Technical report

Better Healthcare with Data Mining

Philip Baylis
Shared Medical Systems Limited, UK

Table of contents

Abstract

This paper illustrates data mining will enable clinicians and managers to find valuable new patterns in data, leading to potential improvement of resource utilization and patient health. As the patterns are based on recent clinical practice, they represent the ultimate in evidence-based care.

This paper briefly introduces the PASW Modeler* data mining system, which incorporates advanced machine learning technologies that extract complex interrelationships and decision-making rules from the data.

Introduction

Healthcare generates mountains of administrative data about patients, hospitals, bed costs, claims, etc. Clinical trials, electronic patient records and computer supported disease management will increasingly produce mountains of clinical data. This data is a strategic resource for healthcare institutions.

With the advent of data warehousing techniques, specific areas of interest may be investigated more thoroughly. Products such as INFoCOM from Shared Medical Systems, which is a clinically-based data warehouse product designed for use throughout a hospital, bring the potential for specialized information production to the clinicians and managers desktop through the use of clinical workstations and Executive Information Systems (EIS).

Data mining products are designed to take this one stage further. It brings the facility to discover patterns and correlation hidden within the data repository and assists professionals to uncover these patterns and put them to work. Therefore, decisions rest with healthcare professionals, not the information system experts.

The key to successful data mining is to first define the business or clinical problem to be solved. New knowledge is not discovered by the algorithms, but by the user. This paper will prove that knowledge can automatically be obtained by the use of machine learning techniques in the hands of healthcare decision-makers. It demonstrates the use of PASW Modeler analyzing seven attributes collected routinely in UK hospitals.

Inpatient length of stay

The analysis is to look at the contributory factors influencing length-of-stay (LOS) of a patient during a consultant episode—a major component in the cost of inpatient treatment.

The purpose is to identify patterns affecting LOS that may help in the reduction of cost (and potentially reduce patient trauma), based upon the premise that it is possible to reduce costs by seeking to reduce patient LOS.

** PASW Modeler, formerly called Clementine®, is part of SPSS Inc.'s Predictive Analytics Software portfolio.*

Patient data

For the purpose of illustration, the following dataset has been created, based upon the current clinical practice in the UK for surgical cases. This table shows a few example records from this database:

table (480 records)

Table Generate

Consultant	PrimaryDiag	PrimaryProc	SecondaryProc	LOS	age	MethodOfDischarge
OCH	T79.3	S47.2	Z49.8	0.0	31	DMA
SJG	R55.X	X33.9	X33.9	39.0	75	DMA
JL	K60.2	H54.1	H44.4	0.0	47	DMA
LJM	R10.4	X42.8	X40.3	20.0	48	DMA
SJG	K65.0	T34.1		5.0	23	DMA
JN	K62.5	H25.9		4.0	85	DMA
LJM	K35.9	H01.2		5.0	21	DMA
SJG	L05.0	H60.3		2.0	35	DMA
OCH	K62.5	H25.9		4.0	80	DMA
LJM	T18.5	H27.2		4.0	50	DMA
OCH	R10.4	T46.2		26.0	63	Died
JL	C79.5 D	X33.9	X33.9	4.0	65	DMA
JL	C76.3	X33.9		2.0	65	DMA
OCH	T18.5	H27.2		1.0	22	SD
LJM	L02.4	S57.1	Z50.4	9.0	37	SD
VSR	L02.4	S47.2	Z50.4	1.0	35	DMA
RHJ	D12.5	H11.3		28.0	53	DMA
RSP	K63.9	H11.4	Q23.5	7.0	62	IHT
LJM	C20.X	H58.1	H41.2	8.0	51	DMA

Figure 1: Example surgical cases

Detect patterns

LOS is shown in a histogram

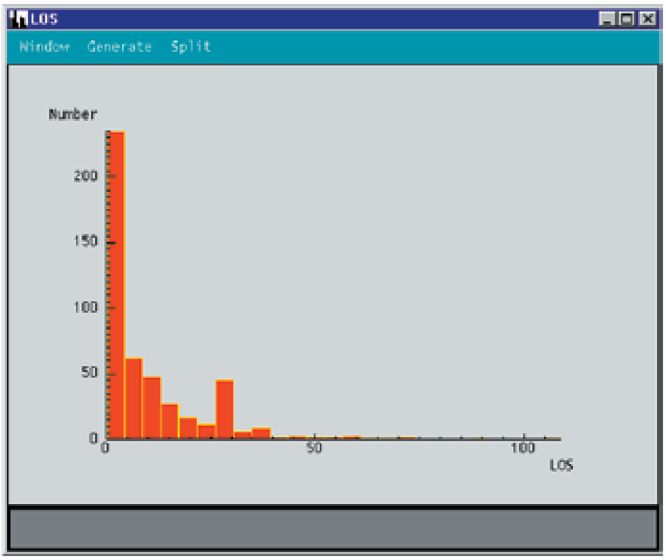


Figure 2: Generate derive mode from an unusually large number of patients with LOS in the 7th bucket

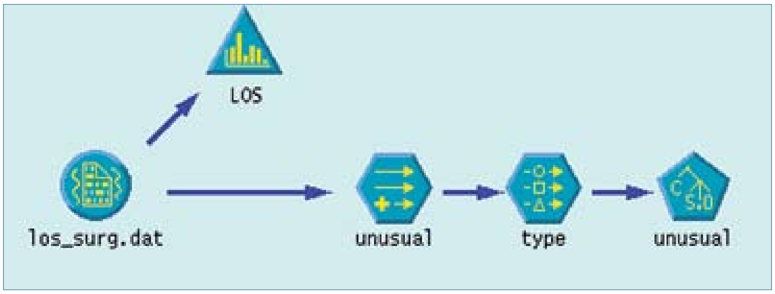
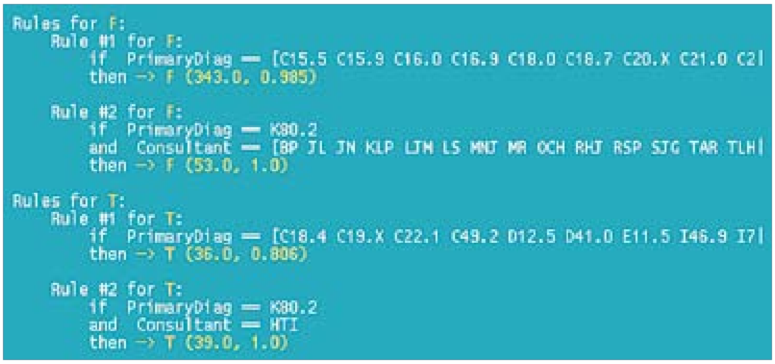


Figure 3: Use rule induction to profile this unusual bucket.

Set of rules from C5 rule induction:



```
Rules for F:
    Rule #1 for F:
        if  PrimaryDiag == [C15.5 C15.9 C16.0 C16.9 C18.0 C18.7 C20.X C21.0 C2|
        then -> F (343.0, 0.985)

    Rule #2 for F:
        if  PrimaryDiag == K80.2
        and  Consultant == [BP JL JN KLP LJM LS MNJ MR OCH RHJ RSP SJG TAR TLH|
        then -> F (53.0, 1.0)

Rules for T:
    Rule #1 for T:
        if  PrimaryDiag == [C18.4 C19.X C22.1 C49.2 D12.5 D41.0 E11.5 I46.9 I7|
        then -> T (36.0, 0.806)

    Rule #2 for T:
        if  PrimaryDiag == K80.2
        and  Consultant == HTI
        then -> T (39.0, 1.0)
```

Figure 4: A profile emerges, describing patients with the same diagnosis K80.2, but an unusual LOS for particular patients treated by consultant HTI.

Lets look at LOS associated with K80.2 for each consultant episode:

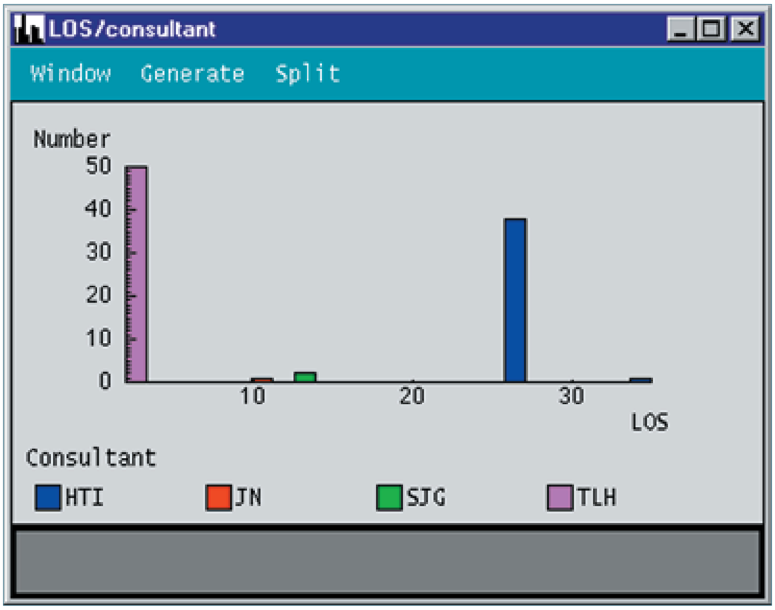


Figure 5: Histogram of LOS by consultant

Figure 6: Here are two distinct bands of LOS, which are exceptionally low and exceptionally high. A derive node is automatically generated. Use rule induction to profile against low/high LOS

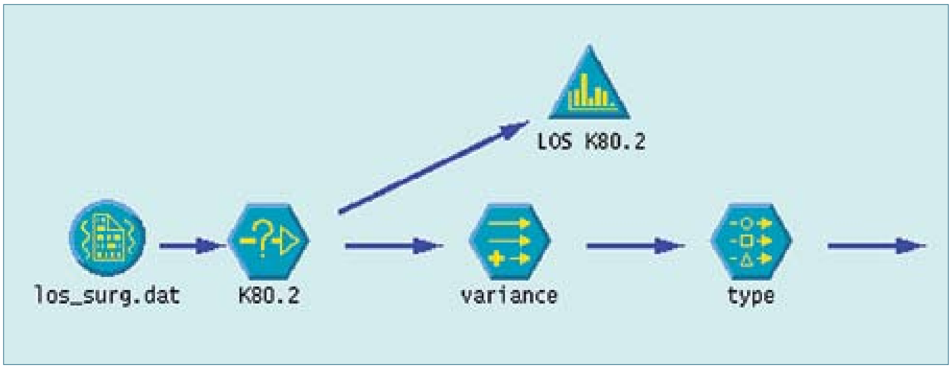


Figure 7: Here the constant is excluded from the analysis

This comes up with a simple rule, showing the significant variance in LOS is governed by the secondary procedure.

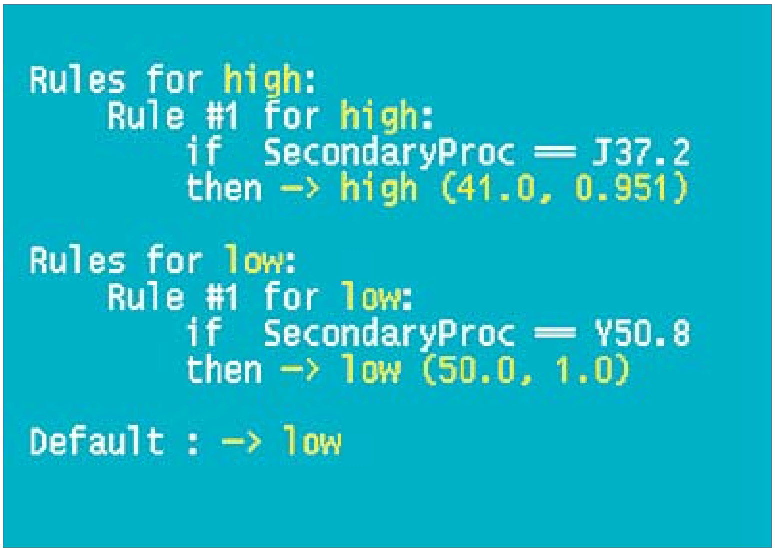


Figure 8: J37.2 Other open operations on Bile Duct/operative Cholangiography (open surgery)

Y50.8 Other Approach through abdominal wall NOS (keyhole surgery)

The Web node is used to find out what the associated primary and secondary procedures are for the primary diagnosis in question.

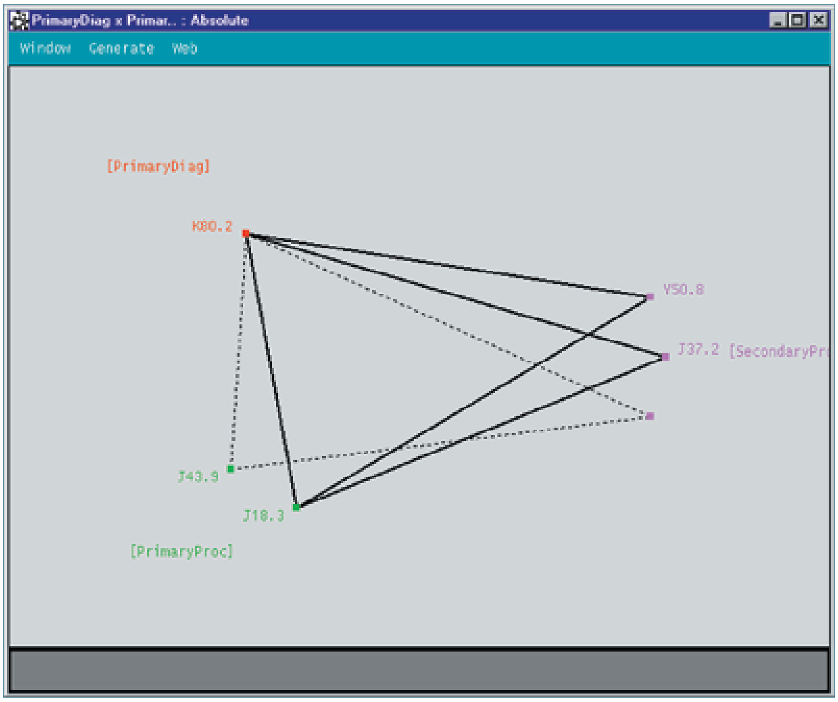


Figure 9: This demonstrates that the secondary procedures in question are strongly associated with primary procedure J18.3 and primary diagnosis K80.2. Generate select node and show LOS distribution.

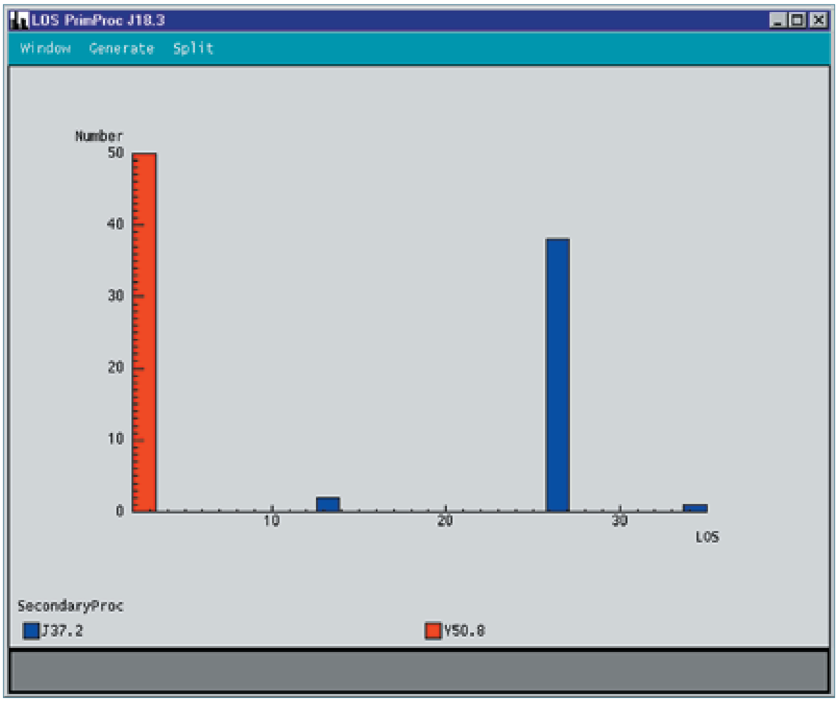


Figure 10: The same primary procedure shows different LOS, depending on secondary procedure.

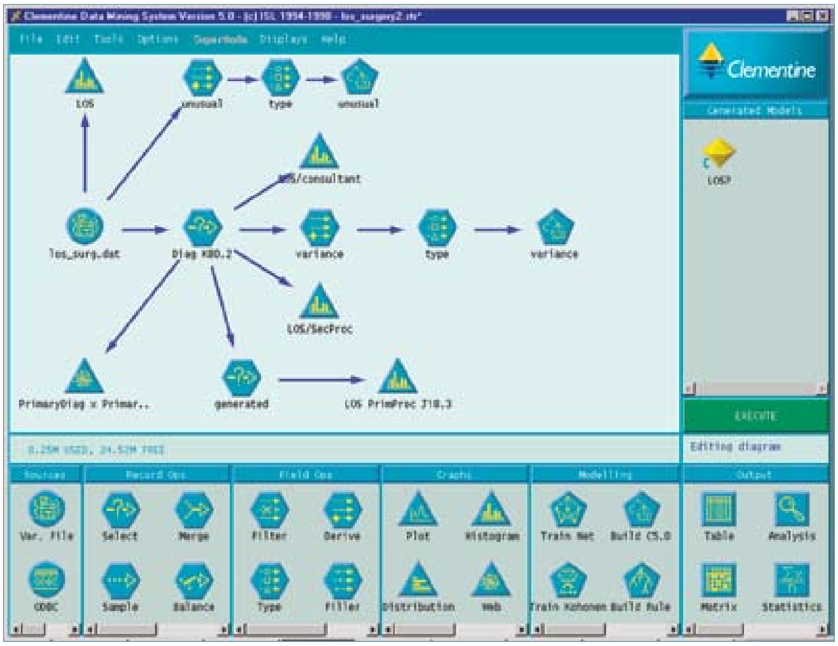


Figure 11: Example of PASW Modeler stream

Discussion

The findings from any data mining exercise can bring patterns to the surface that might otherwise remain undiscovered, and those patterns may suggest alternative ways for treating patients, making better use of resources.

In the described scenario, actual interpretation of these results can only be made by speculation. Using an indirect approach to patient costing—in this case by analyzing length of stay—it may be shown that there is an exceptionally low cost for an emergent technique for surgery, as opposed to the more established techniques for the same primary procedure. However, a hospital is a complex environment and keyhole surgery is likely to extend time spent in the hospital, which would bring further pressures on resource availability.

It would appear from this scenario that there are previously undiscovered patterns which can be induced using data mining techniques. Evidence-based data could stimulate further discussion and investigation by both managers and clinicians working in partnership for the good of the patient and the hospital.

The discussion and further analysis would potentially include the balance of increased institution utilization against reduced utilization of wards and associated support services. Further discriminating factors may then be discovered which could hone the clinical decision, in that there may be other factors that influence the decision to opt for open surgery in preference to keyhole surgery techniques.

Conclusion

The use of data mining has focused on evidence-based patterns from previous patient treatment. In all likelihood, the absence of automated discovery of patterns would leave many questions unasked. These questions, if asked, would benefit not only the resource utilization for patient treatment, but also the health of the patient.

Data mining helps professionals discover these patterns and put them to work. As models are based directly on history, they represent the ultimate in evidence-based care. But technology is no panacea, and professional, ethical and practical issues must be addressed. Decisions must rest with the healthcare professionals, not the information systems experts.

References

Khabaza, T. & Shearer, C. (1985). *Data Mining by Data Owners: Presenting Advanced Technology to Non-Technologists through the Clementine System*. Intelligent Data Analysis '95, Baden-baden.

Quinlan, J. R. (1983). *Learning efficient classification procedures. In Machine Learning: An Artificial Intelligence Approach,* ed. Michalski, Carbonnel & Mitchell. Tioga Press.

Quinlan, J. R. (1993). C4.5: *Programs for Machine Learning.* Morgan Kaufmann.

Shearer, C. (1995). *User-Driven Data Mining Applications*. Unicom Data Mining Seminar, London

About SPSS Inc.

SPSS Inc. (NASDAQ: SPSS) is a leading global provider of predictive analytics software and solutions. The company's predictive analytics technology improves business processes by giving organizations consistent control over decisions made every day. By incorporating predictive analytics into their daily operations, organizations become Predictive Enterprises—able to direct and automate decisions to meet business goals and achieve measurable competitive advantage.

More than 250,000 public sector, academic, and commercial customers rely on SPSS Inc. technology to help increase revenue, reduce costs, and detect and prevent fraud. Founded in 1968, SPSS Inc. is headquartered in Chicago, Illinois. For additional information, please visit **www.spss.com**.

SPSS

To learn more, please visit www.spss.com. For SPSS Inc. office locations and telephone numbers, go to www.spss.com/worldwide.