Data Mining and Decision Systems

08338

Assessed Coursework

Data Mining of Legacy Data

Student Number: 201303310

Stage5: Due • 2pm 14 December 2015 Report

(PDF File with TurnItIn Report)

Date: Tuesday, 22 December 2015

Table of Contents

1. Technique Selection

To select the classifier most suitable for the domain data we will investigate the selected classifiers on their performance across multiple criteria appropriate for the domain data. The criteria are selected by taking the following into account: data volume, space, type and quality. Also the domain, required output and computational aspects will be considered.

Criteria	Trait	Opposite
Build Time	Low	High
Model Understandability	Visible	Hidden
Nominal Data	Good	Poor
Numeric Data	Good	Poor

Table 1. A criteria bi-polar scale table used to determine classifier-criteria performance

As of the table above, four criteria have been chosen to determine which classifiers will be used. The domain data uses both nominal and numeric values so it is important to determine classifier performance for these criteria. Build time and model understandability have been chosen due to the nature of the data mining problem – high risks are involved with errors and duration of data mining.

Criteria	Build Time	Model	Nominal Data	Numeric Data	Sum
Trait	Low (3)	Visible(3)	Good(3)	Good(3)	
Opposite	High(1)	Hidden(1)	Poor(1)	Poor(1)	
MLP	1	1	1	3	6
RBF	2	1	1	3	7
J48	3	3	3	3	12
ID	3	3	3	1	10
LR	2	2	2	3	9
Ridor					
JRip					

Table 2. A table showing classifier performance over different criteria

Table 2 shows that after the different classifier performance have been taken into account, that J48 is the most suited for the domain data used. This is due to having high performance across all criteria. Low build time allows quick results from large data sets, minimizing duration of data mining and potential high risk patients being missed. Model understandability allows easier and more accurate interpretation of data providing the same benefits as the build time criteria. Having good performance for nominal and numeric data eliminates the need to transform data values – however it also gives the option to transform data values to possibly increase accuracy and reduce errors. A health clinic that would use J48 as a classifier would have high performance in data mining – It would take less time to complete with fewer errors, yet also providing good flexibility across data types. Overall this would minimize errors, in turn allowing the health clinic to allocate resources to treating high risk patients first with the confidence knowing that false negatives and positives have been reduced significantly and no patients will be misdiagnosed.

2. Final Data Description

Here table 3 describes the final transformed data used in the data mining problem. It details all the attributes, their classifier and value type, number of values and original values. It details what the data has been transformed in over the course of the data mining to incorporate different classifier requirements E.g. Tertius. required all the values be transformed to numeric. Diabetes, IHD, Hyperextension, Arrhythmia, History and Risk were all relatively straight forward when transformed into numeric values – they became 0 and 1 for no and yes respectively. However Indication was given numeric values for each possible input – 0 for a-f, 1 for cva, 2 for tia and 3 for asx. To convert IPSI and contra to nominal (to increase the performance of rules such as J48 and JRip) a banding technique was used. To get the bands of low, medium and high the mean plus standard deviation of the attribute were used. Low was acquired by Mean – SD, medium was taken as the mean and high was taken as Mean + SD.

Attribute	Classifier Type	Value Type	NumberOfValues	Values	Transformed Nominal	Transformed Numeric	Correlation to Risk
Id	Irrelevant	numeric	979	[110, 180117]	n/a	n/a	-0.080416681
Indication	input	nominal	4	[a-f, cva, tia, asx]	n/a	[0, 1, 2, 3]	0.012883719
Diabetes	input	nominal	2	[yes, no]	n/a	[0, 1]	0.338431232
IHD	input	nominal	2	[yes, no]	n/a	[0, 1]	0.315198571
Hypertension	input	nominal	2	[yes, no]	n/a	[0, 1]	0.412161053
Arrhythmia	input	nominal	2	[yes, no]	n/a	[0, 1]	0.688435865
History	input	nominal	2	[yes, no]	n/a	[0, 1]	-0.00650724
IPSI	input	ordinal	3	[50,99]	[low, medium, high]	n/a	0.369778618
Contra	input	ordinal	3	[10,100]	[low, medium, high]	n/a	0.607676239
Risk	target	nominal	2	[low, high]	n/a	[0, 1]	n/a

Table 3. Table describing the final transformed data

The last column of table 3 is the correlation to risk, this is a number of possible values ranging from negative to positive one, with the high positives numbers showing positives correlation and vice versa. Indication shows a very weak positive correlation to risk showing that it has little (when compared to other attributes) on the risk of mortality of a patient. However Arrhythmia and Contra are shown to have a high positive correlation in turn meaning that their values have a great effect on the risk of mortality. Diabetes, IHD, Hyperextension and IPSI all show middle ground positive correlation with a middle of the road effect on risk. History is an outlier in this respect – almost all of the values are the same in the domain data and it is shown to have the lowest negative correlation; it's removal would not greatly affect the data and may have even increased classifier performance.

3. Classifier Decision Rules

To produce decision rules for the domain data I have used the classifiers J48, JRip and the association rule generator Tertius. These decision rules have then been deployed to classifying patients as high or low risk using the other attributes given. Table 4 shows the high risk rules in a decision table using attribute values that are consistent with the patient test set i.e. any rules which does not conform to the five unknown risk records have been removed.

Id	Indication	Diabetes	IHD	Hypertension	Arrhythmia	History	IPSI	Contra	Risk	Info
JRip-1					yes				high	251.0/6.0
JRip-2				yes			>=85	>=70	high	54.0/0.0
JRip-3			yes				>=70	>=90	high	43.0/1.0
JRip-4			yes				>=90	>=35	high	17.0/0.0
JRip-5		yes						>=50	high	13.0/0.0
JRip-7	asx			yes				>=65	high	7.0/1.0
JRip-9			yes	yes				<=75, >=65	high	3.0/0.0
Tertius-1		yes			yes			high	high	
Tertius-2					yes			high	high	
Tertius-3					yes			high	high	
Tertius-4					yes	yes		high	high	
Tertius-5		yes			yes		high		high	
Tertius-8	asx				yes			high	high	
Tertius-9					yes		high		high	
Tertius-10					yes	no	high		high	
J48-13	tia	no	yes	no	no			>85	high	7.0
J48-16		no	yes	yes	no		>67	>65	high	55.0/2.0
J48-17	cva	no	no	yes	no		>67	>65	high	4.0/1.0
J48-20	asx	no	no	yes	no		>67	>65	high	6.0
J48-23		yes			yes			>35	high	36.0
J48-26	asx				yes			<=40	high	0.0
J48-28					yes			>40	high	237.0/1.0

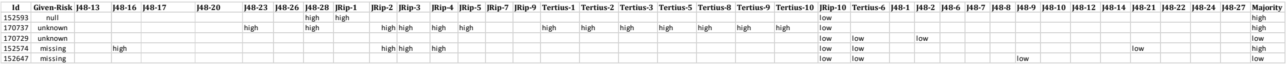
Table 4. High Risk rules in a Decision Table using attribute-values that are consistent with the Patient test set.

In the data-warehouse spreadsheet, a table in the sheet "ConflictTable-Deployed" highlights any contradictory (conflict) rules i.e. rules that disagree with those in table 4.

4. Deployment.

In the domain data given there are five records with risk values unknown, missing or null. As part of the data mining approach these need to be rectified and given values using rules created from the base data. It is important to complete all risk values as these need to be known for the health clinic to act upon these patients. Having unknown risk records could mean patients will not get the treatment they need leading to disastrous consequences.

Table 5 gives what output the rules give for the unknown risk records. These are gathered by first removing rules which do not conform to the records i.e. if the record shows a different indication value to one stated in the rule. Each rule is then applied to the unknown risk records, matching each stated attribute value of the rule to that of the record – if all attribute values match then the record can be stated to be risk high or low for that individual rule. Due to the nature of the data mining problem and the consequences of false negatives (possible mortality of a patient which did not get treatment needed), a low risk will only be applied if no rules give a high risk. Otherwise the majority will be given.



Id	Given-Risk	J48-13	J48-16	J48-17	J48-20	J48-23	J48-26	J48-28	JRip-1	JRip-2	JRip-3	JRip-4	JRip-5	JRip-7	JRip-9	Tertius-1	Tertius-2	Tertius-3	Tertius-5	Tertius-8	Tertius-9	Tertius-10	JRip-10	Tertius-6	J48-1	J48-2	J48-6	J48-7	J48-8	J48-9	J48-10	J48-12	J48-14	J48-21	J48-22	J48-24	J48-27	Majority
152593	null							high	high														low															high
170737	unknown					high		high		high	high	high	high			high	high	high	high	high	high	high	low															high
170729	unknown																						low	low		low												low
152574	missing		high							high	high	high											low	low										low				high
152647	missing																						low	low						low								low

Table 5. A table showing how the classifier decision rules have been deployed to the five unknown risk records

There are many issues with deployment when used in the real world and in this example the health clinic. First there lies the problem of what classifiers to use - each classifier needs to be incorporated into the system; needs to be updated. The classifiers have to be examined on rule accuracy and improvement each time they are used to be sure that the deployed outputs are the most accurate possible.

The DSS needs to be evaluated itself to determine ease of use, hardware, cost effectiveness, discourse, quality of decision advice, performance and design time. The DSS has to be evaluated and compared in its decision to that of a human expert (when given the same information). To evaluate the typical criteria include correctness, efficiency, friendliness and performance.

Does the health clinic need a DSS? Examples of where it does include – high data volume causing low performance due to human inability to perform all tasks, experts needed in environments hostile to humans, human experts are more expensive than a DSS system and unavailability or shortage of human experts and hence danger of corporate goals being compromised.

However, after the DSS has been evaluated, it does have a plethora of benefits e.g. It can be distributed throughout the organisation (the expert can everywhere at once), ability to work with incomplete and uncertain information, cost reduction, increased output and improved quality of output.

Overall there are many issues with using a DSS system, including many ethical issues (using a computer program over a human expert), but when used suitably and alongside human experts it can give the health clinic more tools to diagnose patients.

5. References

Dunham, M. H., Data Mining, Prentice-Hall, 2002.

Fox, J., Glasspool, D., Patkar, V., Austin, M., Black, L., South, M., Robertson, D. and Vincent, C. (2010) 'Delivering clinical decision support services: there is nothing as practical as a good theory' in J Biomed Inform, United States: 831-43.

Hall, M., Frank, E., Holmes, G., Pfahringer,B., Reutemann, P. and Witten, I.H. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Haykin, S., Neural Networks : A Comprehensive Foundation, Prentice-Hall, 1999.

Inc, S. (2000) CRISP -DM 1.0, Chicago, Ill.: SPSS Inc.

Mitchell, T.M., Machine Learning, McGraw-Hill, 1997

Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.

Weka Home Page, http://www.cs.waikato.ac.nz/ml/weka/ Last Accessed 20 November 2014.

Witten, I.H., Frank, E. and Hall, M.A. Data Mining: Practical Machine Learning Tools and Techniques (3/e), Morgan Kaufmann, 2011