Validating Expert System Rule Confidences Using Data Mining of Myocardial Perfusion SPECT Databases

CD Cooke, CA Santana, TI Morris[§], L DeBraal[§], C Ordonez[§], E Omiecinski[§], NF Ezquerra[§], EV Garcia

Emory University & Georgia Institute of Technology[§], Atlanta, GA, USA

Abstract

Our goal with this study was to use data mining techniques, applied to imaging and textual patient databases, to validate the confidences (certainty factors) of the heuristic rules in our previously described Expert System, PERFEX™. A relational database combining textual and imaging information was generated from 655 patients who had undergone both stress/rest myocardial perfusion SPECT and coronary angiography. Initial data mining was concentrated on heuristic rules involving myocardial perfusion defects and the LAD vascular territory. The results show the robustness of the expert system, and furthermore show that data mining of large databases combining textual and imaging information can be used to validate and potentially improve the confidence levels associated with heuristic rules in expert systems.

1. Introduction

Data mining, also known as knowledge discovery, is defined as the automated discovery of previously unknown, nontrivial, and potentially useful information from databases. This information comes in the form of statements that describe the relationship between objects contained in the database, such that each statement is in some sense simpler than enumerating all the relationships between the individual instances of objects [1]. For example, in a database of patients, that includes locations of perfusion defects and measures of coronary stenoses, each record represents the relationship between an individual patient and his clinical variables. A statement such as "patients with defects in the inferior wall often have a significant Right Coronary Artery stenosis", based on the records of the database, conveys information that is implicit and more interesting than listing the perfusion defects and coronary stenoses of all patients. This type of inference is called induction [2]. In other words, data mining infers statements that are supported by the database as opposed to statements that can be proved with respect to the database [3]. Data mining is the process of generating high-level statements that have acceptable certainty and are also interesting from a database of facts.

1.1. Data mining

Consider a database of patient records, where each record is a combination of both textual information and extracted image data for a given patient. Since the data mining algorithm only works with binary variables (specifically, only those binary variables that are "true", or "1"), we must devise a mapping from the database information to binary data. For example, the database field `LAD stenosis` can be converted into a binary variable by assigning a 1 if the stenosis is •50%, and a 0 if it is <50%. Each binary variable is referred to as an `item`. An association rule is an implication of the form $X \Rightarrow Y$, (X implies Y) where X & Y are sets of items, and X & Y have no items in common. Each itemset, or set of items, has an associated measure of statistical significance called support. The support for an itemset X or Y, is defined as the percent of records that contain all of the items in the itemset. In addition, a rule has a measure of its strength called confidence, which is really the conditional probability, $P(Y|X)$, and is defined as the ratio: support($X \cup Y$) / support(X). The higher the confidence, the more certain you can be that the rule is valid. The problem of mining association rules is to generate all rules that have support and confidence greater than some user specified minimum support and minimum confidence thresholds, respectively. This problem can be decomposed into the following sub-problems: generating all itemsets that have support above the user specified minimum (called large or frequent itemsets) and generating all the rules that satisfy the minimum confidence for each frequent itemset generated. Discovering all frequent itemsets and their support is a nontrivial problem, because the number of itemsets increases exponentially as support is lowered. In addition, data mining is generally performed on databases that contain millions of records, with few fields per record. Our application differs in two ways: (1) we have

0276-6547/00 $10.00 © 2000 IEEE

Computers in Cardiology 2000;27:785-788.

combined both textual information and image information into one database, and (2) our database contains hundreds of records with hundreds of fields per record. This makes the process of finding all frequent itemsets and their support even more difficult.

As an example, consider a database of 5 patient records, where each of the fields has already been mapped into a binary variable, as shown in Table 1.

Table 1. Example database of 5 patient records for illustrating data mining techniques, where: Ant = anterior wall perfusion defect, Lat = lateral wall perfusion defect, Inf = inferior wall perfusion defect, LAD = ≤50% LAD stenosis, LCX = ≤50% LCX stenosis, RCA = ≤50% RCA stenosis.

Patient	Ant	Lat	Inf	LAD	LCX	RCA
1	Ant	Lat		LAD		
2		Lat			LCX	
3		Lat	Inf		LCX	RCA
4	Ant			LAD		
5	Ant	Lat		LAD	LCX	

If the minimum support were set to 40%, then the resulting frequent itemsets would be:

Frequent Itemsets	Support
{Lat}	80%
{Ant}	60%
{LAD}	60%
{LCX}	60%
{Ant, LAD}	60%
{Lat, LCX}	60%
{Ant, Lat}	40%
{Lat, LAD}	40%
{Ant, Lat, LAD}	40%

If the minimum confidence is set to 50%, then the rule association, {Ant} ⇒ {LAD} would have a support of 60% (from the table above, the support of {Ant, LAD} is 60%), and the confidence would be 100%: support({Ant, LAD}) / support({Ant}). Some additional rule associations are shown below:

Association	Support	Confidence
{Lat} ⇒ {LCX}	60%	75% (60%/80%)
{Lat} ⇒ {LAD}	40%	50% (40%/80%)
{Ant, Lat} ⇒ {LAD}	40%	100% (40%/40%)
{LAD} ⇒ {Ant, Lat}	40%	67% (40%/60%)
{LAD} ⇒ {Lat}	40%	67% (40%/60%)
{Lat} ⇒ {Ant}	40%	50% (40%/80%)

Note that not all associations make clinical sense (ie, {Lat} ⇒ {Ant}).

1.2. PERFEX

We have previously described our development and validation of an expert system, PERFEX, for interpreting myocardial perfusion scans [4-6]. PERFEX contains 253 heuristic rules that correlate the presence and location of perfusion defects on SPECT studies, with coronary angiography (cath) demonstrated CAD and with expert visual interpretations. Each rule within PERFEX was originally assigned a certainty factor based on the experience of several domain experts (experts in the field of Nuclear Cardiology). Certainty factors are a measure of a rules certainty; they range from –1 (absolutely certain there is no disease) to +1 (absolutely certain there is disease), with the range from –0.2 to +0.2 indicating "unknown" certainty, or indeterminance. When run against 655 patients, with these assigned certainty *factors*, PERFEX demonstrates a significantly lower sensitivity and higher specificity than visual interpretation for identifying the presence and location of CAD vs. Cath, as shown in Table 2.

Table 2. Sensitivities (Sn) and specificities (Sp) for detecting the presence and location of CAD for PERFEX and Visual interpretation verses Cath, for 655 patients.

	CAD		LAD		LCX		RCA	
	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
Visual vs Cath	87*	21*	69	59	61	88*	73	71
PERFEX vs Cath	80*	42*	69	53	68	55*	65	65

*$p < 0.05$

2. Methods

A database was generated from 655 patients, who had previously undergone both stress/rest myocardial perfusion SPECT and coronary angiography. The database consisted of 114 textual variables (including results from the SPECT study and the coronary angiography) from the cardiac databank, plus 64 perfusion variables from the output of the CEqual program: 32 from the stress "blackout" data (the results of comparing the patient's perfusion to a gender matched normal file) and 32 from the reversibility data (the results of comparing the normalized difference between stress and rest, to a gender matched normal file) [7]. The database information was then converted into a format where each database field became several binary variables, as required by the data mining algorithm. This conversion resulted in 476 binary variables for the textual fields and 64 binary variables for the imaging information [8]. If these 540 fields had been used for the data mining algorithm, there would have been a huge set of association rules found (approximately 2^{540}). Therefore,

three methods were used to reduce these possible associations. First, a field-filtering program was used to focus the associations on specific, interesting, database fields. For the purpose of this paper, we focused on 12 fields: nine imaging fields, condensed from the 32 stress blackout variables (see figure 1) and three coronary angiography fields (LAD, LCX and RCA stenoses ≥50%). Second, the association rules were limited to a 10% support and a 40% confidence. Finally, the association rules were limited to those with one antecedent and one consequent. With these constraints, the algorithm (written in C), took less than 1 second to complete, on a Sun Ultra 10 workstation.

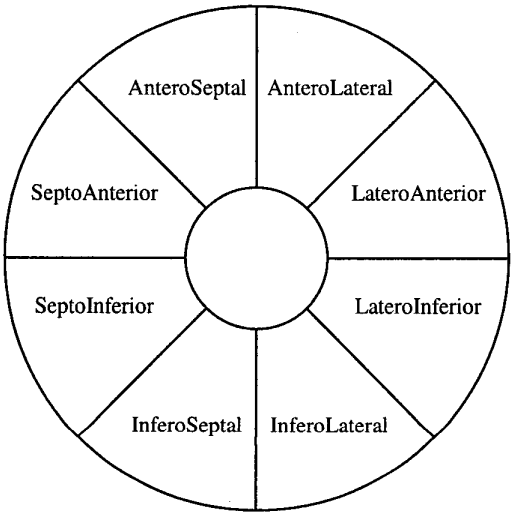

Figure 1. Nine imaging fields used for the data mining.

3. Results

The data mining resulted in 181 association rules, of which 7 involved the LAD, and were part of the 253 heuristic rules within PERFEX. These 7 rules are shown in Table 3 along with their original CF's from PERFEX.

The New CF's were calculated from the confidence using a linear equation, with a 50% confidence set equal to a CF of 0, and a 100% confidence set equal to a CF of 1.0. The certainty of these nine heuristic rules in PERFEX, were changed to the newly calculated CF□s and run against the same 655 patients. When compared to the previous results, there were no statistical differences found in detecting or localizing CAD, as shown in Table 4.

Table 3. Seven heuristic rules from PERFEX, with support (Sup), confidence (Conf) & certainty factors derived from data mining (DM CF), and the original certainty factors derived from the domain experts (PERFEX CF).

Rule	Sup	Conf	DM CF	PERFEX CF
If AnteroLateral then LAD	19%	66%	0.31	0.70
If AnteroSeptal then LAD	17%	72%	0.44	0.80
If SeptoAnterior then LAD	15%	79%	0.58	0.80
If Septonferior then LAD	15%	81%	0.62	0.70
If InferoSeptal then LAD	22%	69%	0.38	0.40
If LateroAnterior then LAD	25%	60%	0.20	0.40
If Apical then LAD	28%	70%	0.39	0.70

Table 4. Sensitivities (Sn) and specificities (Sp) for detecting the presence and location of CAD, for PERFEX vs. Cath and PERFEX vs. Visual interpretation. Original values are from PERFEX using the original CF's derived from the domain experts; Data Mining values are from PERFEX using the CF's derived from the data mining.

	CAD		LAD		LCX		RCA	
	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
Original PERFEX vs Cath	80	41	69	53	68	55	65	65
Data Mining PERFEX vs Cath	79	43	66	56	68	55	65	65
Original PERFEX vs Visual	83	73	77	66	90	70	74	79
Data Mining PERFEX vs Visual	82	74	74	68	89	70	74	79

p = NS

4. Conclusion

We have shown that data mining of large databases of combined textual and imaging information can be used to validate heuristic rules in expert systems.

Acknowledgements

This work was funded in part by grant LM 06726 from the National Library of Medicine.

References

[1] Frawley WJ, Piatetsky-Shapiro G, and Matheus CJ. Knowledge Discovery in Databases: An Overview. In: Piatetsky-Shapiro G and Frawley WJ, editors. Knowledge Discovery in Databases. Menlo Park, CA: AAAI Press: MIT Press, 1991:1-27.

[2] Michalski RS. A Theory and Methodology of Inductive Learning. In: Machine Learning: An Artificial Intelligence Approach. Palo Alto, CA: Tioga Publishing Company, 1983.

[3] Holsheimer M, Siebes A. Data Mining: The Search for Knowledge in Databases. Technical Report CS-R9406, CWI, Amsterdam, The Netherlands, 1993.

[4] Ezquerra N, et al. Interactive, Knowledge-Guided Visualization of 3D Medical Imagery. Future Generation Computer Systems 1999; 15:59-73.

[5] Ezquerra N, Mullick R, Cooke D, Krawczynska E, Garcia E. PERFEX: An expert system for interpreting 3d myocardial perfusion. Expert Systems with Applications, Pergamon Press Ltd., 1993, 6:459-468.

[6] Garcia EV, Cooke CD, Folks RD, Santana CA, Krawczynska EG, Ezquerra NF, Vansant JP, Ziffer JA. Expert system interpretation of myocardial perfusion tomograms: validation using 655 prospective patients. J Nucl Med 1999; 40:126P.

[7] Garcia EV, Cooke CD, Van Train K, Folks RD, Peifer JW, Berman D, DePuey EG, Maddahi J, Alazraki N, Galt JR, Ezquerra NF, Ziffer J. Technical Aspects of Myocardial SPECT Imaging with Tc-99m Sestamibi. Am J Cardiol 1990; 66:23E-31E.

[8] Cooke CD, Ordonez C, Garcia EV, Omiecinski E, Krawczynska EG, folks RD, Santana CA, DeBraal L, Ezquerra NF. Data Mining of Large Myocardial Perfusion SPECT (MPS) Databases to Improve Diagnostic Decision Making. J Nucl Med 1999; 40:292P.

Address for correspondence.

C. David Cooke, MSEE
Emory University Hospital
Division of Nuclear Medicine
1364 Clifton Rd. NE
Atlanta, GA 30322
ccooke@emory.edu