CONF-IRM 2008 Proceedings

International Conference on Information Resources Management (CONF-IRM)

5-2008

Applying Data Mining Techniques to Medical Databases

Fatemeh Hosseinkhah
Howard University Hospital, fatemehhosseinkhah@yahoo.com

Hassan Ashktorab
Howard University Hospital, hashktorab@howard.edu

Ranjit Veen
American University, rveen@yahoo.com

M. Mehdi Owrang
American University, Owrang@american.edu

16F. Applying Data Mining Techniques to Medical Databases

Fatemeh Hosseinkhah
Howard University Hospital
fatemehhosseinkhah@yahoo.com

Hassan Ashktorab
Howard University Hospital
hashktorab@howard.edu

Ranjit Veen
American University
rveen@yahoo.com

M. Mehdi Owrang O.
American University
Owrang@american.edu

Abstract

The data mining techniques such as Neural Network, Naïve Bayes, and Association rules are at present not well explored on medical databases. In this paper, we present and analyze our experimental results on thrombosis medical database by employing data mining tool of XLMiner and using different data mining techniques such as Naive Bayes and Neural Network for classification, Association rules, and Neural Network and K-Nearest Neighbors for prediction. As seen from experiments, some results are common across various mining techniques while others are unique.

Keywords

Data Mining, Medical Database, Data Mining Tool, Data Mining Techniques, Association rules, Neural Networks, Naive Bayes, K-Nearest Neighbors

1. Introduction

Recent advances in medical science have led to revolutionary changes in medical research and technology and the accumulation of large volume of medical data that demands for in-depth analysis. While data analysis and data mining methods have been extensively applied for industrial and business applications, their utilization in medicine and health-care is sparse (Babic, 1999; Brossette et al., 1998; Duhamel et al., 2001; Abidi & Goh, 1998). The data mining techniques such as Neural Network, Naïve Bayes, and Association rules are at present not well explored on medical databases.

In this paper, we present and analyze our experimental results on thrombosis database by employing different data mining techniques such as classification (Neural Network, Naive Bayes), prediction (Neural Network, K-Nearest Neighbors), and Association rules and using data mining tool of XLMiner (XLMiner, 2007; Shmueli et al., 2007)..

2. Medical data

Collagen diseases are auto-immune diseases (Zytkow, 2001) that attack the collagen or other components of connective tissue, such as lupus. Patients generate antibodies attacking their own bodies. In collagen diseases, thrombosis is one of the most important and severe complications, one of the major causes of death. Thrombosis is an increased coagulation of blood that clogs blood vessels (Zytkow, 2001). Domain experts are very much interested in discovering regularities behind patients' observation.

The Thrombosis data set (Table 1) consists of two different tables.

1. TSUM_A.CSV contains basic data of 1232 patients.
2. TSUM_B.CSV consists of results of thrombosis specific tests for 763 patients suspected for thrombosis.

TSUM_A.CSV

Item	Meaning	Remark
ID	identification of the patient	
Sex	Gender	
Birthday	Birthday	YYYY/M/D
Description date	the first date when a patient data was recorded	YY.MM.DD
First date	the date when a patient came to the hospital	YY.MM.DD
Admission	patient was admitted to the hospital (+) or followed at the outpatient clinic (-)	
Diagnosis	disease names	multivalued attribute

3. Medical data mining experiments

We used the random partitioning of the XLMiner and created two mutually exclusive datasets, a training dataset comprising 60% of the total dataset, and a validation dataset of 40%. These are the defaults for partitioning. The training dataset is used to train or build a model. Once a model is built on training data, you need to find out the accuracy of the model on unseen data, the validation dataset.

There are many data mining techniques available for data classification, prediction, association analysis, and data exploration. In our experiments, we used Naïve Bayes and Neural Network of classification techniques, Neural Network and K-Nearest Neighbors of prediction techniques, and Association rules mining techniques (Shmueli et al., 2007; XLMiner, 2007). The goal was to find out if there was any correlation between the results of lab testing and diagnosed level of thrombosis, or whether we could predict the pattern/danger of thrombosis.

3.1 Data mining techniques experiments

The two tables TSUM_A and TSUM_B were joined by the join key (patient ID#). We created a new column (Age) based on birthday for a better discovery. Also, a new column (New admission) was created to represent ordinal values 0 and 1. Classification, Association, and prediction operations were used on the joined data.

TSUM_B.CSV

Item	Meaning	Remark
ID	identification of the patient	
Examination Date	date of the test	YYYY/MM/DD
aCL IgG	anti-Cardiolipin antibody (IgG) concentration	
aCL IgM	anti-Cardiolipin antibody (IgM) concentration	
ANA	anti-nucleus antibody concentration	
ANA Pattern	pattern observed in the sheet of ANA examination	
aCL IgA	anti-Cardiolipin antibody (IgA) concentration	
Diagnosis	disease names	multivalued attribute
KCT	measure of degree of coagulation	
RVVT	measure of degree of coagulation	
LAC	measure of degree of coagulation	
Symptoms	other symptoms observed	multivalued attribute
Thrombosis	degree of thrombosis	0: negative (no thrombosis) 1: positive (the most severe one) 2: positive (severe) 3: positive (mild)

Table 1. Thrombosis dataset.

3.1.1 Naïve Bayes classification algorithm

The output variable was Thrombosis. The rest of the attributes were used as inputs. Based on training model, the prior class probabilities included the attributes and values as ((Class, Prob.), ((0, 0.883333333), (1, 0.088888889), (2, 0.016666667), (3, 0.011111111))). Using the validation data, the observed output, the thrombosis with the highest probabilities, includes the attribute and values as ((Predicted Class, Actual Class, Prob. for class 1, Patient ID), ((1, 1, 0.79354182, 2956679), (1, 1, 0.669181115, 5512586))).

XLminer’s Naïve Bayes operation calculated the overall probability as noted above as well as conditional probabilities for the target output variable with each of the input variable. In our case, the output variable was Thrombosis. After the mining in the validation phase, it showed that the class of thrombosis with the highest probability is Class 1 and patient id# 5512586 and 2956679 have the highest probability 1 having Thrombosis =1 (severe for Thrombosis). The target attribute Thrombosis was defined through each of the input variables and the outcome is shown in Table 2. Only those rows with highest probability are shown below.

	Classes-->							
Input	0		1		2		3	
Variables	Value	Prob	Value	Prob	Value	Prob	Value	Prob
Symptoms	None	1	Thrombo-phlebitis	0.15625	AMI	0.499999999	Thrombo-cytopenia	0.999999998
TSUM_B_diagnosis	SLE	0.58490566	SLE	0.25	SLE	0.666666666	SLE	0.749999998
ANA Pattern	S	0.679245283	S	0.5625	S	0.833333332	S	0.499999999
Diagnosis	SLE	0.893081761	SLE	0.875	SLE	0.999999998	SLE	0.999999998
SEX	F	0.977987421	F	1.0	F	0.999999998	F	0.999999998
RVVT	Neg	0.965408805	Neg	0.5625	Neg	0.833333332	Neg	0.499999999
LAC	Neg	0.949685535	Neg	0.5625	Neg	0.833333332	Neg	0.499999999
RVVT	Pos	0.034591195	Pos	0.4375	Pos	0.166666666	Pos	0.499999999
LAC	Pos	0.050314465	Pos	0.4375	Pos	0.166666666	Pos	0.499999999
ANA	16	0.248427673	16	0.25	16	0.166666666	16	0.499999999

Table 2. Conditional probabilities for the target attribute Thrombosis.

In the following, we have some of the findings on joined tables using Naïve Bayes classification algorithm:

Thrombosis is 3: if Symptom is Thrombocytopenia;
Thrombosis is 2: if the Symptom is AMI;
Thrombosis is 1: if Symptom is Thrombophlebitis;
Thrombosis is 0: if no Symptoms is present;
Thrombosis 3: if FINAL_diagnosis is SLE;
Thrombosis 2: if FINAL_diagnosis is SLE;
Thrombosis 1: if FINAL_diagnosis is SLE;
Thrombosis 0 : if FINAL_diagnosis is SLE is actually lower;
If Sex =”F” then Thrombosis is a very high probability;
If RVVT is negative implies LAC is negative;
If RVVT is positive implies LAC is positive;
If values of ANA (anti nucleus antibody) are high, the chance of Thrombosis increases.

3.1.2 Using Neural Network classification algorithm

The output variable was Thrombosis. The rest of the attributes were used as inputs. Based on training model, the prior class probabilities include the attributes and values as ((Class, Prob.), ((0, 0.883333333), (1, 0.088888889), (2, 0.016666667), (3, 0.011111111))). Using the validation data, the observed output, the thrombosis with the highest probabilities, includes the attribute and values as ((Predicted Class, Actual Class, Prob. for class 1, Patient ID), ((1, 1, 0.656570604 , 2956679), (1, 1, 0.667480204, 5512586), (1,1, 0.774229186, 3541223)).).

In the training phase, the correct class for each record is known (this is termed supervised training), and the output nodes can therefore be assigned "correct" values -- "0.9" for the node corresponding to the correct class, and "0.1" for the others. After the mining in the validation phase, it showed that the class of thrombosis with the highest probability is Class 1 and patient id# 3541223 has the highest probability having Thrombosis =1 (positive for Thrombosis), followed by patient id# 3258822 and patient id# 2956679 in close range.

Additionally, validation output indicated that with higher values for ANA (anti nucleus antibody) the chance of thrombosis increases. Also, higher values for aCL IgG, aCL IgM, aCL IgA are good indicators for thrombosis. Furthermore, if any of the KCT, RVVT, LAC tests is positive, it implies the others are also positive. Positive values of KCT, RVVT and LAC (degree of coagulation) are good indicator of thrombosis.

3.1.3 Using Affinity (Association Rules) algorithm

Using the association rules mining with minimum support of 60 and minimum confidence of 50, 16455 rules have been generated. The output with top 3 confidences of rules is shown below.

Rule 1: IF Diagnosis=SLE THEN Sex=F and Age=38 has a confidence of 100%

Rule 2: IF ANA (anti nucleus antibody) Pattern=S THEN Diagnosis = SLE has a confidence of 90.95%

Rule 3: IF admission= 0, Thrombosis=1, ANA Pattern= S and diagnosis=SLE THEN Sex=F has a confidence of 97.3 %

3.1.4 Using Neural Network Prediction algorithm

The output variable was Thrombosis. The rest of the attributes were used as inputs. The variables that are ignored/not compatible and having non-numeric values are SEX, Diagnosis, ANA Pattern, TSUM_B_Diagnosis, and Symptoms.

Predicted value, Actual value and the difference between them (the Residual) are the measures used to measure the performance of the training and validation data models.
The prediction of validation data for the top 5 rows with highest probability of thrombosis are identified and include the attributes and values as ((Predicted Value, Actual Value, Residual, Patient ID), ((1.116933, 1, -0.116933, 3223244), (0.823696, 1, 0.176304, 3258822), (1.005022, 1, -0.005022, 3541223), (0.30988, 1, 0.69012, 4302591), (0.515827, 1, 0.484173, 4561789))).

It is interesting that patient id 3541223 was also identified using Neural Networks classification algorithm. Further analysis showed that if either of the KCT, RVVT tests are positive; it implies LAC is also positive. Higher values for ANA (anti nucleus antibody) increase the chance of Thrombosis.

After the model is built using the training data set, the model is used to score on the training data set and the validation data set (if exists). Then the data set(s) are sorted using the predicted output variable value. After sorting, the actual outcome values of the output variable are cumulated and the lift curve is drawn as number of cases versus the cumulated value. The baseline is drawn as number of cases versus the average of actual output variable values multiplied by the number of cases. Lift charts are visual aids for measuring model performance (Table 3). They consist of a lift curve and a baseline. The greater the area between the lift curve and the baseline, the better the model is.

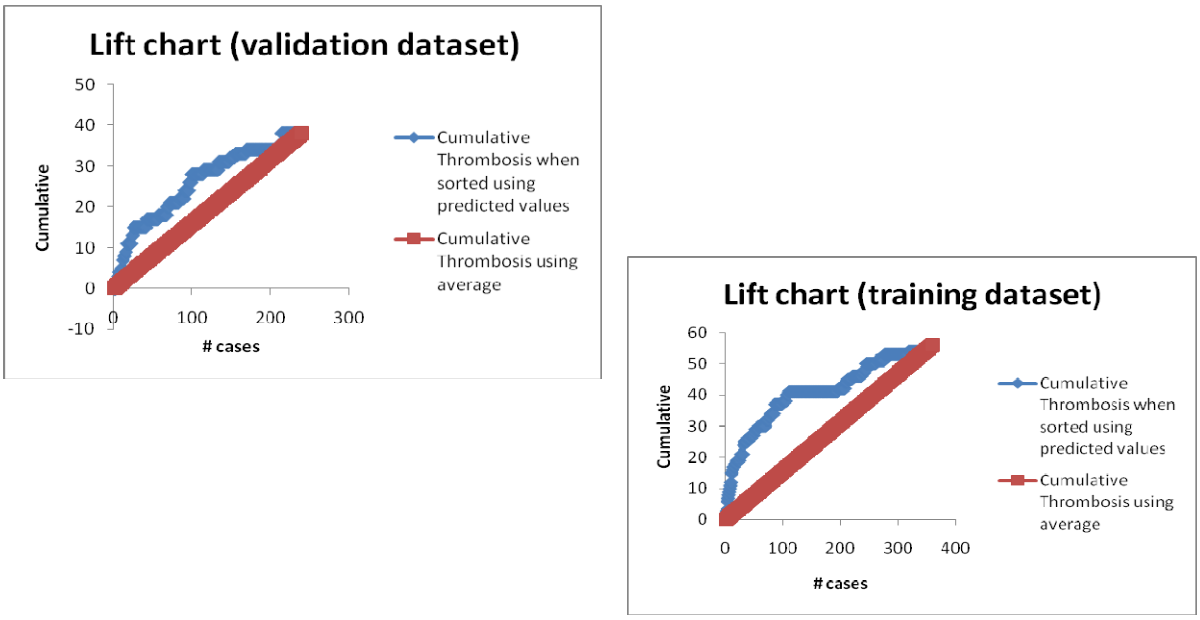

Table 3. Lift charts for training and validation datasets for Neural Network.

In the above lift charts, the training data set has area that can be classified as “average” while the validation dataset depicts the area that is classified as “poor”. Upon investigation, it revealed that the training dataset had representation for Thrombosis 0, 1, 2 and 3 while the validation dataset contained representation for Thrombosis 1 only. The percentage of cases for the entire data set having Thrombosis 0 is 88%, Thrombosis 1 is 8%, Thrombosis 2 is 1.6% and Thrombosis 3 is 1.1 %. In addition, during the mining process, predictions are made against the validation data with prior knowledge from training phase. Many of the predictions match while many do not match, i.e. the Predicted Value and Actual value differ and have a residual value. For example, Predicted value = 0, Actual value = 1 signifying a mismatch.

3.1.5 K-Nearest Neighbors (k-NN) Prediction algorithm

In K-Nearest-Neighbors prediction, the training data set is used to predict the value of a variable of interest for each member of a "target" data set. The input variables are ID, Birthday, Description, Age, First Date, New Admission, Special Examination Date, aCL IgG, aCL IgM, ANA, aCL IgA, KCT, RVVT, LAC and the output variable is Thrombosis. The variables that are ignored/not compatible and having non-numeric values are SEX, Diagnosis, ANA Pattern, TSUM_B_Diagnosis, and Symptoms.

The validation error log for different K includes the attributes and values as ((Value of K, Training RMS Error, and Validation RMS Error), (1, 0, 0.635741037)). The output for the prediction of validation data includes the attributes and values as ((Predicted Value, Actual Value, Residual, Actual#Nearest Neighbors, Patient ID), ((1, 1, 0, 1, 3258822), (1, 1, 0, 1, 4561789), (1, 1, 0, 1, 4563365), (0, 0, 0, 1, 355009), (0, 2, 2, 1, 163109))). XLMiner calculates the RMS error for all values of K and decides that best value of k for which the RMS error is minimum.

Additionally, output showed rows with highest probability of thrombosis are identified. We should note that patient id 3258822 and 4561789 were identified by the Neural Network prediction and Neural Network classification routines. Further analysis showed that if either of the KCT, RVVT tests are positive; it implies LAC is also positive. The lift charts are shown in Table 4.

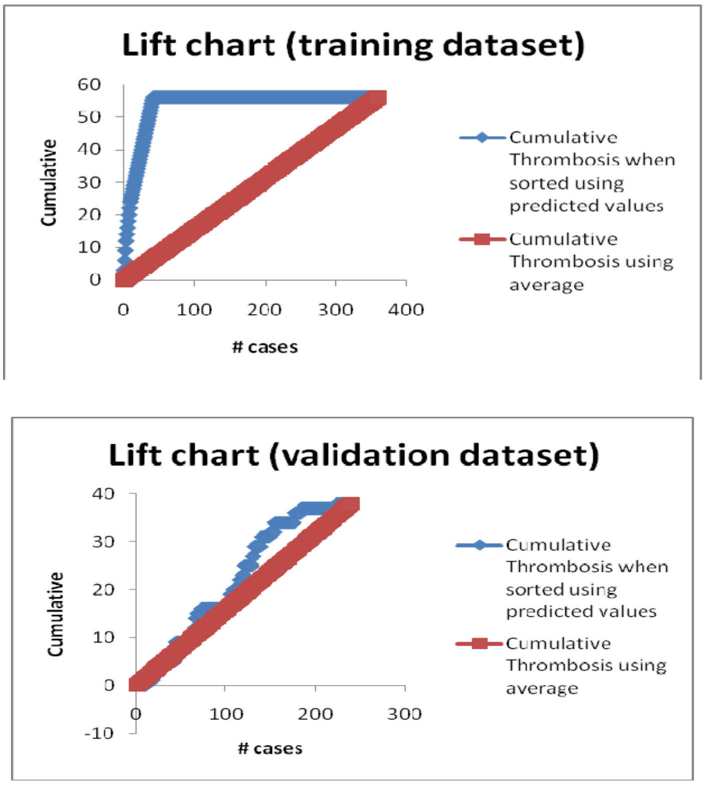

Table 4. Lift charts for training and validation datasets for K-NN.

In the above case for training data, we see the area between the lift curve can be classified as "good". For the validation set, we see a "poor" curve, the predicted value and actual value differ and have a residual value. For example Predicted value = 0, Actual value = 2, Actual Nearest Neighbor = 1 signifying a best value of K =1 and validation data error of 0.65.

3.2 Summary analysis of data mining algorithms

Using Neural Networks classification algorithm, we discovered that patient id# 2956679, 3258822 and 3541223 have the highest probability having Thrombosis =1. Using Naive Bayes classification algorithm we again discovered patient id# 2956679 and 5512586 have the highest probability of having Thrombosis=1 (positive Thrombosis). Using Neural Networks prediction algorithm output showed patient id# 3541223 has the highest probability having Thrombosis= 1. This patient was also identified using Neural Networks classification algorithm. Using K-NN prediction algorithm output showed patient id# 3258822 and 4561789 have the highest probability having Thrombosis =1. These patient ids' were identified by the Neural Network prediction and Neural Network classification routines. The classification algorithms (Naïve Bayes, Neural Networks) gave us a few common results.

Using Affinity algorithm (Association Rules), a total of 16455 rules were generated. A lot of rules did not make sense even though the confidence was up to 100%. As an example, a rule was generated saying all patients seen on 7/23/99 had a high confidence of SLE, which is based on data presented but date may not have any bearing on diagnosis. Finally, as we see from the results, it is not sufficient by just running one operation to correctly predict unknown information. It is advantageous to use 2-3 or more operations and data mining techniques to get a clearer picture and a higher degree of confidence that a particular result was observed by more than one operation. For example, in our experiments, the Naïve Bayes predicted that if values of ANA (anti nucleus antibody) are high, the chance of Thrombosis increases, the Neural Networks predicted Higher values for ANA (anti nucleus antibody) the chance of Thrombosis increases, and the Association rules predicted that IF ANA (anti nucleus antibody) Pattern=S THEN Diagnosis = SLE has a confidence of 90.95%.

Results need to be aggregated from various operations to predict thrombosis with a higher accuracy. Some information like "higher values for aCL IgG, aCL IgM, aCL IgA are good indicators for thrombosis" is observed from Neural Networks only. Whereas the correlation between KCT, RVVT, LAC gives us a higher degree of confidence since it was also produced by Naïve Bayes. The rule IF Diagnosis=SLE THEN Sex=F and Age=38 has a confidence of 100% gives us higher confidence since Naïve Bayes also discovered this information. Hence for predicting degree of Thrombosis, it would add confidence to the prediction if we said that Diagnosis= Thrombophlebitis , Final_diagnosis=SLE, Sex =F , KCT,RVVT,LACV if either of these are positive which is a aggregation of results from Naïve Bayes operation, Neural Networks operation and Association rules.

4. Conclusion

As seen from the experiments, some results are common across various mining techniques while others are unique. The discovered rules appeared to be consistent with the domain experts' views. However, there were several trivial rules generated as well. Mining models, built based on training data, needed to be adjusted in order to get the best performing model. The likelihood of meaningful discovery increases when data is cleansed and transformed. It might be worthwhile to group certain categories (e.g., the field Age) which may improve our chances of discovering less and meaningful rules.

References

Abidi, S. S. R. & Goh, A. Applying knowledge discovery to predict infectious disease epidemics. H. Lee Motoda (Eds.) Lecture notes in artificial intelligence 1531-PRICAI'98: Topics in artificial intelligence, Berlin: Springer Verlag, 1998.

Babic, A., Knowledge discovery for advanced clinical data management and analysis, P. Kokol et al. (Eds.) Medical Informatics Europe'99, Ljubljana, Amsterdam: IOS Press, 409-413, 1999.

Brossette, S.E.; Sprague, A.P.; J. M. Hardin, J.M.; Jones, K.W.T. & Moser, S.A., Association rules and data mining in hospital infection control and public health surveillance. Journal of American Medical Informatics Association, 5(4):373 – 381, 1998.

Duhamel, A.; Picavet, M.; P. Devos, P. & Beuscart, R. (2001). From data collection to knowledge data discovery: A medical application of data mining, V.L. Patel et al (Eds.) 10th World Congress on Medical Informatics (MedInfo'2001), Amsterdam: IOS Press, 2001.

Shmueli, Galit, Patel, Nitin R., Bruce, Peter C., Data mining for business intelligence: concepts, techniques, and applications in Microsoft Office Excel with XLMiner, Wiley-InterScience, Hoboken, N.J., 2007

XLMiner, On Line User Guide, http://www.xlminer.net/, 2007.

Zytkow, Jan M.; Tsumoto, Shusaku & Takabayashi, Katsuhiko, Medical (Thrombosis) Data Description, http://eric.u