Data Mining and Decision Systems

08338

Assessed Coursework

Data Mining of Legacy Data

Student Number: 201305389

Stage5: Due • 2pm 14 December 2015 Report

(PDF File with TurnItIn Report)

Date: Sunday, 13 December 2015

Table of Contents

1. Technique Selection

Classification of data assigns items in a data set to target categories or classes, to achieve an accurate prediction for the target class/record. There are a number of different classifiers that can be used each having situational benefits dependent on the domain data. Each classifier takes on a situational variation of decision trees; A decision tree is a predicative machine-learning model, given the target values of a data set it makes decisions based on various attribute values of the data set. The internal nodes of a decision tree stand for the different attributes of the data, the branches that branch from the nodes stand for the different possible values that the attributes could hold, finally the terminal nodes stand for the final value and the classification of the variable/attribute. J48 Decision tree classifier was the default classifier used throughout the ACW Data Warehouse, it follows a simple algorithm, in order to classify a new item, it first creates a decision tree based on the given attribute values of the available data set. Whenever there is an encounter with a set of records it identifies the attributes that discriminate the various record instances, thus giving an output about the data instance so that the records can be classified. If there are any values within the record that hold no ambiguity, meaning if the data record falls within its own category and has the same value of the target variable the branch is terminated and assigned to the target value that has been obtained. For the reaming records, a new attribute is taken in search for the highest information gain, the end result being either a clear decision of what combination of attributes give an accurate target value or there are no attributes remaining, in which case the target value becomes the majority of the items under the branch.

Criteria	**Task**	**J48**	**Ridor**	**NaiveBayes**	**Logistic**	**Jrip**	**NNge**	**PART**
Accuracy	1	95.90%	95.80%	94.90%	97.20%	97.30%	97.30%	98.00%
RMSE	2	0.1866	0.2049	0.19	0.1527	0.1568	0.1643	0.1364
Sensitivity	3	0.9525	0.9425	0.9125	0.9675	0.975	0.9725	0.9775
Specificity	4	0.963333	0.968333	0.973333	0.975	0.971667	0.973333	0.981667

Table 1. This is a Table for this section using comparison data taken from my ACW Data Warehouse using the Clean2 Data Set consisting of 1000 data records.

The above table data has been acquired through a data set consisting of 1000 data records being ran through Weka under different classifiers in effort to find out what is the most beneficial classifier to achieve the most accurate results for the Health Clinic scenario of the ACW. The first criteria ‘Accuracy’, clear cut criteria that shows the highest value being the most accurate thus being the better classifier. The next criteria being Root Mean Squared Error, which is the risk function corresponding to the expected value of the missing/lost data, relating to the data of the Health Clinic the lower the value the more accurate the output of the classifier. Lastly, Sensitivity and Specificity being calculated through a formula in which the number of True Positive Values are divided between the total of True Positive and False Negative values to gain the Sensitivity of a classifier, similarly the specificity by taking the True Negative and False Positive Values. The closer the Sensitivity and Specificity values are to the value 1 the more accurate the classifier is, with the cleaned data relating to the records of the Health Clinic.

2. Final Data Description

Final Data Description describes the final clean and transformed data – this details all the clean data volume, data attributes and their value ranges, transformed attributes, and their new values ranges. The following 'Table 2' outlines the Final Data and the Transformed Values; the Clean Data taken and Transformed into two types of data sets. Nominal; using a simple formula which takes in the value from the IPSI and Contra (as these attributes are the only relevant attributes that are Numeric) if the value is greater than 85 the value gets transformed into "high" else the value is "low". Numeric; using a similar formula to achieve the opposite result, if the value taken from each of the Nominal Attributes starting with the simple "yes", "no" or "high", "low" values turn into "1" and "0", the Indication Attribute consisted of 1000 (cva), 0100 (a-f), 0010 (asx) and finally 0001 (tia), data set in its Numeric form was then taken and used with the CORREL function within excel to give the correlation to the risk value.

Attribute	Classifier Type	Value Type	# of Values	Values	Transformed Nominal	Transformed Numeric	Correlation to Risk	Comment
Id	Irrelevant	Numeric	1000	110 - 180117	Irrelevant	Irrelevant	Irrelevant	Anonymous patient record identifier: Should be unique values.
Indication	Input	Nominal	cva(268) a-f(342) asx(140) tia(250)	cva, a-f, asx, tia	Irrelevant	1000, 0100, 0010, 0001	0.008985619	What type of Cardiovascular event triggered the hospitalization? Low Correlation.
Diabetes	Input	Nominal	1000	yes, no	Irrelevant	1, 0	0.32535041	Some effects present, if yes to Diabetes strong indicator of risk.
IHD	Input	Nominal	1000	yes, no	Irrelevant	1, 0	0.302922655	Some effects present, not strong indicator of risk.
Hypertension	Input	Nominal	1000	yes, no	Irrelevant	1, 0	0.409662511	Good/Average indicator of risk.
Arrhythmia	Input	Nominal	1000	yes, no	Irrelevant	1, 0	0.680745693	Good chance of high risk if yes to Arrhythmia.
History	Input	Nominal	1000	yes, no	Irrelevant	1, 0	-0.00650724	Negative Correlation.
IPSI	Input	Numeric	1000	50 - 99	high, low	Irrelevant	0.507568049	Average indicator of risk.
Contra	Input	Numeric	1000	10 - 100	high, low	Irrelevant	0.663695496	More chance of high risk the higher percentage of Contra.
Risk	Target/Output	Nominal	1000	high, low	Irrelevant	1, 0	Irrelevant	Is the patient at risk (Mortality)?

Table 2. This is a Table for this section taken from my ACW Data Warehouse using the Final Transformed Data.

3. Classifier Decision Rules

Classifier Decision Rules, using 3 classifiers J48, PART and Tertius within Weka produces the decision rules for classifying patients as High or Low Risk. 'Table 3' is the table taken from my DSS Worksheet within my ACW Data Warehouse, outlines all the Rules dedicated to the 5 different Records that have an unknown Risk.

Id	Indication	Diabetes	IHD	Hypertension	Arrhythmia	History	IPSI	Contra	Risk	J48	PART	Tertius
152593	CVA	no	no	yes	yes	no	80	80	null	J48-5	PART-14	
170737	Asx	yes	yes	no	yes	no	95	100	Unknown		PART-1	
											PART-2	
170729	CVA	no	yes	yes	no	no	80	20	Unknown	J48-6	PART-5	
										J48-7	PART-15	
											PART-16	
152574	TIA	no	yes	yes	no	no	95	100			PART-7	
											PART-9	
152647	ASx	no	yes	no	no	no	80	75		J48-15	PART-17	Tertius-1
										J48-21	PART-18	Tertius-2
											PART-19	Tertius-3
												Tertius-4
												Tertius-5
												Tertius-6
												Tertius-7
												Tertius-8
												Tertius-9
												Tertius-10
												Tertius-11
												Tertius-12
												Tertius-13
												Tertius-14
												Tertius-15
												Tertius-16
												Tertius-17
												Tertius-18
												Tertius-19
												Tertius-20
												Tertius-21
												Tertius-22
												Tertius-23
												Tertius-25

Table 3. This is a Table for this section using data taken from my ACW Data Warehouse.

The above ‘Table 3’ consists of Rules uniquely related to the data records, these rule sets have been adapted and chosen to the different records by me and come with a very high number of conflicts when searched for all possible conflicts related. The following ‘Table 4’ gives a very brief overlook of the conflicting rule sets with the 5 unknown risk records, due to limited space within this document I was required to only include the initial 3 conflicting rule sets and only part of the ‘Tertius’ Rules; I have gone through every possible rule set confliction related to the unknown risk records and have listed them all within the table of the ACW Data Warehouse therefore the full contents of ‘Table 4’ can only be found there.

Id	Indication	Diabetes	IHD	Hypertension	Arrhythmia	History	IPSI	Contra	Risk	ID Matched	Conflict Set
J48-5					yes			>40	high	152593	PART-1, PART-14, PART-18…
J48-15		no	yes		no		>67	>65	high	152647	J48-21, PART-17, PART-18…
PART-1					yes			>40	high	170737	PART-2, PART-7, PART-9…
PART-2		yes						>35	high	170737	PART-1, PART-7, PART-9…
PART-7							>90		high	152574	J48-15, PART-9, PART-17…
PART-9			yes				>67	>85	high	152574	J48-15, PART-7, PART-17…
PART-16				yes	no			<=75	high	170729	J48-6, J48-7, PART-5…
PART-18							<=85		high	152647	J48-15, J48-21, PART-17…
Tertius-13					yes				high	152647	J48-15, J48-21, PART-17…
Tertius-21						no		high	high	152647	J48-15, J48-21, PART-17…
Tertius-23								high	high	152647	J48-15, J48-21, PART-17…
J48-6	cva	no	yes	yes	no			<=65	low	170729	J48-7, PART-5, PART-15…
J48-7		no	yes	yes	no			<=65 <=25	low	170729	J48-6, PART-5, PART-15…
J48-21		no		no	no			>65 <=85	low	152647	J48-15, PART-17, PART-18…
PART-5								<=25	low	170729	J48-6, J48-7, PART-15…
PART-14				yes	yes				low	152593	J48-5, PART-1, PART-18…
PART-15					no			<=65	low	170729	J48-6, J48-7, PART-5…
PART-17			yes		no				low	152647	J48-15, J48-21, PART-18…
PART-19									low	152647	J48-15, J48-21, PART-17…
Tertius-1		no			no			low	low	152647	J48-15, J48-21, PART-17…
…	…	…	…	…	…	…	…	…	…	…	…
Tertius-22		no		no	no				low	152647	J48-15, J48-21, PART-17…
Tertius-25								low	low	152647	J48-15, J48-21, PART-17…

*Table 4. This is a Table for this section using *PARTIAL* data taken from my ACW Data Warehouse – The full table can be found in the ConflictTable-Deployed Worksheet.*

4. Deployment.

Using the Decision and Conflict Tables deployed within the ACW Data Warehouse the decision rules can be worked down to get uniquely identified to each risk unknown data record and come up with 'Table 5' which only shows the partial table as the original table is too large to fit in this document and can therefore be found in the DSS Worksheet of the ACW Data Warehouse. However, the end results of these rule sets being placed on the data records can be seen in the majority result of the 'Table 5'.

Id	Given-Risk	J48-5	J48-6	J48-7	J48-15	J48-21	PART-1	PART-2	PART-5	…	Tertius-21	Tertius-22	Tertius-23	Tertius-25	Majority
152593	null	high								…					high
170737	Unknown						high	high		…					high
170729	Unknown		low	low					low	…					low
152574										…					high
152647					high	low				…	high	low	high	low	low

*Table 5. This is a Table for this section using *PARTIAL* data taken from my ACW Data Warehouse – The full table can be found in DSS Worksheet.*

The most prominent rule set being the Tertius rules covering almost every unknown Data Record with each of the rules made due to such a broad set of values for each attribute and thus causing huge conflict however due to the large amount of similarities the rules can be discarded for the purpose of sticking to a unique record and become viable for the majority vote of the record to get the risk, based on those decisions. The other two rule sets fall pretty selectively to the different data records and again show the accuracy through the majority vote of which risk under which rule within the little data available within the unknown data records.

Due to the large amount of classifiers available there will be a number of different classifiers available to achieve the end results of the unknown risk data records, however using the process I have gone through and gathering all the information based from the size of classifiers that I ran the data through can consider different issues for the Health Clinic data, by potentially choosing classifiers that don't conflict with the rule sets as much as the chosen ones in this scenario did. Also banding the Nominal Data as Numeric could yield an increase in performance of the process rather than going through the Nominal Data as in the scenario, Numeric Set of the Clean Data has been accomplished with the Warehouse and was used to gather correlation information on the entire data set and their risk collating to each of the relevant attributes. More complicated approach would generate improvement metrics for the rules and decision using that metrics would be an alternate approach, using the decision tables and spotting the conflicts through ordering the rules.

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

Appendix A

This appendix shows some extra detail on the missing/unknown risk data records.

Id	Indication	Diabetes	IHD	Hypertension	Arrhythmia	History	IPSI	Contra	Risk
152593	CVA	no	no	yes	yes	no	80	80	null
170737	Asx	yes	yes	no	yes	no	95	100	Unknown
170729	CVA	no	yes	yes	no	no	80	20	Unknown
152574	TIA	no	yes	yes	no	no	95	100	
152647	ASx	no	yes	no	no	no	80	75	

Appendix B

This appendix shows some extra detail on data cleaning and classifier performance.

Classifier	Data	RMSE	Accuracy	TP	FP	TN	FN	Sum	Sensitivity	Specificity
j48	BaseData-All	0.1638	93.94%	377	27	585	34	1023	0.917275	0.955882
Ridor	BaseData-All	0.1927	92.57%	589	52	360	22	1023	0.963993	0.873786
NNge	BaseData-All	0.1363	96.29%	590	15	397	21	1023	0.96563	0.963592
Jrip	BaseData-All	0.1444	95.60%	590	22	390	21	1023	0.96563	0.946602
J48	BaseData	0.223	94.61%	380	24	585	31	1020	0.924574	0.960591
J48	Clean0	0.228	94.19%	375	25	581	34	1015	0.91687	0.958746
J48	Clean1	0.1866	95.90%	381	22	578	19	1000	0.9525	0.963333
J48	Clean2	0.1866	95.90%	381	22	578	19	1000	0.9525	0.963333
Ridor	Clean2	0.2049	95.80%	377	19	581	23	1000	0.9425	0.968333
NaiveBayes	Clean2	0.19	94.90%	365	16	584	35	1000	0.9125	0.973333
Logisitc	Clean2	0.1527	97.20%	387	15	585	13	1000	0.9675	0.975
Jrip	Clean2	0.1568	97.30%	390	17	583	10	1000	0.975	0.971667
NNge	Clean2	0.1643	97.30%	389	16	584	11	1000	0.9725	0.973333
PART	Clean2	0.1364	98.00%	391	11	589	9	1000	0.9775	0.981667
J48	Clean2 Nominal	0.2128	94.70%	374	27	573	26	1000	0.935	0.955