World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:1, No:11, 2007

Analysis of Medical Data using Data Mining and Formal Concept Analysis

Anamika Gupta, Naveen Kumar, and Vasudha Bhatnagar

***Abstract*—**This paper focuses on analyzing medical diagnostic data using classification rules in data mining and context reduction in formal concept analysis. It helps in finding redundancies among the various medical examination tests used in diagnosis of a disease. Classification rules have been derived from positive and negative association rules using the Concept lattice structure of the Formal Concept Analysis. Context reduction technique given in Formal Concept Analysis along with classification rules has been used to find redundancies among the various medical examination tests. Also it finds out whether expensive medical tests can be replaced by some cheaper tests.

***Keywords*—**Data Mining, Formal Concept Analysis, Medical Data, Negative Classification Rules.

I. Introduction

MEDICAL history data consists of large number of tests required to diagnose a particular disease. After studying case history of several patients, it has been found that some of the tests are redundant. Also it has been found that some expensive tests can be replaced by cheaper tests. Our paper proposes a method to find out the redundancies among the tests. The idea used has been taken from context reduction of Formal Concept Analysis and classification rule technique of data mining.

Data mining refers to extracting information from very large databases [5]. Classification and association are the two mechanisms to represent the extracted information. Association rules are the rules of the type A->B where A and B are sets of attributes (items). Each association rule has a support and confidence measure associated with it. Support of x% means that x number of transactions have A and B together. Confidence of y% means that y number of transactions having A must have B.

Classification rules refers to rules where consequent part of the rule is a class. Several algorithms have been proposed to find association rules [1] [2] and classification rules [9] [10]. There are a few algorithms proposed to find classification rules based on association rules [6] [7]. Classification based on Association (CBA) rules gives more accurate results. But rules produced are more in number as compared to traditional methods of classification. In the medical domain we are interested in high accuracy, so we are following CBA method. In the medical examination tests, we are interested in positive as well as negative result of the test. [3] gives a method of finding classification rules based on both positive and negative association rules. We are using the technique mentioned in [3] to find the classification rules and then by using context reduction technique we are finding out the redundant attributes i.e. the medical examination tests.

Context reduction refers to reduction of the database in terms of objects (rows) or attributes (columns). Formal Concept Analysis introduces a novel technique for reduction of database that has been explained in this paper.

Manuscript received June 4, 2005.

Anamika Gupta is a Ph.D. Student in the Department of Computer Science, Delhi University, India, doing research in the field of Data Mining and Formal Concept Analysis.

Naveen Kumar is working as a reader in the Department of Computer Science, Delhi University, India.

Vasudha Bhatnagar is working as a Lecturer in Department of Computer Science, Delhi University, India.

II. Background Knowledge

A. Formal Concept Analysis

Bernhard Ganter *et al.* has defined Formal Concept Analysis as a field of applied mathematics based on the mathematization of concept and conceptual hierarchy and thereby it activates mathematical thinking for conceptual data analysis and knowledge processing [4]. Following basic definitions have been taken from [4] which has been used throughout the paper.

A formal context K= (G,M,I) consists of two sets G and M and a relation I between G and M. The elements of G are called the objects and the elements of M are called the attributes of the context. For a set A⊆G of objects A'={mεM | gIm for all gεA} (the set of all attributes common to the objects in A). Correspondingly, for a set B of attributes we define B' = {gεG | gIm for all mεB} (the set of objects common to the attributes in B). A formal concept of the context (G,M,I) is a pair (A,B) with A⊆G,B⊆M, A'=B and B'=A. A is called the extent and B is the intent of the concept (A,B). ζ(G,M,I) denotes the set of all concepts of the context (G,M,I).

If (A1,B1) and (A2, B2) are concepts of a context, (A1,B1) is called a subconcept of (A2,B2), provided that A1⊆A2 (which is equivalent to B2⊆B1). In this case, (A2,B2) is a superconcept of (A1,B1) and we write (A1,B1) ≤ (A2,B2). The relation ≤ is called the hierarchical order of the concepts. The set of all concepts of (G,M,I) ordered in this way is called the concept lattice of the context (G,M,I).

An ordered set V := (V,≤) is a lattice, if for any two elements x and y in V the supremum x∨y and the infimum

$x \wedge y$ always exist. V is called a complete lattice, if the supremum $\vee X$ and the infimum $\wedge X$ exist for any subset X of V. Every complete lattice V has a largest element, $\vee X$, called the unit element of the lattice, denoted by 1_v. Dually, the smallest element 0_v is called the zero element.

For an element v of a complete lattice V we define

$v_* = \vee \{x \varepsilon V \mid x<v\}$

and $v^* = \wedge \{x \varepsilon V \mid v<x\}$

we call v $\vee$-irreducible, if v is not equal to v_* i.e. if v can not be represented as the supremum of strictly smaller elements. In this case, v_* is the unique lower neighbor of v. Dually we call v $\wedge$-irreducible if v is not equal to v^*.

A. Context Reduction

A context (G,M,I) is called clarified if for any objects g,h ε G from g' = h' it always follows that g=h and correspondingly, m'=n' implies m=n for all m,n ε M.

A clarified context (G,M,I) is called row reduced, if every object concept is $\vee$-irreducible and column reduced, if every attribute concept is $\wedge$-irreducible. A context, which is both row-reduced and column-reduced is reduced.

This definition means that every finite context can be brought into a reduced form without changing the structure of the concept lattice, and the latter is unique. We first clarify the context, i.e. we merge objects with the same intents and attributes with the same extents. Then we delete all objects, the intent of which can be represented as the intersection of other object intents, and correspondingly all attributes, the extent of which is the intersection of other attribute extents..

III. Algorithm

A. Algorithm for redundancy detection

Step 1: Find the extents of all attributes and their negations. Extent of an attribute "a" is denoted as " a' ". This requires one database scan of the entire database.

Step 2: For a given attribute "a", find out whether a' $\cup$ ($\neg$a)' $\subseteq$ (CL)'or a' $\cup$ ($\neg$a)' $\subseteq$ ($\neg$CL)' where CL stands for Class Label. If yes then attribute "a" is redundant otherwise not.

Step 3: For given attributes "a" and "b", find out that if a' = b' then whether ($\neg$a)' = ($\neg$b)' holds or not. If yes then one of the attribute "a" or "b" is redundant.

Step 4: Find the attributes whose extents are intersection of extents of other attributes. These attributes are the probable candidates for redundancy.

Step 5: For probable candidate attribute "a" which is intersection of "b" and "c", find out whether $a \Lambda b \Lambda c \Rightarrow$ CL and $\neg a \Lambda b \Lambda c \Rightarrow$ CL or $a \Lambda b \Lambda c \Rightarrow \neg$CL and $\neg a \Lambda b \Lambda c \Rightarrow \neg$CL.

Step 6: If yes then "a" is redundant w.r.t attributes "b" and "c" else "a" is not redundant.

Here we do not generate all classification rules. We generate rules for only the probable candidates for redundancy. This reduces the time required to find redundancy.

B. Complexity

The algorithm described above requires just one database scan of the entire database to find the extents of all attributes. The algorithm finds out the probable candidates for reduction and does not consider all the possible subsets of attribute sets, thus making it non-exponential. This reduces the complexity of the algorithm.

C. *Example*

We are considering the medical data for which the medical examination test has either the positive result or negative result. We can illustrate the redundancy detection process with an example. Let us assume following data has been given to us.

TABLE I
ORIGINAL CONTEXT

Patient ID	Test a		Test b		Test c		Test d		Test e		Test f		Diagnosis (Class Label)	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-
ID1				x				x						x
ID2			x		x		x			x		x		x
ID3	x		x		x		x		x		x		x	
ID4	x		x		x		x		x		x		x	
ID5		x	x		x			x				x	x	
ID6		x											x	

Find the intersection of extent of each attribute and its negation with the class label. Context obtained after this will be as follows:

TABLE II
CONTEXT AFTER INTERSECTION WITH CLASS LABEL

Patient ID	Test a		Test b		Test c		Test d		Test e		Test f		Diagnosis (Class Label)	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-
ID1														x
ID2														x
ID3	x		x		x		x		x		x		x	
ID4	x		x		x		x		x		x		x	
ID5		x	x		x			x				x	x	
ID 6		x											x	

Intersection of extent of each attribute and its negation with the negation of class label will give following context:

TABLE III
INTERSECTION WITH NEGATION OF CLASS LABEL

Patient ID	Test a		Test b		Test c		Test d		Test e		Test f		Diagnosis (Class Label)	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-
ID1								x						x
ID2			x		x					x		x		x
ID3													x	
ID4													x	
ID5													x	
ID6													x	

We find classification rules from the above context and perform the context reduction process. For finding the classification rules, we have used the technique of [3].

Following classification rules based on positive and negative association rules have been generated:
CR-1. a ⇒ CL Patient ID = {ID3, ID4} (Ref. Table 2)
CR-2. ¬a ⇒ CL Patient ID = {ID5, ID6} (Ref. Table 2)
CR-3. bΛcΛd ⇒ CL Patient ID = {ID3, ID4} (Ref. Table 2)
CR-4. bΛcΛ¬d ⇒ CL Patient ID = {ID5} {Patient ID1 does not have positive result of b and c} (Ref. Table 2)
CR-5. bΛcΛe ⇒ CL Patient ID = {ID3, ID4} (Ref. Table 2)
CR-6. bΛcΛ¬e ⇒ ¬CL Patient ID = {ID2} (Ref. Table 3)
CR-7. bΛcΛf ⇒ CL Patient ID = {ID3, ID4} (Ref. Table 2)
CR-8. bΛcΛ¬f ⇒ ¬CL Patient ID = {ID2} (Ref. Table 3)
CR-9. bΛcΛ¬f ⇒ CL Patient ID = {ID6} (Ref. Table 2)
According to context reduction, tests d, e, f are redundant w.r.t positive result of test b and test c since extents of tests d,e,f are intersection of extents of test b and c. But in context of medical data, where a test is identified by its result, we cannot say that test d, e, f are redundant unless we analyze the classification rules of these attributes. We observe the following:

1. According to CR-1 and CR-2, both attribute "a" and negation of "a" imply class label. This means that presence of "a" does not have any effect on the class label, so attribute "a" is redundant.
2. Test "d" is probable candidate for redundancy since it is intersection of test "b" and "c". We can mark "d" as redundant if and only if bΛcΛd ⇒ CL and bΛcΛ¬d ⇒ CL. This means that if test b and test c has positive results then test d is redundant. If test d is an expensive test then it can be replaced by test b and test c.
3. Test "e" is probable candidate for redundancy since it is intersection of "b" and "c". According to CR-5 and CR-6, bΛcΛe ⇒ CL, bΛcΛ¬e ⇒ ¬CL. This implies test e is not redundant. In fact test e is very important because in case of positive results of test b and test c, test e decides whether diagnosis of the disease will be positive or negative.
4. Attribute "f" is probable candidate for redundancy since it is intersection of "b" and "c". But here we observe that bΛcΛf ⇒ CL, bΛcΛ¬f ⇒ ¬CL for some cases and bΛcΛ¬f ⇒ CL for some cases. This means that we cannot comment on redundancy of the test.

We can make the following conclusion on the basis of above observations:

1. If a' ∪ (¬a)' ⊆ (CL)'or a' ∪ (¬a)' ⊆ (¬CL)' then attribute "a" is redundant.
2. If c = a ∩ b and (a'∧b'∧c') ⊆ CL', (a'∧b'∧(¬c)') ⊆ CL' or if c = a ∩ b and (a'∧b'∧c') ⊆ (¬CL)', (a'∧b'∧(¬c)') ⊆ (¬CL)' then attribute "c" is redundant with respect to attributes "a" and "b".
3. If a' = b' and (¬a)' = (¬b)' then one of the attribute is redundant.

IV. Case Study Description

We experimented on Heart Diseases dataset available at UCI Site [8]. We considered the following attributes:

TABLE IV
HEART DISEASE DATASET

S.No.	Attribute	Purpose	Range
1	age	Age in years	Not Considered
2	sex	Sex of patient	Not Considered
3	cp	chest pain type	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
4.	trestbps	resting blood pressure(in mm Hg on admission to the hospital)	1-500
5.	chol	serum cholesterol in mg/dl	1-50
6.	fbs	fasting blood sugar > 120 mg/dl	0,1
7.	restecg	resting electrocardiographic results	0: normal 1: having ST-T wave abnormality - T wave inversions and/or ST elevation or depression of > 0.05 mV 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8.	thalach	maximum heart rate achieved	1-500
9.	Exang	exercise induced angina	0,1
10.	oldpeak	ST depression induced by exercise relative to rest	0-10
11.	slope	the slope of the peak exercise ST segment	1,2,3
12.	ca	number of major vessels (0-3) colored by fluoroscopy	0,1,2,3
13.	thal	Heart rate	3: normal 6: fixed defect 7: reversible defect
14	num	diagnosis of heart disease	0: < 50% diameter narrowing 1: > 50% diameter narrowing in any major vessel, number represent the vessel.

Each attribute value is obtained by a medical examination test. Attributes thalach, Exang, oldpeak, slope are result of the medical test TMT. We have run our algorithm on the heart disease dataset and found the following classification rules:

1. if (TMT = +ve) and (other attributes are +ve or –ve) then diagnosis is +ve
2. if (TMT = -ve) and (fbs = +ve) and (chol = +ve) and (restecg = +ve) then diagnosis is +ve
3. if (TMT = -ve) and (fbs = -ve) and (chol = +ve) and (restecg = +ve) then diagnosis is +ve
4. if (restecg = +ve) and (TMT = -ve) and (chol = +ve) then diagnosis is +ve
5. if (restecg = -ve) and (TMT = -ve) and (chol = +ve) then diagnosis is +ve
6. if (trestbps = +ve) and ((chol = +ve) or (restecg = +ve)) and (TMT = -ve) then diagnosis is +ve
7. if (trestbps = -ve) and ((chol = +ve) or (restecg = +ve)) and (TMT = -ve) then diagnosis is +ve

Observing above Classification rules we can find the following redundancies:

1. According to classification rule 1, if TMT gives positive result then all other tests are redundant.
2. According to classification rule 2 and 3, if TMT gives negative result and if chol and restecg gives positive result then fbs is redundant.
3. According to classification rule 4 and 5, if TMT gives negative result, then if chol gives positive result, then restecg is redundant.
4. According to classification rule 6 and 7, if TMT gives negative result, then if either of chol or restecg gives positive result then trestbps is redundant.

These results help in deciding the number of medical tests required for diagnosis. Above results can be summarized as: Perform TMT test. If result is +ve then diagnosis is +ve. If result is –ve then perform chol test. If chol test gives +ve result then don't perform trestbps test (redundancy 4), restecg test (redundancy 3), fbs test (redundancy 2).

V. Conclusion

This paper applies the techniques of classification in data mining and context reduction in Formal concept Analysis on medical data and finds out the redundancies among the medical examination tests prescribed for diagnosis of a disease. Currently the technique works on binary data, which means that a test can have either positive result or negative result. In future we will experiment on quantitative values of the test, which means working on multi-valued context.

Acknowledgment

The authors extend their special thanks to Poonam Mittal, a postgraduate student in the department, for her support in implementation and testing of the software.

.

References

[1] Rakesh Aggarwal, R. Srikant, *Fast Algorithms for Mining Association Rules*, VLDB-94.
[2] Rakesh Aggarwal, Tomasz Imielinski, Arun Swami, *Mining Association Rules between set of items in Large Databases*, Proc. ACM SIGMOD Conference management of data, 1993.
[3] Anamika Gupta, Naveen Kumar, Vasudha Bhatnagar, *Mining Classification rules based on Positive and Negative Association Rules using Concept Lattice*, Submitted for publication, 2005.
[4] Bernhard Ganter, Rudolf Wille, *Formal Concept Analysis*, Mathematical Foundations, Springer.
[5] Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.
[6] Bing Liu, Wynne Hsu, Yiming Ma, *Integrating Classification and association Rule Mining*, In proceedings of KDD-98, 1998.
[7] W.Li, J. Han, J.Pei, *CMAR: Accurate and efficient classification based on multiple class-association rules*, IEEE International Conference on Data Mining (ICDM'01) San Jose, California, 2001.
[8] Merz, C.J, and Murthy, P. 1996, *UCI repository of machine learning database* [ftp://ftp.ics.uci.edu/pub/machine-learning-databases/].
[9] Quinlan, J.R. *Induction of decision tree*, Machine Learning, 1986.
[10] Quinlan, J.R. 1992, *C4.5: Program for machine learning*, Morgan Kaufmanns.