Extracting Accident Event Aviso Rules by Soft Computing Techniques in the Healthcare Industry

You-Shyang Chen
Department of Information Management,
Hwa Hsia Institute of Technology,
New Taipei City, Taiwan, ROC
E-mail: ys_chen@cc.hwh.edu.tw

***Abstract*—The hospitalized safety environment of patients is a priority-concerned by the healthcare services for ensuring patient safety and healthcare quality during inpatient admission in hospitals. Particularly, the applications of accident event aviso system (AEAS) and knowledge discovery for identifying accident event aviso category (AEAC) will have assistance to determine potential hazards and reduce medical errors to enhance patient safety and healthcare quality issues in such a complexity healthcare system of hospitals and are vitally important in the healthcare industry. Thus, this work proposes a hybrid procedure to highlight such an interesting issue. The proposed procedure constitutes five components, expert feature-screening approach, the C4.5 algorithm of decision tree, a cumulative probability distribution approach, the LEM2 algorithm of rough sets, and a rule filter technique. An experimental dataset was retrieved from the AEAS of existing hospital databases. The results from implemented experiments indicate that the proposed procedure is capable of effects to remove a redundant attribute, achieve an improved classification performance, offer tools based on the knowledge of aviso system, and provide sufficient information to solving AEAC problems encountered by patients and hospitals, thereby benefiting interested parties.**

Keywords-Accident Event Aviso System (AEAS); decision tree; Cumulative Probability Distribution Approach (CPDA); rough set-based classifier

I. INTRODUCTION

To promote the construction of hospitalized safety environment is a priority-concerned by healthcare industry to ensure healthcare safety and quality for patients during treatment period. One of the goals for patient safety and healthcare quality is to increase detection in recording error aviso events to decrease risk of harm to patients. For patient safety, the number of events to error detection reported is important issues because high reporting rate implies the low risk level of the reported events that can be avoided in the near future. Kaplan et al. [1] showed that a severeness level of accident event is taken as an index of addressing fault management when high rates of detection are occurred. Thus, developing an accident event aviso system (AEAS) to record such an error practices will be useful and helpful to determine a problem of hazards for improving the safety of patients. The valuable issue is accordingly created by identifying the core determinants of such a reporting system for improving the healthcare quality, and the core determinants will be reliable and valid for interested parties. With this view, the influencing variables on medical errors should be studied, and the related issues should also be highlighted. Research on improving models for solving the AEAC problems would be useful for the following reasons. (1) The healthcare services attract the focus of both practitioners and researchers seeking healthcare benefit; however, artificial intelligence (AI) tools are seldom used in accident event aviso researches to generate an understood decision rule when they are compared to statistical techniques. (2) Although rough set theory (RST) has animated numerous studies and has great advancements, but applying the RST to address problems of applications in constructing hybrid classification models is a new trial for identifying accident event aviso category (AEAC) presented in the healthcare industry. Therefore, the study is motivated by expected promising results on such an issue. This study aims to build a hybrid procedure to process objectively the related classification problems of AEAC to enhance the quality of healthcare services and to reduce medical errors, and creates its knowledge-based rules to evaluate the strength of existing evidence for patient safety. The generated decision rules provide explicit guidelines that can be played into clinical treatment forms or can make decisions of medical services for involved healthcare parties. The objectives of this study are as follows: (1) Construct a hybrid procedure to predict AEAC in the healthcare industry; (2) apply expert feature-screening methods to address objectively core determinants when determining AEAC; (3) evaluate the classification performance in the proposed procedure; and (4) provide meaningful rules to interested parties for achieving specific objectives.

II. LITERATURE REVIEW

A. Patient Safety and its Accident Event Aviso System

Almost 98,000 Americans died due to avoidable medial errors each year [2]. Making a safety environment to avoid medial errors for patients is major concerns in hospitals of the healthcare industry. The medial errors contain three different categories, drug error, human error, and systemic problem errors. Thus, one of the best way to

improving the information quality of the patient safety environment can design an event aviso system [3] to record best practices for reducing medial errors and building the ability of lowing future medical errors. An accident event aviso system records a tangible evidence of medical professional efforts to improve the healthcare quality and low preventable medical faults for protecting patients being harmed in the professional healthcares of physicians. Developing such an aviso system to insure that the safety information of patients can be analyzed and employed to identify risk hazards and safety practices will be a requirement toward constructing the development of patient safety information standards into existing medical databases, and it can be taken to construct a satisfied mediator of managing medical errors [4] to achieve the goals of decreasing risk of harm to patients.

B. Rough Set Theory

Rough set theory, RST, models computational methods to process uncertain data, hybrid data, and vague concepts of classification problems of various class data and has become useful tools for addressing knowledge systems of making decisions [5]. The RST used for classification outperforms statistical methods and is a research topic for both researchers and practitioners, and the application fields are marked in various domains, including energy [6], finance [7], mathematics learning [8], credit rating [9], and medicine [10,11]. In the implementation of the rough sets (RS), relational databases are begun with objects table of attributes as well as attribute values of objects. One attribute is the decisional attribute; in other words, the remaining attributes are the conditional attributes [5,12]. Applying the concepts of equal classes to the partitioned training examples based on a specified criterion, the RS addresses the data problems of incompletion, vagueness, and uncertainty. The partition members are formally presented by a unary set-theoretic operator or a successor function for the lower and upper approximation spaces from which both possible and definite rules are derived easily. Definitely, not a clear-cut boundary is defined in imprecise and vague data sets. The RS classifier refuses a certain boundary of given set and has an implication of every set, which can be determined roughly by using the lower and upper approximations. The details of definitions and equations of RS are referred to the studies of Pawlak [5,13].

C. Rule Induction Methods

The rule induction methods of the RS were presented in learning from examples based on RS (LERS) system [14]. The LERS algorithm brings on a decision rule set from given real examples to classify a new example by using the induced rule set. Furthermore, the learning from examples module, version 2 algorithm, which is abbreviated LEM2, from the LERS of a data mining system [15] can be used to symbolic attributes. The algorithm of LEM2 calculates local covering of each concept to generate a decision rule set derived from a decisional table. The index of quality of each deduced rule is computed by using specified functions of rule quality according to the measurement of coverage, consistency, and support to identify the rule strength. The decision rule is constituted by the following three factors, including support, specificity, and strength [16]. Generally, this LEM2 algorithm is frequently used in a rule induction option of the LERS system [17] in the data-mining practice. Therefore, the rule induction of LEM2 algorithm is applied into this study.

D. Rule Filter

The rule set generated by RS classifier always contains very large numbers of various rules [18], and it is useful to specify a filter rule technique explicitly in situations that the generated numerous rules low the classification abilities of the deduced rule set for that some decision rules may be superfluous or a poor quality. Given the drawbacks mentioned-above, an algorithm of filtering decision rule is effective to cut down the number of generated decision rules [19]. Filters define the criteria that must be satisfied by an event before a rule is run. That is, the solution of filtering rule is calculated by the quality indices of the decision rules from a generated rule set.

E. Decision Tree C4.5 Algorithm

Decision tree classifier is a tree structure liking flowchart that each internal node is a test on attributes, each branch is the test outcome, and each leaf node is a class distribution [20]. In the learning system of decision tree, iterative dichotomiser 3 (ID3) is the algorithm used to generate a decision tree and based on information theory proposed by Quinlan [21]. The C4.5 algorithm is an extended of the ID3 algorithm invented by Quinlan [22] to process the issues that are not dealt with by ID3 algorithm, avoiding the over-fitting data, lowing error pruning on trees, ruling post-pruning on trees, handling a continuous attribute, selecting an appropriate measurement of attribute selection, processing missing attribute values in the training data, and increasing computational efficiency. The further details of the decision tree C4.5 algorithm can be referenced to the studies of Quinlan [21,22].

F. Cumulative Probability Distribution Approach

A distribution of probability on the real example is identified by using the probability of forming half-open intervals notated by p(a, b] or F(b)−F(a) if a < b. Following that, the distribution of probability in a real randomly valued variable $\boldsymbol{X}$ can be characterized completely by its cumulative distribution function (CDF) [23]. A cumulative probability distribution approach, CPDA, is a discrete method according to the CDF. To discretize observations into the requested number of given intervals based on characteristics of normal distribution of the data is an objective means. Furthermore, the experiments of simulation offer convinced evidence that the 30 sample data are sufficient for overcoming skewness of population distributions and give a normal approximation distribution [24]. The implementing procedure of the CPDA technique is divided into four

steps. First, run a normal distribution test for the given experimental data set; second, determine the argument ***U*** is defined over a universe of discourse; third, define the length of intervals and construct the function of a membership; finally, fuzzify observations. The definitions and equations of the CPDA technique can be referenced by the studies [25-28].

III. EXPERIMENTAL SETTINGS

This study proposes a hybrid classification procedure, which constitutes the five component techniques or methods: feature screening, decision tree C4.5, CPDA, RST, and, a rule filter. The detailed algorithms of the proposed procedure by using a practical data set extracted from AEAS are implemented as follows:

Step 1: Data selection. Thirty-three pre-selected attributes are first characterized by the labeled AEAS data set and are determined in a unified format from patient-centered views covered by the period 2004–2007 based on experiential knowledge of the medical expert (please see the Acknowledgements section). The 33 attributes are 'Event no.,' 'Notifier grade,' 'Notifier position,' 'Notifier capacity,' 'Starting date,' 'Seniority,' 'Date of birth,' 'Notifier age,' 'Education,' 'Level,' 'Patient name,' 'Medical record no.,' 'Patient age,' 'Patient gender,' 'Sickroom (or Ward),' 'Division,' 'Happen time,' 'Notify time,' 'Time difference,' 'Happen location,' 'Related personnel,' 'Time interval,' 'Reason,' 'Cause,' 'Recommendation,' 'Notify phase,' 'Notify type,' 'Patient address,' 'Patient tel.,' 'ID no.,' 'Nationality,' 'Medical care type,' and 'Category.' The last 'Category' is the decision-attribute (or called Class), and the others are belonging to the conditional-attributes.

Step 2: Feature screening. Due domain experts can follow his intuition, judgment, or knowledge, or uses rules built in the subject domain to stipulate the conditional and decisional attributes, and the expert experiences yield more sensible selections on the attributes than the automatic feature-selection approach. Thus, this feature-screening step follows the expert recommendation to select essential attributes for classifying AEAC from AEAS data set. As a result, one decisional-attribute AEAC and six conditional attributes are remained in the AEAS data set that the 927 instances are experimented. Consequently, the seven conditional-attributes are Age, Gender, Division, Location, Phase, Type, and Category. The decisional attribute are merged into three classes, including drug event, human event, and others. Except for the continuous attribute Age, all the attributes are symbolic data.

Step 3: Data discretization. Automatic data discretization method is used to process conditional attributes. Primarily, this step runs a Lilliefors test [25] for normality. In the test results, it is proven that the experimental data set is appropriate for the CPDA. This step is divided into two substeps to partition the given data.

Substep 3-1: Decision tree C4.5 algorithm is run to perform a 10-fold cross-validation for discretizing the continuous condition attribute Age to obtain the cutoff points, and two thresholds on 36 and 59 are obtained.

Substep 3-2: The CPDA technique is accordingly applied to discretize the continuous attribute Age into three linguistic values, L_1 (low), L_2 (medium), and L_3 (high), according to the previous running results on implementing the decision tree C4.5 algorithm. Subsequently, Table I lists the interval values of three linguistics on the conditional-attribute Age by using the CPDA technique. Figure 1 shows the membership functions of Age attribute based on the three bounds, lower, midpoint, and upper, in Table I, and Table II lists the membership degrees and corresponding values of the three linguistics on the Age attribute. The value of maximum membership in Table II is used to make values of the above linguistics.

TABLE I. THE LINGUISTIC INTERVAL VALUES OF THE CONTINUOUS ATTRIBUTE AGE BY USING THE CPDA TECHNIQUE

Linguistic value	Linguistic interval		Universe of discourse *U*			
	PLB	***PUB***	***Lower***	***Midpoint***	***Upper***	***Interval length***
L_1	0.00	0.33	0.00	21.84*	43.67	43.67#
L_2	0.17	0.67	31.78	47.27	62.77	30.99
L_3	0.50	1.00	53.22	85.69	96.00	42.78

*(0.00+43.67)/2=(21.84), and #(43.67-0.00)=43.67

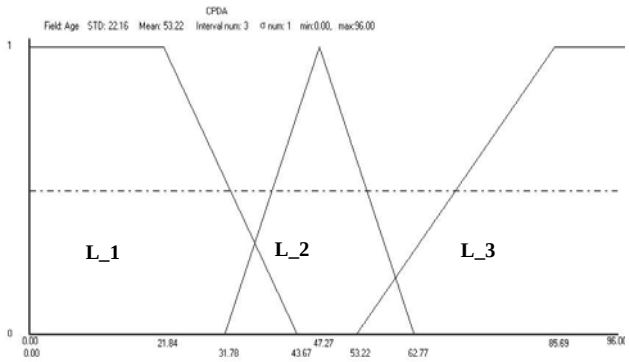

Figure 1. Membership function of the continuous attribute Age by using the CPDA technique

TABLE II. THE DEGREES OF MEMBERSHIP AND CORRESPONDING VALUES OF LINGUISTICS ON THE CONTINUOUS ATTRIBUTE AGE

Age value	L_1	L_2	L_3	Linguistic value
32	0.53*	0.01	0	L_1
61	0	0.11	0.24	L_3
21	1	0	0	L_1
47	0	1	0	L_2
17	1	0	0	L_1
⋮	⋮	⋮	⋮	⋮

*(43.67-32.00)/(43.67-21.84)=0.53

Step 4: Rule induction. In the generated linguistic values from the Step 3, this step applies the LEM2 algorithm of RS to deduce a knowledge-based decision rule set from the training (66% of the experimental data) data set. In total, 89 rules are extracted. For the limited

space, only the Rule 1 is expressed. For example, Rule 1: (Type=Web) & (Location=Dispensary) => (Category=Drug). This expression indicates that if type of reporting is from web site and happening location is in dispensary. The event aviso category then is belonging to drug event. The remaining testing (34%) data set is then experimented and the accuracy rate is acquired.

Step 5: Rule filtering. The algorithm of filtering rules from the RS classifier conducts a filtering procedure that eliminates decision rules with the support threshold below 2 for the reason that only one real example supports this rule. Accordingly, the rule set remains 38 decision rules, and the classification accuracy rate is refined.

Step 6: Performance evaluations and comparisons. To confirm the classification performance of this proposed procedure, the AEAS data set is separated into two parts of sub-datasets randomly again: the 66% training data set and the 34% testing data set. Experiments are made with the repetition of 66%/34% random separation by using five classification techniques—including decision tree-C4.5 [22], logistic [29], Bayes net [30], multilayer perceptron (MLP) [31], and the proposed procedure. Consequently, Table III describes the experimental results using the five methods with the seven selected attributes included in the AEAS data set.

TABLE III. THE COMPARISON RESULTS OF DIFFERENT METHODS IN THE AEAS DATA SET

Model	Accuracy (Rank)
Decision tree-C4.5 [22]	90.32% (2)
Logistic [29]	87.65% (5)
Bayes net [30]	89.57% (3)
MLP [31]	89.03% (4)
The proposed procedure	98.71% (1)

IV. ANALYTICAL RESULTS AND MANAGEMENT IMPLICATIONS

The study further examines the implied analytical results and the extracted decision rules to dig meaningful information hidden in the experimental data set as follows:

1) A meaningful phenomenon in related expert feature-screening is explored: The study provides an exploration of the effect on a good example for using the expert recommendation of selecting conditional attributes. This study further offers evidence that the reduced number and complexity of irrelative condition-attributes is effective to improve accuracy of classification by the expert feature-screening method in advance.

2) Certain model attributes are redundant for classifying event aviso category: Based on the expert recommendation, some superfluous attributes are emerged or can be removed from the experimental data set. The expert feature-screening result of the proposed procedure is helpful to remove the irrelevant/redundant attributes that are not correlated with classifying event aviso category for the given input data and for diminishing computational costs.

3) The proposed procedure performs well in classifying AEAC: The proposed procedure performs well in classification accuracy listed in Table III in the AEAS data set. The information implies that the proposed procedure is an effective alternate applied for classifying AEAC in the healthcare industry.

4) Knowledge-based decision rules are generated: The proposed procedure applies the algorithm of RS LEM2 to derive a meaningful decision rule set and to support the form of 'if…then' decision rule set that can be taken as a knowledge-based healthcare service system for understanding the related hidden information of accident event aviso systems and for providing intelligently powerful explanations for interested parties.

V. CONCLUSIONS

The study has proposed an effectual hybrid classification approach to solve AEAC problems of patient safety and healthcare quality issues in the healthcare industry. The proposed procedure implements the following five components and functions from an intelligent perspective, including an expert recommendation-based feature-screening approach, the decision tree C4.5 algorithm, the CPDA, (4) the RS LEM2 algorithm, and a rule filter. Two key directions are accordingly concluded with a real AEAS data set. (1) The analytical results indicate the proposed procedure has satisfactory resilts in the AEAS data set, and thus, outperforms the other listed methods. (2) Especially, the proposed procedure is constructed by a rule-based classifier of RS and is applicable for knowledge discovery in a complicated professional field like the healthcare systems. Although the proposed procedure performs well, further experiments and improvements are still necessary. For example, use other data sets into the proposed procedure for further analysis in various industries to handle different classification problems.

ACKNOWLEDGMENT

The author would like to much appreciate Dr. Cheng-Yi Hsu (a medical doctor) for his useful background knowledge, which was very helpful in preparing this study. The author also would wish to thank the National Science Council of the Republic of China, Taiwan, ROC, for financially supporting this research under Contract No. NSC 101-2221-E-146-002.

REFERENCES

[1] H. Kaplan, J. B. Battles, T. W. Van der Schaaf, C. E. Shea, and S. Q. Mercer, "Identification and classification of the causes of events in transfusion medicine," Transfusion, vol. 38, no. 11-12, 1998, pp. 1071–1081.

[2] L. Kohn, J. Corrigan, and M. Donaldson, "To err is human: building a safer health system, Washington," DC: Institute of Medicine, Committee on Quality of Healthcare in America, 2000.

[3] H. J. Murff, V. L. Patel, G. Hripcsak, and D. W. Bates, "Detecting adverse events for patient safety research: a review of current

methodologies," Journal of Biomedical Informatics, vol. 36, no. 1–2, 2003, pp. 131–143.

[4] D. Zapf and J. T. Reason, "Introduction: Human errors and error handling," Applied Psychology: An International Review, vol. 43, 1994, pp. 427–432.

[5] Z. Pawlak, "Rough sets," Informational Journal of Computer and Information Sciences, vol. 11, no. 5, 1982, pp. 341–356.

[6] S.K. Chong, M. M. Gaber, S. Krishnaswamy, and S. W. Loke, "Energy conservation in wireless sensor networks: a rule-based approach," Knowledge and Information Systems, vol. 28, 2011, pp. 579–614.

[7] Y. S. Chen and C. H. Cheng, "A soft-computing based rough sets classifier for classifying IPO returns in the financial markets," Applied Soft Computing, vol. 12, no. 1, 2012, pp. 462–475.

[8] Y. S. Chen and C. H. Cheng, "Assessing mathematics learning achievement using hybrid rough set classifiers and multiple regression analysis," Applied Soft Computing, vol. 13, no. 2, 2013, pp. 1183–1192.

[9] Y. S. Chen and C. H. Cheng, "Hybrid models based on rough set classifiers for setting credit rating decision rules in the global banking industry," Knowledge-Based Systems, vol. 39, 2013, pp. 224–239.

[10] S. Karthik, A. Priyadarishini, J. Anuradha, and B. K. Tripathy, "Classification and rule extraction using rough set for diagnosis of liver disease and its types," Advances in Applied Science Research, vol. 2, no. 3, 2011, pp. 334–345.

[11] Y. S. Chen and C. H. Cheng, "Application of rough set classifiers for determining hemodialysis adequacy in ESRD patients," Knowledge and Information Systems, vol. 34, no. 2, 2013, pp. 453–482.

[12] S. Tan, X. Cheng, and H. Xu, "An efficient global optimization approach for rough set based dimensionality reduction," International Journal of Innovative Computing, Information and Control, vol. 3, no. 3, 2007, pp. 725–736.

[13] Z. Pawlak, "Rough sets," theoretical aspects of reasoning about data, The Netherlands: Kluwer, Dordrecht, 1991.

[14] J. W. Grzymala-Busse, "LERS—a system for learning from examples based on rough sets, In: Slowinski R (ed) Intelligent decision support," Kluwer Academic Publishers, Dordrecht, 1992, pp. 3–18.

[15] J. W. Grzymala-Busse, "A new version of the rule induction system LERS," Fundamenta Informaticae, vol. 31, no. 1, 1997, pp. 27–39.

[16] J. W. Grzymala-Busse, W. J. Grzymala-Busse, and L. K. Goodwin, "Coping with missing attribute values based on closest fit in preterm birth data: a rough set approach," Computational Intelligence: An International Journal, vol. 17, no. 3, 2001, pp. 425–434.

[17] J. W. Grzymala-Busse, "MLEM2 rule induction algorithms: with and without merging intervals," Studies in Computational Intelligence, vol. 118, 2008, pp. 153–164.

[18] H. Sakai and M. Nakata, "On rough sets based rule generation from tables," International Journal of Innovative Computing, Information and Control, vol. 2, no. 1, 2006, pp. 13–31.

[19] H. S. Nguyen and S. H. Nguyen, "Analysis of stulong data by rough set exploration system (RSES), in: Berka P (ed.) Proceedings of the ECML/PKDD Workshop 2003 Discovery Challenge," 2003, pp. 71–82.

[20] J. Han and M. Kamber, "Data mining: concepts and techniques," San Francisco: Morgan Kaufmann Publishers, 2001.

[21] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, 1986, pp. 81–106.

[22] J. R. Quinlan, "C4.5: Programs for machine learning," CA: Morgan Kaufmann, San Mateo, 1993.

[23] P. J. Acklam, "An algorithm for computing the inverse normal cumulative distribution function," Available from http://home.online.no/~pjacklam/notes/invnorm/, 2004.

[24] J. L. Devore, "Probability and statistics for engineering and the sciences," Duxbury, Belmont, 2004.

[25] G. E. Dallal and L. Wilkinson, "An analytic approximation to the distribution of Lilliefors' test for normality," The American Statistician, vol. 40, 1986, pp. 294–296.

[26] H. J. Teoh, C. H. Cheng, H. H. Chu, and J. S. Chen, "Fuzzy time series model based on probabilistic approach and rough set rule induction for empirical research in stock markets," Data and Knowledge Engineering, vol. 67, 2008, pp. 103–117.

[27] Math Works Incorporation, "Internet Communication," Available from http://www.mathworks.com/help/toolbox/stats/normcdf.html.

[28] L. A. Zadeh, "Fuzzy sets," Information Control, vol. 8, 1965, pp. 338–353.

[29] S. le Cessie and J. C. van Houwelingen, "Ridge estimators in logistic regression," Applied Statistics, vol. 41, 1992, pp. 191–201.

[30] K. P. Murphy, "Bayes Net ToolBox," Technical report, MIT Artificial Intelligence Labortary, Available from http://www.ai.mit.edu/~murphyk/, 2002.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in: Rumelhart, D.E. et al., (Eds.), Parallel Distributed Processing, MIT Press, Cambridge, MA, 1986, pp. 318–362.