Research Journal of Recent Sciences ISSN 2277-2502
Vol. **3(3),** 72-77, March **(2014)** *Res.J.Recent Sci.*

A Model in Recommender Systems for Disease Diagnosis Using Combined Method

Anaram Yaghoobi Notash[1], Behrouz Minaei[2] and Afshin Salajegheh[1]
[1]Islamic Azad University south Tehran Branch, Tehran, IRAN
[2]Iran University of Science and Technology, Tehran, IRAN

Available online at: www.isca.in, www.isca.me
Received 17th July 2013, revised 7th August 2013, accepted 6th October 2013

Abstract

Diagnosis of dieses is always of the concerns for physicians. Since wrong diagnosis of diseases especially in diseases leading to surgery would have unpleasant consequences, it was attempted to offer a model of data mining models so that physicians are aided in diagnosis of diseases. To this end, considering some disease have very similar symptoms and there is the highest probability of wrong diagnosis by the physician about them, data mining is used for an appropriate solution. Hence, 6 diseases were selected with are treated by surgery and have similar symptoms in diagnosis. After medical data collection in 550 patients and purification of 50 cases, the number of patients was reduced to 500. Then data were divided into 5 groups by reviewing medical literature and recognizing important of attributes in symptoms of diseases. This division was based on the fact that a group of data has different impact on the designed model compared to the other group. This database was implemented in Clementine software in categorization, clustering and partitioning methods and the best method was evaluated. Then a combined method (including partitioning and categorization and also categorization and clustering and partitioning) was developed as the final model. The final model in both combined methods offers the best diagnosis for aid of physician with evaluation percentage 96.99.

Keyword: Recommender systems, categorization algorithm, clustering algorithm, disease, partitioning.

Introduction

Proper diagnosis in different diseases in medical sciences including surgery is one of the major concerns of specialized physicians. Despite of advances in various sciences both in medical sciences and computer and extraction of laws related to diseases, yet physicians perform surgeries on the patients with wrong diagnosis probability concerning disease with high similarity in terms of diagnostic symptoms and they find proper diagnosis within the surgery. Data mining is a common technique which is applied in various fields including in medical sciences as well as in diagnosis of diseases. Data mining starts its work with the intention of aiding the physician so that it orders data extracted from patients' records and derives some rules for achieving the goal using related algorithms[1-5]. By investigating 6 surgery-related diseases with high error probability by physicians, it was attempted to extract some rules so that proper disease is diagnosed using its optimal performance and the physician is aided to have the best result with lowest error rate for the new patient by pursuing extracted tree process. Disease included appendicitis, complicated ovarian cysts, ectopic pregnancy¸ ureteral stone, perforation of duodenal ulcer¸ acute Cholecystitis.

Following extraction of related data, it was concluded that extracted models can be made by elimination of some data which do not help model progress. To this end, the models are developed in 5 groups of selected data. These groups include: i. All data, ii. Selected data regardless of initial symptoms, iii. Selected data regardless of the whole tests, iv. Selected data regardless of routine tests such as blood and urine tests, v. Data related to just tests, age and gender.

Description: Statistical population which was used for testing designed model included emergency and non-emergency patients in surgery unit in mentioned diseases in Sina and Atieh Hospitals during 2010-2012.

In collecting related data, data for 550 patients were entered into respective database which was reduced to 500 patients after purification. In addition, considering initial data extracted about initial symptoms in these diseases, 84 attributes were included in database. It was reduced to 34 useful attributes following purification. Both reductions were due to lack of data written by physician in the patient's record as well as wrong diagnosis and/or lack of test paper. Finally a collection of 500 patients with 34 useful, proper and perfect attributes included the final statistical population.

Data Preparation: In order to reduce the number of attributes and size of decision making tree, data preprocessing and selection of features is used so that better rules are obtained for developing the tree. Thus preprocessing was run on 84 attributes within following steps and 34 attributes were selected. i. Initial selection step is fields in selection of the most important data which were filtered in model creation step, and elimination of data which were entered by continuous values in one attribute and by discrete values in the other attribute with the same content[6-8]. In this step, also data which didn't provide useful information for the diagnosis were eliminated. For example, fields which were not mentioned by the physician in 80% of the records or fields with fixed values in all diseases and fields including personal information such as record number and fields which were used for pathologic diagnosis. Following entering disease features for 550 patients, a limited number of cases (50 cases) with negative pathologic results were eliminated and 30 cases which their pathologic results were present in 6 diseases, data were kept in database by updating disease results.

Data Test Using Partitioning Method

Another test which can be used for diagnosis of model accuracy is partitioning model. In this method, a percentage of data are used for data training and the remaining part is used for test. It is known as Test and Train method[3].

Partitioning and Categorization Method

In this work it is attempted to extract the best combined algorithm for better diagnosis, thus in this part it was decided to use 5 top algorithms for obtaining better results. 5 algorithms included C5 categorization, Bayesian algorithm, SVM (Sub Vector Machine) algorithm, neural networks and logistic regression model. As it was found the best method is when all data are used for categorization. Thus, for using partitioning method, 25% of data were defined for test and 75% for data training. Then 5 categorizations methods were used. For example, the model designed in figure 1 shows the model in SVM algorithm.

Output of Partitioning and Categorization Combined Methods

For showing the output as decision tree, combined method with C5 algorithm was used. Output 1 in Apendix indicates tree algorithm. Figure 2 indicates major variables in formation of the tree.

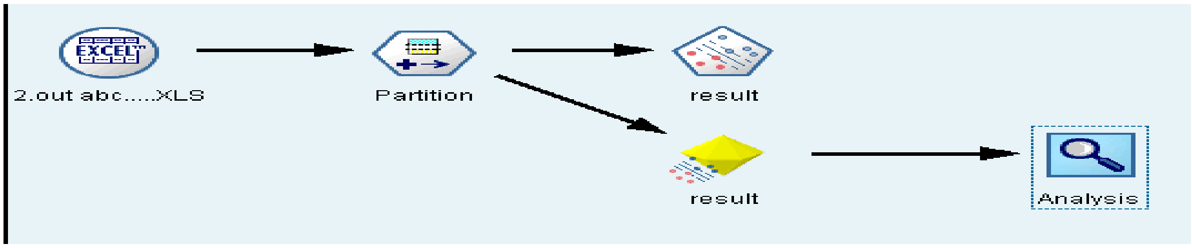


Figure-1
Designed model in partitioning and SVM combined method

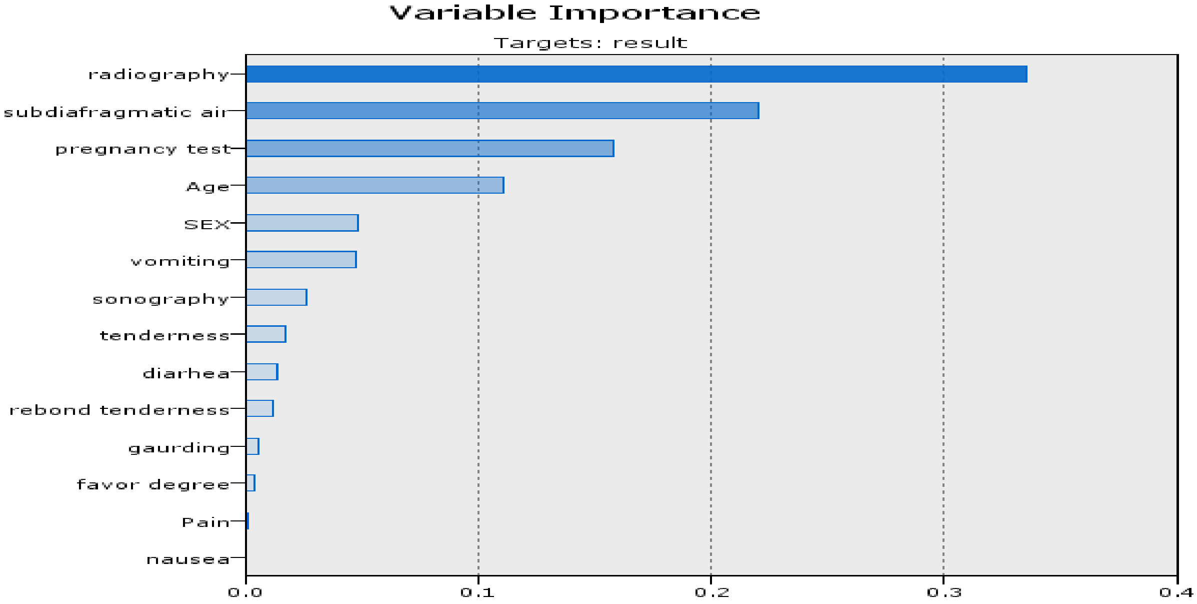


Figure-2
Important variables after partitioning in categorization C5 model

Evaluation of Partitioning and Categorization Model

Models are compared in table 1. As it can be observed, the best method is SVM. Also interference matrix and its evaluation are shown in table 2.

Table-1
Comparison of 5 algorithms of categorization and partitioning

Model	Evaluation result
C5	93.98
Logistic regression	91.73
SVM	96.99
Bayesian algorithm	89.72
Neural network	87.97

As it can be seen in table 1, the best method is partitioning combined method with SVM algorithm with 96.99% evaluation. By interference tables it can be concluded that in the related algorithm, what number of appendicitis (for example) were correctly diagnosed and what number of other dieses were (wrongly) diagnosed.

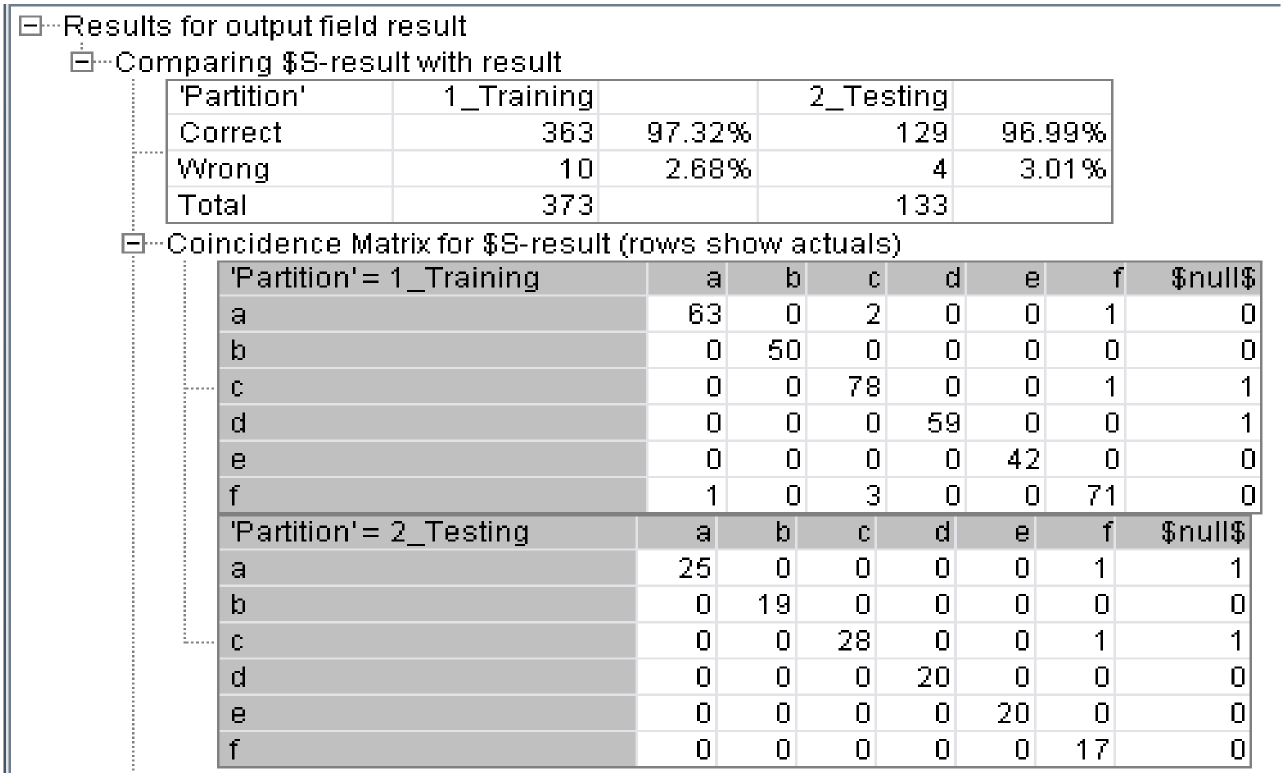
Designed Model in Partitioning, Categorization and Clustering Method: In this model, 3 methods including partitioning, categorization and clustering were combined. Partitioning method used 75% training data and 25% test data, categorization method used 5 selective algorithms and clustering method used K-means algorithm. K-means algorithm was selected because it had the best evaluation in previous models. Categorization algorithms include C5 categorization algorithms, Bayesian algorithm, SVM (Sub Vector Machine) algorithm, neural networks and logistic regression model. Designed model in figure 3 shows combined algorithms with k-means clustering and SVM algorithm.

Output of Partitioning, Categorization and Clustering Model: Final tree output in combined method with neural network algorithm is shown in Output 2. This output shows the main attributes in related algorithm.

Output 3 is part of the output in combined method with C5 algorithm. It shows the final tree.

Evaluation of Partitioning, Categorization and Clustering Model: Model evaluation with interference matrix in combined methods of partitioning algorithm and k-means with SVM categorization algorithm is shown in table 3. Table 4 shows comparison of 5 models.

Table-2
Interference matrix and evaluation of combined method with SVM



Results for output field result

Comparing $S-result with result

'Partition'	1_Training		2_Testing	
Correct	363	97.32%	129	96.99%
Wrong	10	2.68%	4	3.01%
Total	373		133	

Coincidence Matrix for $S-result (rows show actuals)

'Partition' = 1_Training	a	b	c	d	e	f	$null$
a	63	0	2	0	0	1	0
b	0	50	0	0	0	0	0
c	0	0	78	0	0	1	1
d	0	0	0	59	0	0	1
e	0	0	0	0	42	0	0
f	1	0	3	0	0	71	0

'Partition' = 2_Testing	a	b	c	d	e	f	$null$
a	25	0	0	0	0	1	1
b	0	19	0	0	0	0	0
c	0	0	28	0	0	1	1
d	0	0	0	20	0	0	0
e	0	0	0	0	20	0	0
f	0	0	0	0	0	17	0

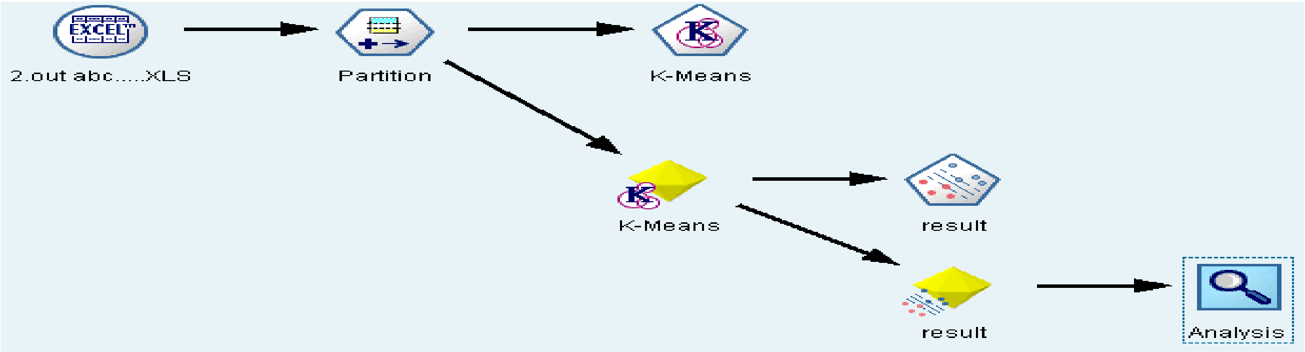
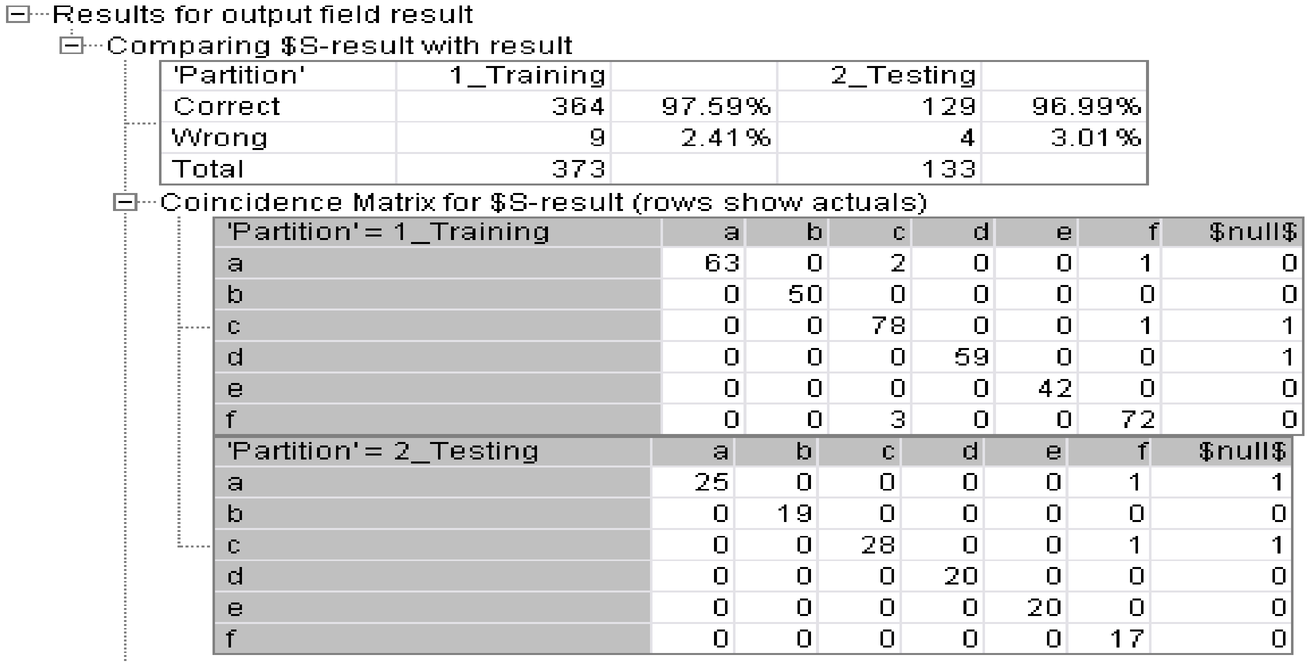


Figure-3
Designed model in partitioning, clustering, and SVM categorization methods

Table-3
Evaluation of combined algorithm method with partitioning and K-means with algorithm SVM



Results for output field result

Comparing $S-result with result

'Partition'	1_Training		2_Testing	
Correct	364	97.59%	129	96.99%
Wrong	9	2.41%	4	3.01%
Total	373		133	

Coincidence Matrix for $S-result (rows show actuals)

'Partition' = 1_Training	a	b	c	d	e	f	$null$
a	63	0	2	0	0	1	0
b	0	50	0	0	0	0	0
c	0	0	78	0	0	1	1
d	0	0	0	59	0	0	1
e	0	0	0	0	42	0	0
f	0	0	3	0	0	72	0

'Partition' = 2_Testing	a	b	c	d	e	f	$null$
a	25	0	0	0	0	1	1
b	0	19	0	0	0	0	0
c	0	0	28	0	0	1	1
d	0	0	0	20	0	0	0
e	0	0	0	0	20	0	0
f	0	0	0	0	0	17	0

Table-4
Comparison of 5 models

Model	Evaluation
Neural Network	84.96
C5	93.98
Logistic regression	89.47
SVM	96.99
Bayesian algorithm	89.47

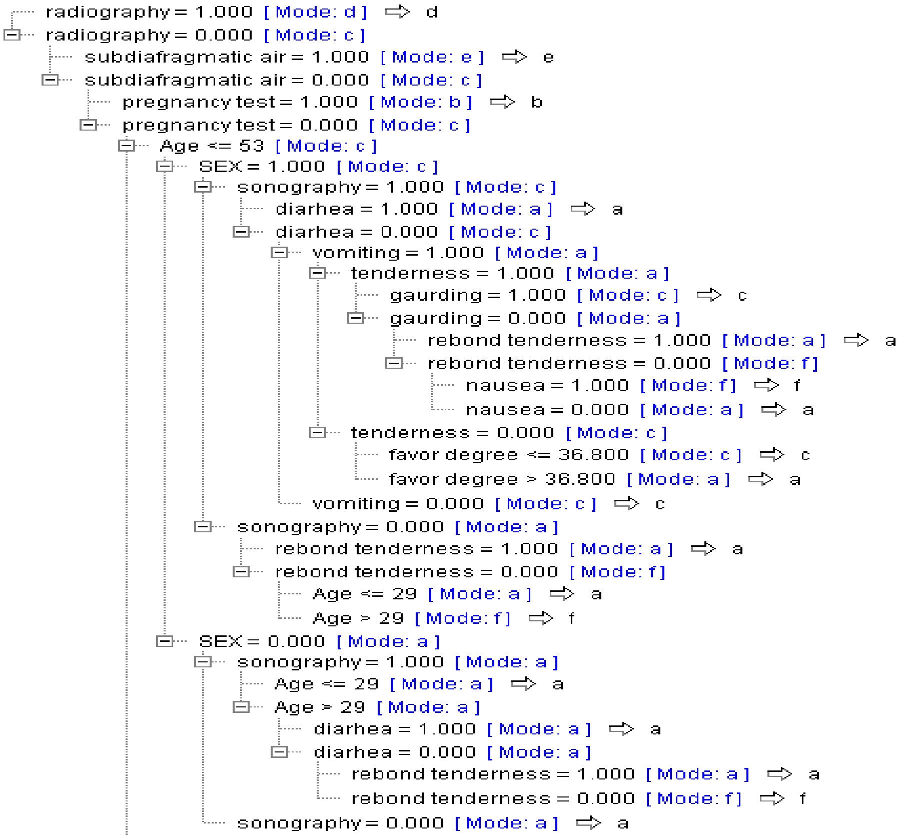
Table 5 is a comparison table which indicates two combined methods of partitioning and categorization with partitioning, categorization and clustering.

Table-5
Comparison of two combined methods

Model	Partitioning, categorization	Partitioning, Categorization, clustering
Neural network	87.97	84.96
C5	93.98	93.98
Logistic regression	91.73	89.47
SVM	96.99	96.99
Bayesian algorithm	89.72	89.47

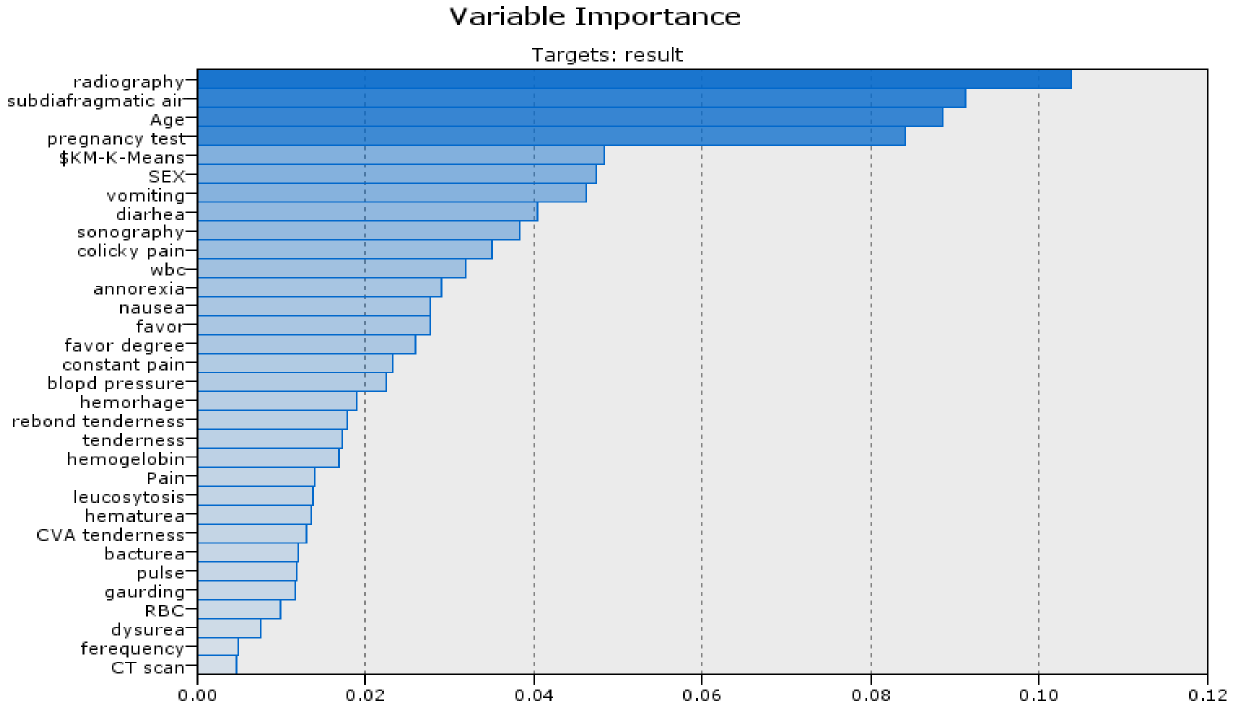
Conclusion

In this paper, a combined method (including partitioning and categorization and also categorization and clustering and partitioning) was developed as the final model. As it can be observed, the best evaluation is in the method which is derived from combined method of partitioning and clustering and SVM categorization and it was selected as the final model.

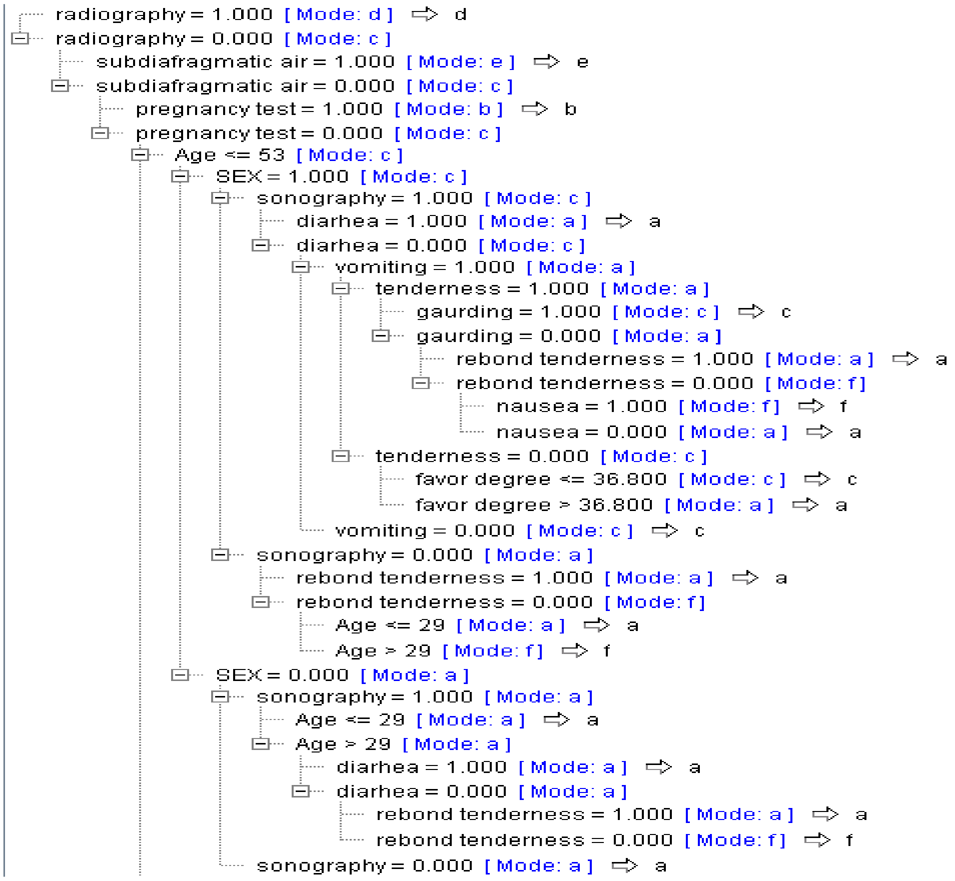


```
radiography = 1.000 [ Mode: d ] ⇨ d
radiography = 0.000 [ Mode: c ]
    subdiafragmatic air = 1.000 [ Mode: e ] ⇨ e
    subdiafragmatic air = 0.000 [ Mode: c ]
        pregnancy test = 1.000 [ Mode: b ] ⇨ b
        pregnancy test = 0.000 [ Mode: c ]
            Age <= 53 [ Mode: c ]
                SEX = 1.000 [ Mode: c ]
                    sonography = 1.000 [ Mode: c ]
                        diarhea = 1.000 [ Mode: a ] ⇨ a
                        diarhea = 0.000 [ Mode: c ]
                            vomiting = 1.000 [ Mode: a ]
                                tenderness = 1.000 [ Mode: a ]
                                    gaurding = 1.000 [ Mode: c ] ⇨ c
                                    gaurding = 0.000 [ Mode: a ]
                                        rebond tenderness = 1.000 [ Mode: a ] ⇨ a
                                        rebond tenderness = 0.000 [ Mode: f ]
                                            nausea = 1.000 [ Mode: f ] ⇨ f
                                            nausea = 0.000 [ Mode: a ] ⇨ a
                                tenderness = 0.000 [ Mode: c ]
                                    favor degree <= 36.800 [ Mode: c ] ⇨ c
                                    favor degree > 36.800 [ Mode: a ] ⇨ a
                            vomiting = 0.000 [ Mode: c ] ⇨ c
                    sonography = 0.000 [ Mode: a ]
                        rebond tenderness = 1.000 [ Mode: a ] ⇨ a
                        rebond tenderness = 0.000 [ Mode: f ]
                            Age <= 29 [ Mode: a ] ⇨ a
                            Age > 29 [ Mode: f ] ⇨ f
                SEX = 0.000 [ Mode: a ]
                    sonography = 1.000 [ Mode: a ]
                        Age <= 29 [ Mode: a ] ⇨ a
                        Age > 29 [ Mode: a ]
                            diarhea = 1.000 [ Mode: a ] ⇨ a
                            diarhea = 0.000 [ Mode: a ]
                                rebond tenderness = 1.000 [ Mode: a ] ⇨ a
                                rebond tenderness = 0.000 [ Mode: f ] ⇨ f
                    sonography = 0.000 [ Mode: a ] ⇨ a
```

Appendix 1
Tree algorithm



Appendix 2
The main attributes in related algorithm



```
radiography = 1.000 [ Mode: d ]  ⇨ d
radiography = 0.000 [ Mode: c ]
    subdiafragmatic air = 1.000 [ Mode: e ]  ⇨ e
    subdiafragmatic air = 0.000 [ Mode: c ]
        pregnancy test = 1.000 [ Mode: b ]  ⇨ b
        pregnancy test = 0.000 [ Mode: c ]
            Age <= 53 [ Mode: c ]
                SEX = 1.000 [ Mode: c ]
                    sonography = 1.000 [ Mode: c ]
                        diarhea = 1.000 [ Mode: a ]  ⇨ a
                        diarhea = 0.000 [ Mode: c ]
                            vomiting = 1.000 [ Mode: a ]
                                tenderness = 1.000 [ Mode: a ]
                                    gaurding = 1.000 [ Mode: c ]  ⇨ c
                                    gaurding = 0.000 [ Mode: a ]
                                        rebond tenderness = 1.000 [ Mode: a ]  ⇨ a
                                        rebond tenderness = 0.000 [ Mode: f ]
                                            nausea = 1.000 [ Mode: f ]  ⇨ f
                                            nausea = 0.000 [ Mode: a ]  ⇨ a
                                tenderness = 0.000 [ Mode: c ]
                                    favor degree <= 36.800 [ Mode: c ]  ⇨ c
                                    favor degree > 36.800 [ Mode: a ]  ⇨ a
                            vomiting = 0.000 [ Mode: c ]  ⇨ c
                    sonography = 0.000 [ Mode: a ]
                        rebond tenderness = 1.000 [ Mode: a ]  ⇨ a
                        rebond tenderness = 0.000 [ Mode: f ]
                            Age <= 29 [ Mode: a ]  ⇨ a
                            Age > 29 [ Mode: f ]  ⇨ f
                SEX = 0.000 [ Mode: a ]
                    sonography = 1.000 [ Mode: a ]
                        Age <= 29 [ Mode: a ]  ⇨ a
                        Age > 29 [ Mode: a ]
                            diarhea = 1.000 [ Mode: a ]  ⇨ a
                            diarhea = 0.000 [ Mode: a ]
                                rebond tenderness = 1.000 [ Mode: a ]  ⇨ a
                                rebond tenderness = 0.000 [ Mode: f ]  ⇨ f
                    sonography = 0.000 [ Mode: a ]  ⇨ a
```

Appendix 3
The final tree

References

1. Jannach D., Zanker M., Felfernig A. and Friedrich G., Recommender systems: an introduction, Cambridge University Press **(2010)**

2. Schafer J.B., Frankowski D., Herlocker J. and Sen S., Collaborative filtering recommender systems, In *The adaptive web* (pp. 291-324), Springer Berlin Heidelberg, **(2007)**

3. Chen A.Y.A. and McLeod D., Collaborative filtering for information recommendation systems, *Encyclopedia of Data Warehousing and Mining, Idea Group,* **(2005)**

4. Su X. and Taghi M. Kh., A survey of collaborative filtering techniques, *Advances in artificial intelligence,* **4(1),** 1-19 **(2009)**

5. Schafer J.B., The application of data-mining to recommender systems, *Encyclopedia of data warehousing and mining*, **1**, 44-48 **(2006)**

6. Schafer J.B., Frankowski D., Herlocker J. and Sen S., Collaborative filtering recommender systems, In *The adaptive web,* 291-324, Springer Berlin Heidelberg **(2007)**

7. Herlocker J.L., Konstan J.A., Terveen L.G. and Riedl J.T., Evaluating collaborative filtering recommender systems, *ACM Transactions on Information Systems (TOIS)*, **22(1),** 5-53 **(2004)**

8. Walker A., Recker M.M. Lawless K. and Wiley D., Collaborative information filtering: A review and an educational application, *International Journal of Artificial Intelligence in Education*, **14(1),** 3-28 **(2004)**