Experienced Physicians and Automatic Generation of Decision Rules from Clinical Data

William Klement[1], Szymon Wilk[1,3], Martin Michalowski[2] and Ken Farion[3]

[1] MET Research Group, University of Ottawa, Canada.
[2] Adventium Labs, Minneapolis MN 55401, USA.
[3] Institute of Computing Science, Poznan University of Technology, Poland.
[4] Division of Emergency Medicine, Children's Hospital of Eastern Ontario, Canada.

Abstract. Clinical Decision Support Systems embed data-driven decision models designed to represent clinical acumen of an experienced physician. We argue that eliminating physicians' diagnostic biases from data improves the overall quality of concepts, which we represent as decision rules. Experiments conducted on prospectively collected clinical data show that analyzing this filtered data produces rules with better coverage, certainty and confirmation. Cross-validation testing shows improvement in classification performance.

1 Introduction

The functionality of clinical decision support systems (CDSS) relies on their embedded decision models that represent knowledge acquired from either data or domain experts. Data-driven models are created to acquire knowledge by deriving relationships between data features and decision outcome. In medical domains, while the data describes patients with a clinical condition, the decision indicates a diagnostic outcome. These diagnostic decisions are normally transcribed from patient charts and are verified for correctness, e.g. by a follow-up. Traditionally, a verified outcome forms the gold standard (GS) used in the analysis of the decision models. In this paper, we assume a data-driven approach but argue that the reliance in the analysis on the GS may skew the resulting relationships. Our goal is to show that the use of cases, where the experienced physician (EP) makes correct diagnoses according to the verified patient outcome (GS), results in better decision models. To acquire unbiased clinical knowledge, we argue that it is essential to eliminate records where EP decisions do not match GS prior to constructing a decision model from data.

In clinical domains, patient records represent instances of a relationship, between attribute values, describing the status of patients' health and diagnostic decisions made by the physician. Several studies establish that EP makes good clinical decisions, particularly when dealing with critical cases. However, EPs often err in favor of caution and tend to over-diagnose patients who are relatively healthy. Therefore, their decisions are characterized by high sensitivity and lower specificity. From a decision making perspective, low specificity if not controlled introduces noise in the data. This can be evident by discrepancies between decisions made by EP and those established by the GS.

Our approach in identifying unbiased clinical knowledge in the data represents a departure from the established common practice where relationships are sought between patient attribute values and class labels (diagnoses) derived from the GS. These decision models are constructed from data collected by EPs but the diagnostic outcome is established from the GS. This practice is skewed because resulting models represent knowledge distorted by erroneous characteristics associated with the EP's decision models. As a remedy, we propose to eliminate EPs' biases from the data. Cases for which GS and EP decisions match constitute "correct decisions", are particularly valuable, and provide "sound" clinical knowledge. Classical methods construct their decision models from all available records regardless of the correctness of their EP decisions. We argue that such model construction may introduce bias in discovered knowledge, and we propose to focus the analysis on cases with correct EP decisions only.

This paper aims to demonstrate the value of these "correct" cases in the context of knowledge acquisition from patient data. We apply our analysis to two clinical domains, the diagnosis of pediatric abdominal pain (AP) and pediatric asthma exacerbation (AE). To this extend, our experiment shows that filtering the data, based on the match between EP decisions and the GS, produces crisp knowledge. Naturally, we must clarify what form of knowledge we extract, and how we measure its quality. To represent knowledge, we exploit concepts of rough set theory that represent it by a set of minimal decision rules [9, 8]. We apply the MODLEM algorithm [11, 12] to generate these decision rules. For evaluation, we employ several metrics to assess the quality of rules based on their structure and their performance on data. They include; the number of generated decision rules, length, coverage, certainty, and confirmation and are reviewed in [4].

The paper is organized as follows. Section 2 discusses clinical domains used for our experiments, and section 3 describes basic principles behind rough sets, the MODLEM algorithm used for rule generation, and rule evaluation metrics. Experimental design and results are discussed in section 4, and conclusions can be found in section 5.

2 The role of EP's expertise

Clinical decision-making is a complex process influenced by a verity of uncertain factors and should include the integration of clinical expertise [5]. Information technology solutions have been commonly considered as decision support mechanisms to provide clinicians with appropriate information while making clinical decisions. Such solutions include Clinical Decision Support Systems (CDSS) which have increasingly captured the attention of the medical community in recent years. A CDSS is defined as *"any program designed to help healthcare professionals make clinical decisions"* [7]. CDSS provide information in three widely accepted categories including; information management, focusing attention on specific health events, and patient-specific recommendations. The latter helps physicians make two types of decisions. The first is diagnostic where the focus is set on the patients underlying health condition, and the second type

deals with patient management with regards to what treatment plan is most appropriate for the patient. Despite the varying techniques for extracting expert knowledge, patient-specific CDSS decision models almost always reflect clinician expertise with the embedding of the knowledge of the 'best practice'. Obviously, the knowledge of an EP is vast. It has been documented that a physician is considered an expert after 10 years of training [10] who is able to summarize information and to develop a complex network of knowledge [1].

The focus of our research is to enhance and to support the process of acquiring expert knowledge from patient-specific data, we are able to capture it by filtering the data according to the EP's correct practice. Such knowledge can be used in the construction of decision models ready for integration into CDSS. To this extent, we wish to exploit decisions made by the EPs which reflect their clinical acumen. When comparing their decisions to the verified patient outcomes, the GS, physician's diagnostic biases become clear. Investigating the circumstances of these biases is a difficult task as they can be caused by multitude of factors including differing expertise of physicians [5]. To account for the diagnostic biases, we propose to rely only on correct EP decisions, and therefore we consider data for those patients where EP decisions match the GS.

AE data was collected as a part of a study conducted at the Children Hospital of Eastern Ontario (CHEO), and it includes patients who visited the hospital emergency department (ED) experiencing asthma exacerbation. In the ED, a patient is repeatedly evaluated by multiple clinicians at variable time intervals. This information is documented and collected prospectively for each patient. The resulting patient records contain information about history, nursing, physician triage assessment, and reassessment information collected approximately 2 hours after triage. Records in the AE data set are assigned to one of two outcome classes: mild or other severity of exacerbation. The verified severity of exacerbation is used as a GS. The dynamic nature of asthma exacerbation and the collection of assessments over time would lend itself naturally to a temporal representation for analysis of data. However, inconsistencies in data recording meant it was not possible to incorporate a temporal aspect into the analysis.

The AP data is also collected in the ED of CHEO and includes patients who have serious conditions, mostly appendicitis, who require surgery. However, most records describe benign causes. Before a cause can be found, symptoms often resolve without complications so that a definitive diagnosis is not possible during the ED visit. Therefore, choosing the correct triage plan is an important proxy [2] and we use it as a class label. This triage plan may involve discharging the patient, continuing observation, or asking for a specialty consultations. In our AP data, these outcomes are transformed into binary values indicating whether a patient requires specialist consultation. As with the AE data, a GS was established from verified patient outcomes.

Table 1. The contingency table of a decision rule "$if X$ *then* Y".

	Y	$\overline{Y}$	
X	a	b	r_1
$\overline{X}$	c	d	r_2
	c_1	c_2	

3 Generating and assessing decision rules

Based on the mathematical model of rough set theory [9], we generate a minimal set of decision rules using the MODLEM algorithm described in [13]. These decision rules represent knowledge extracted from data, and we assess their quality using several measures presented in [4] and discussed later in this section. Our objective is to show that analysis conducted on cases where EP decisions are correct result in better rules, and therefore higher quality knowledge, than those performed on data with the GS being the class label.

Rough set analysis rely on an information table which contains data points (examples) described by a finite set of attributes. Such table becomes a decision table when we are able to identify a set of condition attributes C and relate them to a set of decisions D. From a decision table, we can induce decision rules of the form "*if* $\cdots$*, then* $\cdots$". We now describe an intuitive illustration appropriate for our domains which appears in [4]. Given a data sample describing patients and their diseases, the set of signs and symptoms $S = \{s_1, \cdots, s_n\}$ contains their condition attributes and a set of diseases $D = \{d_1, \cdots, d_m\}$ as their decision attributes. A decision rule has the form "*if symptoms* $s_i, s_j, \cdots, s_w$ *appear, then there is disease* d_v" with $s_i, s_j, \cdots, s_w \in S$ and $d_v \in D$.

The MODLEM algorithm [13] is designed to induce such decision rules based on the idea of *sequential covering* to generate a *minimal set* of decision rules for every decision concept. A decision concept may be the decision class or a rough approximation of the decision class in the presence of inconsistent examples. The objective of this minimal set of decision rules is to cover all the positive examples in the positive class without covering any negative examples. A benefit of using MODLEM lies in its ability to process numerical attributes without discretization. In addition, this algorithm has been shown to produce effective and efficient single classification models [12]. The process of generating the set of minimal decision rules is iterative. For every decision class, the MODLEM algorithm repeatedly builds decision rules to cover examples in that class, then, it removes examples covered by this rule from the data. This process continues until all examples in the class are covered, and the "*best*" rules are selected according to a chosen criterion, e.g. class entropy. For more detailed description of the MODLEM algorithm, we refer to [13].

Comparing the characteristics and performance of decision rules has long been a subject of research. In this paper, we utilize several classical rule evalua-

tion measures including the rule confirmation measure, all of which, are reviewed in details in [4]. We also illustrate calculations and present interpretations of metrics used in our experiments. For simplicity, let "*if* X *then* Y" be a decision rule where X is a subset of conditions and Y is a decision class. Applying this decision rule to data produces entries that populate the contingency Table 1. Essentially, this table depicts counts of examples that are covered by all possible combinations of either side of the decision rule. In the data, while there are a examples that satisfy the set of conditions X whose decision is Y, examples that fail conditions X and their decision is $\overline{Y}$[5] are depicted by d. Similarly, b is the number of examples that meet conditions X but their decision is $\overline{Y}$, and finally, c is the count of examples that fail conditions X but their decision is Y. While c_1 and c_2 represent the column summations, r_1 and r_2 are the row summations. Their interpretations are simple; the column summations show the number of examples whose decision class is Y or $\overline{Y}$ respectively, and the row summations indicate the number of examples that satisfy X or $\overline{X}$ also respectively.

Rule evaluation measures are well established and fall into two main categories; the first involves assessing the structure of the rule, and the second relates to their performance. While the former is based primarily on the length of the rule, the latter includes rule *coverage, certainty,* and *confirmation*. These measures are discussed in [4], and we compute their values for each rule by counting entries in Table 1, then, we substitute their values in equations 1, 2, and 3.

$$coverage(X, Y) = \frac{a}{c_1} \tag{1}$$

$$certainty(X, Y) = \frac{a}{r_1} \tag{2}$$

$$confirmation(X, Y) = \frac{ad - bc}{ad + bc + 2ac} \tag{3}$$

While higher coverage values depict the strength of the rule, a high value of rule certainty indicates higher confidence. In addition, several measures have been proposed to asses rule confirmation. However, we use the $f()$ measure presented in [4], which quantifies the degree to which the observed evidence supports for, or against, a given hypothesis. The findings in [4] show its effectiveness.

To assess the quality of knowledge extracted from data in the form of decision rules, we first consider characteristics describing this set. Such characteristics include the number of decision rules, the number of conditions, and the average length of a rule with the associated standard deviation. Such characteristics reflect the complexity of the concept in the sense that, while complex concepts may have more rules, these rules tend to be longer because they include more conditions. This is consistent with simpler, more effective rules having fewer conditions, thus they are shorter in length and there are fewer of them. From a performance assessment perspective, concepts which are described by fewer rules, show greater coverage, produce higher levels of certainty, and have better confirmation are considered of better quality.

[5] $\overline{X}$ is $\neg X$, the complement of X.

Table 2. Characteristics of the clinical data sets.

		GS Outcome			EP Decisions	
Data	Examples	Positives	Negatives	Ratio	Positives	Negatives
AE_{all}	240	131	109	55%	136	72
AE_{corr}	140	90	50	64%	90	50
AP_{all}	457	48	409	11%	55	402
AP_{corr}	422	34	388	8%	34	388

4 Experimental design and results

The objective of our experiment is to demonstrate that the quality of knowledge, acquired from clinical data, is improved by including only those patient cases for which the EP makes correct decisions as indicated by the verified patient outcome, the GS. Furthermore, the experiment shows that the performance of the associated classification model can also be improved using the proposed filtering of patient records. We conduct our experiment in two phases. In the first phase, the knowledge acquisition phase, we generate a minimal set of decision rules and record their characteristics for analysis. The second phase uses 10-fold cross-validation runs repeated 5 times to evaluate the performance of the decision models after filtering the training data.

Data sets used in the experiment and their characteristics are listed in Table 2. The subscript *all* for AP and AE indicates that all patient records are analyzed (non-filtered). Similarly, the subscript *corr* indicates filtered data sets, where class labels correspond to EP decisions that match GS after eliminating mismatching records. Examining Table 2 reveals that while the class distribution of the AE data is almost balanced, it is not so for the AP data where the ratio of positive examples is less than 11%. Therefore, we use an under-sampling technique to balance it by randomly selecting, without replacement, an equal number of examples in both classes to retain the complete set of positive examples. A data set after under-sampling is labeled APS. This set is processed in two settings; the first consists of all examples that are randomly under-sampled and is indicated by APS_{all}. The second is denoted by APS_{corr} and includes an under-sample set of cases for which EP decision are correct, i.e. we under-sample data resulting from EP-based filtering.

A pairwise comparison of the number of examples on the *all* rows to those on the *corr* rows of Table 2, respectively, shows that EP makes more correct decisions in the AP domain than in the AE. In both domains, however, EPs over-diagnose patients as having a positive condition more often than indicated by the GS. This is seen by comparing the number of positive EP decisions for both *all* and *corr* rows in both domains.

Table 3. Characteristics of resulting concepts.

	Both Classes			+ Class			– Class		
Data	Cond.	Rules	Ave. Length	Cond.	Rules	Ave. Length	Cond.	Rules	Ave. Length
AE_{all}	199	50	3.98 ±1.19	92	23	4.00 ±1.00	107	27	3.96 ±1.34
AE_{corr}	13	6	2.17 ±0.98	5	3	1.67 ±1.16	8	3	2.67 ±0.58
AP_{all}	163	38	4.29 ±1.51	99	21	4.71 ±1.65	64	17	3.77 ±1.15
AP_{corr}	83	24	3.46 ±1.14	37	10	3.70 ±1.25	46	14	3.29 ±1.07
APS_{all}	91	25	3.64 ±0.95	51	14	3.64 ±1.08	40	11	3.64 ±0.81
APS_{corr}	19	8	2.38 ±0.52	9	4	2.25 ±0.50	10	4	2.50 ±0.58

4.1 Discussion

We begin discussion by considering characteristics recorded for each concept extracted from the data. A concept is represented by a set of decision rules for which we show the number of conditions, the number of rules, and the average rule length with its standard deviation for both and for individual classes. In this order, these values are shown in the columns of Table 3. Examining these characteristics on individual classes shows that the rules are almost evenly distributed on the positive and on the negative class. Individually, they have an almost equal number of conditions, number of rules, and average rule length.

An important observation points to the fact that values recorded for data sets with *corr* index are smaller than those recorded for the non-filtered data sets. For all three data sets, AE, AP, and APS, the set of decision rules generated form data containing correct EP decisions results in fewer conditions, fewer rules, and shorter average rule length with a lower standard deviation. This observation remains consistent whether we consider the set of decision rules describing both classes or individual classes. This suggests that using data with correct EP decisions produces concepts with less complexity and possibly ones with higher quality.

Results of evaluating the performance of these decision rules are shown in Table 4. The pairwise comparison of average values for each performance measure reveals that they remain unchanged or increase when rules are generated form data containing correct EP decisions. With such filtering, the average rule coverage increases dramatically for the AE data. For the AP data, the use of under-sampling allows the average rule coverage a higher increase than that obtained without under-sampling. This is seen when we compare the difference in average coverage values of APS_{all} and APS_{corr} against that for AP_{all} and AP_{corr}. This is attributed to the imbalanced class distribution of the AP data.

The average certainty on the AE domain achieves its maximum value of 1 and remains unaffected by the elimination of cases with incorrect EP decisions, see average certainty values for AE_{all} and AE_{corr}. A similar statement can be made for the average rule confirmation in the AE domain in the same table. On the

Table 4. Assessing the performance of resulting decision rules.

Measure	Data	Both classes	+ Class	– Class
Coverage	AE_{all}	0.055 ±0.044	0.062 ±0.052	0.048 ±0.036
	AE_{corr}	0.534 ±0.311	0.507 ±0.362	0.560 ±0.330
	AP_{all}	0.117 ±0.153	0.068 ±0.051	0.176 ±0.209
	AP_{corr}	0.158 ±0.204	0.176 ±0.130	0.145 ±0.248
	APS_{all}	0.137 ±0.132	0.131 ±0.128	0.145 ±0.143
	APS_{corr}	0.423 ±0.209	0.338 ±0.098	0.507 ±0.271
Certainty	AE_{all}	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
	AE_{corr}	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
	AP_{all}	0.955 ±0.157	0.929 ±0.208	0.988 ±0.030
	AP_{corr}	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
	APS_{all}	0.969 ±0.104	0.964 ±0.134	0.974 ±0.053
	APS_{corr}	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
Confirmation	AE_{all}	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
	AE_{corr}	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
	AP_{all}	0.914 ±0.219	0.953 ±0.165	0.867 ±0.269
	AP_{corr}	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
	APS_{all}	0.937 ±0.210	0.928 ±0.270	0.949 ±0.104
	APS_{corr}	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000

AP data and regardless of using under-sampling, the average rule certainty and the average confirmation are both improved by our filtering, their average values for AP_{corr} and APS_{corr} are increased over AP_{all} and APS_{all} respectively. Such results lead to the conclusion that generating decision rules using examples with correct EP decisions enhances the coverage, the certainty and the confirmation of the rules. However, the standard deviation increases for the rule coverage measure in both domains. This is not surprising because data with cases of correct EP decisions are always smaller than their respective original sets, their sizes are shown in Table 2, but the average values of rule certainty and confirmation achieve their maximum of 1 with a standard deviation of 0. Clearly, our filtering helps the model achieve high certainty and strong confirmation.

Our final results are given in Table 5, which shows the average sensitivities, specificities, accuracies, and the geometric means[6] of sensitivities and specificities resulting from testing the classification performance of the decision rules. The testing method relies on the 10-fold cross-validation repeated 5 times for which the above averages are recorded. Results for AE data show a clear gain in sensitivity with a loss in specificity, accuracy and geometric mean. Clearly,

[6] The geometric mean measure is used for imbalanced class distributions [6]

Table 5. Classification performance of resulting decision rules.

Data	Sensitivity	Specificity	Accuracy	Geometric Mean†
AE_{all}	0.7024 ± 0.1510	0.5930 ± 0.1634	0.7158 ± 0.0865	0.6341 ± 0.1059
AE_{corr}	0.7908 ± 0.1268	0.5070 ± 0.1309	0.6825 ± 0.0803	0.6243 ± 0.1029
AP_{all}	0.4930 ± 0.2231	0.9640 ± 0.0264	0.9252 ± 0.0322	0.6619 ± 0.1952
AP_{corr}	0.5877 ± 0.2236	0.9526 ± 0.0368	0.9305 ± 0.0310	0.7284 ± 0.1696
APS_{all}	0.7913 ± 0.2246	0.7856 ± 0.0525	0.8127 ± 0.0406	0.7777 ± 0.1296
APS_{corr}	0.7470 ± 0.2353	0.8560 ± 0.0642	0.8559 ± 0.0509	0.7909 ± 0.1396

† Entries are averaged over 5 runs of 10-fold cross validation.

examples with correct EP decisions set their focus on the positive class.

For the AP data, the classification performance improves in principle. Balancing the AP_{all} data by under-sampling, to produce APS data, improves the classification performance with the exception of specificity (0.96 to 0.79). However, under-sampling the EP-filtered data, which produces the APS_{corr} data, recovers the specificity (0.79 to 0.86). Consequently, combining our filtering approach with sampling techniques must be done with care.

Given that the positive class represents an acute medical condition, the need for a specialist consult for AP and the pronounced asthma exacerbation for AE, the resulting sensitivity values show that our decision rules produce a reasonable classification performance. The latter can be improved by conducting a comprehensive experiment to select an appropriate data mining method.

5 Conclusions

Data-driven knowledge acquisition techniques used to extract knowledge describing EP decision making is a complex process which involves various factors. This paper shows that capturing knowledge in the form of decision rules from examples of correct EP decisions results in a better description of knowledge. This is exemplified by reduced complexity characterized with fewer, shorter rules. The performance of these rules is also enhanced with better coverage, higher certainty, and increased confirmation. With enhanced quality of knowledge, the classification performance is shown to improve with increased sensitivity.

6 Acknowledgment

The authors acknowledge the support of the Natural Sciences and Engineering Council of Canada. The second author also acknowledges support of the Polish Ministry of Science and Higher Education (grant N N519 314435).

References

1. Arocha, J.F., Wang, D., Patel, V.: Identifying reasoning strategies in medical decision making: A methodological guide. Biomedical Informatics **38(2)** (2005) 154-171.
2. Farion, K., Michalowski, W., Rubin, S., Wilk, S., Correll, R., Gaboury, I.: Prospective evaluation of the MET-AP system providing triage plans for acute pediatric abdominal pain. Int. Journal of Medical Informatics, **77(3)** (2008) 208-218.
3. Farion, K., Michalowski, W., Wilk, S., O'Sullivan, D., Matwin, S.: A tree-based decision model to support prediction of the severity of asthma exacerbations in children. Journal of Medical Systems, 2009 (forthcoming).
4. Greco, S., Pawlak, Z., and Slowinski, R.: Can Bayesian confirmation measures be useful for rough set decision rules? Engineering App. of AI **17** (2004) 345-361.
5. Hine, M.J., Farion, K., Michalowski, W., Wilk, S.: Decision Making By Emergency Room Physicians And Residents: Implications for the Design of Clinical Decision Support Systems. International Journal of Healthcare Information Systems and Informatics **4(2)** (2009) 17-35.
6. Kubat, M., Holte, R.C., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning, **30(2-3)** (1998) 195-215.
7. Musen, M.A., Sahar, Y., Shortliffe, E.H.: Clinical decision support systems. Medical Informatics, Computer applications in healthcare and biomedicine. ($n^n d$ edition) New York: Springer 574-609.
8. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences **11** (1998) 341-356.
9. Pawlak, Z.: Rough Sets – Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Boston, 1991.
10. Prietula, M.J., Simon, H.: The experts in your midst. Harvard Business Review **67(1)** (1989) 120-124.
11. Stefanowski, J.: The rough set based rule induction technique for classification problems. In proceedings of Sixth European Conference on Intelligent Techniques and Soft Computing EUFIT98, 1998, pp. 109-113.
12. Stefanowski, J.: Algorithms of rule induction for knowledge discovery. Habilitation Thesis published as Series Rozprawy **361**, Poznan University of Technology Press, (2001) Poznan (in Polish).
13. Stefanowski, J.: On combined classifiers, rule induction and rough sets. Transactions on Rough Sets, **6** (2007) 329-350.
14. Wilk, S., Slowinski, R., Michalowski, W., Greco, S.: Supporting triage of children with abdominal pain in the emergency room. European Journal of Operational Research **160** (2005) 696-709.

This article was processed using the LaTeX macro package with LLNCS style