Research Article

Open Access CrossMark

Application of K-means technique in data mining to cluster hemodialysis patients

Abstract

Hemodialysis patients with end stage renal disease are often hospitalized due to infection, cardiovascular syndromes, or cancer. Through interviews with the doctors at a well-known hospital for kidney disease, a list of important variables including the clinical lab test results, the demographic and the socio-economic information was developed. After finalizing the list with the doctors, the needed data of 50 patients was provided anonymously so that real data can be used for this research. The review of literature reveals that no other research considered all of these measures used in this work simultaneously. In the next stage, the data was processed and prepared to be used as the input data for the implementation of data mining techniques. The clustering of patients via the K-Means technique is explained here. The results obtained from the data mining model in this research can be useful for the decision makers and doctors. Implementing the model seems promising and analyzing the results proved this point and delivered interesting grouping of patients. Also, it can be claimed that this research is a good prologue for future researches.

Keywords: hemodialysis, clustering, data mining, forecasting, hospitalization, health care engineering

Volume 2 Issue 2 - 2017

Reza Ghodsi,[1,2] Shiva Bagheri Marani,[3] Abbas Keramati[4]

[1]Central Connecticut State University, USA
[2]Department of Industrial Engineering, University of Tehran, Iran
[3]Central Connecticut State University, USA
[4]Department of Industrial Engineering, University of Tehran, Iran

Correspondence: Reza Ghodsi, Associate Professor, Central Connecticut State University, USA, Email ghodsi@ccsu.edu

Received: October 28, 2017 | **Published:** March 31, 2017

Abbreviations: DSS, decision support system; ESRD, end stage renal disease

Introduction

Currently, there is no treatment for patients at end stage renal disease (ESRD) other than kidney transplant or dialysis. Patients who go three times a week to hospital for the dialysis of their blood face enormous problems. One of these problems is the hospitalization of these patients due to infection, cardiovascular syndromes, and other disorders like cancer. To reduce the aftermath of hemodialysis, to increase the comfort and life time of these patients, and also to lessen the additional costs for patients and the health care system, it is necessary to do prevention. Hence, precautions necessary to decrease the probability of hospitalization due to the above disorders must be instructed by doctors to the patient and all clinical staff. If a tool, as a decision support system (DSS), can be provided to physicians such that they know their patients more, cluster them in appropriate groups, and forecast the probability of their hospitalization and the cause of hospitalization then they will be able to instruct better preventive measures to the patients and their caregivers which is a great help to the patient and the health care system of any country.

The main purpose of this research is helping the dialysis patients with the end stage renal disease (ESRD). At this stage, the kidney functions below 15cc/min which means that the kidney filters less than 15cc of blood within one minute. These patients have to undergo hemodialysis 3 times a week at a hospital or clinic. Each time they spent about 4 hours connected to the dialysis machine. These patients face many challenges and problems that might affect their life severely. Pain, limitations in social life, travel and work, frequent visits to clinics and hospitals and thus more exposure to contagious diseases, depression and isolation from society, huge costs are few of these problems. One other problem is frequent hospitalization because these patients might experience infection, cardiovascular syndromes, and other disorders like cancer. The frequent hospitalization and such complications not only is life threatening, but also have huge impact on the economics of both the patient and the health care system. It hurts the patient the most but also has a considerable negative impact on the health care system consuming resources.

To avoid frequent hospitalization, many remedies and actions might be possible to consider. Each of these actions can be from different science and knowledge perspectives. From the perspective of medicine, research can be done to invent better remedies and procedures. The other perspective could be through the invention of better devices and technology for dialysis by biomedical engineers such as better AV fistula. Another perspective can be through health care engineering by creating a decision support system (DSS) that physicians can use as a help in their decision about what instruction to provide to the patients and caregivers. The goal of this research is to use data mining techniques to help create such a decision support tool for medical doctors dealing with ESRD patients in future. The need for such tool is justified if one observes the abundance of parameters and variables which a doctor must consider to make a decision and provide proper instructions. Hence, as the first step of this researches all of the related and important variables had to be specified. Four groups of variables are proposed and considered as follows.

Demographic variables

Age, gender, birth place, race, and nationality are considered.

Laboratory variables

Table 1 illustrates the list of the different lab variables that has been considered based on the interviews done with medical doctors specialized in kidney disease in Iran. However, the choice of lab variables could vary based on common practice in different countries. Whether other lab variables should be considered or not is beyond the expertise of the authors and this can be debated among medical experts to finalize a complete list. It should also be noted that the normal range provided is based on the interview with specialist in Iran as well. This range might vary based on each country and needs to be tuned if this work is to be referred elsewhere. Table 2 shows the frequency of the different lab test. This table is also based on the interviews done with medical doctors specialized in kidney disease in Iran. Thus, it could vary based on common practice in different countries.

Table 1 Laboratory Variables

Lab variables	Unit	Normal range
WBC	Qty/microliter	4.5 - 9.0
RBC	Qty/microliter	3.9 - 5.8
Hb	g/dl	12.0 - 15.8
Hct	g/dl	38.8-46.4
FBS	mg/dl	70 - 115
BUN	%	7_21
cr	mg/dl	0.6-1.3
ca	mg/dl	8.6-10.3
P	mg/dl	2.0-5.0
AlkP	Iu/l	64-306
Na	meq/l	135-145
k	meq/l	3.5-5
Fer	ng/mg	14-165
iPTH	pg/ml	10_65
Iron	mg/dl	35-168
Chol	mg/dl	Disirable<200 Borderline=200-240 High>240
TG	mg/dl	Disirable<200 Borderline=200-400 High>400
TIBC		250-450
BilT		0.1-1.2
BilD		<0.2
HDL		>35
LDH		<160
CRP		<10
HBsAb		Positive/Negative
HBsAg		Positive/Negative
HCVAb		Positive/Negative
HIVAB		Positive/Negative

Table 2 The periods for some of the laboratory variables

Periodical lab test	Variables to be tested
Monthly	Bun-Creatinine-K-Na-Ca -P- Hb
Quarterly	Albumin-FBS-Cholesterol-Triglyceride
Semi-annually	HBS Ag – Ferritin – Iron – TiBC – PTH
Annually	HCV Ab – HBS Ab

Economic variables

Income (or) family income, number of children.

Social variables

Job position or occupation, marriage status and education.

It should be noted that hemodialysis test data is time series. To determine the status of the end stage renal disease for any patient, the change of the lab result data over time has to be considered. However, since the goal of this research is not to determine the status of the disease, this work uses the average values for clinical lab test results only. The other useful data for each patient is the number of previous hospitalization and its cause. The general practice in such health care facilities is usually such that most of these data about each patient is already recorded and kept. What is the other use of such data is beyond the purpose of this research. The goal here is to use this data and the hospitalization history of the patients and the data mining techniques to inference new knowledge. The abundance of variables and the large number of patients makes it very difficult, if not impossible, for the human to draw any sort of conclusion. Thus, the use of data mining here is very reasonable and a valid tool. Computers have given us the ability to process huge database and to write codes able to take advantage of the huge historical data. Artificial intelligence and heuristic algorithms, artificial neural networks, and data mining algorithms are a few approaches to name here. Liao and Chu[1] present a decade review from 2000 to 2011 of data mining techniques and applications.

The objective in this research is to use the data to cluster the patients. Data mining contains different algorithms for clustering and classification of the data. The reason that clustering methods are considered in this work but not any of the classification methods is that the classes (groups) of patients are not known in advance. Due to the huge variety of patients and the abundance of variables, it is not easy for any expert to group the patients into certain clusters. There are lots of clustering methods such as K-Means, ANN (Artificial Neural Networks), DBScan, Cobweb, expectation maximization, hierarchical clustering, and density based cluster. In this work however, only K-Means is used due to simplicity and experimenting with other techniques is left for future research. For the medical doctors and the clinical staff a tool that suggests that the individual patient has a possible high risk in hospitalization and due to what cause (infection, cardiovascular syndromes, and cancer) is useful. The suggestion will be not certain and it does not need to be. It is only to increase the level of precaution for each patient based on the data of that patient and the trend of that data. It is rather an alerting system. This can act as a decision support system (DSS) tool and will not advise or prescribe anything. The remedies and how to treat each individual patient is still based on the physician's decision. This DSS will only point out to the doctor that the patient can be clustered into what group and that the patient might have a future higher risk in which type of hospitalization. The user (medical doctors) should be notified that this is an uncertain forecast drawn from the database just to increase the level of alertness.

There are other similar works reported in the literature. Montani et al.[2] propose a case-base retrieval to support the treatment of ESRD patients. Kusiak et al.[3] predict survival time for ESRD patients using data mining. Hu et al.[4] provide predictive factors for failure of ESRD in earthquake. Hurst etal.[5] discuss predictors and associated outcomes of ESRD patients. Noia et al.[6] present an ESRD predictor based on artificial neural network. Modi et al.[7] report a population-based study of ESRD patients in India. Mullins et al.[8] use data mining and clinical data from 667000 patients to gain some insight. Martin etal.[9] Propose a self-organizing map as a new way to screen the satisfaction of ESRD patients. Hoxworth et al.[10] discuss infections associated in dialysis centers. DMan etal.[11] Provide a novel strategy for reducing infections in dialysis. Bellazzi et al.[12] do a temporal data mining for the quality assessment of hemodialysis services. Bellazzi et al.[13] implement an automated system for monitoring adherence to hemodialysis treatment. Sacchi et al.[14] Perform data mining with temporal abstraction. Titapiccolo et al.[15] use artificial intelligence models to evaluate the cardiovascular risk in hemodialysis patients. Yeh etal[16] predict hospitalization of hemodialysis patients using data mining techniques. They use fewer variables than discussed in the research at hand. Also, their study is based in Taiwan and the normal range of the laboratory variables differ from the current common practice referred in the research at hand. There are also papers regarding the clustering of the symptoms for hemodialysis patients such as the works by Jablonski.[17–19] Several symptoms like shortness of breath and joint pain are discussed but the implication is not to use these data for the prediction of the future of the patient regarding hospitalization due to infection, cardiovascular syndromes, or cancer. In general, to the best of the authors' knowledge none of the available research considers all of the four groups of variables presented in this work.

Proposed models

As mentioned before the objective of this research is to use a data mining technique to cluster the patients. Using these clusters, a physician can recognize the cluster of any new patient. The K-Means model developed for clustering of patients will be discussed in this paper. Prior to discussion of the model, the provided data has to be processed and prepared which is a common and necessary practice for any data mining technique.

Data processing

The demographic, social, economic and laboratory data have to be sorted and prepared for use. Samples of data prior to and after processing are illustrated in the (Tables 3-5). The Tables 3 (demographic and socio-economic data) and Table 4 (laboratory data) are data prior to processing. Next, data is processed and changed to the format shown in the Table 5. Table 7 in the appendix is the processed data for all 50 patients. The lab specimens are collected and tested on the days the patient come for hemodialysis. For example, as shown in Table 4, patients 1 to 5 have test results in November and December 2012. Hence, each row of the results is coded with the month and year. Hb-11-12 and Hb-12-12 are Hb results for November and December 2012 respectively. If the test result of a lab variable for a patient was not available (e.g, the FBS-12-12 for all patients 1 to 5 and FBS-11-12 for patients 3 and 5) it is considered as zero in the tables and then in the coding, the median of the normal range is used. Using the median from all patient data neutralizes the effect of the missing data. It is necessary to consider a nominal value for some of the collected data as per Table 5 in order to input to the data mining model. Table 5 shows a processed data sample. The 5 patient shown in (Table 3) (Table 4) are hospitalized only due to infection or cardiac problems. But in Table 5, the patient with identification 15 (ID=15) is hospitalized 5 times due to cancer.

Table 3 Sample demographic and socio-economic data of 5 patients prior to processing

	Patient	**Patient**	**Patient**	**Patient**	**Patient**
	One	Two	Three	Four	Five
Gender	F	M	F	M	M
Date of birth	1956	1949	1965	1985	1993
Birth place	Tehran	Tehran	Tehran	Tehran	Kabul
Nationality	Iranian	Iranian	Iranian	Iranian	Afghan
Education	High School	Master	Elementary	High School	Illiterate
Marriage status	Married	Married	Divorcee	Single	Single
Number of children	4	3	3	0	0
Vocation	Own Business	Retired	Housekeeper	Own Business	Housekeeper
High hypertension	N	N	Y	Y	Y
Diabetes	N	Y	N	N	N
Biopsy	N	N	N	N	N
Urologic	N	N	Y	N	N
Transplant	N	N	N	N	N
Hemodialysis	N	Y	N	N	N
Peritoneal	N	N	N	N	N
Heart disease	N	N	N	N	Y
Infection disease	N	Y	N	Y	N
Other disease/surgery	Y	Y	N	Y	Y
Smoking habit	N	Y	N	Y	N
Addiction	N	N	N	N	N
Cause of renal problem	Unknown	Diabetes	High Hypertension	High Hypertension /Polycystic	Cold
Vascular access	Fistula	Fistula	Fistula	Fistula	Fistula
Hospitalization	Y	N	Y	Y	Y
Cause of hospitalization (1=Infection, 2=Cardiac)	1	1	2	2	1
No. of hospitalization	3	2	4	1	3
Income	Medium	Strong	Weak	Strong	Weak

Citation: Ghodsi R, Marani SB, Keramati A. Application of K-means technique in data mining to cluster hemodialysis patients. *Int Rob Auto J.* 2017;2(2):34–39. DOI: 10.15406/iratj.2017.02.00013

Table 4 Sample lab data of 5 patients prior to processing

Patients	One	Two	Three	Four	Five
Hb-12-12 (Dec 2012)	11.6	13.2	10.3	12.3	10.8
HCT-12-12	35.8	42.6	31.6	37.4	32.9
FBS-12-12					
PreBUN-12-12	87	55	59	77	97
Cr-12-12	13	11.5	7.5	10.8	15.1
Ca-12-12	7.2	9.4	8.8	8.6	9
P-12-12	7.4	5.2	4.5	7.1	7
AlkP-12-12	290	280	232	345	185
Na-12-12	142	144	145	141	139
K-12-12	4.9	4.2	4.4	6.7	6.7
Hb-11-12 (Nov 2012)	11.5	11.8	9.6	11	11.6
HCT-11-12	35.9	38.4	29.9	33.6	36
FBS-11-12	77	241	Not available	227	Not available
PreBUN-11-12	92	83	77	81	111
PostBUN-11-12	29	27	19	23	27
Cr-11-12	12.2	11.3	11.8	9	14.8
Ca-11-12	8.4	9.3	8.9	9.2	9.7
P-11-12	6.9	5.8	5.5	6	6.8
Na-11-12	143	143	143	141	138
K-11-12	6.5	6	6.1	5.1	6
PostK-11-12	3.8	4.9	4	3.9	4
Fer-11-12	277	42	401	368	188
Iron-11-12	500	29	58	73	95
TIBC-11-12		296	268	270	286
iPTH-11-12	212	473	318	662	24

Table 5 For each of the patients, the assigned cluster is reported as per the following

Cluster 1:The possibility of hospitalization based on infection might be higher.
Cluster 2:The possibility of hospitalization based on cardiovascular syndrome might be higher.
Cluster 3:The possibility of hospitalization based on cancer or other disorders might be higher.

K-means algorithm

i. Arbitrarily choose k objects from dataset as the initial cluster centers;

ii. Loop

iii. Assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

iv. Update the cluster means, that is, calculate the mean value of the objects for each cluster;

v. Return to step 2 until no change.

The K-Means algorithm divides the patients into K number of clusters. First an arbitrary k is chosen. The average of the data in each cluster is considered the center of each cluster. Next K vectors X from the data (here it means the data of K patients) are considered as K cluster centers. The remainder of X vectors for the other patients is assigned to one of the clusters based on the least minimum Euclidian distance between them and the center of the K clusters. After all members of each cluster are specified, a new center is found for each cluster by taking the average of the components of the members. For example, the first component of the vectors for each member belonging to that cluster are summed and divided over the number of the members of the cluster. The member which has the closest Euclidian distance from this vector of averages is then chosen as a new center. Once, all new centers are specified, the other vectors are assigned as before. This loop will continue till the cluster centers do not change anymore. It means that one condition stops the loop which is that all cluster centers stay the same in the same iteration.

In order to select a K value at the start of the algorithm, there are several possible policies. Sometimes clustering is performed based on a specific purpose. For example, if the goal of the research at hand

Citation: Ghodsi R, Marani SB, Keramati A. Application of K-means technique in data mining to cluster hemodialysis patients. *Int Rob Auto J.* 2017;2(2):34‒39. DOI: 10.15406/iratj.2017.02.00013

is to have a guess on the cause of hospitalization for the patient of concern, then the number three is a good assumption for K. This is because three types of hospitalization are considered that are due to infection, cardiovascular disorder and cancer. This is the chosen policy in this research. Other policies can also be considered that can be considered in future research. The chosen policy is implemented in this research as there is a clear purpose to divide the patients into the three mentioned clusters. The K-Means algorithm and the considered policy are coded using MATLAB. The results are reported and discussed next.

Experimentation

To do some experimentation using read data, after finalizing the list of required data, a questionnaire was developed. Using the information of the filled questionnaire by the patients or their caregiver and the laboratory test data, a combined data file was prepared. Then, the data was processed and prepared and saved as a Data.xlsx file to be used by the program. Since, at this stage of the research, the goal is to validate the concept only and not to develop the final tool for the physicians, the provided data of 50 anonymous patients should suffice. Table 5, given in the appendix, shows the data file based on the information of the 50 patients. The total number of columns is 33. The first column (ID) is the identification number for each patient (for this work 1 to 50). The other 32 columns are the decision variables (The X vector has 32 components) for a physician to consider deciding and making suggestions. This shows the challenge that a physician faces to consider in mind all of these variables simultaneously. Thus, the model proposed in this research can be useful as an extra alerting system. Of course, the physicians do not rely on this clustering completely and solely. They still have to use all data (specially the lab data) to prescribe medicine and provide instructions based on their expertise. The proposed clustering will only alert the physician of a possible type of hospitalization as an uncertain forecast. The result of the code using the data of Table 5 is shown in Table 6. For each of the patients, the assigned cluster is reported as per the following:

Table 6 The result of clustering 50 patients

Patient ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Cluster #	1	2	1	2	2	1	3	2	1	2	1	1	2	2	2	2	2	2	1	3	2	2	3	2	3
Patient ID	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Cluster #	2	2	3	2	2	3	1	1	1	1	2	2	1	1	1	1	2	1	2	2	3	2	2	2	1

Conclusion

This research proposed a clustering model to forecast the possible type (cause) of future hospitalization of ESRD patients. The three considered causes are infection, cardiovascular syndrome and cancer. Thus, the clustering model incorporates cluster 1 (Patients with infections as high risk for admission), cluster 2 (Those with cardiovascular risk) and cluster 3 (those with cancer risk). Table 6 shows that the proposed model is able to assign each patient to a distinct cluster. Now this information must be verified over time which due to time and limited access was not possible within this research. Other data mining techniques like artificial neural networks (ANN) can be used if available historical data exists to verify whether the clustering has helped or not. This is left for future research. The proposed clustering model only suggests the possibility of hospitalization in future for these patients that must be considered by the medical staff. Then they can try to prevent it by adding extra effort in monitoring each patient according to that cluster. This information should not be a reason to lessen the routine treatment of the patients in any way and is solely to add some extra efforts.

The proposed model is strictly to be used by the physician. It is important to note that this clustering is not replacing the expertise of the medical doctors or the clinical staff. The physician has to prescribe and provide instructions based on his or her expertise or whatever his profession guidelines are and not based on the outcome of this clustering. As any other forecasting model, this clustering model is an uncertain alerting system and is only suggesting to the physician to pay an added attention that the patient might be hospitalized based on a certain type although other type of hospitalization are still possible. The program is considering 32 the patient's variables and suggesting a cluster. The doctor, however, still has to check the validity of this suggestion. The clustering is like an added brainstorming. It is kind of asking the doctor to see what he/she thinks about this possibility of hospitalization for the patient. The medical staffs are sometimes overwhelmed with lots of information especially on a busy day. It is helpful to have this clustering model as it invites the doctor to think twice revisiting the data of 32 variables. The list of variables used in this research can be considered in other similar researches. Furthermore, other policies for the choice of K in the proposed K-Means algorithm can be explored. Although, similar research can be carried out in many countries but the variables to consider and their normal range need to be revisited and fine-tuned based on the instructions of the medical doctors in that country. Other type of data-mining techniques such as decision tree or artificial neural networks can also be tested. Experimental design and analysis can be extended to compare this method to other clustering methods.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

References

1. Liao S, Chu P, Hsiao P. Data mining techniques and applications-A decade review from 2000 to 2011. *Expert Syst Appl*. 2012;39:11303–11311.

2. Montani S, Portinale L, Leonardi G, et al. Case-based retrieval to support the treatment of end stage renal failure patients. *Artif Intell Med*. 2006;37(1):31–42.

3. Kusiak A, Dixon B, Shah S. Predicting survival time for kidney dialysis patients: a data mining approach. *Comput Biol Med*. 2005;35:311–327.

4. Hu Z, Zeng X, Fu P, et al. Predictive factors for acute renal failure in crush injuries in the Sichuan earthquake. *Injury*. 2012;43(5):613–618.

5. Hurst FP, Jindal R M, Fletcher J J, et al. Incidence, predictors and associated outcomes of renal cell carcinoma in long-term dialysis patients. *Urology*. 2011;77(6):1271–1276.

6. Noia TD, Ostuni V C, Pesce F, et al. An end stage kidney disease predictor based on an artificial neural networks ensemble *Expert Syst Appl*. 2013;40(11):4438–4445.

7. Modi GK, Jha V. The incidence of end-stage renal disease in India: a population-based study. *Kidney Int*. 2006;70(12):2131–2133.

8. Mullins IM, Siadaty MS, Lyman J, et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Comput Biol Med*. 2006;36(12):1351–1377.

9. Martín Guerrero J D, Marcelli D, Soria-Olivas E, et al. Self-Organising Maps: A new way to screen the level of satisfaction of dialysis patients. *Expert Syst Appl*. 2012;39:8793–8798.

10. Hoxworth T, Reese SM. Healthcare-associated infections in colorado dialysis treatment centers. *Am J Infect Control*. 2011;39(5):E130–131.

11. Lindberg C, Downham G, Buscell P, et al. Embracing collaboration: A novel strategy for reducing bloodstream infections in outpatient hemodialysis centers. *Am J Infect Control*. 2013;41(6):513–519.

12. Bellazzi R, Larizza C, Magni P, et al. Temporal data mining for the quality assessment of hemodialysis services. *Artif Intell Med*. 2005;34(1):25–39.

13. Bellazzi R, Sacchi L, Caffi E, et al. Implementation of an automated system for monitoring adherence to hemodialysis treatment: A report of seven years of experience. *Int J Med Inform*. 2012;81(5):320–331.

14. Sacchi L, Larizza C, Combi C, et al. Data mining with temporal abstractions: learning rules from time series”. *Data Min Knowl Discovery*. 2007;15(2):217–247.

15. Titapiccolo J I, Ferrario M, Cerutti S, et al. Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients. *Expert Syst Appl*. 2012;40:4679–4686.

16. Yeh J, Wu T, Tsao Ch. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decis Support Syst*. 2011;50:439–448.

17. Jablonski A. The multidimensional characteristics of symptoms reported by patients on hemodialysis. *Nephrol Nurs J*. 2007;34(1):29–38.

18. Thong MSY, van Dijk S, Noordzij M, et al. Symptom clusters in incident dialysis patients: Associations with clinical variables and quality of life. *Nephrol Dial Transplant*. 2009;24(1):225–230.

19. Yu IC, Huang JY, Tsai YF. Symptom cluster among hemodialysis patients in Taiwan. *Appl Nurs Res*. 2012;25(3):190–196.

Citation: Ghodsi R, Marani SB, Keramati A. Application of K-means technique in data mining to cluster hemodialysis patients. *Int Rob Auto J.* 2017;2(2):34–39. DOI: 10.15406/iratj.2017.02.00013