In 15th IEEE Symposium on Computer Based Medical Systems (CBMS'2002), Maribor (Slovenia).

Symbolic Exposition of Medical Data-Sets: A Data Mining Workbench to Inductively Derive Data-Defining Symbolic Rules

Syed Sibte Raza Abidi
Faculty of Computer Science, Dalhousie University, Halifax B3H 1W5, Canada
Email: sraza@cs.dal.ca

Kok Meng Hoe
School of Computer Science, Universiti Sains Malaysia, 11800 Penang, Malaysia

Abstract

The application of data mining techniques upon medical data is certainly beneficial for researchers interested in discerning the complexity of healthcare processes in real-life operational situations. In this paper we present a methodology, together with its computational implementation, for the automated extraction of data-defining CNF symbolic rules from medical data-sets comprising both annotated and un-annotated attributes. We propose a hybrid approach for symbolic rule extraction which features a sequence of methods including data clustering, data discretization and eventually symbolic rule discovery via rough set approximation. We present a generic data mining workbench that can generate cluster/class-defining symbolic rules from medical data, such that the resultant symbolic rules are directly applicable to medical rule-based expert systems.

1. Introduction

One of the tangible outcomes of the incumbent information revolution in the medical/healthcare sector is the generation and accumulation of voluminous data, originating from diverse tasks such as clinical practices, hospital and resource administration, clinical trials, medical research and so on. To make sense of such large volumes of medical data, numerous inductive data analysis techniques in the realm of '*Knowledge Discovery from Databases*' (KDD) and more so '*Data Mining*' (DM) have been successfully applied to medical data to derive useful, operative knowledge [1, 2, 3].

Mining medical data has now become a standard practice in understanding the underlying dynamics of a complex healthcare processes. For instance, a medical scientist studying the effects of bacterial organisms against a regime of antibiotics in a given environment will in some cases mine 'bacterial epidemic' data by using stochastic data mining models to understand the causal relationship between 'bacterial-antibiotic sensitivity [3]. Effective mining of healthcare data enables the discovery and characterization of strong regularities inherent within the data, which can otherwise be employed as a concise human-comprehensible generalization of the data-set, that has implications towards medical decision-support.

A functional DM solution is expected to provide a *non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* [4]. The featured work focuses on providing 'cluster-defining' knowledge for a priori defined data classes or

inductively derived data clusters. The basis of our work is a hybrid approach for the automated extraction of cluster-defining *Conjunctive Normal Form* (CNF) symbolic rules from medical data-sets. We have implemented a generic *Symbolic Rule Extraction Workbench* (see Figure 1) that can generate cluster (and class)-defining symbolic rules from medical data—comprising symbolic, discrete and continuous-valued attributes—such that the emergent rule-sets are directly applicable to medical rule-based expert systems [5]. An important aspect of the proposed approach is that the generated decision-rules can potentially be (100-X)% accurate of (100-Y)% with unknown outcomes, where X and Y could be relatively small values. From a healthcare diagnostic support perspective, we believe that the availability of cluster/class-defining knowledge—in terms of symbolic rule-sets—is highly desirable as it can provide interesting insights into the complex inter-relationships between the various attributes of large medical data-sets.

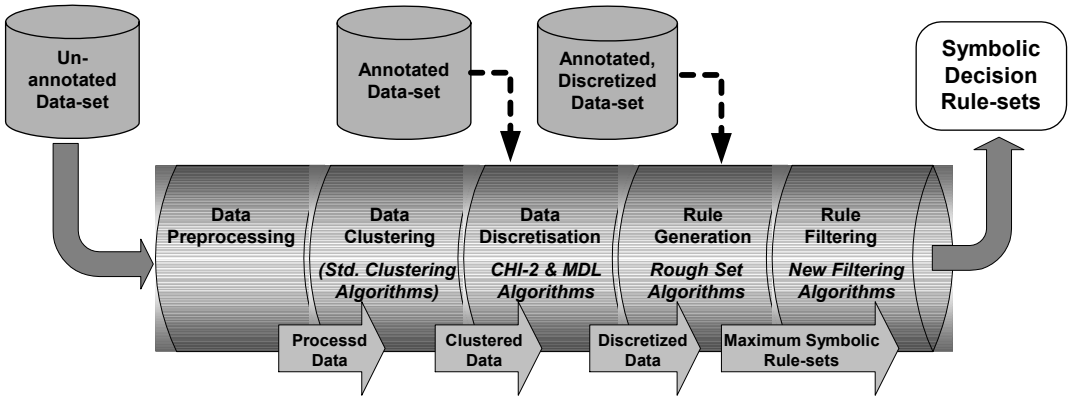

Figure 1. The functional architecture of our symbolic rule extraction workbench

2. Problem statement

Data clustering is an explorative data mining task that involves the distribution of data, in an inductive manner, into a number of clusters such that data items within a cluster are similar in some respect and unlike those from other classes. The underlying assumption of data clustering, or any other data mining task, is that the data in question is not totally random and that there exist some 'hidden' patterns or concepts which can both be revealed by the clustering effort or form the basis for grouping data-points/items into higher-level and consolidated groups of data-items—i.e. clusters or classes.

From a user perspective, data clustering aims (a) to increase understanding of the data under investigation by providing some description of the underlying structures or concepts represented by objects within the data; and (b) to reduce or summarize the data by grouping similar data objects together into categories or classes. Such grouping (or clusters) are ubiquitous in the way humans process and understand information, hence one of the motivations for using clustering algorithms is to construct categories or taxonomies.

Notwithstanding the efficacy of data clustering techniques, novice data mining practitioners—such as medical practitioners, healthcare managers and administrators—find it rather difficult to interpret the emergent data clusters for practical purposes. What is implied here is that the data clustering output is typically a 'visual explanation' of the data topology and it does not provide value-added 'deductive' knowledge—most preferably in a symbolic formalism such as deductive rules—describing both the structure of the emergent clusters and the cluster membership principles. Hence, the motivation for the extraction of cluster/class-defining symbolic rules from, both un-annotated and annotated, medical data-sets.

3. Problem solution: A methodology for symbolic rule extraction

We have postulated a hybrid approach that allows for the generation of data-defining deductive symbolic rules from medical data. We are able to handle the following medical data: (a) *un-annotated data*—i.e. the data-set comprises an undifferentiated collection of multi-component data-vectors $S = \{\mathbf{X}_i : \ i \in [1, n]\}$, for which the classification attribute $c(\mathbf{X}_i) = \alpha$ for $\alpha \in [1, k]$ is unknown.; (b) *annotated data*—i.e. the data-set comprises a priori class label for each data-vector, thus implying that the data has already been classified into distinct classes; (c) *non-discretized data*—i.e. the data comprises ordinal or continuous valued attributes; and (d) *discretized data*—i.e. the data attributes are properly discretized into well defined intervals. We anticipate that medical data may originate as a combination of the above types, hence we have implemented mechanisms to handle a variety of medical data. Table 1 illustrates the different data-types that we can handle and the underlying sequence of processes that need to be performed to produce symbolic rules from the said data-set.

Table 1: Data types and the associated processing tasks

	Discretized	**Non-Discretized**
Un-Annotated	(1) Data Clustering (2) Symbolic Rule Generation	(1) Data Clustering (2) Data Discretization (3) Symbolic Rule Generation
Annotated	(1) Symbolic Rule Generation	(1) Data Discretization (2) Symbolic Rule Generation

We have formulated a methodology for symbolic rule generation that leverages the individual effectiveness of various intelligent data analysis mechanisms: (1) cluster formation via unsupervised clustering algorithms, (2) data-set simplification and attribute selection via attribute discretization, and (3) symbolic rule extraction via rough set approximation. Symbolic rule extraction is achieved via a sequence of phases, as described below:

3.1. Phase-1: Data clustering

Given an un-annotated data-set, phase-1 involves the unsupervised partitioning of a data-set into *k* clusters using the popular *K-Means* data clustering algorithm. Each emergent data cluster is assigned a particular class label, and each cluster is distinguished by its representative data objects which are normally data vectors from the data-set. In essence, the net outcome of phase-1 is an annotated copy of the original data-set where every data object is accompanied by its classification information.

3.2. Phase-2: Data discretization

Medical data is available in a variety of types—for instance data for a patient with a high-blood pressure problem may include the patient's sex (binary), age (non-bounded integer), diastolic and systolic blood pressure (non-bounded integer), and whether the patient has a family history of such problems (categorical). However, most existing machine learning, rule induction and statistical techniques used for data mining can only be applied to categorical values as they are ineffective when faced with an infinite range of integer or real-valued attributes. One possible solution to this problem is to partition the numeric attributes into a number of intervals or bins and converts each interval into a categorical value. The process of partitioning continuous attributes into categories is called *discretization*.

Given a non-discretized data-set, phase-2 involves the supervised transformation of ordinal/continuous-valued attributes to discrete-valued attributes—i.e. to reduce the domain of values of an attribute to a small number of attribute-value ranges—where each interval can be identified by a label or symbol (shown in Figure 2). The motivation for phase-2 is driven by the fact that ordinal/continuous valued attributes are proven to be rather unsuitable for the extraction of concise symbolic rules. Note that the antecedents of rules extracted from discretized data will consist of descriptors with *attribute-interval* pairs instead of *attribute-value* pairs. More attractively, the data discretization phase not only reduces the complexity and volume of the data-set, but also serves as an attribute filtering mechanism whereby attributes that are deemed to have a minimal impact on class distinction and specification are eliminated.

In our work, we employ two data discretization methods: (1) statistical discretization via Chi-2 [6] and (2) class information entropy reduction via MDL partitioning [7].

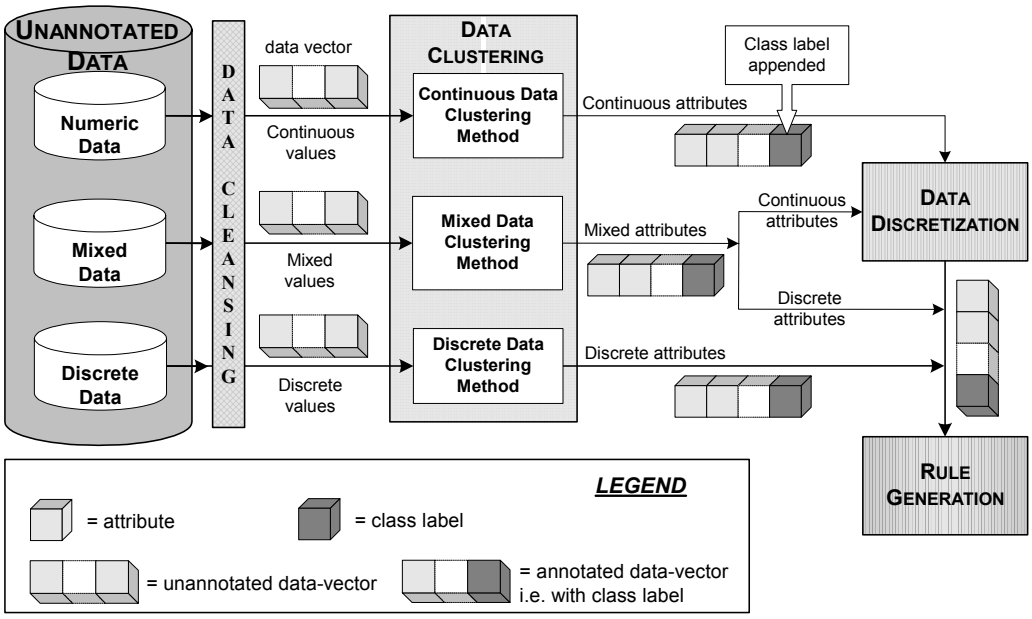

Figure 2. The input-output relationship for data discretization

3.3. Phase 3: Symbolic rule generation

Given a 'clustered' and 'discretized' data-set, phase-3 involves the inductive search for regularities or patterns in the data-set in the form of descriptive symbolic rules. The emergent symbolic rules elucidate class-membership principles and complex inter-relationships between the various data attributes. We use rough set approximation—an interesting alternative to a variety of symbolic rule extraction methods [8]—to derive symbolic rules. Given an annotated data-set of discrete values represented as a decision system, our proposed rule generation method involves the following four steps.

Step 1 - Systematic Data Partitioning: We randomly split the data-set into two separate partitions: i) the training set which comprises not less than 70% of the data-set, and ii) the testing set which comprises the remainder of the data-set. Rules are derived from the training set and, subsequently, evaluated by the testing set. We argue that, splitting the data into two disjoint sets allows for a more impartial evaluation of the performance of the emergent rule-set. Since random partitioning the data-set might result in a training set that is highly

representative (fortunate) or highly unrepresentative (unfortunate) of the entire data-set, we use the *k-fold cross validation* method to systematically split the training set into *k* folds.

Step 2 - Dynamic Reducts Computation: Rule generation via rough sets involves the computation of dynamic reducts (minimal attribute-sets that can adequately distinguish particular data-vectors from the rest of the data vectors in the same data-set). In rough sets, an annotated data-set with *N* objects and *M* number of attributes is known as a decision table with *N* rows and each row having *M* columns. To find reducts from the decision table, the table must be reduced as follows: (a) *vertical information reduction* reduces the number of rows (or objects) of the table by removing redundant objects which cannot be differentiated from each other with reference to their attributes and values; (b) *horizontal information reduction* reduces the number of columns (or attributes) of the table by eliminating attributes which are unnecessary in preserving the classification information of the entire data-set. The final outcome of both information reduction steps is a minimal set of attributes called the *reduct set*. However, the search for reducts is an NP-hard problem. To overcome this potential bottleneck in reduct computation, we use an efficient *reducts approximation* method which is based on *genetic algorithms* [9].

Step 3 - Succinct Rule Generation: This step involves the selection of dynamic reducts of the shortest possible length followed by the generation of symbolic rules from the selected dynamic reducts. In our work, we do not filter or exclude reducts from the reduct set prior to rules generation, because in this case the reducts filtering criterion will then be based only on objects in the training data-set. For a broader outlook, we have devised a novel *rule generation strateg*y (SHRED) [5] that involves: (1) the selection of dynamic reducts that have a short length and (2) the generation of rules that satisfy a user-defined accuracy level.

Step 4 - Rule Filtering: This step comprises a sequence of operations to filter out low quality rules from the rule-set generated in the previous step. We observed that, the rough sets-mediated rule-set usually contains a large number of distinct rules, thereby limiting the classification capabilities of the rule-set as some rules are redundant or of 'poor quality'. To achieve an efficient rule filtering activity, we have devised a novel rule filtering heuristic termed as MEDIATOR [5]—a simple rule filtering solution based on the computed quality indices of rules in a rule-set. The quality index of each rule is computed using a particular rule quality function, which determine the strength of a rule based on the measures of *support*, *consistency* and *coverage*. MEDIATOR is a stepwise filtering algorithm that uses the *testing accuracy* and the *median of rule quality indices* to ensure that "high-quality" rules are retained and "low-quality" rules are progressively eliminated from the rule-set. Functionally, MEDIATOR takes as input three user-specified parameters: (a) an accuracy level, η; (b) type of rule quality function, *RQF*; and (c) a fluctuation rate, θ where $\theta \in [0, 0.1]$. Given η, MEDIATOR performs a stepwise elimination of rules with low quality indices and produces a near-minimal subset of rules that satisfies η. The quality index of rules is determined by the chosen *RQF*.

4. Experimental results

In this paper we present experimental results based on two medical datasets: The *Wisconsin Breast Cancer (WBC)* and *New Thyroid Gland (NTG)* datasets. After the successful clustering of the datasets, we (1) discretize the continuous data values into meaningful intervals—i.e. nominal values and (2) perform insignificant attribute elimination. In the next step, symbolic rules were generated from the discretized data-set to explicate the underlying structure of the derived clusters. For pragmatic reasons, the rules discovered were moderated based on two rule-filtering criteria: (1) Left Hand Side (LHS) length and (2) RHS support. Finally, a filtered rule-set was selected based on two criteria: (1) accuracy of the filtered rules when compared

with the testing data; and (2) the number of rules in the rule-set—ideally we seek less than 30 rules in the rule-set. Table 2(a &b), show exemplar rule-sets generated from the two data-sets.

Table 2a. Exemplar rule-set generated for the NTG dataset. The LHS length = 2. Rule 1 and rule 6 defining class 1 and 2, respectively, can be read as:
IF Basal_TSH(1) ∧ T3_resin(3) THEN Class 1
IF Serum_Thyrox(3) ∧ T3_resin(1) THEN Class 2

No	Attributes		Class	Support
1	Basal_TSH(1)	T3_resin(3)	1	87
2	Serum_Thyrox(2)	T3_resin(3)	1	78
3	Basal_TSH(1)	T3_resin(1)	2	16
4	Basal_TSH(1)	T3_resin(4)	1	15
5	Serum_Thyrox(2)	T3_resin(4)	1	14
6	Serum_Thyrox(3)	T3_resin(1)	2	12
7	Basal_TSH(2)	T3_resin(4)	3	8
8	Basal_TSH(1)	T3_resin(2)	1	7
9	Serum_Thyrox(2)	T3_resin(2)	1	6
10	Serum_Thyrox(3)	T3_resin(3)	1	6
11	Basal_TSH(1)	T3_resin(5)	3	7
12	Serum_Thyrox(1)	T3_resin(4)	3	7

Table 2b. Exemplar rule-set generated for the WBC dataset. The LHS length = 5. Legend is: bn=bare_nuclei, bc=bland_chromatin, ct=clump_thickness, nn=normal_nucleoli, sez=sing_epi_cell_sz, ucp=uni_cell_shape, ucz=uni_cell_size.

No	Attributes					Class	Support
1	bn(1)	bc(1)	ct(1)	Nn(1)	sez(1)	0	245
2	bn(1)	bc(1)	ct(1)	nn(1)	ucp(1)	0	243
3	bn(1)	bc(1)	ct(1)	sez(1)	ucp(1)	0	231
4	bn(1)	ct(1)	sez(1)	ucp(1)	ucz(1)	0	231
5	bn(1)	ct(1)	nn(1)	sez(1)	ucp(1)	0	229
6	bn(1)	bc(1)	nn(1)	sez(1)	ucp(1)	0	228
7	bn(3)	bc(2)	nn(2)	sez(2)	ucp(2)	1	62
8	bn(3)	ct(2)	sez(2)	ucp(2)	ucz(2)	1	46
9	bn(3)	ct(1)	sez(2)	ucp(2)	ucz(2)	1	45
10	bn(3)	bc(2)	ct(1)	sez(2)	ucp(2)	1	42
11	bn(3)	bc(2)	ct(2)	sez(2)	ucp(2)	1	40
12	bn(3)	ct(1)	nn(2)	sez(2)	ucp(2)	1	39

5. Concluding remarks

In conclusion we will like to point out that the proposed sequential application of multiple techniques—i.e. data-vector clustering, data discretization, attribute selection and finally rough set approximation—for knowledge extraction via symbolic rule generation, appears to be a sound and pragmatic methodology for understanding the complexities of medical data-sets. Furthermore, we believe that the featured work provides an alternate inductive approach to supplement medical knowledge-bases with data-mediated medical knowledge.

6. References

[1] A. Duhamel, M. Picavet, P. Devos & R. Beuscart, "From data collection to knowledge data discovery: A medical application of data mining", V.L. Patel et al (Eds.) 10th World Congress on Medical Informatics (MedInfo'2001), Amsterdam: IOS Press, 2001.

[2] A. Babic, "Knowledge discovery for advanced clinical data management and analysis", P. Kokol et al. (Eds.) Medical Informatics Europe'99, Ljubljana, Amsterdam:IOS Press, 1999.

[3] S. S. R. Abidi & A. Goh, "Applying knowledge discovery to predict infectious disease epidemics". H. Lee & H. Motoda (Eds.) Lecture notes in artificial intelligence 1531- PRICAI'98: Topics in artificial intelligence, Berlin:Springer Verlag, 1998.

[4] U. Fayyad, "Advances in knowledge discovery and data mining", Cambridge: MIT Press, 1996.

[5] S. S. R. Abidi, K. M. Hoe, A. Goh, "Analyzing data clusters: A rough set approach to extract cluster defining symbolic rules, Fisher, Hand, Hoffman, Adams (Eds.) Lecture Notes in Computer Science: Advances in Intelligent Data Analysis, 4th Intl. Symposium, IDA-01. Springer Verlag: Berlin, 2001.

[6] H. Liu, R. Setiono, "Chi2: Feature selection and discretization of numeric attributes", Proceedings of 7th International Conference on Tools with Artificial Intelligence, Washington D.C., 1995.

[7] R. Kohavi, M. Sahami, "Error-based and entropy-based discretization of continuous features", Proc. 2nd International Conference on Knowledge Discovery and Data Mining, 1996, pp. 114-119.

[8] Z. Pawlak, "Rough Sets", Lin T.Y., Cercone N. (Eds.) Rough sets and data mining: Analysis of imprecise data, Dordrecht:Kluwer Academic Publishers, 1997, pp. 3-7.

[9] J. Wróblewski, "Finding minimal reducts using genetic algorithms", Proc. 2nd Annual Joint Conference on Information Sciences, Wrightsville Beach, North Carolina, 1995, pp.186-189.