Ensemble based Cost-Sensitive Feature Selection for Consolidated Knowledge Base Creation

Syed Imran Ali
Department of Computer Science & Engineering,
Kyung Hee University
Yongin-si, Gyeonggi-do, Korea
imran.ali@oslab.khu.ac.kr

Sungyoung Lee
Department of Computer Science & Engineering,
Kyung Hee University
Yongin-si, Gyeonggi-do, Korea
sylee@oslab.khu.ac.kr

***Abstract*—This is paper proposes a knowledge construction system. The key objective of the system is to extract knowledge from structured data which is generally available in the form of electronic medical records (EMR). In this regard, the main focus of the research is to design and develop a domain-independent system that is capable of assisting the domain expert(s) in gaining non-trivial insights from the underlying EMR data. It is important to note that most of the research in the domain of cost-sensitive feature selection relies on black-box models which only provide a prediction of a final class label. Whereas, the goal of this research is to acquire insights for domain experts such as chronic kidney disease classification. This goal is achieved by designing and developing a knowledge construction system that is based on a two-stage methodology. Stage one deals with identifying salient cost-sensitive features in the EMR data, whereas, stage-two deals with consolidating knowledge (i.e. in the form of production rules) from a set of interpretable machine learning models. Finally, in order to demonstrate the efficacy of the system a chronic kidney disease case study is adopted.**

***Keywords*— Data driven system, Feature Selection, Interpretable Machine Learning models, Ensemble Models, Decision Tree Models.**

I. INTRODUCTION

Knowledge discovery in data (KDD) is a well-known methodology for extracting non-trivial insights from data [1]. KDD methodology is based on a number of steps where each step adds some value to the overall goal of gaining insights from the data. Insights can be in the form of a set of visualizations which show important trends e.g. financial time-series data, insights can also be in the form of automated decision making through computational models e.g. decision models for medical fraud detection, and finally these insights can be in the form of a set of plain rules which are of the format IF *X* THEN *Y*, in order to glean insights from large amounts of data for identifying salient patterns which are otherwise non-available. This research deals with assisting human domain experts in understanding important relationships are which are readily apparent in the data stored as electronic medical records [2]. A high level pictorial presentation of an overall system is shown in Fig. 1. There are three key processes in the design of the system. 'Data Processing' effectively deals with the ETL process. 'Information Processing' specifically deals with data preprocessing and salient feature selection tasks, whereas, 'Knowledge Creation' task deals with building multiple machine leaning models and finally consolidating the salient rules from those models, which are then provided to domain experts. In order to consolidate rules from different knowledge representations, a simple to understand and evaluate formalism is adopted such as production rules based on first-order logic.

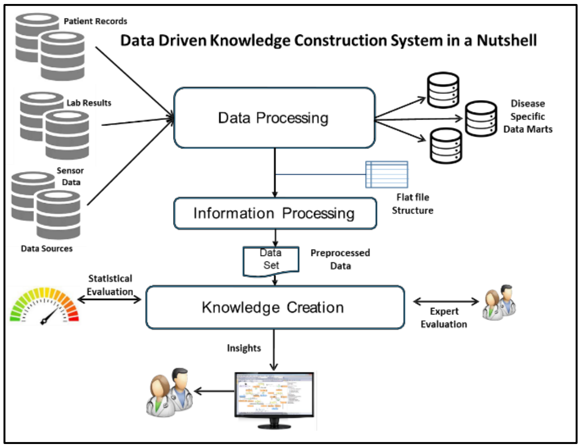

Figure 1. Idea diagram of the proposed system

The main goal of **"Data Processing"** task is to acquire patients' data from an array of different repositories such as patient's lab test results stored in different locations, perform data transformation in order to bring all the information under uniform headings, remove duplications, perform aggregations, etc. Data consistency check is also performed at this stage along with data type validation.

"Information Processing" task deals with identifying and correcting issues due to missing values in the EMR data, data normalization, and data discretization. In this entire process, the most important task is to select salient features. Since, data are acquired from different sources therefore it is expected to contain such features which may not be beneficial of machine learning algorithms such as medical record numbers, name of patients, registration/visit date, etc. Along with it, the data redundancy issue is also tackled with feature selection task. In this regard we propose an ensemble based feature selection method which takes account of both features importance in the dataset along with feature cost, and tends to find a trade-off solution. This approach is adopted in present research in order to empower the domain expert to factor in cost of constructing a decision model.

Lastly, **"Knowledge Consolidation"** goal is to provide a user with a set of interpretable machine learning models and then subsequently translate those models into production rules. Furthermore, the translated models are evaluated for inconsistencies in the rule such as in the following case where a set of rules are extracted from four different models. In example set 1, Rule R1 and R2 are in conflict with each other and in order to inforce consistency, either of the two will be filtered out [3]. Although enforcing the consistency is optional and will be performed on explicit expert's decision.

Likewise, in example set 2, both rules R3 and R4 can be used in different contexts. For example, R4 will be selected in case expert has enforced maximum specificity requirement, and R3 will be selected for such cases as which require rule-subsumption constraint [4], also explicitly enforced by the domain expert.

$\boldsymbol{Example\ Set\ 1}$:

$[\boldsymbol{M1}]\ R1: \boldsymbol{IF}\ < Predicate - A > \wedge < Predicate - B > \boldsymbol{THEN}\ < Class - X1 >$

$[\boldsymbol{M2}]\ R2: \boldsymbol{IF}\ < Predicate - A > \wedge < Predicate - B > \boldsymbol{THEN}\ < Class - X2 >$

$\boldsymbol{Example\ Set\ 2}$:

$[\boldsymbol{M3}]\ R3: \boldsymbol{IF}\ < Predicate - A > \boldsymbol{THEN}\ < Class - X3 >$

$[\boldsymbol{M4}]\ R4: \boldsymbol{IF}\ < Predicate - A > \wedge < Predicate - B > \wedge < Predicate - C > \boldsymbol{THEN}\ < Class - X3 >$

As it can be seen from aforementioned examples that knowledge consolidation from multiple intendent models is a challenging but an important task.

The main contributions of this paper are:

- An end-to-end design of a data driven knowledge construction methodology for EMR healthcare data repositories
- An ensemble based salient feature selection method which accounts for both statistical feature important and external factors such as cost of acquiring data for the feature
- A knowledge consolidation method which combines the results of a set of interpretable machine learning models in uniform manner (i.e. production rules).

The next section, is on ensemble feature selection. It describes in detail the feature selection process along with specific techniques employed for creating an ensemble method for feature selection and cost incorporation. Section III. deals with knowledge consolidation process using a set of machine learning models for uniform knowledge base creation. A detailed case study is presented in section IV for chronic kidney disease patients. The case study demonstrates how an expert can benefit from the cost-effectiveness the final solution. Section V concludes the paper along with future research directions.

II. ENSEMBLE-BASED FEATURE SUBSET SELECTION

Feature selection is one of the important tasks in a machine learning. The main objective of this task is to select a set of salient features which capture most of the information contained in a dataset [5,6]. Feature selection is an important technique to addresses issues such as 'curse of dimensionality' which arises in domain containing high dimensional datasets. In an EMR data, a patient record may contain some demographic information along with a set of features containing information about signs & symptoms, measurements from laboratory test results and final diagnosis performed by a physician. Hence, each patient record may contain a large number of features [7].

Furthermore, feature selection is divided into semantic preserving methods such as those which tend to select a subset of an original feature set. While non-semantic preserving methods tend to transform the original dataset such as Principal Component Analysis [8]. In this research, since we are concerned with the actual semantics present in the healthcare dataset for knowledge creation therefore, semantic preserving feature subset selection methods are selected. In this regard we have relied on three filter based methods which compute univariate feature score which in turn can be used for feature ranking. In order to find important of a feature in a given dataset we are using Symmetric

Uncertainty, Chi-Squared statistic, and filter method called Relief [9].

A large number of feature selection methods use information-theoretic measures for univariate feature selection. In this regard, symmetric uncertainty is widely used variants of normalized mutual information. The application of symmetric uncertainty in feature is in terms of information exchange between two feature vectors. In a univariate case, one of the vectors is an independent feature such as “age” of a patient, while the other vector is the dependent variable i.e. class variable such as “diagnosis”. This measure quantifies the mutual dependence of two variables as shown in Eq. (1).

$$SU(A,B) = 2\left[\frac{MI(A,B)}{H(A)+H(B)}\right] \qquad (1)$$

Where $MI(A,B)$ is the mutual information between feature A and feature B, and entropy of feature A and B is computed by $H(A) and\ H(B)$, respectively.

Chi-square statistics is used to compare expectations with that of original observed data. In feature selection, this test is used to evaluate the nature of relationship between two variables. Using *observed* and *counted* statistics one can test the independence of whether a strong correlation exists between an intendent variable and a dependent variable or not. Chi-square is computed as given in Eq. (2).

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} \qquad (2)$$

Where O_i denotes observed values for an instance ‘i’, and E_i represents expected values.

Third ranker approach is based on Relief algorithm, which provides a feature score based on their interactions and thereafter the provided scores can be subsequently used for of generating features ranks. Relief algorithm tends to compute a feature vector W according to Eq. (3).

$$W_i = W_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \qquad (3)$$

Where ‘*nearHit’* refers to the closet same-class instances and ‘*nearMiss’* refers to closet instances from other classes, and $(x_i - nearHit_i)$ denotes the Euclidean distance.

Based on these three aforementioned techniques, we have proposed an ensemble based feature selection method which takes into account different ranking approaches in order to generate a robust feature ranking as shown in Fig. 2.

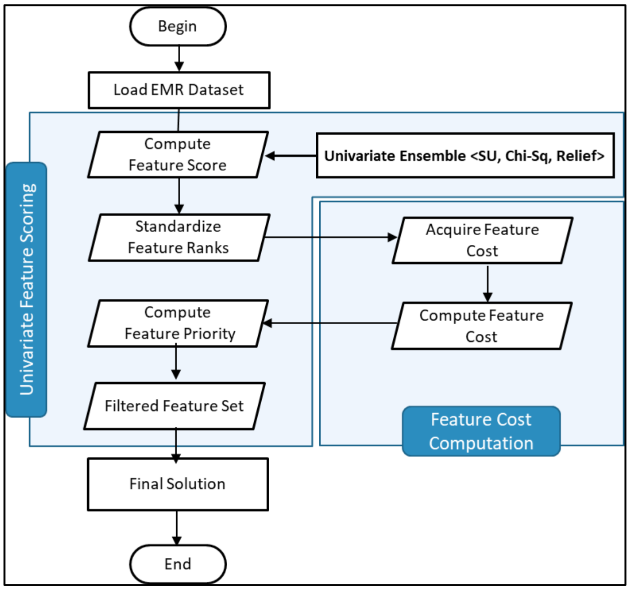

Figure 2. Ensemble-based Feature Selection

- *Compute Feature Score* is the first step in which three different feature scoring methods are applied in order to get multiple feature scores.
- *Standardize Feature Ranks* is performed by combining different scores in order to generate a single ranking. In this regard, for each feature we compute percentile of the feature in its ranked list. For example, “Age” is the best predictor among 10 features based on all three measures then combined score of “Age” feature would be as given in Eq. (4):

$$FeatureScore_{age} = \sum_{i=1}^{K} \frac{(\#Features - Rank_{age})}{\#Features} \qquad (4)$$

Where “K” is number of scoring measures i.e. 3 in our case, “#Features” denote total number of features in a dataset, and “Rank$_{age}$” refers to rank of feature age in a given scoring measure.

$$FinalFeatureScore_{age} = \frac{FeatureScore_{age}}{K} \qquad (5)$$

It is important to note that Eq. (5) normalizes the final features score values i.e. $\forall\ FeatureScore_i\ \epsilon\ (0,1)$.

- *Acquire Feature Cost* deals with expert provided meta-information about the feature under consideration which may influence the inclusion/exclusion decision regarding the feature. We are considering three types of cost factors i.e. monetary cost, availability cost, and risk cost.

- *Compute Feature Cost* deals with normalizing the cost factor between 0 and 1 for each type of the cost factor. Since cost factors may vary due to different measurement units therefore normalization would bring a uniformity to the cost factor.
- *Compute Feature Priority* deals with two type of information i.e. statistical importance of a feature which is obtained from scoring measures and applicability of a feature from cost-effectiveness perspective. Hence, a final rank is computed based on the importance of a feature as well as its cost.
- *Filtered Feature Set* tasks deals with selecting a threshold value for a cut-off point for selecting a final subset of features. This threshold value would vary for different datasets, therefore, based on each dataset's characteristics a specific threshold value is selected.
- *Final Solution* is the subset of features obtained from preceding stage, and the same is provided to knowledge consolidation process for building multiple machine learning models.

One of the important considerations in this research is accounting for the cost factor. In this research we are considering three different types of cost factors. Each cost factor has its associated context e.g. availability cost factor takes into account the ease of obtaining values for a certain medical test. In this regard, if the medical test can easily be performed and evaluated then the availability cost will be 0 and if the medical test can't be performed and hence it is highly unlikely that data for a given medical test would be readily available then the cost value will be 1, for all other cases, the value will be between 0 and 1 based on the judgement of the medical expert. Total cost of a feature "i" is given by Eq. (6)

$$FC_i = \frac{Mcost_{fi}}{tot_{Mcost}} + \frac{Acost_{fi}}{tot_{Acost}} + \frac{Rcost_{fi}}{tot_{Rcost}} \qquad (6)$$

Where *Mcostfi*, *Acostfi*, and *Rcostfi* are the feature i's monetary cost, availability cost and risk cost, respectively. Administration of a medical facility may provide actual cost values for both monetary and availability, while risk cost can be obtained from a medical expert. It is important to note that we normalize values obtained from Eq. (6) between 0 and 1.

III. KNOWLEDGE CONSOLIDATION

A desirable subset of features is obtained in the preceding step. In the knowledge consolidation step, a set of interpretable machine learning models are trained on a given subset of features. Each model is independently built in parallel. Afterwards, each model is translated into a uniform format i.e. first-order logic for production rules. At this stage the consolidated knowledge base is provided to the expert for inspection. An expert can provide an input test case, against which a set of three options are provided:

- Identify rules which are triggered for the test case and also detect the anomalies such as contradiction among rules
- Retrieve a set of most generalized rules for the test case (i.e. enforce the subsumption property of the knowledge base)
- Retrieve a set of most specific rules for the test case (i.e. enforce the maximum specificity property)

A step-by-step process of knowledge consolidation is shown in Fig. 3. A set of heterogeneous models are selected, in order to introduce diversity in the knowledge base.

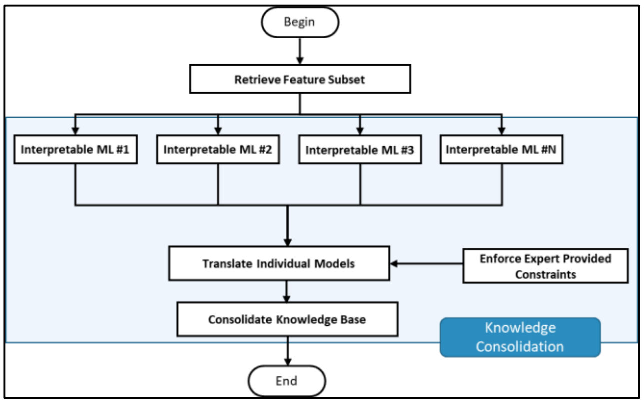

Figure 3. Knowledge Consolidation Process

It is important to note that on the user interface, the domain expert is provided with a set of options to select the desired classification model. The pool of classification models consists of such algorithm as which provide the final model either in the form of a decision tree or a decision list. Furthermore, in case of an ensemble model such as Random

Forrest, rules are extracted from its best decision tree among "n" generated decision trees. Since, classification models prefer shorter models over the large ones, following the Occam's Razor, therefore, a set of multiple different classification models are generated for creating the final consolidated knowledge base. Knowledge base created from multiple models is more robust in capturing important non-trivial patterns in the dataset. As a default case we use a set of five classifiers i.e. C4.5, CART, RIPPER, PART, and Ridor [9].

IV. CASE STUDY

Chronic kidney disease case study is adopted for demonstrating the consolidated knowledge base creation methodology. We have obtained a publicly available chronic kidney disease dataset [10]. This dataset contains 400 patients' data, each contains information about 24 independent features and 1 binary dependent feature. Moreover, out of 400 patients, 250 patients are diagnosed with chronic kidney disease. TABLE I provides description of the dataset along with a sample monetary cost of obtaining data for each feature [11]. Please note that both F-score and C-score are computed using Eq. (5) and Eq. (6), respectively.

TABLE I. Chronic Kidney Disease dataset description

FID	Feature Name	Data Type	F- Score	C-Score
1	Age	Numeric	0.25	0.00
2	Blood Pressure (mm/Hg)	Number	0.08	0.01
3	**Specific Gravity**	**Nominal**	**0.93**	**0.01**
4	**Albumin**	**Nominal**	**0.81**	**0.08**
5	Sugar	Nominal	0.44	0.06
6	Red Blood Cells	Binary	0.44	0.11
7	Pus Cells	Binary	0.61	0.09
8	Pus Cells Clumps	Binary	0.43	0.09
9	Bacteria	Binary	0.26	0.13
10	Blood Glucose Random	Numeric	0.42	0.11
11	Blood Urea	Numeric	0.18	0.09
12	Serum Creatinine	Numeric	0.08	0.09
13	Sodium	Numeric	0.08	0.02
14	Potassium	Numeric	0.10	0.15
15	Hemoglobin	Numeric	0.76	0.01
16	Packed Cell Volume	Nominal	0.61	0.01
17	White Blood Cell Count	Numeric	0.18	0.09
18	**Red Blood Cell Count**	**Numeric**	**0.89**	**0.09**
19	**Hypertension**	**Binary**	**0.90**	**0.01**
20	**Diabetes Mellitus**	**Binary**	**0.83**	**0.08**
21	CAD	Binary	0.38	0.15
22	Appetite	Nominal	0.71	0.01
23	Pedal Edema	Nominal	0.65	0.01
24	Anemia	Nominal	0.54	0.09

A comparison between F-core and C-sore of the features is given in Fig. 4. As it can be observed that according to the F-score, which represents combined ranking scores of feature importance measures, feature number 3, Specific Gravity, is the most important one, represented by blue line. Furthermore, orang line represents the combined cost of feature. Both these lines assist the domain expert in selecting a subset of features. An expert may select top 5 features while keeping the accumulated cost into account, since after that the cost remains constant and then again starts growing while the worth of the individual features is in constant decline. And since at feature point 13 both the lines cross each other therefore, a number of different subset can be selected till this point, after that the feature cost keeps on increasing while adding little overall value to the dataset.

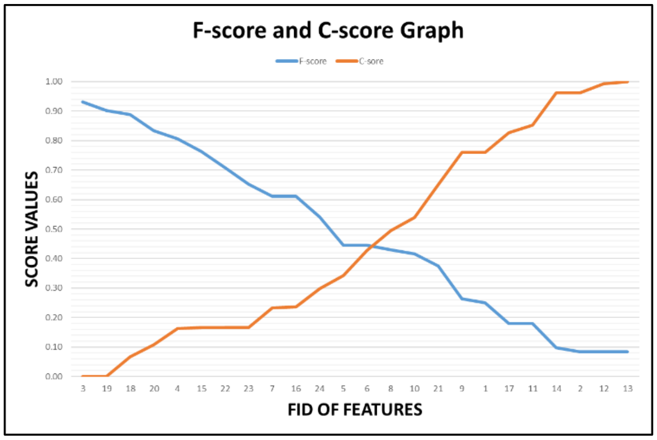

Figure 4. Comparison between F-score and C-score of Features

Combined percentile rank of the selected features is shown in Fig. 5. Based on these features we will a set of interpretable classification models are created. These models are used for generating the final consolidated knowledge base.

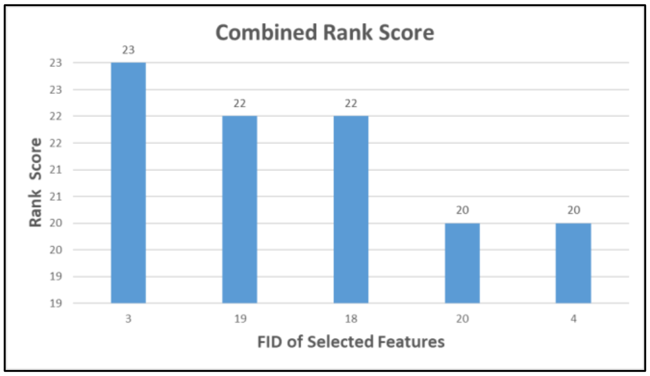

Figure 5. Combined percentile rank score of the selected features

A comparison is drawn between interpretable model construction for original dataset, 24 features, and reduced dataset based on 5 features. As it can be seen in TABLE II, the reduced dataset has retained most of the important characteristics of the original dataset while reducing the over cost of the acquiring the data. The original dataset has 1.0 C-score i.e. in terms of combined cost score, while the reduced dataset has 0.26 C-sore. Which shows a lot of reduction in terms of the overall cost of data acquisition. Hence, the ensemble based cost-sensitive feature selection technique is capable of identifying such features as important yet cost-effective. The consolidated knowledge base contains a set of rules which are composed of both important and cost-effective features.

TABLE II. Comparison between full feature set and selected feature set on Model Construction

Algorithm	**Full Feature Set**		**Selected Feature Set**	
	Accuracy	**#Rules**	**Accuracy**	**#Rules**
C4.5	93.33	14	90	17
CART	91.66	6	91.66	5
RIPPER	93.33	4	92.5	4
PART	91.66	7	91.66	6
Ridor	90	7	90.833	6
Average	91.99	7.6	91.33	7.6

IV. CONCLUSION

In this study we have proposed a cost-effective methodology for consolidated knowledge creation, which leverages existing machine learning techniques such as filter methods for feature selection and interpretable classification models in order to provide non-trivial insights to the domain expert from electronic medical records data. Most of the studies in the domain of chronic kidney disease classification have focused on creating models which optimize performance measure. This research demonstrates the importance of taking domain related meta-information into account in order to address the practicality aspect of the solution. In this regard, cost of generating a model is generally not taken into consideration. With the help of the case study we have demonstrated how both performance and cost perspectives can be addressed in order to create a more robust yet cost-effective solution.

In future, we would like to explore more powerful feature selection techniques such as evolutionary computational approaches for selecting a subset of features. Moreover, a de-centralized model of cost-sensitive feature selection can also be investigated for cases where EMR data is persisted in a central repository.

ACKNOWLEDGMENTS

This research was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-00655).

REFERENCES

[1] Couceiro, Miguel, and Amedeo Napoli. "Elements About Exploratory, Knowledge-Based, Hybrid, and Explainable Knowledge Discovery." International Conference on Formal Concept Analysis. Springer, Cham, 2019.

[2] Ismail, Ahmed, Abdulaziz Shehab, and I. M. El-Henawy. "Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations." In Security in Smart Cities: Models, Applications, and Challenges, pp. 27-45. Springer, Cham, 2019.

[3] Hempelmann, Christian F., et al. "An entropy-based evaluation method for knowledge bases of medical information systems." Expert Systems with Applications 46 (2016): 262-273.

[4] Movshovitz-Attias, Dana, et al. "Discovering subsumption relationships for web-based ontologies." Proceedings of the 18th International Workshop on Web and Databases. ACM, 2015.

[5] Zhang, Li, et al. "Feature selection using firefly optimization for classification and regression models." Decision Support Systems 106 (2018): 64-85.

[6] Fan, Jianqing, and Runze Li. "Statistical challenges with high dimensionality: Feature selection in knowledge discovery." arXiv preprint math/0602133 (2006).

[7] Gronsbell, Jessica, et al. "Automated feature selection of predictors in electronic medical records data." Biometrics 75.1 (2019): 268-277.

[8] Jensen, Richard, and Qiang Shen. Computational intelligence and feature selection: rough and fuzzy approaches. Vol. 8. John Wiley & Sons, 2008.

[9] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[10] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.