterion CMC as shown in Eq. (12). W_i^{equal} represents the weight for each meta-criterion under the condition in which MC(I), MC(N), and MC(C) have the same priority. W_I^{equal}, W_N^{equal}, and W_C^{equal} have the normalized values in the ratio of 1:1:1. W_i^{inequal} represents the condition under which the highest priority is placed upon MC(I). W_I^{inequal}, W_N^{inequal}, and W_C^{inequal} have normalized values in the ratio of 1:0:0.

$$\begin{aligned}\mathrm{CMC} &= \frac{1}{2}\{\mathrm{CMC}^{\mathrm{equal}} + \mathrm{CMC}^{\mathrm{inequal}}\}\\ &= \frac{1}{2}\left\{\sum_{i=I,N,C} W_i^{\mathrm{equal}} \times \mathrm{MC}(i) + \sum_{i=I,N,C} W_i^{\mathrm{inequal}} \times \mathrm{MC}(i)\right\}\end{aligned} \tag{12}$$

where $W_I^{\mathrm{equal}} = 1/(1+1+1)$, $W_N^{\mathrm{equal}} = 1/\ (1+1+1)$, $W_C^{\mathrm{equal}} = 1/(1+1+1)$, $W_I^{\mathrm{inequal}} = 1/\ \ (1+0+0)$, $W_N^{\mathrm{inequal}} = 0/(1+0+0)$, and $W_C^{\mathrm{inequal}} = 0/(1+0+0)$.

There is another issue that should be considered. Is the comprehensive meta-criterion CMC robust to the values of the meta-criterion weights? Instead of the true values of weights that are domain-dependent, user-dependent, and frequently unknowable, the values logically derived from two extreme conditions are used for the calculation of CMC. The true values of the weights and the true value of CMC exist between the values of weights and the value of CMC under one extreme condition and those under the other extreme condition. If the value of CMC calculated by Eq. (12) does not differ greatly from the values of CMC under the two extreme conditions, the calculated value probably reflects the true value of CMC with the true values of weights. We therefore examine the robustness of CMC, which can be interpreted as the robustness of the rank of objective measures provided with the values of CMC, by comparing CMC with CMC$^{\mathrm{equal}}$ and CMC$^{\mathrm{inequal}}$. Concretely speaking, we estimate the robustness of CMC based on the following metrics: the number of objective measures maintained their high/low ranks and the rank correlation coefficient of objective measures between the condition of CMC and the conditions of CMC$^{\mathrm{equal}}$ and CMC$^{\mathrm{inequal}}$.

It is not easy to precisely evaluate the absolute usefulness of objective measures because the absolute usefulness depends on the application domain. From an engineering viewpoint it will be more useful to evaluate a generic usefulness for medical domains than to evaluate a specific usefulness only to meningitis. Thus, we estimate the lower boundaries of meta-criteria assuming a random allocation of quality labels to rules, and evaluate the relative usefulness of objective measures to the lower boundaries. We explain the lower boundaries of PR-based meta-criteria MC(I) and MC(N) in the next paragraph. We set the lower boundary of the correlation-based meta-criterion MC(C) to zero, which means no correlation. We can calculate the lower boundary of the comprehensive meta-criterion CMC by substituting the values of the lower boundaries of meta-criteria MC(I), MC(N), and MC(C) into Eq. (12).

Eq. (13) defines the lower boundary of a PR-based meta-criterion. The notation ${}_a\mathbf{C}_b$ denotes the number of combinations when choosing b rules out of a rules given by $(a!/b!(a-b)!)$. N_R is the total number of rules, and N_{label} is the number of rules to which a medical expert gave the quality label $label$. This is also the number of rules given the quality label $label$ by an objective measure. The quantity ${}_{N_R}\mathbf{C}_{N_{label}}$ represents the total number of estimation events that are generated by the random allocation of quality labels. The first term in Eq. (13) is the probability that the estimation succeeds for i rules out of N_{label} rules. The second term is the value of a meta-criterion. Thus, Eq. (13) is the expectation of the meta-criterion value of a random allocation. We can thus obtain the lower boundary of MC(I) for label **I**, while we can obtain the lower boundary of MC(N) for the disjunction of labels **NU** $\vee$ **NI**.

Lower limit($label$)

$$= \sum_{i=0}^{N_{label}} \frac{{}_{N_{label}}\mathbf{C}_i\ {}_{(N_R - N_{label})}\mathbf{C}_{(N_{label}-i)}}{{}_{N_R}\mathbf{C}_{N_{label}}} \times \left(2\frac{i}{N_{label}}\frac{i}{N_{label}}\right) \Big/ \left(\frac{i}{N_{label}} + \frac{i}{N_{label}}\right) \qquad (13)$$

In the quantitative analysis, we calculated the lower boundaries and actual values of the meta-criteria and the comprehensive meta-criterion for each objective measure according to the calculation described above. These values appear on the right-hand side of Figs. 2 and 3. We sorted the objective measures into descending order of the values of the comprehensive meta-criterion. Those values of meta-criteria and comprehensive meta-criterion that are larger than the lower boundaries have a plus sign (+) appended to the right.

With regard to CMC in Fig. 2, 20 objective measures out of 40 (50%) obtained a +, and the estimation performance of these measures is better than that of a random allocation. The top five are UncNeg, Accuracy, RR, Peculiarity, and Lift. It is confirmed that their rank orders are robust to the values of meta-criterion weights. Among the objective measures ranked in the top five for CMC, four and five measures are also ranked in the top five for CMC$^{\mathrm{equal}}$ and CMC$^{\mathrm{inequal}}$, respectively. The rank correlation coefficients between the rank for CMC and the ranks for CMC$^{\mathrm{equal}}$ and CMC$^{\mathrm{inequal}}$ are 0.9720 and 0.9444, respectively. The CMC of top five objective measures are more than 2.0 times the lower boundary, that is, these objective measures are at least 2.0 times more likely to correctly identify interesting rules than a random allocation. It is thus suggested that these measures may be effective in estimating the interest of a medical expert.

The bottom five are Support, Recall, Jaccard, GOI, F-M, and Coverage. It is confirmed that their rank orders are robust to the values of meta-criterion weights. Among the objective measures ranked in the bottom five for CMC (six measures including ties are ranked in), six and six measures are also ranked in the bottom five for CMC$^{\mathrm{equal}}$ and CMC$^{\mathrm{inequal}}$, respectively. The CMC of bottom five objective measures are less than 0.4 times the lower boundary, that is, these objective measures are at most 0.4 times as likely to correctly identify interesting rules as a random allocation. It is thus suggested that these measures may not be effective in estimating the interest of a medical expert, although this cannot be determined with complete certainty under the limited conditions of the present experiment.

With regard to CMC in Fig. 3, 25 objective measures out of 40 (63%) obtained a +, and the estimation performance of these measures is better than that of a random allocation. The top five are Peculiarity, Accuracy, χ^2-M1, Kappa, and UncNeg. It is confirmed that their rank orders are robust to the values of meta-criterion weights. Among the objective measures ranked in the top five for CMC, five and five measures are also ranked in the top five for $\text{CMC}^{\text{equal}}$ and $\text{CMC}^{\text{inequal}}$, respectively. The rank correlation coefficients between the rank for CMC and the ranks for $\text{CMC}^{\text{equal}}$ and $\text{CMC}^{\text{inequal}}$ are 0.9423 and 0.9796, respectively. The CMC of the top five objective measures are more than 1.9 times the lower boundary, and so it is suggested that they may be effective in estimating the interest of a medical expert.

The bottom five are Prevalence, GOI, Coverage, PSI, and Support. It is confirmed that their rank orders are robust to the values of meta-criterion weights. Among the objective measures ranked in the bottom five for CMC, five and five measures are also ranked in the bottom five for $\text{CMC}^{\text{equal}}$ and $\text{CMC}^{\text{inequal}}$, respectively. The CMC of bottom five objective measures are less than 0.6 times the lower boundary, and so it is suggested that they may not be effective in estimating the interest of a medical expert.

Next, we discuss the comprehensive tendency of objective measures for binary and multivalued classes with respect to Figs. 2 and 3, and Table 3. UncNeg, Accuracy, Peculiarity, χ^2-M1, and RR showed a high estimation performance for both binary and multivalued classes, and this increases the likelihood that they are useful. Conversely, Prevalence, GOI, Coverage, Support, and PSI showed a low performance for both classes, and this increases the likelihood that they are not useful. From Figs. 2 and 3, it may be observed that objective measures of similar rank exhibit similar patterns of white, gray, and black. This indicates that the rule evaluation of the medical expert was consistent throughout the evaluation process, and thus agrees with the medical expert's comment that he preliminarily generated a fixed hypothesis and evaluated the rules based on the hypothesis.

Summarizing the above discussion, we have that for a situation in which a medical expert is proceeding from a fixed hypothesis in seeking to discover medical knowledge, the top ranked objective measures, namely UncNeg, Accuracy, Peculiarity, χ^2-M1, and RR can be considered to be useful to recommend possibly interesting rules to the medical expert.

We observed the process of a medical expert evaluating rules, and realized that the process comprises two phases. The first phase is the generation of a hypothesis. In the phase of hypothesis generation, a medical expert considers a variety of different ideas for hypotheses, and selects a promising

Table 3 Summary of the results of Experiments I and II. A plus symbol '+' to the right of a CMC value indicates that the value is larger than the lower boundary of CMC

Experiment I: meningitis			Experiment II: hepatitis			Overall		
Rank	Measure	Averaged CMC	Rank	Measure	Averaged CMC	Rank	Measure	Averaged CMC
1	UncNeg	0.5376 +	1	Accuracy	0.6008 +	1	Accuracy	0.5685 +
2	Accuracy	0.5361 +	2	RR	0.5964 +	2	χ^2-M1	0.5540 +
3	Peculiarity	0.5328 +	3	χ^2-M1	0.5840 +	3	RR	0.5536 +
4	χ^2-M1	0.5240 +	4	F-M	0.5725 +	4	UncNeg	0.5521 +
5	RR	0.5108 +	4	Jaccard	0.5725 +	5	Peculiarity	0.5412 +
6	Lift	0.5058 +	6	Lift	0.5700 +	6	Lift	0.5379 +
7	K-M	0.4259 +	7	Recall	0.5697 +	7	Kappa	0.4837 +
8	YLI1	0.4096 +	8	CST	0.5688 +	8	CST	0.4657 +
9	Kappa	0.4060 +	9	UncNeg	0.5665 +	9	YLI1	0.4262 +
10	AV	0.4034 +	10	Kappa	0.5614 +	10	J-M	0.4136 +
31	CSI	0.1517	31	CF	0.3125	31	Credibility	0.2620
32	Recall	0.1516	32	Precision	0.3102	32	PSI	0.2591
33	F-M	0.1485	33	Credibility	0.3021	33	MI	0.2567
33	Jaccard	0.1485	34	KI	0.2659	34	Support	0.2395
35	BI	0.1367	35	MI	0.2640	35	Coverage	0.2233
36	PSI	0.1336	36	ϕ	0.2142	36	ϕ	0.2222
37	Support	0.1304	37	GOI	0.1755	37	LC	0.1706
38	Coverage	0.1283	38	LC	0.1744	38	BI	0.1496
39	GOI	0.0943	39	BI	0.1625	39	GOI	0.1349
40	Prevalence	0.0862	40	Prevalence	0.1085	40	Prevalence	0.0973
	Lower limit	0.2654		Lower limit	0.3619		Lower limit	0.3136

hypothesis at an abstract level. The second phase is the confirmation of the hypothesis. In the hypothesis confirmation phase, the medical expert confirms the reliability of the hypothesis at a concrete level. It is implied that a sequential arrangement of these two phases, that is hypothesis generation and hypothesis confirmation, can support the evaluation process of a medical expert, and enhance medical KDD.

4. Experiment II: evaluation of interestingness measures in the research domain of hepatitis

4.1. Experimental conditions

Following a procedure similar to that of Experiment I, we conducted Experiment II in the domain of hepatitis research. We used the rules and the rule evaluation results by a medical expert from a medical KDD study on hepatitis [10].

In a previous study [10], Ohsaki et al. discovered rules that describe the change of hepatitis symptoms in a clinical dataset [30]. They adopted a typical time-series data mining framework [31]. This framework cuts out subsequences from the time sequence of medical test results via a time window, and extracts representative patterns by clustering. It then forms graph-based prognosis classification rules, which predict the present combination of representative patterns based on the past combination of representative patterns by classification learning. They applied EM and K-Means algorithms to clustering using Euclidean distance for distance calculation, and C5.0, which is the commercial version of C4.5, to classification learning using information gain ratio for data division.

Ohsaki et al. divided the mining process into the two phases of hypothesis generation and hypothesis confirmation. A medical expert (not the medical expert of Experiment I) evaluated the obtained rules. The first mining, corresponding to the hypothesis generation phase, derived some rules that supported the medical expert to generate a new hypothesis on hepatitis symptoms. The second mining, corresponding to the hypothesis confirmation phase, derived some rules that supported the medical expert to raise the reliability of the hypothesis. Fig. 5 shows the examples of rules that contributed to hypothesis generation and confirmation. The medical expert was interested in these rules, and judged that they are medically valuable.

The evaluation process by the medical expert was the same as that in Experiment I. Here, we briefly note the expertise of the medical expert at the time of rule evaluation [10]. The medical expert was a medical doctor and an associate professor at a national university hospital and had a 7-year career in clinical internal medicine and medical informatics. He had been the chief evaluator of data-mined rules provided from all research groups in the MEXT Japan project "Implementation of Active Mining in the Era of Information Flood" [32] for 3 years. As a result of rule evaluation, the medical expert evaluated 12 out of 30 rules in the first mining, and 8 out of 21 rules in the second mining, as **I**. The evaluation process of the objective measures is also the same as that in Experiment I.

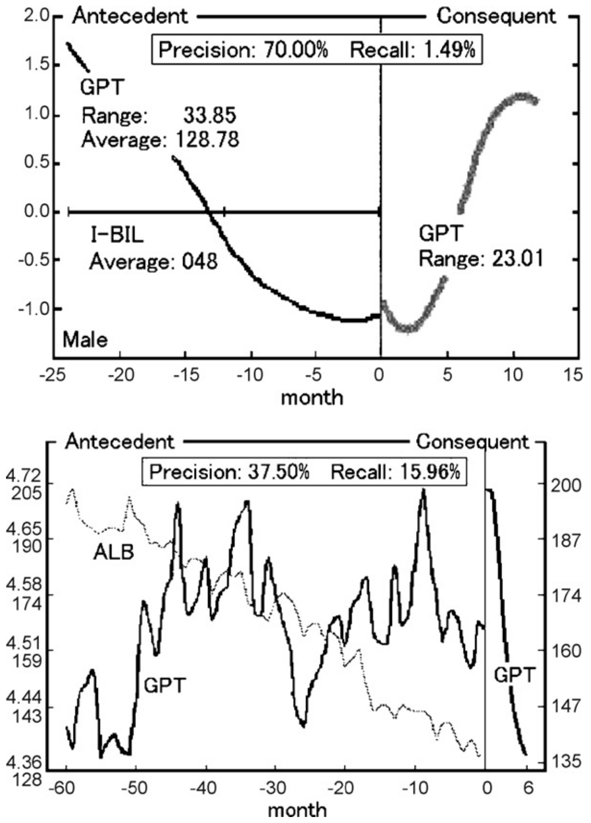
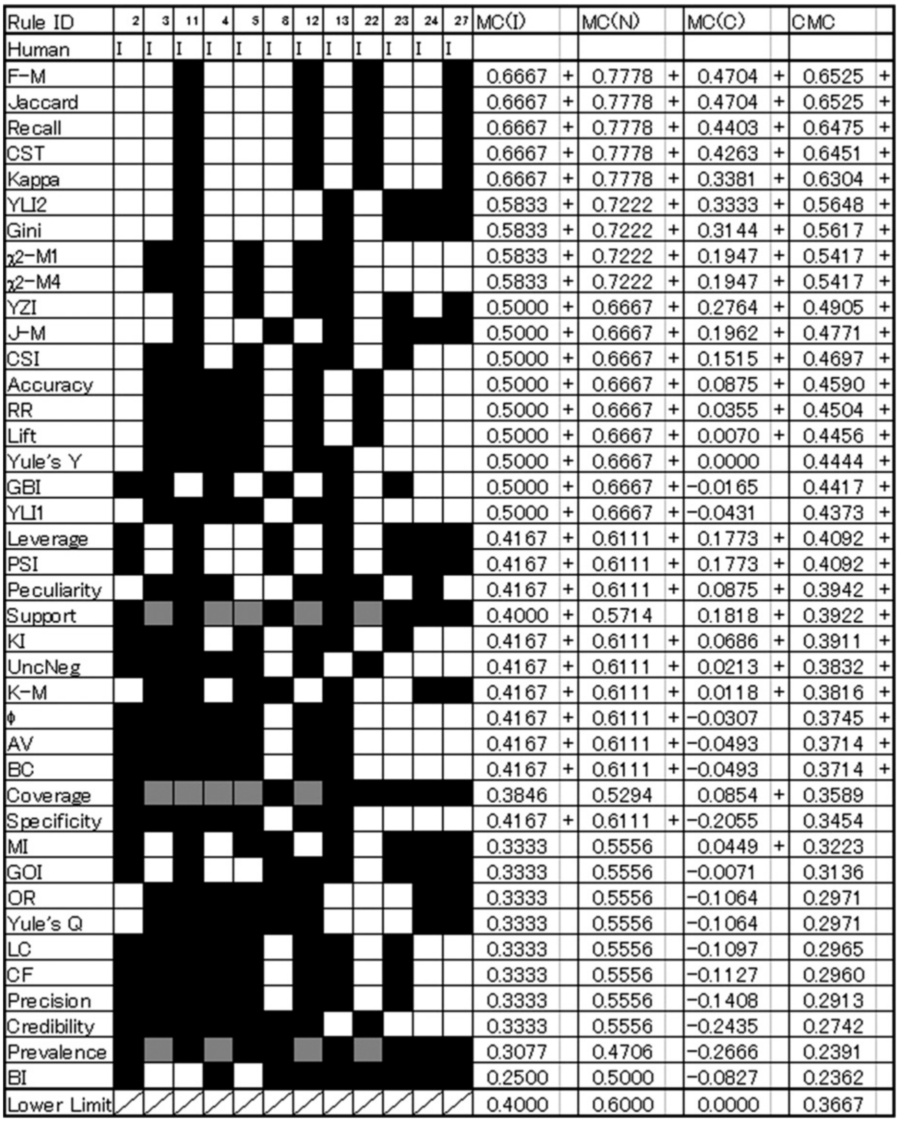


Figure 5 An example of rules on hepatitis prognosis to which a medical expert showed interest. The upper part of the figure corresponds to the first mining. This shows that the value of GPT, which is a necessary medical test to understand the symptoms of hepatitis, changes over a period of around 3 years. The lower part of the figure corresponds to the second mining, and more clearly shows the 3-year-period described above.

4.2. Results, analysis, and discussion

Fig. 6 shows the experimental results for the first mining, while Fig. 7 shows those for the second mining. Note that we omitted the data of **NI** and **NU** due to space limitations. We examine the extent to which rule evaluations made by each objective measure agree with those made by the medical



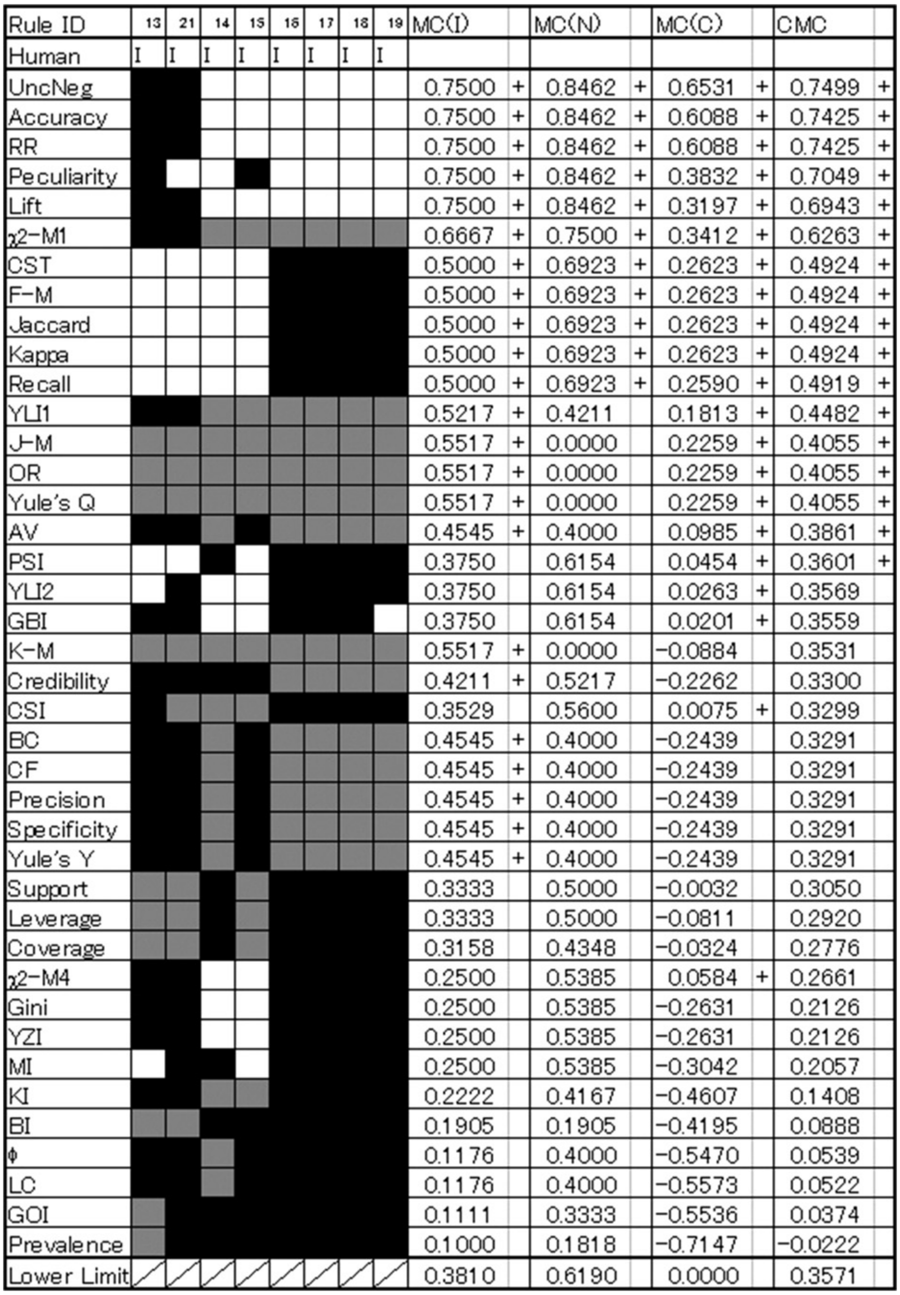
Rule ID	2	3	11	4	5	8	12	13	22	23	24	27	MC(I)		MC(N)		MC(C)		CMC	
Human	I	I	I	I	I	I	I	I	I	I	I	I								
F−M	□	□	■	□	□	□	■	□	■	□	□	■	0.6667	+	0.7778	+	0.4704	+	0.6525	+
Jaccard	□	□	■	□	□	□	■	□	■	□	□	■	0.6667	+	0.7778	+	0.4704	+	0.6525	+
Recall	□	□	■	□	□	□	■	□	■	□	□	■	0.6667	+	0.7778	+	0.4403	+	0.6475	+
CST	□	□	■	□	□	□	■	□	■	□	□	■	0.6667	+	0.7778	+	0.4263	+	0.6451	+
Kappa	□	□	■	□	□	□	■	□	■	□	□	■	0.6667	+	0.7778	+	0.3381	+	0.6304	+
YLI2	□	□	■	□	□	□	□	■	□	■	■	■	0.5833	+	0.7222	+	0.3333	+	0.5648	+
Gini	□	□	■	□	□	□	□	■	□	■	■	■	0.5833	+	0.7222	+	0.3144	+	0.5617	+
χ2−M1	□	■	■	□	■	□	■	■	□	□	□	□	0.5833	+	0.7222	+	0.1947	+	0.5417	+
χ2−M4	□	■	■	□	■	□	■	■	□	□	□	□	0.5833	+	0.7222	+	0.1947	+	0.5417	+
YZI	□	□	■	□	■	□	■	■	□	■	□	■	0.5000	+	0.6667	+	0.2764	+	0.4905	+
J−M	□	□	■	□	□	■	□	■	□	■	■	■	0.5000	+	0.6667	+	0.1962	+	0.4771	+
CSI	□	■	■	□	■	□	■	■	□	■	□	□	0.5000	+	0.6667	+	0.1515	+	0.4697	+
Accuracy	□	■	■	■	■	□	■	□	■	□	□	□	0.5000	+	0.6667	+	0.0875	+	0.4590	+
RR	□	■	■	■	■	□	■	□	■	□	□	□	0.5000	+	0.6667	+	0.0355	+	0.4504	+
Lift	□	■	■	■	■	□	■	□	■	□	□	□	0.5000	+	0.6667	+	0.0070	+	0.4456	+
Yule's Y	□	■	■	■	■	□	■	■	□	□	□	□	0.5000	+	0.6667	+	0.0000		0.4444	+
GBI	■	■	□	■	□	■	□	■	□	■	□	□	0.5000	+	0.6667	+	−0.0165		0.4417	+
YLI1	□	■	■	■	■	□	■	■	□	□	□	□	0.5000	+	0.6667	+	−0.0431		0.4373	+
Leverage	■	□	■	□	□	■	□	■	□	■	■	■	0.4167	+	0.6111	+	0.1773	+	0.4092	+
PSI	■	□	■	□	□	■	□	■	□	■	■	■	0.4167	+	0.6111	+	0.1773	+	0.4092	+
Peculiarity	□	■	■	■	□	□	■	■	■	□	■	□	0.4167	+	0.6111	+	0.0875	+	0.3942	+
Support	■	▨	■	▨	▨	■	▨	■	▨	■	■	■	0.4000	+	0.5714		0.1818	+	0.3922	+
KI	■	■	■	□	■	□	■	■	□	■	□	□	0.4167	+	0.6111	+	0.0686	+	0.3911	+
UncNeg	■	■	■	■	■	□	■	□	■	□	□	□	0.4167	+	0.6111	+	0.0213	+	0.3832	+
K−M	□	■	■	□	■	■	□	■	□	□	■	■	0.4167	+	0.6111	+	0.0118	+	0.3816	+
ϕ	■	■	■	■	■	□	■	■	□	□	□	□	0.4167	+	0.6111	+	−0.0307		0.3745	+
AV	■	■	■	■	■	□	■	■	□	□	□	□	0.4167	+	0.6111	+	−0.0493		0.3714	+
BC	■	■	■	■	■	□	■	■	□	□	□	□	0.4167	+	0.6111	+	−0.0493		0.3714	+
Coverage	■	▨	▨	▨	▨	■	▨	■	■	■	■	■	0.3846		0.5294		0.0854	+	0.3589	
Specificity	■	■	■	■	■	□	■	■	□	□	□	□	0.4167	+	0.6111	+	−0.2055		0.3454	
MI	■	□	■	□	■	■	□	■	□	■	■	■	0.3333		0.5556		0.0449	+	0.3223	
GOI	■	□	■	□	□	■	■	■	□	■	■	■	0.3333		0.5556		−0.0071		0.3136	
OR	□	■	■	■	■	■	■	□	□	□	■	■	0.3333		0.5556		−0.1064		0.2971	
Yule's Q	□	■	■	■	■	■	■	□	□	□	■	■	0.3333		0.5556		−0.1064		0.2971	
LC	■	■	■	■	■	□	■	■	□	■	□	□	0.3333		0.5556		−0.1097		0.2965	
CF	■	■	■	■	■	□	■	■	□	■	□	□	0.3333		0.5556		−0.1127		0.2960	
Precision	■	■	■	■	■	□	■	■	□	■	□	□	0.3333		0.5556		−0.1408		0.2913	
Credibility	■	■	■	■	■	■	■	□	■	□	□	□	0.3333		0.5556		−0.2435		0.2742	
Prevalence	■	▨	■	▨	■	■	▨	■	▨	■	■	■	0.3077		0.4706		−0.2666		0.2391	
BI	■	□	□	■	□	■	■	■	■	■	■	■	0.2500		0.5000		−0.0827		0.2362	
Lower Limit	/	/	/	/	/	/	/	/	/	/	/	/	0.4000		0.6000		0.0000		0.3667	

Figure 6 The results of Experiment II for the first mining. The configuration is the same as Fig. 2.

expert, both in a qualitative and quantitative manner as in Experiment I.

With regard to the CMC in Fig. 6, 28 objective measures out of 40 (70%) obtained a +, and their estimation performance is better than that of a random allocation. The top five are F-M, Jaccard, Recall, CST, and Kappa. It is confirmed that their rank orders are robust to the values of meta-criterion weights. Among the objective measures ranked in the top five for CMC, five and five measures are also ranked in the top five for $\mathrm{CMC}^{\mathrm{equal}}$ and $\mathrm{CMC}^{\mathrm{inequal}}$, respectively. The rank correlation coefficients between the rank for CMC and the ranks for $\mathrm{CMC}^{\mathrm{equal}}$ and $\mathrm{CMC}^{\mathrm{inequal}}$ are 0.9894 and 0.9577, respectively. The CMC of the top five objective measures are more than 1.7 times the lower boundary, that is, these objective measures are at least 1.7 times more likely to correctly identify interesting rules than a random allocation. It is thus suggested that they may be effective in estimating the interest of a medical expert.

The bottom five are BI, Prevalence, Credibility, Precision, and CF. It is confirmed that their rank orders are robust to the values of meta-criterion weights. Among the objective measures ranked in the bottom five for CMC, five and five measures are also ranked in the bottom five for $\mathrm{CMC}^{\mathrm{equal}}$ and $\mathrm{CMC}^{\mathrm{inequal}}$, respectively. The CMC of bottom five objective measures are less than 0.8 times the lower boundary, that is, these objective measures are at most 0.8 times as likely to correctly identify interesting rules as a random allocation. It is thus suggested that they may not be effective in estimating the interest of a medical expert, although this cannot be determined with complete certainty under the limited conditions of present experiment.



Rule ID	13	21	14	15	16	17	18	19	MC(I)		MC(N)		MC(C)		CMC	
Human	I	I	I	I	I	I	I	I								
UncNeg	■	■	□	□	□	□	□	□	0.7500	+	0.8462	+	0.6531	+	0.7499	+
Accuracy	■	■	□	□	□	□	□	□	0.7500	+	0.8462	+	0.6088	+	0.7425	+
RR	■	■	□	□	□	□	□	□	0.7500	+	0.8462	+	0.6088	+	0.7425	+
Peculiarity	■	□	□	■	□	□	□	□	0.7500	+	0.8462	+	0.3832	+	0.7049	+
Lift	■	■	□	□	□	□	□	□	0.7500	+	0.8462	+	0.3197	+	0.6943	+
χ2-M1	■	■	▨	▨	▨	▨	▨	▨	0.6667	+	0.7500	+	0.3412	+	0.6263	+
CST	□	□	□	□	■	■	■	■	0.5000	+	0.6923	+	0.2623	+	0.4924	+
F-M	□	□	□	□	■	■	■	■	0.5000	+	0.6923	+	0.2623	+	0.4924	+
Jaccard	□	□	□	□	■	■	■	■	0.5000	+	0.6923	+	0.2623	+	0.4924	+
Kappa	□	□	□	□	■	■	■	■	0.5000	+	0.6923	+	0.2623	+	0.4924	+
Recall	□	□	□	□	■	■	■	■	0.5000	+	0.6923	+	0.2590	+	0.4919	+
YLI1	■	■	▨	▨	▨	▨	▨	▨	0.5217	+	0.4211		0.1813	+	0.4482	+
J-M	▨	▨	▨	▨	▨	▨	▨	▨	0.5517	+	0.0000		0.2259	+	0.4055	+
OR	▨	▨	▨	▨	▨	▨	▨	▨	0.5517	+	0.0000		0.2259	+	0.4055	+
Yule's Q	▨	▨	▨	▨	▨	▨	▨	▨	0.5517	+	0.0000		0.2259	+	0.4055	+
AV	■	■	▨	■	▨	▨	▨	▨	0.4545	+	0.4000		0.0985	+	0.3861	+
PSI	□	□	■	□	■	■	■	■	0.3750		0.6154		0.0454	+	0.3601	+
YLI2	□	■	□	□	■	■	■	■	0.3750		0.6154		0.0263	+	0.3569	
GBI	■	■	□	□	■	■	■	□	0.3750		0.6154		0.0201	+	0.3559	
K-M	▨	▨	▨	▨	▨	▨	▨	▨	0.5517	+	0.0000		−0.0884		0.3531	
Credibility	■	■	■	■	▨	▨	▨	▨	0.4211	+	0.5217		−0.2262		0.3300	
CSI	■	▨	▨	▨	■	■	■	■	0.3529		0.5600		0.0075	+	0.3299	
BC	■	■	▨	■	▨	▨	▨	▨	0.4545	+	0.4000		−0.2439		0.3291	
CF	■	■	▨	■	▨	▨	▨	▨	0.4545	+	0.4000		−0.2439		0.3291	
Precision	■	■	▨	■	▨	▨	▨	▨	0.4545	+	0.4000		−0.2439		0.3291	
Specificity	■	■	▨	■	▨	▨	▨	▨	0.4545	+	0.4000		−0.2439		0.3291	
Yule's Y	■	■	▨	■	▨	▨	▨	▨	0.4545	+	0.4000		−0.2439		0.3291	
Support	▨	▨	■	▨	■	■	■	■	0.3333		0.5000		−0.0032		0.3050	
Leverage	▨	▨	■	▨	■	■	■	■	0.3333		0.5000		−0.0811		0.2920	
Coverage	▨	▨	■	▨	■	■	■	■	0.3158		0.4348		−0.0324		0.2776	
χ2-M4	■	■	□	□	■	■	■	■	0.2500		0.5385		0.0584	+	0.2661	
Gini	■	■	□	□	■	■	■	■	0.2500		0.5385		−0.2631		0.2126	
YZI	■	■	□	□	■	■	■	■	0.2500		0.5385		−0.2631		0.2126	
MI	□	■	■	□	■	■	■	■	0.2500		0.5385		−0.3042		0.2057	
KI	■	■	▨	▨	■	■	■	■	0.2222		0.4167		−0.4607		0.1408	
BI	▨	▨	■	■	■	■	■	■	0.1905		0.1905		−0.4195		0.0888	
ϕ	■	■	▨	■	■	■	■	■	0.1176		0.4000		−0.5470		0.0539	
LC	■	■	▨	■	■	■	■	■	0.1176		0.4000		−0.5573		0.0522	
GOI	▨	■	■	■	■	■	■	■	0.1111		0.3333		−0.5536		0.0374	
Prevalence	▨	■	■	■	■	■	■	■	0.1000		0.1818		−0.7147		−0.0222	
Lower Limit	/	/	/	/	/	/	/	/	0.3810		0.6190		0.0000		0.3571	

Figure 7 The results of Experiment II for the second mining. The configuration is the same as Fig. 2.

With regard to the CMC in Fig. 7, 17 objective measures out of 40 (43%) obtained a +, and the estimation performance of these measures is better than that of a random allocation. The top five are UncNeg, Accuracy, RR, Peculiarity, and Lift. It is confirmed that their rank orders are robust to the values of meta-criterion weights. Among the objective measures ranked in the top five for CMC, five and five measures are also ranked in the top five for $\mathrm{CMC}^{\mathrm{equal}}$ and $\mathrm{CMC}^{\mathrm{inequal}}$, respectively. The rank correlation coefficients between the rank for CMC and the ranks for $\mathrm{CMC}^{\mathrm{equal}}$ and $\mathrm{CMC}^{\mathrm{inequal}}$ are 0.9254 and 0.9364, respectively. The CMC of the top five objective measures are more than 1.9 times the lower boundary, and it is thus suggested that they may be effective in estimating the interest of a medical expert.

The bottom five are Prevalence, GOI, LC, ϕ, and BI. It is confirmed that their rank orders are robust to the values of meta-criterion weights. Among the objective measures ranked in the bottom five for CMC, five and five measures are also ranked in the bottom five for $\mathrm{CMC}^{\mathrm{equal}}$ and $\mathrm{CMC}^{\mathrm{inequal}}$, respectively. The CMC of bottom five objective measures are less than 0.3 times the lower boundary, and it is thus suggested that they may not be effective in estimating the interest of a medical expert.

Next, we discuss the comprehensive tendency of objective measures in the first and second mining with respect to Figs. 6 and 7, and Table 3. Although the estimation performance of objective measures changed slightly with the mining phase, Accuracy, RR, χ^2-M1, F-M, and Jaccard showed a relatively high estimation performance on average, and this increases the likelihood that they are useful. On the other hand, Prevalence, BI, LC, GOI, and ϕ showed a relatively low estimation performance on average, and this increases the likelihood that they are not useful.

Objective measures of similar rank showed similar patterns of white, gray, and black in Fig. 7. The patterns for these measures are different in Fig. 6; however, with the patterns taking the form of a fine mosaic. This indicates that the rule evaluation by the medical expert was unstable in the first mining, but came to be stable in the second mining. This result agrees with the comment of the medical expert that he evaluated the rules while changing the evaluation criteria in his mind, and tried to derive a solid hypothesis from more vague hypotheses by gradually stabilizing the evaluation criteria.

Summarizing the above discussion, it is implied that the support for a medical expert to actively diverge and converge his/her thinking from various viewpoints, together with the support for the medical expert to explicitly recognize his/her implicit interest, can enhance medical KDD. It is also implied that the combinational use of objective measures with a complementary relationship may result in better estimation performance when objective measures have greatly differing estimation characteristics.

5. General discussion

5.1. Estimation performance of objective measures

In the averaged CMC rank for Experiments I and II in Table 3, the top five are Accuracy, χ^2-M1, RR, UncNeg, and Peculiarity. Their CMC are more than 1.7 times the lower boundary, that is, these objective measures are at least 1.7 times more likely to correctly identify interesting rules than a random allocation. These objective measures were superior throughout Experiments I and II. It thus seems more likely that they are effective in estimating the interest of medical experts. The bottom five are Prevalence, GOI, BI, LC, and ϕ. Their CMC are less than 0.7 times the lower boundary, that is, they are at most 0.7 times as likely to correctly identify interesting rules as a random allocation. These objective measures were inferior throughout Experiments I and II, and it thus seems more likely that they are potentially of little use in these medical domains. The total trend implies that really interesting rules are prescriptive to a certain degree (although the prescriptiveness is not extremely high) by specifying the implicit features of these rules with the combinational use of objective measures, especially highly ranked objective measures.

The top-ranked is Accuracy, which is the sum of Support and UncNeg. As shown in Table 2, the more instances for which both of the antecedent and the consequent hold, the higher evaluation values Support gives to the rule. Meanwhile, the more instances for which both the antecedent and the consequent do not hold, the higher evaluation values UncNeg gives to the rule. The experimental results presented in Table 3 show that the estimation performance of UncNeg is high, whereas that of Support is not particularly high. In short, it can be surmised that the medical experts regard rules with low risks more favorably than rules with high returns.

Among the top five, only Peculiarity represents the uniqueness of a rule, the others, by contrast, represent the correctness of the rule. We used GBI and Peculiarity in the present experiments as objective measures for uniqueness. As is shown in Table 2, the more different the attributes of a rule are from those of other rules, the higher evaluation values GBI gives to the rule. The more different the attribute values of a rule are from those of other rules, the higher evaluation values Peculiarity gives to the rule. Table 3 shows that the estimation performance of Peculiarity is high, whereas that of GBI is not particularly high. In short, it can be surmised that the medical experts placed importance on not only correctness, but also uniqueness, and that, for the purposes of uniqueness judgment criteria, they favored the differences in the attribute values of rules over the differences in the attributes of rules.

Here, we compare the experimental results of the present research with the results of other related studies. Ohsaki et al. [6] and Carvalho et al. [7] examined the relationship between objective measures and real human interest and reported that there were no clear winners among objective measures. Ohsaki et al. [6] used only one dataset, and consequently could not generalize the experimental results. On the other hand, Carvalho et al. [7] used different types of datasets and found that the estimation performance of objective measures changed depending on the datasets. The present research considered only the clinical domain and found comparatively general trends whereby some objective measures were useful in estimating the interest of medical experts. The difference in meta-criteria is also partially responsible for the difference in the experimental results between the present research and other related studies. Ohsaki et al. [6] and Carvalho et al. [7] used Recall and correlation coefficient, respectively, as a meta-criterion.

The appearance of consistency in estimation performance suggests the following possibilities. Although the experimentation is not perfectly uniform between Experiments I and II, similar trends in the top and bottom ranked objective measures were observed. It is thought that the trend will appear more clearly if the experimentation is more uniform. About the commonality of domains, the meningitis and hepatitis domains had some criteria of rule quality in common. About the individuality of medical experts, the medial experts conducted rule evaluation in a similar way. About the propriety of meta-criteria, the meta-criteria in the present study rather reflected the trend of estimation performance than using Recall or correlation coefficient at least.

The above discussions imply two research perspectives from the prescriptiveness: One is to establish proper meta-criteria and to find the general trends of estimation performance for many datasets. The other is to utilize objective measures adaptively to individually different interests in the human-related steps in KDD. The next section deepens discussions on the latter perspective.

5.2. Utilization of objective measures

We obtained the following knowledge through Experiments I and II. *Knowledge 1*: Several objective measures and their combinations can estimate the interest of a medical expert to some degree even if they lack inherent medical semantics. *Knowledge 2*: A medical expert follows a process of thinking in which he/she thinks of many ideas in the phase of hypothesis generation, and then narrows down his/her thoughts in the phase of hypothesis confirmation. From *Knowledge 1* and *2*, it is suggested that medical KDD outcomes may be limited in a framework in which a mining algorithm fixes a few prescribed objective measures in its learning process and automatically generates rules.

A new framework seems to be more useful, in which a mining algorithm generates rules using a small number of simple objective measures, and then a user interface of post-processing presents the rules and/or highlights some possibly interesting rules to a medical expert using flexible combinations of objective measures. A semi-automatic structure that allows an active invention of a medical expert would be suitable to such a new framework, so that the thinking of the medical expert, which may change during the medical KDD process, will be reflected in the process.

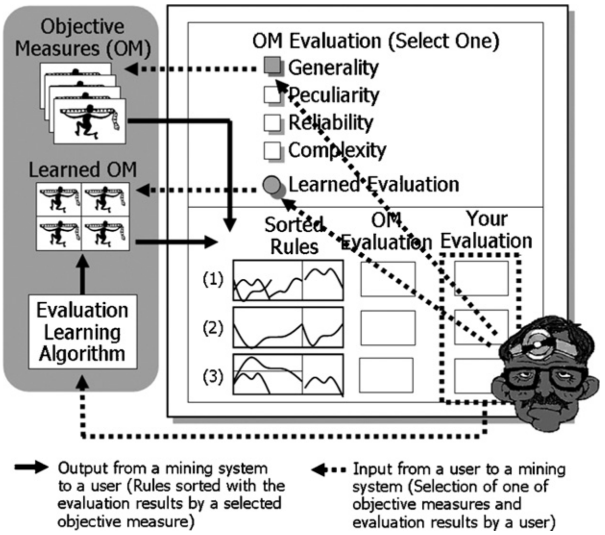

Objective Measures (OM)
Learned OM
Evaluation Learning Algorithm
OM Evaluation (Select One)
Generality
Peculiarity
Reliability
Complexity
Learned Evaluation
Sorted Rules
OM Evaluation
Your Evaluation
(1)
(2)
(3)
Output from a mining system to a user (Rules sorted with the evaluation results by a selected objective measure)
Input from a user to a mining system (Selection of one of objective measures and evaluation results by a user)

Figure 8 Proposed user interface supporting knowledge creation.

From the discussion above, we propose a user interface that supports a user in discovering (more like creating) new knowledge as shown in Fig. 8. We expediently call it medical KDD support user interface. Solid lines in the figure represent output from the mining system to the user, while dotted lines represent input to the mining system from the user. The user clicks a select button on the display, and can select a desirable objective measure from the library of objective measures. The user interface sorts and presents the rules based on the rule evaluation results by the selected objective measure. This function, *Function 1*, will recommend possibly interesting rules to a user and helps the user explicitly change his/her viewpoint. It is expected that giving a variety of viewpoints to the user promotes hypothesis generation.

In addition to the above, it would also be useful to learn the rule evaluation results obtained by a user in the interactive process. We can regard the rule evaluation results obtained by the objective measures as attribute values (the values of explanatory variables), and those obtained by a user as classes (the values of an independent variable). A supervised learning algorithm embedded in the user interface learns the relationship between these values and classes, and constructs the model of real human interest. In order to gradually improve the model, it updates and memorizes the model as a new objective measure every time the user evaluates a rule. The user can evaluate and sort rules using not only the objective measures prepared in the library, but also the learned new objective measure. This function, *Function 2*, will help a user in explicitly recognizing their implicit interest. It is expected that making the user recognize his/her own interest promotes hypothesis confirmation.

We have already developed a prototype medical KDD support user interface and made a part of the library of objective measures open to the public as a free, open-source program [33]. At the present time, we are examining the usability and efficiency of *Function 1* as implemented in this prototype through trial use by medical experts [34—36]. With *Function 2*, a new problem arose concerning the amount of training data, because the number of rule evaluation results (i.e., training data) is strictly limited due to the problem of user fatigue. We are currently estimating the minimum acceptable number of training data points by examining the relationship between the quantity of training data and the accuracy of the learned model of real human interest [34—36].

6. Conclusions and future work

We examined the usefulness of rule interestingness measures in discovering really interesting rules through experiments using rules obtained from clinical datasets on meningitis and hepatitis, along with rule evaluation results obtained from medical experts. More specifically, we examined the performance of interestingness measures to estimate the interest of medical experts based on the degree of agreement between the rule evaluation results obtained from the interestingness measures, and those obtained from the medical experts. As a result, it was suggested that the interestingness measures, accuracy, chi-square measure for one quadrant, relative risk, uncovered negative, and peculiarity, are useful in estimating the interest of medical experts. It was also found that it should be possible to promote medical KDD by the constructing a library and combinational utilization of interestingness measures through the interaction between a medical expert and a mining system. Future work will comprise the examination of the usefulness of the user interface supporting knowledge creation that utilizes interestingness measures, and also the practical application of this interface to other medical domains.

Acknowledgments

This research was supported by a Grant-in-Aid for Scientific Research on Priority Area (B) (No. 18300054), a Grant-in-Aid for Young Scientists (B) (No. 17700162), and a Grant-in-Aid for Scientific Research (C) (No. 19500132) from MEXT Japan, and by the Doshisha Univ. Academic Frontier Research Project.

References

[1] Lavrac N, Flach P, Zupan B. Rule evaluation measures: a unifying view. In: Dzeroski S, Flach P, editors. Proceedings of the 9th international workshop on inductive logic programming ILP-1999. Lecture notes in artificial intelligence, vol. 1634. Berlin: Springer; 1999. p. 174—85.

[2] Yao YY, Zhong N. An analysis of quantitative measures associated with rules. In: Zhong N, Zhou L, editors. Proceedings of the 3rd Pacific-Asia conference on knowledge discovery and data mining PAKDD-1999. Lecture notes in computer science, vol. 1574. Berlin: Springer; 1999. p. 479—88.

[3] Hilderman RJ, Hamilton HJ. Knowledge discovery and measure of interest. Boston: Kluwer Academic Publishers; 2001. p. 1—97.

[4] Tan PN, Kumar V, Srivastava J. Selecting the right interestingness measure for association patterns. In: Hand D, Keim D, Ng R, editors. Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining KDD-2002. New York: ACM Press; 2002. p. 32—41.

[5] Fürnkranz J, Flach P. ROC 'n' rule learning—towards a better understanding of covering algorithms. Mach Learn 2005;58: 39—77.

[6] Ohsaki M, Kitaguchi S, Okamoto K, Yokoi H, Yamaguchi T. Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In: Boulicaut JF, Esposito F, Giannotti F, Pedreschi D, editors. Proceedings of the 15th European conference on machine learning and the 8th European conference on principles and practice of knowledge discovery in databases ECML/PKDD-2004. Lecture notes in artificial intelligence, vol. 3202. Berlin: Springer; 2004. p. 362—73.

[7] Carvalho DR, Freitas AA, Ebecken N. Evaluating the correlation between objective rule interestingness measures and real human interest. In: Jorge A, Torgo L, Brazdil P, Camacho R, Gama J, editors. Proceedings of the 16th European conference on machine learning and the 9th European conference on principles and practice of knowledge discovery in databases ECML/PKDD-2005. Lecture notes in artificial intelligence, vol. 3731. Berlin: Springer; 2005. p. 453—61.

[8] Geng L, Hamilton HJ. Interestingness measures for data mining—a survey. ACM Comput Surveys 2006;38(3) [article 9].

[9] Hatazawa H, Abe H, Komori M, Tachibana Y, Yamaguchi T. Knowledge discovery support from a meningoencephalitis dataset using an automatic composition tool for inductive applications. In: Terano T, Nishida T, Namatame A, Tsumoto S, Ohsawa Y, Washio T, editors. Post-Proceedings of the joint JSAI-2001 workshop on new frontiers in artificial intelligence. Lecture notes in artificial intelligence, vol. 2253. Berlin: Springer; 2002. p. 500—7.

[10] 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. In: Tsumoto S, Yamaguchi T, Numao M, Motoda H, editors. Proceedings of the 1st international workshop on active mining AM-2002 in the 2nd IEEE international conference on data mining ICDM-2002. Washington, DC: IEEE; 2002. p. 97—102.

[11] Freitas AA. On rule interestingness measures. Knowl Syst J 1999;12(5/6):309—15.

[12] Freitas AA. Data mining and knowledge discovery with evolutionary algorithms. Berlin: Springer; 2002 . p. 27—31.

[13] McGarry K. A survey of interestingness measures for knowledge discovery. Knowl Eng Rev 2005;20(1):39—61.

[14] Silberschatz A, Tuzhilin A. On subjective measures of interestingness in knowledge discovery. In: Fayyad U, Uthurusamy R, editors. Proceedings of the 1st ACM SIGKDD international conference on knowledge discovery and data mining KDD-1995. Cambridge: AAAI/MIT Press; 1995 . p. 275—81.

[15] Silberschatz A, Tuzhilin A. What makes patterns interesting in knowledge discovery systems. IEEE Trans Knowl Data Eng 1996;8(6):970—4.

[16] Padmanabhan B, Tuzhilin A. A belief-driven method for discovering unexpected patterns. In: Agrawal R, Stolorz P, Piatetsky-Shapiro G, editors. Proceedings of the 4th ACM SIGKDD international conference on knowledge discovery and data mining KDD-1998. Cambridge: AAAI/MIT Press; 1998. p. 94—100.

[17] Sahara S. On incorporating subjective interestingness into the mining process. In: Kumar V, Tsumoto S, Zhong N, Yu PS, Wu X, editors. Proceedings of the 2nd IEEE international conference on data mining ICDM-2002. Washington, DC: IEEE; 2002. p. 681—4.

[18] Yao H, Hamilton HJ. Mining itemset utilities from transaction databases. Data Knowl Eng J 2006;59:603—26.

[19] Klementtinen M, Mannila H, Ronkainen P, Toivonen H, Verkamo AI. Finding interesting rules from large sets of discovered association rules. In: Grossman D, Gravano L, Zhai CX, Herzog O, Evans DA, editors. Proceedings of the 3rd international conference on information and knowledge management CIKM-1994. New York: ACM Press; 1994 . p. 401—7.

[20] Klementtinen M, Mannila H, Toivonen H. A data mining methodology and its application to semi-automatic knowledge acquisition. In: Hameurlain A, Tjoa AM, editors. Proceedings of the 8th international conference on database and expert systems applications DEXA-1997. Berlin: Springer; 1997. p. 670—7.

[21] Liu B, Hsu W, Chen S, Mia Y. Analyzing the subjective interestingness of association rules. Intell Syst J 2000;15(5): 47—55.

[22] Liu B, Hsu W, Mia Y. Identifying non-actionable association rules. In: Provost F, Srikant R, editors. Proceedings of the 7th ACM SIGKDD international conference on knowledge discovery and data mining KDD-2001. New York: ACM Press; 2001. p. 329—34.

[23] Padmanabhan B, Tuzhilin A. On characterization and discovery of minimal unexpected patterns in rule discovery. IEEE Trans Knowl Data Eng 2006;18(2):202—16.

[24] Tsumoto S. Guide to the meningoencephalitis diagnosis data set. In: Dataset guideline of international workshop of KDD challenge on real-world data KDD-challenge-2000 in the 6th ACM SIGKDD international conference on knowledge discovery data mining KDD-2000; 2000. URL http://www.slab.dnj.ynu.ac.jp/challenge2000/menin.html [accessed April 1, 2001].

[25] Abe H, Yamaguchi T. Constructive meta-learning with machine learning method repository. In: Orchard B, Yang C, Ali M, editors. Proceedings of the 17th international conference on industrial and engineering applications of artificial intelligence and expert systems IEA/AIE-2004. Lecture notes in artificial intelligence, vol. 3029. Springer: Berlin; 2004. p. 502—11.

[26] Arikawa S, Shinohara A, editors. Progress in discovery science—final report of the Japanese discovery science project. Lecture notes in artificial intelligence, vol. 2281. Berlin: Springer; 2002. p. 1—684.

[27] Davis J, Goadrich M. The relationship between precision—recall and ROC curves. UW-CS Technical Report, TR1551; 2006.

[28] Rijsbergen C. Information retrieval, Oxford: Butterworth-Heinemann; 1979. http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html [accessed July 23, 2007].

[29] Brazdil BP, Soares C. A comparison of ranking methods for classification algorithm selection. In: Mántaras RL, Plaza E, editors. Proceedings of the 11th European conference on machine learning ECML-2000. Lecture notes in artificial intelligence, vol. 1810. Berlin: Springer; 2000 . p. 63—74.

[30] Tsumoto S. Discovery challenge—a collaborative effort in knowledge discovery from databases. In: Proceedings of the dataset guideline of discovery challenge in the 13th European conference on machine learning and the 6th European conference on principles and practice of knowledge discovery in databases ECML/PKDD-2002; 2002. http://lisp.vse.cz/challenge/ecmlpkdd2002/index.html [accessed July 23, 2007].

[31] Das G, King-Ip L, Heikki M, Renganathan G, Smyth P. Rule discovery from time series. In: Agrawal R, Stolorz P, Piatetsky-Shapiro G, editors. Proceedings of the 4th ACM SIGKDD international conference on knowledge discovery and data mining KDD-1998. Cambridge: AAAI/MIT Press; 1998. p. 16—22.

[32] Motoda H, editor. Active mining—new directions of data mining. Amsterdam: IOS Press; 2002. p. 1—302.

[33] Abe H. COIN (calculating objective indices for data mining results);http://sourceforge.jp/projects/coin/ [accessed July 23, 2007] 2005.

[34] Abe H, Ohsaki M, Yokoi H, Yamaguchi T. Implementing an integrated time-series data mining environment based on temporal pattern extraction methods—a case study of an interferon therapy risk mining for chronic hepatitis. In: Washio T, Sakurai A, Nakajima K, Takeda H, Tojo S, Yokoo M, editors. Post-proceedings of the joint JSAI-2005 workshop on new frontiers in artificial intelligence. Lecture notes in artificial intelligence, vol. 4012. Berlin: Springer; 2005. p. 425—35.

[35] Abe H, Tsumoto S, Ohsaki M, Yamaguchi T. A rule evaluation support method with learning models. In: Han J, Wah BW, Raghavan V, Wu X, Rastogi R, editors. Proceedings of the 5th IEEE international conference on data mining ICDM-2005. Washington, DC: IEEE; 2005 . p. 549—52.

[36] Abe H, Tsumoto S, Ohsaki M, Yamaguchi T. Evaluating a rule evaluation support method based on objective rule evaluation indices. In: Ng WK, Kitsuregawa M, Li J, Chang K, editors. Proceedings of the 10th Pacific-Asia conference on knowledge discovery and data mining PAKDD-2006. Lecture notes in computer science, vol. 3918. Berlin: Springer; 2006. p. 509—19.

[37] Chen M, Zheng A, Lloyd J, Jordan M, Brewer E. Failure diagnosis using decision trees. In: Kephart J, Parashar M, Das R, Sunderam V, editors. Proceedings of the 1st IEEE international conference on autonomic computing ICAC-2004. Washington, DC: IEEE; 2004. p. 36—43.

[38] Huang Y, Shekhar S, Xiong H. Discovering colocation patterns from spatial data sets: a general approach. IEEE Trans Knowl Data Eng 2004;16(12):1472—85.

[39] Bayardo R, Agrawal R, Gunopulos D. Constraint-based rule mining in large, dense databases. Data Mining Knowl Discov J 2000;217—40.

[40] Knowledge discovery in databases. In: Piatetsky-Shapiro G, Frawley WJ., editors. Ch. Discovery, analysis and presentation of strong rules. Cambridge: AAAI/MIT Press; 1991. p. 229—48.

[41] Advances in knowledge discovery and data mining. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy, R. (editors), Ch. Explora: a multipattern and multistrategy discovery assistant. Cambridge: AAAI/MIT Press; 1996. p. 249—71.

[42] Ali K, Manganaris S, Srikant R. Partial classification using association rules. In: Heckerman D, Mannila H, Pregibon D, Uthurusamy R, editors. Proceedings of the 3rd international conference on knowledge discovery and data mining KDD-1997. Cambridge: AAAI/MIT Press; 1997 . p. 115—8.

[43] Brin S, Motwani R, Silverstein C. Beyond market baskets: Generalizing association rules to correlations. In: Peckham J, editor. Proceedings of the 16th ACM SIGMOD international conference on Management of Data SIGMOD-1997. New York: ACM Press; 1997. p. 265—76.

[44] Jaccard P. Nouvelles recherches sur la distribution florale. Bull Soc Vaudoise Sci Nat 1908;44:223—70 [in French].

[45] Fleiss J. Statistical methods for rates and proportions. New York: John Wiley and Sons; 1981. p. 1—352.

[46] Gray B, Orlowska ME. CCAIIA: Clustering categorical attributes into interesting association rules. In: Wu X, Ramamohanarao K, Korb KB, editors. Proceedings of the 2nd Pacific-Asia conference on knowledge discovery and data mining PAKDD-1998. Lecture notes in computer science, vol. 1394. Berlin: Springer; 1998. p. 132—43.

[47] Aggrawal C, Yu P. A new framework for itemset generation. In: Proceedings of the 17th ACM SIGACT—SIGMOD—SIGART symposium on principles of database systems PODS-1998. New York: ACM Press; 1998. p. 18—24.

[48] Gini C. Variability and mutability—contribution to the study of statistical distributions and relations. Studi Economico-Giuridici dell' Univ di Cagliari 1912;3:1—158 [in Italian].

[49] Hamilton HJ, Shan N, Ziarko W. Machine learning of credible classifications. In: Sattar A, editor. Proceedings of the 10th Australian joint conference on Artificial Intelligence AUS-AI-1997. Lecture notes in computer science, vol. 1342. Berlin: Springer; 1997. p. 330—9.

[50] Goodman LA, Kruskal WH. Measures of Association for Cross Classifications. Berlin: Springer; 1979. p. 1—146.

[51] Morimoto Y, Fukuda T, Matsuzawa H, Tokuyama T, Yoda K. Algorithms for mining association rules for binary segmentations of huge categorical databases. In: Gupta A, Shmueli O, Widom J, editors. Proceedings of the 24th international conference on very large databases VLDB-1998. San Fransisco: Morgan Kaufmann Publishers; 1998. p. 380—91.

[52] Knowledge discovery in databases. In: Piatetsky-Shapiro G, Frawley WJ, editors. Ch. Rule induction using information theory. Cambridge: AAAI/MIT Press; 1991; p. 159—76.

[53] Yao J, Liu H. Searching multiple databases for interesting complexes. In: Lu H, Motoda H, Liu H, editors. Proceedings of the 1st Pacific-Asia conference on knowledge discovery and data mining PAKDD-1997. Singapore: World Scientific Publishing Company; 1997. p. 198—210.

[54] Gago P, Bento C. A metric for selection of the most promising rules. In: Zytkow JM, Quafafou M, editors. Proceedings of the 2nd European conference on principles of data mining and knowledge discovery PKDD-1998. Lecture notes in artificial intelligence, vol. 1510. Berlin: Springer; 1998. p. 19—27.

[55] Zhong N, Yao YY, Ohshima M. Peculiarity oriented multi-database mining. IEEE Trans Knowl Data Eng 2003;15(4): 952—60.