

Knowledge-based Systems and Interestingness Measures: Analysis with Clinical Datasets

Jabez J. Christopher¹, Khanna H. Nehemiah¹ and Kannan Arputharaj²

¹Ramanujan Computing Centre, Anna University, Chennai, India

²Department of Information Science and Technology, Anna University, Chennai, India

Knowledge mined from clinical data can be used for medical diagnosis and prognosis. By improving the quality of knowledge base, the efficiency of prediction of a knowledge-based system can be enhanced. Designing accurate and precise clinical decision support systems, which use the mined knowledge, is still a broad area of research. This work analyses the variation in classification accuracy for such knowledge-based systems using different rule lists. The purpose of this work is not to improve the prediction accuracy of a decision support system, but analyze the factors that influence the efficiency and design of the knowledge base in a rule-based decision support system. Three benchmark medical datasets are used. Rules are extracted using a supervised machine learning algorithm (PART). Each rule in the ruleset is validated using nine frequently used rule interestingness measures. After calculating the measure values, the rule lists are used for performance evaluation. Experimental results show variation in classification accuracy for different rule lists. Confidence and Laplace measures yield relatively superior accuracy: 81.188% for heart disease dataset and 78.255% for diabetes dataset. The accuracy of the knowledge-based prediction system is predominantly dependent on the organization of the ruleset. Rule length needs to be considered when deciding the rule ordering. Subset of a rule, or combination of rule elements, may form new rules and sometimes be a member of the rule list. Redundant rules should be eliminated. Prior knowledge about the domain will enable knowledge engineers to design a better knowledge base.

ACM CCS (2012) Classification: Information systems
→ Information systems applications → Decision support systems → Expert systems

Keywords: knowledge base, decision support systems, rule-based classification, rule list, interestingness measures

1. Introduction

Computer-aided diagnosis has become an inevitable technique in most hospitals and medical centers. Several frameworks and systems such as expert system [1], medical diagnostic system [2], integrated healthcare systems [3], emergency medical decision support system (MDSS), clinical decision support system [4], clinical recommender system [5], prediction system, and various other machine learning models [6-9], are used for efficient diagnosis and prognosis. These systems are preferred more than human experts due to factors such as efficient response, consistency, target-specificity, permanence, and sometimes even multi-domain functionality. Decision support systems are cost-effective [10] and capable of modeling the different levels of clinician's decision making strategies [11]. Apart from medical diagnosis, machine learning techniques and systems are used for various purposes in diverse areas of engineering [12-14], management [15], [16] and science [17-19].

A medical decision support system can be a non-knowledge-based or knowledge-based system [20]. A non-knowledge-based system uses computational intelligence and machine learning techniques to acquire information from data, whereas a knowledge-based system has well defined information organization and manipulation schemes. Fuzzy inferencing system and neural network can be considered as examples for the former, whereas an expert system and rule-based decision support system can be con-

sidered as examples for the latter. A knowledge-based system consists of three components: a modeling subsystem, processing or inferencing subsystem and a knowledge base. The central component of the knowledge-based system is the knowledge base (KB). The knowledge base is domain specific. The quality, consistency, integrity and efficient usage of the KB profoundly affect the performance of these systems. If health information technology is going to transform healthcare, a deeper understanding of the complex dynamics underlying the system implementation and application is needed. Since poor clinical decisions may lead to adverse effects, there is a need for efficient and effective decision support systems.

The representation and manipulation of knowledge in a KB is still a major research issue. The main objective of Knowledge Discovery in Databases (KDD) is to extract interesting patterns. This notion of interest highly depends on the domain of application and user's objectives. The user may not be a data mining expert, but rather an expert in the domain being mined. The task of medical knowledge discovery and knowledge validation [21] is challenging and also important as it serves as the input for decision making and prediction.

Decision making is a crucial process in medical diagnosis, and data occupies a central part for this purpose. In fact, most medical care activities involve gathering, analyzing and using data for various purposes. Data provide the basis for categorizing the problems a patient may be having, or for identifying subgroups within a population of patients. Amateur clinicians and doctors can get the required needed additional information regarding what actions should be taken to gain a greater understanding of a situation.

Experience, practice and caring for individual patients bestow medical students and clinicians with special skills and enhanced judgment over time. But in the field of research, researchers develop and validate new clinical knowledge of general applicability by formal analysis of the data collected from large numbers of patients. Thus, clinical data are not only used for diagnosis and prognosis, but also to support clinical research through the aggregation and statistical analysis of observations gathered from clinical trials, experience and populations of patients.

The contribution of the experimental study presented in this paper is novel and significant in the following ways: first, the concept of rule list is introduced and its difference from a ruleset is explained. Second, the significance of a rule list and the impact of rule interestingness measures on a rule list are discussed. These concepts are experimentally verified over benchmark medical datasets. Finally, we provide a few guidelines which knowledge engineers may consider for rule base design and construction.

The organisation of this paper is as follows: the following section elaborates the terms and concepts used in this work. Explanation about the interestingness measures used in this work is also presented. Section 3 highlights the related work, their results and limitations. Section 4 describes the implementation and experimental results in detail. The observations are discussed in Section 5, followed by the conclusion and scope for future work.

2. Background Knowledge

A clinical dataset D can be represented as a set S having row vectors (R_1, R_2, \dots, R_m) and column vectors (C_1, C_2, \dots, C_n) . Each record can be represented as an ordered n -tuple of clinical and laboratory attributes $(A_{i_1}, A_{i_2}, \dots, A_{i_{(n-1)}}, A_{i_n})$ for each $i = 1, 2, \dots, m$, where the last attribute (A_{i_n}) for each i represents physician's diagnosis to which the record $(A_{i_1}, A_{i_2}, \dots, A_{i_{(n-1)}})$ belongs. Each attribute of an element in S that is A_{ij} for $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n - 1$ can either be a categorical (nominal) or numeric (real or integer) type.

Data mining is defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from databases" that is used to make crucial business decisions [22]. From the definition we can infer that the mined knowledge should be interesting and new; furthermore, the process is significant and should not be obvious. Rather than simple computations, complex processing is required to uncover the patterns that are buried in the data. The mined patterns should suit data other than the training data, and, at the same time, should be interesting. The knowledge should be useful and comprehensible. The patterns extracted should catalyse the process of diagnosis

and prognosis. At the same time they should be simple and understandable to the user.

Taking all of the above statements into consideration, it is essential to review and analyse the construction, representation and working of the knowledge base which is the heart of a knowledge-based system.

2.1. Knowledge Base and Rule Base

Mined knowledge can be represented as propositional logic, first order logic, semantic nets, frames, rules or a combination of these methods [23]. In most knowledge-based systems IF-THEN rules are used for knowledge representation. Rules are easy to understand and interpret.

A KB is defined as a collection of facts, relationships and rules which embody current expertise in a particular area [24]. A knowledge base may contain associations, mined patterns, frequent itemsets or knowledge represented as images and signals.

A knowledge base in which knowledge is represented in the form of classification rules is known as a rule base. A classification rule has at least one conditional attribute-value pair in its antecedent and at most one predefined class as its consequent. An association rule can have more than one attribute-value pair in its consequent [25].

A classification rule in IF-THEN format can be generalised as an implication of the form

$$\langle C_1 \text{ op } A_{11}, C_2 \text{ op } A_{21}, C_i \text{ op } A_{ij}, \dots, C_{n-1} \text{ op } A_{m,n-1} \rangle \rightarrow \langle C_n = A_{mn} \rangle$$

where C_i is a conditional attribute, A_{ij} is a value that belongs to the domain of C_i , C_n is the class attribute, A_{mn} is a class value, and op is a relational operator such as =, < or >. The value of ' i ' denotes the attribute index and the value of ' j ' denotes the number of values in the domain of the attribute C_i . This form of representation of rules, Antecedent \rightarrow Consequent ($A \rightarrow C$), is highly comprehensible for the system designer.

A classifier learns from training data and stores learned knowledge into classifier parameters, such as the weights, as in a neural network. However, it is difficult to interpret the knowledge into an understandable format using these

classifier parameters. Hence, it is desirable to extract IF-THEN rules to represent valuable information in data.

Several data mining techniques are available to extract knowledge from data. Generally, a prediction model is generated using supervised or unsupervised machine learning techniques. Probabilistic classification, associative classification, neural networks, decision trees are examples for the former, and clustering is a classical example for the latter. These predictive models can be valuable tools in medicine. They can be used to assist in decision support, diagnosis and prognosis determination. Precise diagnosis of diseases allows subsequent investigations, treatments and interventions to be delivered in a well organized manner. The most common way for generating classification rules is to generate a decision tree [26] and extract rules from the tree. But this approach led to the common subtree problem [27]. The first rule extraction technique from neural networks was proposed by Gallant [28]. C4.5 [29] and RIPPER [30] are yet other rule induction strategies that perform a global optimization procedure on an initial ruleset to obtain an optimal ruleset. Keedwell et al. developed a system in which a genetic algorithm is used to search for rules in the Neural Network input space [31].

Most of the works in literature contemplate on rule extraction and rule evaluation. Rule ordering and its impact on the classifier is still a less explored area of research. Rule mining algorithms tend to generate a large number of rules. Some authors eliminate redundant rules [32],[33], while others evaluate and order the rules using interestingness measures [34]. In this paper, we focus on the latter path: the use of interestingness measures to arrange the rules. Many studies compared the different objective measures reported in the literature according to several points of view [35],[36]. These articles have highlighted some of the interestingness measures, their properties, and relationships with other measures.

But the real problem with all these rule generation schemes is that they tend to over fit the training data and do not generalize well to independent test sets, particularly on noisy and inconsistent data. To generate rule sets for such data, it is necessary to have some way of measuring the real worth of individual rules. The standard approach to evaluate the worth of rules

is to compute the error rate on an independent set of instances held back from the training set. Alternatively, interestingness measures can be used to examine rule quality.

2.2. Ruleset and Rule List

A Ruleset is a structure that provides an unordered collection of rule objects. Each rule object is of the form $A \rightarrow C$. A ruleset does not allow duplicates. A Rule list is a structure that provides ordered and indexed collection of rule objects which may contain duplicates. Each rule object is associated with a Rule index. A rule map is a structure that contains each rule object in the form of key-value pairs. The rule antecedent forms the key and the consequent corresponds to the value. A rule map is an ordered and unique collection. Knowledge about appropriate use of these structures for design of a KB is important for knowledge engineers.

2.3. Interestingness Measures

Interestingness measures provide a universal and sensible approach to automatically identify interesting rules. They are very useful for filtering and ranking the rules presented to the user. In classification rule mining, interestingness measures can be applied for three tasks. First, they can be used during the rule extraction process as heuristics to select attribute-value pairs that are going to be included in the classification rules. Second, they can be used as a measure to weigh the goodness of a rule and rank them. Third, interestingness measures can be used to prune uninteresting and unfit elements from the rule base.

Rule interestingness has both an objective and a subjective aspect. The former can be regarded as a data-driven approach, and the latter, as a user-driven approach. Subjective measures evaluate the rules based on the previous knowledge and experience of the data. Objective measures use the data from which the information is extracted, to evaluate rule interestingness. Our work focuses on the objective aspect of rule interestingness. In practice, we suggest that both objective and subjective approaches should be used to select interesting rules. For instance, objective approaches can be used as an initial filter to select potentially interesting

rules, while subjective approaches can then be used as a final filter to select truly interesting rules. However, the purpose of final evaluation of the results of classification rule mining is usually to measure the predictive accuracy of the whole ruleset on testing data. The prediction quality and efficiency depend on the size and goodness of the entire rule set rather than of a single rule. In this work we have used some well known, frequently used measures to illustrate the impact of choice of interestingness measures on classification accuracy.

3. Literature Review

There are numerous works in literature which attempt to evaluate rule interestingness measures. Some works try to establish a relationship between these measures as well. This section brings to light some of the recent and major works carried out.

Tan et al. [37] proposed a method to categorize interestingness measures based on a specific dataset and properties. In their method, the mined rules are ranked by the users, and the measure that has the most similar ranking results for these rules is selected for further use. They concluded that application of different measures may lead to significantly unlike orderings of patterns. Their work ranks different rule interestingness measures and does not use any classifier evaluation measures. Based on their comparative study of twenty one interestingness measures, several groups of consistent measures can be identified, but their method is not directly applicable if the number of rules in the rule base is vast.

Lenca et al. suggested an approach where marks and weights are assigned to each property of a measure that the user considers to be important [38]. The user is not required to rank the mined patterns. Rather, he is required to identify the desired properties and specify their importance for a particular application. The quality measures and methods, considered in their study, evaluate only the individual quality of rules. They do not evaluate the quality of the whole set of rules. Furthermore, the work is confined to associative classification and not a generalized approach.

Vaillant et al. proposed a method where interestingness measures are clustered into groups

[39]. The clustering method is based on either the properties of the measures or rulesets generated by experiments on datasets. Each cluster represents a similarity value (distance) between the two measures on the specified ruleset. Similarity is calculated on the rankings of the two measures on the ruleset. The authors have experimented with twenty measures on ten benchmark datasets. However, the impact of the choice of interestingness measures on rule base design and organisation for the medical domain is not highlighted.

Ohsaki et al. in their research explore the performance of conventional rule interestingness measures and discuss their practicality for supporting KDD through human-system interaction in medical domain [40]. Their results indicate that the measures can predict really interesting rules at a certain level and that their combinational use will be useful. Even though their work focuses on the usefulness of rule interestingness measures for the medical domain, the impact of rule evaluation measures on rulesets is not elaborated.

S. Dreiseitl et al. in their work compared the performance of four single-rule ranking algorithms with the performance of a multi-rule ranking algorithm [41]. They used the rule ranking algorithm proposed by Vinterbo et al. [42]. They concluded that a multi-variate rule ranking algorithms perform better than the single-rule ranking algorithms. Their work does not emphasize the impact of rule interestingness measures and the impact of the choice of these measures on the classifier.

These authors define the properties that are associated with the measures and compare the measures experimentally to determine the correlation between different interestingness measures. All the authors convey a similar message: there is no best measure; each domain and problem has a different best set of measures. Therefore, before the pattern mining algorithm is applied to a determined domain, it is necessary to select the correct set of measures. To the best of our knowledge, no systematic comparison of classification ruleset evaluation has so far been published in the literature. This work aims to fill this gap by providing a comprehensive investigation of the relative advantages of different approaches by using a wrapper approach for rulelist evaluation, specifically for medical datasets.

4. Materials and Methods

The steps involved in this work are illustrated in Figure 1. The initial step is acquisition of the medical dataset. The datasets used in this work are all benchmark datasets obtained from the UCI Repository. Holdout approach [48] was used to split the data into training and testing sets. This approach is chosen for the sake of computational simplicity, as this work is only intended to show the relative change in classification accuracy for different rule interestingness measures. The training set is used for rule extraction and the testing set is used for performance evaluation of the classifier.

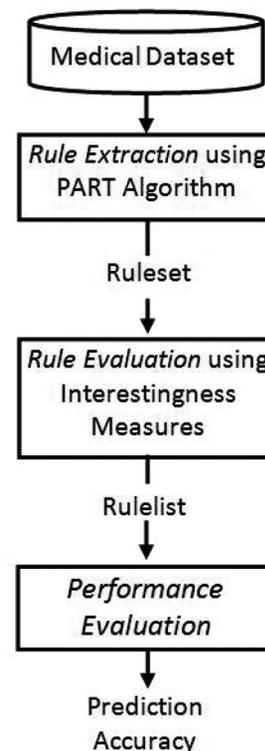


Figure 1. Steps in Estimating the Performance of Interestingness Measures.

4.1. Dataset Description

In this work, experiments were conducted on three benchmark medical datasets, namely, Pima Indians Diabetes dataset, Wisconsin Breast Cancer Dataset and Cleveland Heart Disease dataset.

The datasets were obtained from the UCI machine learning repository [51]. The Diabetes dataset consists of 768 instances out of which 268 are tested positive. A description about the attributes and their values is given in Table 1. The Breast Cancer Dataset comprises 699 instances out of which 458 are benign. A description about the attributes and their values is given in Table 2. The Heart Disease Dataset consists of 303 instances spread across five class labels. The description of the attributes is presented in Table 3.

Table 1. Description of Pima Indians Diabetes dataset.

ATTRIBUTE	DESCRIPTION	VALUE
Preg	Number of pregnancies	[0–17]
Plas	Plasma glucose concentration in an oral glucose tolerance test	[0–199]
Pres	Diastolic blood pressure	[0–122]
Skin	Triceps skin fold thickness	[0–99]
Insu	2-Hour serum insulin	[0–846]
Mass	Body mass index	[0–67]
Pedi	Diabetes pedigree function	[0–2.45]
Age	Age of an individual	[21–81]
class	Tested positive / negative	(0,1)

Table 2. Description of Breast Cancer dataset.

ATTRIBUTE	DESCRIPTION	VALUE
Clump Thickness	Assesses if cells are mono- or multi-layered.	1 – 10
Uniformity of Cell Size	Evaluates the consistency in size of the cells in the sample.	1 – 10
Uniformity of Cell Shape	Estimates the equality of cell shapes and identifies marginal variances.	1 – 10
Marginal Adhesion	Quantifies how much cells on the outside of the epithelium tend to stick together.	1 – 10
Single Epithelial Cell Size	Relates to cell uniformity, determines if epithelial cells are significantly enlarged.	1 – 10
Bare Nuclei	Presence and size of nuclei.	1 – 10
Bland Chromatin	Rates the uniform “texture” of the nucleus in a range from fine to coarse.	1 – 10
Normal Nucleoli	Determines whether the nucleoli are small and barely visible or larger, more visible, and more plentiful.	1 – 10
Mitoses	Describes the level of mitotic (cell reproduction) activity.	1 – 10
Class	2- benign, 4- malignant	(2, 4)

Table 3. Description of Heart Disease dataset.

ATTRIBUTE	DESCRIPTION	VALUE
Age	Age in years	[29, 77]
Sex	Sex of subject	[female, male]
Cp	Chest pain type	[typ_angina, asympt, non_anginal, atyp_angina]
Trestbps	Resting blood sugar	[94.0, 200.0]
Chol	Serum cholesterol	[126.0, 564.0]
Fbs	Fasting blood sugar	[true, false]
Restecg	Resting ECG result	[left_vent_hyper, normal, st_t_wave_abnormality]
Thalach	Maximum heart rate achieved	[71.0, 202.0]
Exang	Exercise induced angina	[no, yes]
Oldpeak	ST depression induced by exercise relative to rest	[0.0, 6.2]
Slope	Slope or peak exercise ST segment	[up, flat, down]
Ca	Number of major vessels coloured by fluoroscopy	[0.0, 3.0]
Thal	Defect type	[fixed_defect, normal, reversable_defect]
Class	Narrowing Diameter percentage	< 50, > 50, > 50_2, > 50_3, > 50_4

4.2. Rule Extraction

Rules are extracted from the data using PART algorithm. The PART algorithm [49] builds a J48 tree [50] and then extracts the path with the highest coverage to form the first rule. The instances covered by this rule are removed from the training data and the process is then repeated to generate the second rule. Hence a set of J48 trees are grown on increasingly smaller subsets of the training data. The algorithm is a combination of two major paradigms for rule generation: extracting rules from decision trees and separate and conquer rule learning technique.

A decision tree with branches reduced to undefined sub-trees is called a partial decision tree. To generate such a partial decision tree, the tree-building and tree-pruning operations are combined in order to find a minimal subtree that cannot be further simplified. Once this subtree has been found, tree-building ceases and a single rule is read off. It splits a set of examples recursively into a partial tree. The first step chooses a test and divides the examples into subsets accordingly. The algorithm makes this choice in exactly the same way as C4.5 [29]. Then the subsets are expanded in order of their average entropy, in a non-decreasing order. This continues recursively until a subset is expanded into a leaf, and then continues further by backtracking. However, as soon as an internal node appears, which has all its children expanded into leaves, pruning begins: the algorithm checks whether that node is better replaced by a single leaf. The decision made using the rules obtained from the PART algorithm is exactly the same as C4.5. We have used the rules generated from PART algorithm for its compactness and ease of interpretation. Furthermore, the generated rules require no post-processing. Table 4 gives the rules generated for Heart Disease dataset. The rules for the Diabetes dataset and Breast Cancer dataset are given in Table 5 and Table 6 respectively.

Sample Rule Description for Heart Disease Dataset.

R1: IF patient has angina induced by exercise AND asympt chest pain AND age is between 54 – 57 THEN Narrowing Diameter percentage is > 50 .

R7: IF slope of ST segment of ECG is upward AND ST depression induced by exercise relative to rest is between 0 to 0.5 AND Resting

blood sugar level is between 139 – 141 THEN Narrowing Diameter percentage is < 50 .

R13: IF ST depression induced by exercise relative to rest is between 0 to 0.5 AND Resting blood sugar level is between 127 – 129.5 THEN Narrowing Diameter percentage is < 50 .

Sample Rule Description for Diabetes Dataset.

R1: IF Plasma glucose concentration in an oral glucose tolerance test is between 79.6 – 99.5 THEN the patient is tested negative.

R7: IF Body mass index is between 33.5 – 40.2 AND Diabetes pedigree function value is between 0 – 0.312 AND age is between 27 – 33 years THEN the patient is tested negative.

R10: IF Body mass index is between 33.5–40.2 AND age is between 21 – 27 AND Diastolic blood pressure 73 – 85.4 mm Hg.

R17: IF Age is between 21 – 27 AND 2-Hour serum insulin is 169.2 – 253.8 U/ml AND Plasma glucose concentration in an oral glucose tolerance test is between 99.5 – 119.4.

R26: IF Body mass index is 26.84 – 33.5 AND Triceps skin fold thickness 0 – 9.9 mm AND Diastolic blood pressure is 61 – 73.2 mm Hg.

Table 4. Ruleset obtained for Heart Disease dataset.

Rule No.	Antecedent (A)	Consequent (C)
R1	exang = yes AND cp = asympt AND age = 54.5 – 56.5	> 50
R2	exang = yes AND cp = asympt	> 50
R3	sex =female	< 50
R4	slope = up AND trestbps = 129.5 – 131	< 50
R5	slope = up AND trestbps = 119 – 121	< 50
R6	slope =down	< 50
R7	slope = up AND oldpeak = 0–0.05 AND trestbps = 139 – 141	< 50
R8	oldpeak = 0.25 – 0.45	< 50
R9	slope =flat	> 50
R10	trestbps = 111 – 119	< 50
R11	trestbps = 94 – 109	< 50
R12	trestbps = 137 – 139	< 50
R13	oldpeak = 0–0.05 AND trestbps = 127 – 129.5	< 50

Table 5. Ruleset obtained for Diabetes dataset.

Rule No.	Antecedent (A)	Consequent (C)
R1	plas = (79.6 – 99.5)	tested_negative
R2	plas = (159.2 – 179.1)	tested_positive
R3	plas = (179.10)	tested_positive
R4	mass = (20.13 – 26.84)	tested_negative
R5	plas = (59.7 – 79.6)	tested_negative
R6	age = (0 – 27)ANDskin = (19.8 – 29.7)	tested_negative
R7	mass = (33.55 – 40.26) AND pedi = (0 – 0.3122) AND age = (27 – 33)	tested_negative
R8	mass = (26.84 – 33.55) AND pres = (48.8 – 61)	tested_negative
R9	mass = (33.55 – 40.26) AND plas = (139.3 – 159.2)	tested_positive
R10	mass = (33.55 – 40.26) AND age = (0 – 27) AND pres = (73.2 – 85.4)	tested_negative
R11	mass = (33.55 – 40.26) AND pres = (85.4 – 97.6)	tested_negative
R12	mass = (46.97 – 53.68)	tested_positive
R13	skin = (9.9 – 19.8)	tested_negative
R14	age = (45 – 51) AND pres = (73.2 – 85.4)	tested_positive
R15	age = (51 – 57)	tested_positive
R16	age = (57 – 63)	tested_negative
R17	age = (0 – 27) AND insu = (169.2 – 253.8) AND plas = (99.5 – 119.4)	tested_negative
R18	age = (27 – 33) AND insu = (84.6 – 169.2)	tested_positive
R19	age = (27 – 33) AND insu = (0 – 84.6) AND pres = (73.2 – 85.4)	tested_negative
R20	age = (27 – 33)	tested_positive
R21	mass = (40.26 – 46.97) AND insu = (0 – 84.6)	tested_positive
R22	mass = (40.26 – 46.97) AND insu = (84.6 – 169.2)	tested_negative
R23	mass = (33.55 – 40.26) AND pedi = (0 – 0.3122)	tested_negative
R24	mass = (33.55 – 40.26) AND preg = (0 – 1.7) AND age = (0 – 27)	tested_positive
R25	mass = (26.84 – 33.55) AND skin = (29.7 – 39.6)	tested_negative
R26	mass = (26.84 – 33.55) AND skin = (0 – 9.9)ANDpres = (61 – 73.2)	tested_positive
R27	mass = (33.55 – 40.26)	tested_positive
R28	mass = (26.84 – 33.55) AND pedi = (0 – 0.3122) AND skin = (0 – 9.9) AND age = (0 – 27)	tested_positive
R29	mass = (0 – 6.71)	tested_negative
R30	mass = (26.84 – 33.55) AND pedi = (0.3122 – 0.5464) AND preg = (0 – 1.7)	tested_negative
R31	mass = (26.84 – 33.55) AND skin = (19.8 – 29.7)	tested_positive
R32	mass = (26.84 – 33.55) AND pres = (85.4 – 97.6) AND plas = (99.5 – 119.4)	tested_negative
R33	mass = (13.42 – 20.13)	tested_negative
R34	pres = (73.2 – 85.4)	tested_negative

Table 6. Ruleset obtained for Breast Cancer dataset.

Rule No.	Antecedent (A)	Consequent (C)
R1	Cell_Size_Uniformity =1	benign
R2	Cell_Shape_Uniformity =10	malignant
R3	Cell_Shape_Uniformity =4	malignant
R4	Cell_Shape_Uniformity =3	malignant
R5	Cell_Shape_Uniformity =5	malignant
R6	Cell_Shape_Uniformity =2	benign
R7	Cell_Shape_Uniformity =6	malignant
R8	Cell_Shape_Uniformity =7	malignant
R9	Cell_Shape_Uniformity =8	malignant
R10	Cell_Shape_Uniformity =1	benign

4.3. Rule Evaluation

In this work we have used some well known, frequently used measures to illustrate the impact of choice of interestingness measures on classification accuracy. Nine classification rule interestingness measures are used for generating rule lists. The definition of the measures and the representation are presented below.

N – Total number of instances in the dataset

A – Rule antecedent

C – Rule consequent

P – Probabilistic scale

n_A – Number of instances that contain A

n_C – Number of instances that contain C

n_{AC} – Number of instances that contain both A and C

Support

The support for a rule can be considered as the number of transactions in the dataset that satisfy the union of items in the consequent and antecedent of the rule. This measure was initially applied for mining market-basket data type transactions, for association rules [43].

$$\text{Support}(A \rightarrow C) = P(A \wedge C) = \frac{n_{AC}}{N}$$

While confidence is a measure of the rule's strength, support corresponds to its statistical significance. Besides statistical significance,

another motivation for support constraints comes from the fact that we are usually interested only in the rules with support above some minimum threshold. If the support is not large enough, it means that the rule is not worth consideration or that it is simply less preferred. The exponential search space of a rule is reduced because of the downward closure property of this measure.

Confidence

Confidence is an estimate of the probability of observing the consequent given the antecedent [43]. Confidence value ranges from 0 to 1.

$$\text{Confidence}(A \rightarrow C) = P\left(\frac{C}{A}\right) = \frac{n_{AC}}{n_A}$$

For rules with the same confidence, the rule with the highest support is preferred. The rationale is that the estimate for confidence is more reliable.

Laplace

Laplace is a confidence estimator that takes support into account, becoming more negative as the support of the antecedent decreases [44]. It ranges within [0, 1].

$$\text{Laplace}(A \rightarrow C) = \frac{n_{AC} + 1}{n_A + 2}$$

Lift

Confidence or Laplace alone may not be sufficient to assess the descriptive interest of a rule. Rules with high confidence may occur by chance. Such spurious rules can be detected by determining whether the antecedent and the consequent are statistically independent. Lift is intended for this purpose and is defined as follows:

$$\text{Lift}(A \rightarrow C) = \frac{P\left(\frac{C}{A}\right)}{P(C)} = \frac{Nn_{AC}}{n_{AC}n_C}$$

Lift measures how far from independence the antecedent and the consequent are [45]. It ranges from zero to infinity. Values close to 1 imply that the antecedent and the consequent are independent and the rule is not interesting. Values far from 1 indicate that the evidence of the antecedent provides information about the consequent. Lift measures co-occurrence only and is symmetric.

Coverage

Coverage is also known as antecedent support. It measures how often a rule $A \rightarrow C$, is applicable in a database irrespective of the consequent [35].

$$\text{Coverage}(A \rightarrow C) = P(A) = \frac{n_A}{N}$$

Leverage

Leverage is used to measure how much more counting is obtained from the co-occurrence of the antecedent and consequent from independence or the expected value [46]. It ranges from $[-0.25, +0.25]$.

$$\begin{aligned} \text{Leverage}(A \rightarrow C) &= P\left(\frac{C}{A}\right) - P(A) \times P(C) \\ &= \frac{n_{AC}}{n_A} - \frac{n_A}{N} \times \frac{n_C}{N} \end{aligned}$$

Sensitivity

Sensitivity, also known as Recall, is defined as the chance of the entire rule to occur, given that the consequent of the rule has already occurred. This measure was first used in information retrieval and text mining applications [47], but can be used for evaluating rules too.

$$\text{Sensitivity}(A \rightarrow C) = P\left(\frac{A}{C}\right) = \frac{n_{AC}}{n_C}$$

Prevalence

The commonness and generality of a rule is expressed by Prevalence. It is the probability of

the consequent of the rule.

$$\text{Prevalance}(A \rightarrow C) = P(C) = \frac{n_C}{N}$$

Specificity

Specificity is generally used for classifier evaluation. In the context of rule evaluation, it is defined as the ratio of the records that are not covered by the rule as a whole, to the records that are not covered by the antecedent of the rule.

$$\text{Specificity}(A \rightarrow C) = P\left(\frac{\neg C}{\neg A}\right)$$

4.4. Performance Evaluation

Each rule is evaluated using nine interestingness measures that are most frequently used in medical informatics. The ruleset is now ordered based on the values of different interestingness measures. This gives nine rule lists. Each rule list corresponds to a prediction model (classifier). When a test record is to be classified, the classifier assigns the class label of the first-best matching rule in the rule list. If none of the rules match, then the default class label is assigned. The goodness of the model is evaluated using the test data. The variation in accuracy, across the nine measures, for the Heart Disease dataset, Diabetes dataset and the Breast Cancer dataset are also presented in Table 7, Table 8 and Table 9 respectively.

Table 7. Rule lists and prediction accuracy for Heart Disease dataset.

MEASURE	RULELIST	ACCURACY (%)
Support	R9, R3, R2, R11, R5, R4, R10, R1, R8, R12, R6, R7, R13	73.927
Confidence	R1, R5, R2, R13, R8, R11, R4, R7, R12, R3, R9, R10, R6	81.188
Laplace	R1, R5, R2, R8, R13, R11, R4, R3, R12, R7, R9, R10, R6	81.188
Coverage	R9, R3, R2, R10, R6, R11, R4, R5, R12, R8, R1, R7, R13	73.927
Leverage	R3, R10, R6, R11, R4, R5, R12, R8, R7, R13, R9, R2, R1	72.937
Lift	R9, R2, R3, R11, R5, R4, R10, R1, R12, R8, R6, R7, R13	73.927
Recall	R3, R11, R5, R4, R10, R12, R8, R6, R7, R13, R9, R2, R1	74.221
Prevalence	R9, R2, R3, R11, R5, R4, R10, R1, R12, R8, R6, R7, R13	73.927
Specificity	R9, R13, R7, R6, R12, R8, R10, R4, R5, R11, R1, R3, R2	73.927

Table 8. Rule lists and prediction accuracy for Diabetes dataset.

MEASURE	RULELIST	ACCURACY (%)
Support	R34,R1, R4, R27, R13, R6, R23, R20, R2, R3, R25, R8, R5, R31, R19, R21, R15, R7, R9, R11, R30, R26, R16, R18, R33, R10, R14, R12, R24, R22, R29, R17, R32, R28	70.964
Confidence	R33, R17, R5, R6, R1, R7, R4, R32, R13, R30, R3, R12, R29, R2, R8, R19, R10, R23, R25, R9, R26, R34, R15, R16, R14, R11, R22, R18, R21, R28, R27, R20, R31, R24	78.225
Laplace	R33, R5, R6, R1, R17, R4, R7, R13, R3, R32, R30, R2, R12, R29, R8, R19, R10, R23, R25, R9, R34, R26, R15, R16, R14, R11, R22, R18, R21, R28, R27, R20, R31, R24	78.385
Coverage	R34, R27, R1, R4, R20, R13, R23, R6, R31, R2, R25, R3, R8, R21, R24, R5, R15, R11, R18, R19, R9, R26, R16, R14, R7, R30, R10, R22, R33, R12, R29, R32, R28, R17	67.839
Leverage	R34, R1, R4, R13, R23, R6, R25, R8, R5, R11, R19, R16, R7, R30, R10, R22, R33, R29, R32, R17, R27, R20, R31, R2, R3, R21, R24, R15, R18, R9, R26, R14, R12, R28	72.786
Lift	R27, R34, R1, R4, R20, R13, R2, R6, R3, R23, R31, R21, R25, R15, R8, R9, R5, R26, R18, R14, R19, R24, R12, R7, R11, R30, R16, R33, R10, R22, R29, R32, R17, R28	67.448
Prevalence	R27, R34, R1, R4, R20, R13, R2, R6, R3, R23, R31, R21, R25, R15, R8, R9, R5, R26, R18, R14, R19, R12, R24, R7, R11, R30, R16, R33, R10, R22, R29, R17, R32, R28	67.448
Specificity	R17, R32, R22, R29, R10, R16, R33, R11, R30, R7, R19, R5, R8, R25, R23, R6, R13, R4, R1, R34, R28, R12, R24, R14, R18, R26, R9, R15, R21, R31, R3, R2, R20, R27	72.786

Table 9. Rule lists and prediction accuracy for Breast Cancer dataset.

MEASURE	RULELIST	ACCURACY (%)
Support	R1, R10, R2, R6, R5, R3, R8, R9, R7, R4	94.420
Confidence	R2, R10, R1, R9, R8, R5, R7, R6, R3, R4	94.420
Laplace	R10, R1, R2, R9, R8, R5, R7, R6, R3, R4	94.420
Coverage	R1, R10, R6, R2, R4, R3, R5, R8, R7, R9	94.420
Leverage	R1, R10, R6, R2, R4, R3, R5, R8, R7, R9	94.420
Lift	R1, R10, R2, R3, R5, R8, R6, R7, R9, R4	94.420
Recall	R1, R10, R2, R6, R5, R3, R8, R9, R7, R4	94.420
Prevalence	R1,R10,R2,R3,R5,R8,R6,R7,R9,R4	94.420
Specificity	R1,R4,R2,R3,R5,R8,R6,R7,R10,R9	94.420

5. Results and Discussion

This section discusses the results, issues to be addressed in rule base organization and some solutions and guidelines that would enable KB developers and researchers to design efficient rule-based classifiers for clinical diagnosis and decision support.

Each measure gives rise to new rule lists. From the experiment it can be observed that the accuracy of the classifier is propositional to the position of the rules in the rule list. When rules

are listed based on Confidence or Laplace measure, the longer rules are placed at the beginning of the list. These rules boost the prediction accuracy of that rule list.

From the experiment, it can be inferred that rule length needs to be considered when deciding the rule ordering. Subset of a rule, or combination of rule elements, may form new rules, and sometimes may be a member of the rule list too. In this scenario, the ordering of rules is important as in the case of the Heart Disease and Diabetes datasets. When the antecedent of the rules contains a distinct element, and if the

rules are short, then the rule ordering does not play a major role.

Considering the Breast Cancer dataset, all the rules contain at most one attribute. The rule length of all the rules is equal. The classification accuracy remains stable for all rule lists. In such situations, the structure and content of the rules can be modified for optimized performance.

In medical diagnosis and prognosis, decision making is a critical task and is of prime importance. Small errors and mistakes may end up in undesirable consequences. Each and every stage of design and construction of a knowledge base needs proper guidelines. Choosing a rule generation strategy, choice of rule interestingness measures and deciding the final prediction model, which is to be deployed, are also issues of prime importance. Knowledge engineers, Data Mining professionals and experts from the application domain need to work in unanimity to ensure the construction of an efficient and effective KB.

6. Conclusion

Knowledge-based systems should be capable of producing meaningful and useful results. This work presents an experiment over medical datasets that highlights the importance of the organization of contents in a knowledge base (rule base). To reduce the number of mined results, interestingness measures have been used for various kinds of patterns. These measures are generally used relative to some threshold which depends on the situation and the application where a particular measure is applied. The primary objective of this work is not to improve the accuracy of a classification system, but to show the effect of rule interestingness measures and rule ordering on rule base design. Each interestingness measure has its own pros and cons. For example, the disadvantage of support is that items (rules) that occur very infrequently in the data set are often pruned although they would still be interesting and potentially valuable. Confidence is sensitive to the frequency of the consequent (C) of the rule in the database. Lift is susceptible to noise in small databases. Rare itemsets with low probability which occur a few times (or only once) can produce enormous lift values. Use of an appropriate mea-

sure depends on the domain of application and also on the designer of the knowledge-based systems. The performance of the knowledge-based systems has a profound impact on clinical decision making. Deterioration in accuracy can affect diagnosis and prognosis processes. It may result in adverse effects too. Hence, a well-designed rule base of a knowledge-based system will enhance the process of clinical diagnosis and decision making. Choosing interestingness measures that reflect real human interest remains an unsolved issue. Vast experimentation on more clinical datasets, involving more interestingness measures and diverse rule extraction strategies, may yield useful and novel findings. This work can still be extended over temporal medical data, in order to improve the design of a knowledge base for systems that manipulate time-series data.

References

- [1] A. Rabelo jr. *et al.*, "An Expert System for Diagnosis of Acute Myocardial Infarction with ECG Analysis", *Artificial Intelligence in Medicine*, vol. 10, no. 1, pp. 75–92, 1997. [http://dx.doi.org/10.1016/S0933-3657\(97\)00385-0](http://dx.doi.org/10.1016/S0933-3657(97)00385-0)
- [2] J. Piecha, "The Neural Network Selection for a Medical Diagnostic System Using an Artificial Data Set", *CIT. Journal of Computing and Information Technology*, vol. 9, no. 2, pp. 123–32, 2004. <http://dx.doi.org/10.2498/cit.2001.02.03>
- [3] V. Mantzana and M. Themistocleous, "Identifying and Classifying Benefits of Integrated Healthcare Systems Using an Actor-Oriented Approach", *CIT. Journal of Computing and Information Technology*, vol. 12, no. 4, pp. 265–78, 2004. <http://dx.doi.org/10.2498/cit.2004.04.01>
- [4] J. H. Eom *et al.*, "AptaCDSSE: A Classifier Ensemble-based Clinical Decision Support System for Cardiovascular Disease Level Prediction", *Expert Systems with Applications*, vol. 34, no. 4, pp. 2465–2479, 2008. <http://dx.doi.org/10.1016/j.eswa.2007.04.015>
- [5] L. Duan *et al.*, "Healthcare Information Systems: Data Mining Methods in the Creation of a Clinical Recommender System", *Enterprise Information Systems*, vol. 5, no. 2, pp. 169–181, 2011. <http://dx.doi.org/10.1080/17517575.2010.541287>
- [6] H. Debbi and M. Bourahla, "Generating Diagnoses for Probabilistic Model Checking Using Causality", *CIT. Journal of Computing and Information Technology*, vol. 21, no. 1, pp. 13–22, 2013. <http://dx.doi.org/10.2498/cit.1002115>

- [7] S.-J. Zheng *et al.*, “A genetic fuzzy radial basis function neural network for structural health monitoring of composite laminated beams”, *Expert Systems with Applications*, vol. 38, no. 9, pp. 11837–11842, 2011. <http://dx.doi.org/10.1016/j.eswa.2011.03.072>
- [8] C.-F. Wu *et al.*, “A Functional Neural Fuzzy Network for Classification Applications”, *Expert Systems with Applications*, vol. 38, no. 5, pp. 6202–6208, 2011. <http://dx.doi.org/10.1016/j.eswa.2010.11.049>
- [9] K. Vijaya *et al.*, “Fuzzy Neuro-Genetic Approach for Predicting the Risk of Cardiovascular Diseases”, *International Journal of Data Mining, Modelling and Management*, vol. 2, no. 4, pp. 388–402, 2010.
- [10] C.-L. Chi *et al.*, “A Decision Support System for Cost-Effective Diagnosis”, *Artificial Intelligence in Medicine*, vol. 50, no. 3, pp. 149–161, 2010. <http://dx.doi.org/10.1016/j.artmed.2010.08.001>
- [11] J. J. Ross *et al.*, “A Hybrid Hierarchical Decision Support System for Cardiac Surgical Intensive Care Patients. Part II”, *Clinical Implementation and Evaluation. Artificial Intelligence in Medicine*, vol. 45, no. 1, pp. 53–62, 2009. <http://dx.doi.org/10.1016/j.artmed.2008.11.010>
- [12] J. Mehedi and M. K. Naskar, “A Fuzzy Based Distributed Algorithm for Maintaining Connected Network Topology in Mobile Ad-Hoc Networks Considering Freeway Mobility Model”, *Journal of Computing and Information Technology*, vol. 20, no. 2, pp. 69–84, 2012. <http://dx.doi.org/10.2498/cit.1001736>
- [13] P. Korošec *et al.*, “A Multi-Objective Approach to the Application of Real-World Production Scheduling”, *Expert Systems with Applications*, vol. 40, pp. 5839–53, 2013. <http://dx.doi.org/10.1016/j.eswa.2013.05.035>
- [14] I. Svalina *et al.*, “An Adaptive Network-based Fuzzy Inference System (ANFIS) for the Forecasting: The Case of Close Price Indices”, *Expert Systems with Applications*, vol. 40, pp. 6055–63, 2013. <http://dx.doi.org/10.1016/j.eswa.2013.05.029>
- [15] C. Cassie, “Marketing Decision Support Systems”, *Industrial Management & Data Systems*, vol. 97, no. 8, pp. 293–296, 1997. <http://dx.doi.org/10.1108/02635579710195000>
- [16] M. E. Bounif and M. Bourahla, “Decision Support Technique for Supply Chain Management”, *CIT. Journal of Computing and Information Technology*, vol. 21, no. 4, pp. 255–68, 2014. <http://dx.doi.org/10.2498/cit.1002172>
- [17] A. Karci and M. Demir, “Estimation of Protein Structures by Classification of Angles Between A-Carbons of Amino Acids Based on Artificial Neural Networks”, *Expert Systems with Applications*, vol. 36, no. 3, pp. 5541–5548, 2009. <http://dx.doi.org/10.1016/j.eswa.2008.06.118>
- [18] L.-C. Wu *et al.*, “An Expert System to Predict Protein Thermostability Using Decision Tree”, *Expert Systems with Applications*, vol. 36, no. 5, pp. 9007–9014, 2009. <http://dx.doi.org/10.1016/j.eswa.2008.12.020>
- [19] Z. Yang *et al.*, “BioPPIExtractor: A Protein-Protein Interaction Extraction System for Biomedical Literature”, *Expert Systems with Applications*, vol. 36, no. 2, pp. 2228–2233. <http://dx.doi.org/10.1016/j.eswa.2007.12.014>
- [20] E. S. Berner, *Clinical decision support systems: theory and practice*. Springer Science & Business Media, 2007, pp. 6–7. <http://dx.doi.org/10.1007/978-0-387-38319-4>
- [21] M. M. Abbasi and S. Kashiyarndi, “Clinical Decision Support Systems: A Discussion On Different Methodologies Used in Health Care”, 2006.
- [22] U. M. Fayyad *et al.*, “KnowledgeDiscovery and Data Mining: Towards A Unifying Framework”, In *KDD*, vol. 96, pp. 82–88, 1996.
- [23] H. M. Brooks, “Expert Systems and Intelligent Information Retrieval”, *Information Processing & Management*, vol. 23, no. 4, pp. 367–382, 1987. [http://dx.doi.org/10.1016/0306-4573\(87\)90023-9](http://dx.doi.org/10.1016/0306-4573(87)90023-9)
- [24] A. J. Meadows, *The Origins of Information Science*, vol. 1. London: Taylor Graham and the Institute of Information Scientists, 1987.
- [25] A. A. Freitas, “Understanding the Crucial Differences between Classification and Discovery of Association Rules: A Position Paper”, *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 1, pp. 65–69, 2000. <http://dx.doi.org/10.1145/360402.360423>
- [26] J. R. Quinlan, “Induction of Decision Trees”, *Machine Learning*, vol. 1, no.1, pp. 81–106, 1986. <http://dx.doi.org/10.1023/A:1022643204877>; <http://dx.doi.org/10.1007/BF00116251>
- [27] G. Bagallo and D. Haussler, “Boolean Feature Discovery in Empirical Learning”, *Machine Learning*, vol. 5, no. 1, pp. 71–99. <http://dx.doi.org/10.1023/A:1022611825350>; <http://dx.doi.org/10.1007/BF00115895>
- [28] S. I. Gallant, “Connectionist Expert Systems”, *Communications of the ACM*, vol. 31, no. 2, pp. 152–169, 1988. <http://dx.doi.org/10.1145/42372.42377>
- [29] J. R. Quinlan, *C4.5: Programs for Machine Learning*, vol. 1. Morgan Kaufmann, 1993.
- [30] W. W. Cohen, “Fast Effective Rule Induction”, In *ICML*, vol. 95, pp. 115–123, 1995.
- [31] E. Keedwell *et al.*, “Evolving Rules from Neural Networks Trained on Continuous Data”, In *Proceedings of the 2000 IEEE Congress on Evolutionary Computation*, vol. 1, pp. 639–645. <http://dx.doi.org/10.1109/cec.2000.870358>
- [32] M. J. Zaki, “Generating Non-Redundant Association Rules”, In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 34–43. <http://dx.doi.org/10.1145/347090.347101>

- [33] M. Z. Ashrafi *et al.*, “A New Approach of Eliminating Redundant Association Rules”, In *Database and expert systems applications*, Springer Berlin Heidelberg, pp. 465–474, 2004. <http://dx.doi.org/10.1007/978-3-540-30075-545>
- [34] P. Lenca *et al.*, “On Selecting Interestingness Measures for Association Rules: User Oriented Description and Multiple Criteria Decision Aid”. *European Journal of Operational Research*, vol. 184, no. 2, pp. 610–626, 2008. <http://dx.doi.org/10.1016/j.ejor.2006.10.059>
- [35] L. Geng and H. J. Hamilton, “Choosing the Right Lens: Finding What is Interesting in Data Mining”, In *Quality Measures in Data Mining*, Berlin Heidelberg: Springer, 2007, pp. 3–24.
- [36] P.-N. Tan *et al.*, “Selecting the Right Interestingness Measure for Association Patterns”, In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 32–41.
- [37] P.-N. Tan *et al.*, “Selecting the Right Objective Measure for Association Analysis”, *Information Systems*, vol. 29, no. 4, pp. 293–313, 2004. [http://dx.doi.org/10.1016/S0306-4379\(03\)00072-3](http://dx.doi.org/10.1016/S0306-4379(03)00072-3)
- [38] P. Lenca *et al.*, “Association Rule Interestingness Measures: Experimental and Theoretical Studies” In *Quality Measures in Data Mining*, Berlin Heidelberg: Springer, pp. 51–76, 2007.
- [39] B. Vaillant *et al.*, “A Clustering of Interestingness Measures”, In *Discovery Science*, Berlin Heidelberg: Springer, pp. 290–297, 2004. <http://dx.doi.org/10.1007/978-3-540-30214-823>
- [40] M. Ohsaki *et al.*, “Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis”, In *Knowledge Discovery in Databases: PKDD*, Berlin Heidelberg: Springer, pp. 362–373, 2004. <http://dx.doi.org/10.1007/978-3-540-30116-534>
- [41] S. Dreiseitl *et al.*, “An Evaluation of Heuristics for Rule Ranking”, *Artificial Intelligence in Medicine*, vol. 50, no. 3, pp. 175–180, 2010. <http://dx.doi.org/10.1016/j.artmed.2010.03.005>
- [42] S. A. Vinterbo *et al.*, “Small, Fuzzy and Interpretable Gene Expression Based Classifiers”, *Bioinformatics*, vol. 21, no. 9, pp. 1964–1970, 2005. <http://dx.doi.org/10.1093/bioinformatics/bti287>
- [43] R. Agrawal *et al.*, “Mining Association Rules Between Sets of Items in Large Databases”, *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207–216, 1993. <http://dx.doi.org/10.1145/170036.170072>
- [44] I. J. Good, *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, vol. 30, Cambridge, MA: MIT press, 1965.
- [45] S. Brin *et al.*, “Beyond Market Baskets: Generalizing Association Rules to Correlations”, *ACM SIGMOD Record*, vol. 26, no. 2, pp. 265–276, 1997. <http://dx.doi.org/10.1145/253262.253327>
- [46] G. I. Webb and D. Brain, “Generality is Predictive of Prediction Accuracy”, in *Proceedings of the 2002 Pacific Rim Knowledge Acquisition Workshop*, 2002, pp. 117–130.
- [47] C. J. Van Rijsbergen, *Information Retrieval*, Butterworth, 1997.
- [48] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.
- [49] E. Frank and I. Witten, “Generating Accurate Rule Sets without Global Optimization”, in *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 144–51.
- [50] I. H. Witten *et al.*, *Weka: Practical Machine Learning Tools and Techniques with Java Implementations*, 1999.
- [51] C. L. Blake and C. J. Merz, (1998) *UCI Repository of machine learning databases*, (460). Irvine, CA: University of California, Department of Information and Computer Science, [Online]. Available: <http://www.ics.uci.edu/mllearn/MLRepository.html>

Received: September, 2014

Revised: August, 2015

Accepted: September, 2015

Contact addresses:

Jabez J. Christopher
Ramanujan Computing Centre
Anna University
Chennai, India
e-mail: jabezc@gmail.com

Khanna H. Nehemiah
Ramanujan Computing Centre
Anna University
Chennai, India
e-mail: nehemiah@annauniv.edu

Kannan Arputharaj
Department of Information Science and Technology
Anna University
Chennai, India
e-mail: kannan@annauniv.edu

JABEZ J. CHRISTOPHER received his B.E. in Computer Science and Engineering from Anna University in 2009 and M.E. in Computer Science and Engineering from Anna University in 2011. He is a research scholar, pursuing Ph.D. degree at Ramanujan Computing Centre, Anna University, Chennai, Tamilnadu, India. His research interests are clinical data mining and optimization.

KHANNA H. NEHEMIAH received his B.E. in Computer Science and Engineering from the University of Madras in 1997, M.E. in Computer Science and Engineering from the University of Madras in 1998 and Ph.D. from Anna University, Chennai in 2007. He is currently working as an Associate Professor at Ramanujan Computing Centre, Anna University, Chennai, Tamilnadu, India. His research interests include software engineering, data mining, database management systems, bioinformatics and medical image processing.

KANNAN ARPUTHARAJ received his M.E. in Computer Science and Engineering from Anna University Chennai in 1991 and Ph.D. from Anna University, Chennai in 2000. He is currently working as Professor and Head, in the Department of Information Science and Technology, Anna University, Chennai, Tamilnadu, India. His research interests are artificial intelligence, data mining and database management systems.
