Journal of Engineering and Applied Sciences 12 (14): 3638-3643, 2017
ISSN: 1816-949X

Data Mining Variables and Features Selection for Malaysia Blood Donor's Preference Using Correlation Technique

Nor Syuhada Che Khalid, Burhanuddin Mohd. Aboobaider, Nuzulha Khilwani Ibrahim,
Zahriah Sahri and Mohd. Khanapi Abd. Ghani
Department of Biomedical Computing and Engineering Technologies (Biocore),
Applied Research Group Malaysia, Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal,
Melaka, Malaysia

Abstract: Dataset that was constructed from survey, interview or questionnaires forms may suggest about Leading Features (LFs) from all Member Features (MFs) available and produce many sets of LF and MFs combination. However, which LFs will take priority to extract important information approaches were not clearly determine from past studies. Therefore, these study objectives are to introduce and analyze features arrangement for prediction problem on blood donor's preferences datasets to determine which LFs will take priority to extract information through ranking and simplification. Artificial neural network will be used as prediction algorithm for training, validating and testing. In the end, LFs analysis on features arrangement will become useful to blood bank and health care community or organizer to arrange suitable strategy to attract blood donors and contribute their blood to society, especially for everyday emergency and critical situation for worst condition patients in surgeries, accidents and life threatening illnesses.

Key words: Prediction, blood donor's preferences, features arrangement, data mining, feature selection

INTRODUCTION

Prediction is a renowned problem for data mining with several algorithms or approaches as solution. Generally as one of main components of prediction which are Member Features or attributes (MFs) MFs selection will be applied to select as best or prior features needed and eliminate unwanted or inconsistent features (Jiang and Wang, 2016; Stijare and Kumbhalkar, 2015). This approach may apply to survey or questionnaires dataset too. However, utmost importance of those datasets used on this experiment was not the MFs as attributes, apart from MFs selected as Leading Features (LFs). Therefore, these stidies objectives are to discuss data mining problem of Malaysian blood donor's survey and to introduce arrangement of features as additional research for current feature selection. However, ultimate purpose of feature selection is to improve prediction performance, oppositely features arrangement which is to compare features based on associations or correlations as main purpose from simplification processes, quite similar with features relevancy analysis in correlation approach in term of arranging and scoring with different objectives (Guyon and Elisseeff, 2003; Lei and Liu, 2004).

Blood donor's preferences: Feedback or respond from direct consumer and contributor on particular market are always important to determine their influence and attraction as their motivation. Therefore, this study would like to comprehend what are main preferences of blood donors as main contributors on blood service (Stijare and Kumbhalkar, 2015). Besides that rather than simply extract main preferences from prediction, their positions, quantities and contributions to performance from prediction model, should be considered as main indicator to determine whether their association between candidates of LFs as preferences have higher or lower contribution among each other.

In 2015, a survey is conducted in Malaysia. There are 1504 respondents gave full feedback and all feedbacks were compiled into a dataset of 27 variables. Survey conducted is based on background donation experience and favorite, donation fear and donation motivation of the respondents.

MATERIALS AND METHODS

Data mining problem: Several problems have recognized from dataset collected. First, dataset from survey or

Corresponding Author: Nor Syuhada Che Khalid, Department of Biomedical Computing and Engineering Technologies (Biocore), Applied Research Group Malaysia, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

random dataset used did not contain definite LFs and MFs. Determination of LFs and MFs usually came from dependent and independent variables on scientific experiment or official observation but not all survey collection based on behavior or human preferences measurement have definite dependent and independent variables (Bollen *et al*., 2016). Human desires are especially difficult to measure with mainstream approaches. Secondly, when definite LFs and MFs have not clearly determined another issue has arisen. All variables on dataset must be used as LFs. This situation will produce not effective results because after prediction, all LFs would need more evaluation to recognize which LF stoextract information. This problem will become worse if LFs quantities are bigger. Confusion and complex situation will occur after prediction because no focus on LFs used. However, this study will focus on all binary or two nominal variables as LFs.

Features arrangement: FA has main purposes which is to compare between sets of LFs performances based on MFs associations. Construction of this idea has origin from several techniques combination. Main layout of this this idea is shown as Fig. 1. This idea has 4 steps.

First, step is to assign available MFS AND LFS. For Step 2, bivariate Pearson correlation has applied. It should show association between continuous variables through significant linear as statistic calculation, strength (close or not to straight line which is 1) altogether with direction (positive or negative) of relationship and measure correlations between pair of variables in this study (Gravetter and Wallnau, 2016a, b). Dataset of this study is not causal but independent because it contains feedback from various respondents with diversity of preferences. Two-tailed test will show association: if null hypothesis has population correlation is equal to zero as no association which is H_0: p-value = 0 when alternative hypothesis has population correlation is not equal to zero as has association which is H_1: $\rho \neq 0$. Significance level used is 0.01 (Gravetter and Wallnau, 2016a, b). Sample correlation, r between x and y is mentioned as in Eq. 1:

$$\gamma_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)} \cdot \sqrt{\text{var}(y)}} \quad (1)$$

This study would like to highlight how much of MFs are associated with a particular LF to compare among LFs.

These associations have described graphically on Fig. 2 to show better overview. A better LFs from questionnaires or survey should be prioritized on association too, besides than prediction performance. This will make focus and consideration of information

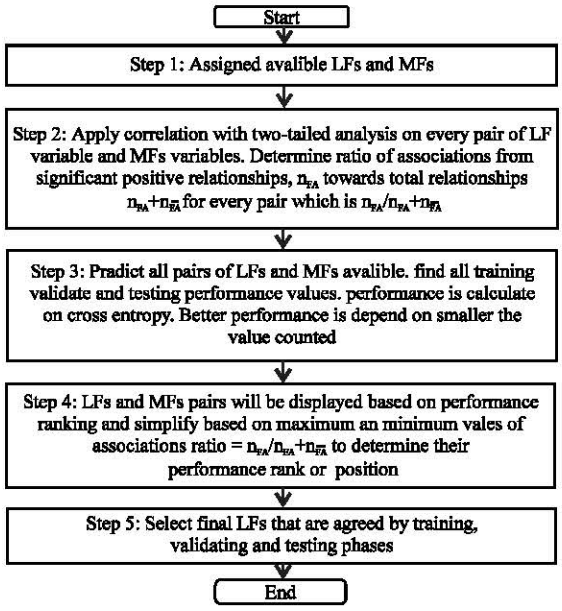

Fig. 1: Steps flow of FA

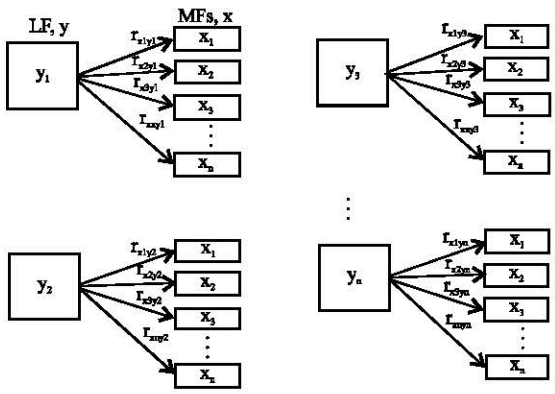

Fig. 2: Relationship between LFs and MFs

extraction on particular group become more accurate because not all high performance prediction will produce useful information because not all better prediction accuracies will origin from higher correlation between class and MFs (Han and Kamber, 2006a). Higher correlation means more respondents or contributors prefer this group or features than others, based on more similar pattern frequently used on this LFs (Han and Kamber, 2006b). Furthermore, when majority preferences are important, greater quantity or increment is prioritize than reduction. Therefore, association needed from Leading Feature (LF) and MFs is significant positive relationship (Han and Kamber, 2006c). As comparison between LFs, ratio or ratio of significant positive relationship toward available relationship is important.

However, this study has applied ratios to make obvious differences with performance values in smaller and ratio number. Association of a LF with MFs will becounted as in Eq. 2 when is number counted from total significant positive relationship or all relationship:

$$\text{Association ratio} = \frac{n_{FA}}{n_{FA} + n_{\overline{FA}}} \quad (2)$$

For example based on (Fig. 2) y_3 has better association ratio than y_1. Why? More r between y_3 with MFs of x are significantly positive than y_1. Highest association ratio does mean primary LF or main factor as certain LF associated most with MFs.

Step 3 is calculation of performance from prediction algorithm applied. Prediction performance would bring clear image of algorithm used. Better performance may suggest better prediction results. However, a survey or questionnaires may produce not too wide on prediction performance gap among LFs. Therefore, smaller or larger number to differentiate better performance may proof difficult. This study has extract performance values from cross entropy. Smaller values may suggest better performance (Boer *et al*., 2005) .

Algoritham: Step 4 of FA: Pseudocode of performance ranks and simplifies LFs based on association ratios:

1.0 start
2.0 Sort table by Cross Entropy (CE%) performance value by decrement order
3.0 Determine (leading features) Lfs from performance comparison suggestio
  3.1 Sort table by Cross Entropy (CE%) performance value by decrement order
  3.2 Check table by every row of significant positive relationships with total relationships percantages for maximum cutter
    3.2.1 Initiate current row as n row
    3.2.2 Determine value inside n row as c
    3.2.3 Compare with value of previous n-m rows
      3.2.3.1 From m are 1, 2, 3,...until n-m = 1, which is first row
        3.2.3.1.1 Determine value inside n-m rows as d
        3.2.3.1.2 Compare valueod of d and c
          3.2.3.1.2.1 If value of d less then c, eliminate that n-m row
          3.2.3.1.2.2 Else, maintain that n-m row
      3.2.3.2 Non-eliminated rows as remaining leading features from highest associations ratio as maxcut features, maxcFs
  3.3 Check table by every row of significant positive relationships with total relationships percantages for maximum cutter
    3.2.1 Initiate current row as n row
    3.2.2 Determine value inside n row as c
    3.3.3 Compare with value of previous n-m rows
      3.3.3.1 From m are 1,2,3,...until n-m = 1, which is first row
        3.3.3.1.1 Determine value inside n-m rows as d
        3.3.3.1.2 Compare valueod of d and c
          3.3.3.1.2.1 If value of d less then c, eliminate that n-m row
          3.3.3.1.2.2 Else, maintain that n-m row
      3.3.3.2 Non-eliminated rows as remaining leading features from highest associations ratio as maxcut features, maxcFs
4.0 Compare maxcF and mincF from trining and validating phase Eliminate non-similar features from maxcFs and mincFs from all phases
5.0 End

Step 4 is ranking and simplifying LFs. When on unclear situation either performance or association is better, both are compulsory information. Therefore, ranking all LF sets performance and simplify to one or several highest associations will become helpful. Ranking performance is based on most perform value until less perform values which are smallest value until biggest value. Simplifying has two forms which are maximum cutter and minimum cutter.

Maximum cutter and minimum cutter purpose is to show and compare performance and quantity of Lfs of highest (maxcF) and lowest (mincF) association Lfs. This is not a best approach to select LFs. However, this approach is better than manual selection and random selection. Step 4 FA algorithm is encoded in pseudocode as shown in algorithm. For this step, extracted maxcFs and mincFs will be ranked and scored to choose final ranking of maxcFs and mincFs. Highest ranking will be scored as 0 and getting lower after that. For Step 5, highest total scores from all phases for selected variables will finalize as first rank for maxcF and mincF.

Experimental procedures: Variable division in this study is given in Fig. 3. This study has conducted based on 5 steps of FA using various tools including neural network tool box. Dataset used already mentioned in part 1.1. Every set of Lfs and Mfs will be divided as shown in Fig. 3.

As mentioned in part 1.2, LFs selected are Q1, Q2, Q6, Q7, Q9, Q10, Q12A until Q12M as shown. For Step 1, pairings of Lfs and their Mfs or Step 1 as shown in Table 1. Basically, 19 variables have selected as LFs.

Variables available on dataset from blood donor's preferences Questionnaires 2015

Variables/descriptions:

Q1: Gender
Q2: Marriage status
Q3: Age
Q4: Donation frequency per year
Q5: Job
Q6: Donation experience
Q7: Donated more than once per year experience
Q8: Favourite donation centre
Q9: Donation fear
Q10: Interested to overcome donation fear
Q11A: Up to date donation
Q11B: Donate frequently
Q11C: More donation volume or capacity is better
Q11D: Longer donation experience
Q12A: Up to date donation
Q12B: Donate frequently
Q12C: High overall donation volume

Table 1: Step 2 of FA: Find Associations ratio of assigned LFs with MFs using pearson's correlation (total relationships are 26 MFs for each LF)

Assigned LFs	Associated MFs (FA)	n_{FA}	Associations ratio
Q1	Q4, Q5, Q6, Q7, Q11D, Q12A, Q12B,Q12C,Q12E, Q12G, Q12K, Q12L	12	0.461538
Q2	8, Q11C, Q12B, Q12H	4	0.153846
Q6	Q1, Q3, Q4, Q5, Q7, Q10, Q11D, Q12A, Q12B, Q12C, Q12D, Q12F, Q12M	13	0.50000
Q7	Q1, Q3, Q4, Q5, Q6, Q10, Q11A, Q11D, Q12A, Q12B, Q12C, Q12D, Q12F, Q12G, Q12L, Q12M	16	0.615385
Q9	Q11B, Q11C	2	0.076923
Q10	Q5, Q6, Q7, Q11B, Q12F, Q12M	6	0.230769
Q12A	Q1, Q5, Q6, Q7, Q11D, Q12B, Q12C, Q12D, Q12E, Q12F, Q12G, Q12I, Q12J, Q12L, Q12M	15	0.576923
Q12B	Q1, Q2, Q4, Q6, Q7, Q11D, Q12A, Q12C, Q12D, Q12E, Q12F, Q12G, Q12I, Q12J, Q12L, Q12M	15	0.576923
Q12C	Q1, Q6, Q7, Q11, Q12A, Q12B, Q12D, Q12E, Q12F, Q12G, Q12I, Q12L, Q12M, Q12N	13	0.50000
Q12D	Q3, Q4, Q6, Q7, Q11A, Q12A, Q12B, Q12C, Q12G, Q12I, Q12J, Q12L, Q12M	13	0.50000
Q12E	Q1, Q12A, Q12B, Q12C, Q12G, Q12H, Q12J, Q12L, Q12M	9	0.346154
Q12F	Q7, Q8, Q10, Q11C, Q12A, Q12B, Q12C, Q12G, Q12L, Q12M	10	0.384615
Q12G	Q1, Q7, Q11A, Q11B, Q12A, Q12B, Q12C, Q12D, Q12E, Q12F, Q12I, Q12J, Q12K, Q12L, Q12M	15	0.576923
Q12H	Q2, Q8, Q11B, Q11C, Q12E, Q12I, Q12J, Q12K, Q12M	9	0.346154
Q12I	Q11B,Q12A, Q12B, Q12C, Q12D, Q12G, Q12H, Q12J, Q12K, Q12L, Q12M	11	0.423077
Q12J	Q3, Q11B, Q12A, Q12D, Q12E, Q12G, Q12H, Q12I, Q12K, Q12L	10	0.384615
Q12K	Q1, Q11B, Q12G, Q12H, Q12I, Q12J, Q12L	7	0.269231
Q12L	Q1, Q6, Q7, Q11A, Q11B, Q12A, Q12B, Q12C, Q12D, Q12E, Q12F, Q12G, Q12I, Q12J, Q12K, Q12M	16	0.615385
Q12M	Q6, Q7, Q10, Q12A, Q12B, Q12C, Q12D, Q12E, Q12F, Q12G, Q12H, Q12I, Q12L	13	0.50000

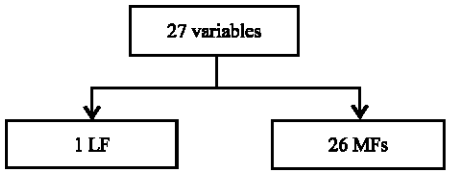

Fig. 3: Variables division as prediction components

Q12D: Longer donation experience
Q12E: Overcome donation fear
Q12F: Appreciate donation benefit
Q12G: Health self-awareness and save another people
Q12H: Utilize blood donation incentives
Q12I: Tend to donate for family or acquaintances
Q12J: Donation motivation from leaders, notable speakers or charismatic figures
Q12K: Donation motivation by celebrities
Q12L: Donating as religious purpose
Q12M: Information announcement medium such as social media

For training, validating and testing on Step 3, prediction part, dataset were distributed randomly into 70, 15 and 15, respectively as described in Fig. 4. Performance on step 4 will be measured based on training, validating and testing performance from cross validation for 30 times dataset has distributed (Han and Kamber, 2006a-c).

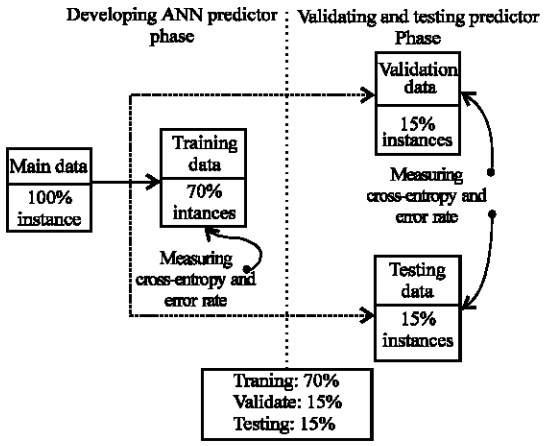

Fig. 4: Dataset distribution

RESULTS AND DISCUSSION

Results of this study have been extracted from Step 2 in Table 2, Step 3 in Table 3 and Step 4 in Table 4 and 5. From Table 2 based on Step 2, highest association ratio is Q12L, donating as religious purpose and lowest one is Q9, fear for blood donation. Step 3 has shown performance through Table 3 from training, validation and testing for assigned LFs.

Highest and lowest performance for all phases have displayed in Table 4 as mentioned in Step 4. Highest performance for training and testing phases are Q12M,

Table 2: Step 3 of FA: average performance of each LF

Assigned LFs	Training CE	Validation CE	Testing CE
Q1	0.1305733	0.1464281	0.1430556
Q2	0.0635397	0.0692312	0.0697423
Q6	0.0874779	0.0963441	0.0966334
Q7	0.0701690	0.0745562	0.0729696
Q9	0.0668982	0.0752321	0.0723663
Q10	0.0729556	0.0813416	0.0809836
Q12A	0.0626258	0.0739334	0.0729415
Q12B	0.0685503	0.0759847	0.0749056
Q12C	0.0601143	0.0692755	0.0680200
Q12D	0.0745969	0.0855252	0.0866120
Q12E	0.0754097	0.0830377	0.0870221
Q12F	0.0729231	0.0815346	0.0826203
Q12G	0.0928991	0.1027297	0.1036560
Q12H	0.0931159	0.1004646	0.1001920
Q12I	0.0615983	0.0676900	0.0709676
Q12J	0.0742589	0.0775098	0.0792379
Q12K	0.0591854	0.0652065	0.0655629
Q12L	0.0686051	0.0819233	0.0782593
Q12M	0.0575995	0.0653847	0.0640224

Table 3: Step 4 F A: ranking based on performance and simplifying LFs by associated MFs (highlighted box, grey for maxcFs, black for mincFs, cross boxes at bottom are black or both cutters)

Ranking	Training CE	Validating CE	Testing CE
1	Q12M	Q12K	Q12M
2	Q12K	Q12M	Q12K
3	Q12C	Q12I	Q12C
4	Q12I	Q2	Q2
5	Q12A	Q12C	Q12I
6	Q2	Q12A	Q9
7	Q9	Q7	Q12A
8	Q12B	Q9	Q7
9	Q12L	Q12B	Q12B
10	Q7	Q12J	Q12L
11	Q12F	Q10	Q12J
12	Q10	Q12F	Q10
13	Q12J	Q12L	Q12F
14	Q12D	Q12E	Q12D
15	Q12E	Q12D	Q12E
16	Q6	Q6	Q6
17	Q12G	Q12H	Q12H
18	Q12H	Q12G	Q12G
19	Q1	Q1	Q1

Table 4: Step 4 FA: scoring maxFs and minFs based on ranking, starting from 0 from highest rank and lower for next ranks

Training CE		Validating CE		Testing CE	
MaxcF	MincF	MaxcF	MincF	MaxcF	MincF
Q12L [0]	Q9 [0]	Q7 [0]	Q9 [0]	Q7 [0]	Q9 [0]
Q7 [-1]	Q10 [-1]	Q12L [-1]	Q10 [-1]	Q12L [-1]	Q10 [-1]
Q12G [-2]	Q12E [-2]	Q12G [-2]	Q12E [-2]	Q12G [-2]	Q12E[-2]
Q1 [-3]	Q12H [-3]	Q1 [-3]	Q12H [-3]	Q1 [-3]	Q12H[-3]
	Q1 [-4]		Q1 [-4]		Q1 [-4]

Table 5: Step 5 FA: finalizing maxFs and minFs from total ranking scores for each chosen variable (inside square brackets], highest score is top ranking

MaxcF	MincF
Q7 [-1]	Q9 [0]
Q12L [-2]	Q10 [-3]
Q12G [-6]	Q12E [-6]
Q1 [-9]	Q12H[-9]
	Q1 [-12]

information medium such as social media but validating phase is Q12K, celebrities. Lowest performance for all phases is Q1, gender.

Table 5 has mentioned about scoring all maxcFs and mincFs for each phases. For Step 5, scores for each variable Vs are counted as shown in Eq. 3 and have inserted into Table 5:

$$\text{Total Vs}_{\text{maxcFor min cF}} = \text{Vs}_{\text{maxcFor min cF Training}} + \text{Vs}_{\text{maxcFor min cF validating}} + \text{Vs}_{\text{maxcFor min cF testing}} \quad (3)$$

From Table 5, all phases have agreed that Q7, multiple donations per year has highest total score as maxcF or main important factor and Q9, donation fear has highest total score as mincF or least important factor. Lowest score for maxcF and mincF variables are definitely similar because of based on association.

Therefore, current last ranks hold unimportant information. However, second last ranks variable for maxcF and mincF are real last ranks. Last rank for variable maxcF is Q12G, health and save people and for variable mincF is Q12H, blood donor's incentives.

CONCLUSION

This study used correlation technique to select blood donor's preferences. However, it can be extended to different techniques or algorithms of prediction. Furthermore, besides than survey and questionnaires dataset, this method may applicable on non-related variables of other dataset too. Suggestion from these results can assist to time of data mining and decision making for uncertain situation for too many variables involvement and for different dataset too. In future, more studies on LF and MFs order and selection problem can get more attention to solve many similar problems from different areas.

ACKNOWLEDGEMENT

This research was funded by Centre of Research and Innovation Management (CRIM) and Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka.

REFERENCES

Boer, P.D.E., S. Mannor and R.Y. Rubinstein, 2005. A tutorial on the cross-entropy method. Ann. Oper. Res., 134: 19-67.

Bollen, K.A., P.P. Biemer, A.F. Karr, S. Tueller and M.E. Berzofsky, 2016. Are survey weights needed? A review of diagnostic tests in regression analysis. Annu. Rev. Stat. Appl., 3: 372-395.

Gravetter, F.J. and L.B. Wallnau, 2016a. Hypothesis Testing. In: Statistics for the Behavioral Sciences, Frederick, J.G. and L.B. Wallnau (Eds.). Cencage Learning, Boston, Massachusetts, pp: 223-226.

Gravetter, F.J. and L.B. Wallnau, 2016b. The Pearson Correlation. In: Statistics for the Behavioral Sciences, Frederick, J.G. and L.B. Wallnau (Eds.). Cencage Learning, Boston, Massachusetts, pp: 489-490.

Guyon, I. and A. Elisseeff, 2003. An introduction to variable and feature selection. J. Machine Learn. Res., 3: 1157-1182.

Han, J. and M. Kamber, 2006a. Data Mining Functionalities: What Kinds of Patterns can be Mined?. In: Data Mining: Concepts and Techniques, Cerra, D. (Ed.). Morgan Kaufmann Publisher, San Francisco, California, pp: 1-23.

Han, J. and M. Kamber, 2006b. From Association Mining to Correlation Analysis. In: Data Mining: Concepts and Techniques, Cerra, D. (Ed.). Morgan Kaufmann Publisher, San Francisco, California, pp: 262-264.

Han, J. and M. Kamber, 2006c. Model-Based Clustering Methods. In: Data Mining: Concepts and Techniques, Cerra, D. (Ed.). Morgan Kaufmann Publisher, San Francisco, California, pp: 433-433.

Jiang, S.Y. and L.X. Wang, 2016. Efficient feature selection based on correlation measure between continuous and discrete features. Inf. Process. Lett., 116: 203-215.

Lei, Y. and H. Liu, 2004. Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res., 5: 1205-1224.

Stijare, J.R. and D.T. Kumbhalkar, 2015. Review article alternatives to human blood transfusion-reality or dream?. Vidarbha J. Intern. Med., 18: 32-39.