Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Mining Health Care Sequences using Weighted Associative Classifier

By Sunita Soni & Dr. O. P. Vyas

Bhilai Institute of Technology, India

Abstract- This paper proposes the general framework for mining sequences from health care database. The database is a relational model consisting of set of temporal records of individual patient consisting of basic information of the patient ie Patient_ID, age, gender etc. the second part is a series of sequences representing the set of treatment given to the patient during regular visit to the doctor and the third part is class label. Similarity search of sequences is performed to convert the database of sequences, to the database of items, so that apriori algorithm can be applied. Weighted association rule mining has been performed to find the frequent sequence of treatment provided to the patient. Classification association rules (CAR) having positive class label as consequent, represents the frequent sequence of treatment given to the patient for successful treatment. With the experimental results, author feels confident in declaring that the framework is feasible in the medical domain.

Keywords: *sequence mining, weighted associative classifier, weighted support, weighted confidence, prediction*

GJCST-C Classification: *H.2.8*

Strictly as per the compliance and regulations of:

Mining Health Care Sequences using Weighted Associative Classifier

Sunita Soni [α] & Dr. O. P. Vyas [σ]

Abstract- This paper proposes the general framework for mining sequences from health care database. The database is a relational model consisting of set of temporal records of individual patient consisting of basic information of the patient ie Patient_ID, age, gender etc. the second part is a series of sequences representing the set of treatment given to the patient during regular visit to the doctor and the third part is class label. Similarity search of sequences is performed to convert the database of sequences, to the database of items, so that apriori algorithm can be applied. Weighted association rule mining has been performed to find the frequent sequence of treatment provided to the patient. Classification association rules (CAR) having positive class label as consequent, represents the frequent sequence of treatment given to the patient for successful treatment. With the experimental results, author feels confident in declaring that the framework is feasible in the medical domain.

Keywords: sequence mining, weighted associative classifier, weighted support, weighted confidence, prediction.

I. Introduction

Time plays a crucial role as patient's care as well as data collection and decision-making activities are performed over time. It is therefore often mandatory to deal with the temporal aspects by deriving useful summaries of the patient's behavior, including physiological signals or measurement time series, and adapting the decisions to the accumulated data and information. The goal of predictive data mining is to derive models that can use patient's historical information to exploit hidden information which will ultimately improve clinical Decision-making [1].

Diagnosis is related to the classification of patients into disease classes or subclasses on the basis of patients' data gathered from regular visit gathered time series. There are a growing number of papers that exploit data mining approaches for clinical prediction purposes. In a clinical context, predictions may support diagnostic, therapeutic, or monitoring tasks. Therapeutic prediction means the choice of the most suitable treatment for the patient.

Time series or temporal sequences; appear naturally in a variety of different domains, from engineering to scientific research, finance and medicine. In healthcare, temporal sequences are a reality for decades; with data originated by complex data acquisition systems like ECG's or even with simple ones like measuring the patient temperature or treatments effectiveness . In the last years, with the development of medical informatics, the amount of data has increased considerably, and more than ever, the need to react in real-time to any change in the patient behavior is crucial. In general, applications that deal with temporal sequences serve mainly to support diagnosis and to predict future behaviors [2].

The ultimate goal of temporal data mining is to discover hidden relations between sequences and subsequences of events. Just to mention few, following prediction can be performed using patient's historical temporal data.

1. Prediction for drug treatment planning or for the prognosis of surgical interventions.
2. Predictions in clinical monitoring are crucial in several contexts, such as in intensive care units (ICUs), which needs continue updating on the basis of the monitoring data.
3. Prediction may range from simply predicting the risk of disease based on the age factor or lifestyle for whole population, to the forecast of consequences of taking a particular drug or treatment for long time. For example drug taken for hypertension for a long time may affect the functioning of kidney.
4. Prediction of chance of any disease or casualty on neonatal based on different symptoms and other information like weight, systematic growth mother's blood group etc.
5. Predicting the risk of chronic disease as a result of another disease. For example diabetic patient having hypertension are more prone to Cardio Vascular Disease.

In this paper we have proposed a general framework to mine prediction rule for the accurate treatment of disease which will ultimately lead to cure of disease. The framework uses the historical data of the patient consisting of sequence of treatment given at regular interval. Further each sequence element may be various pathological test or advanced test results, regular observations like blood sugar, blood pressure etc. and medicines and other treatment recommended to the patient for that time period. The database consists of set of sequence of treatment given to the patient and a class label that defines whether the patient is cured or not.

Author α : Department of computer applications, Bhilai Institute of Technology. e-mail: sunitasoni74@gmail.com

Author σ : Professor, Indian Institute of Information Technology, Allahabad, India.

The major steps of the work proposed are

1. *Representation and modeling:* In this step, sequences of the temporal data are transformed into a suitable form. Every unique sequence is assigned a numeric symbol using step 2 and ultimately the database is converted to form suitable to perform apriori type algorithm.
2. *Similarity Measure:* This step defines the similarity measures between sequences. We are using Euclidian distance measure to find the unique sequences.
3. *Mining Operation:* In this step actual mining operation is performed to extract hidden information. In this framework we are extracting the set of frequent sequences (representing the treatment given to the patient) applied on the patient, which ultimately caused the patient to be cured. Association rule mining is used to find the association among the treatment given, with the given class label, the rules in this step are known as Class Association Rule (CAR).
4. *Prediction:* We use the high confidence CAR rules generated in step 3 to predict the sequence of treatment.

The proposed Framework for sequence mining is shown in figure1.

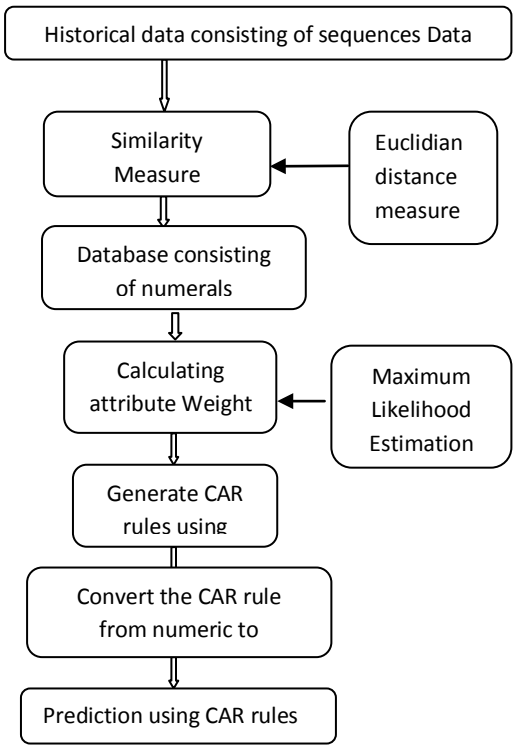

Fig 1 : Sequence Mining framework using WAC.

II. Related Work

a) Medical Prediction Using Temporal Data Mining

Temporal databases consist in databases containing time stamped information. A time stamp can be represented by a valid time, which denotes the time period in which the element information is true in the modeled real world, and/or by a transaction time, which is the time in which the information is stored in the database. Temporal data mining approaches provide the opportunity to address different tasks, such as data exploration, clustering, classification or prediction [3].

In [4] temporal data mining has been applied on the hepatitis temporal database collected at Chiba university hospital between 1982-2001. The database is large where each patient record consists of 983 tests represented as sequences of irregular timestamp points with different lengths. The work presents a temporal abstraction approach to mine knowledge from this hepatitis database.

In [5] visual data mining technique on temporal data is applied for the management of hemodialysis. The approach is based on the integration of 3D and 2D information visualization techniques which offers a set of interactive functionalities. The paper described the main features of IPBC (Interactive Parallel Bar Charts), a VDM system developed to interactively analyze collections of time-series, and showed its application to the real clinical context of hemodialysis.

In [6] temporal data mining techniques has been applied to extract information from temporal health records consisting of a time series of elderly diabetic patients' tests. The first step is to find pattern structures using structural-based search using wavelets. In the second step a value-based search over the discovered patterns using the statistical distribution of data values is performed. In the third step the results from the first two steps is combined to form a hybrid model. The feature of the hybrid model proposed is the expressive power of both wavelet analysis and the statistical distribution of the values. Global patterns have been identified successfully.

In [7] initially a framework is proposed for the definition of methods and tools for the assessment of the clinical performance of hemodialysis (HD) services, on the basis of the time series which has been automatically collected during hemodialysis sessions. For the implementation the method proposed is intelligent data analysis and temporal data mining techniques to gain insight and to discover knowledge on the causes of unsatisfactory clinical results.

In [8] a new kind of temporal association rule and the related extraction algorithm is proposed. An Apriori-like algorithm has been used to search for meaningful temporal relationships among the complex patterns of interest.

In [9] a new algorithm is presented to mining of Temporal Association Rules which has the main innovative feature of handling both events with a temporal duration and events represented by single time points. This new method has been applied to analyze the healthcare administrative data of diabetic patients.

The method is found to be useful to observe frequent health care temporal patterns in a population.

In [10] a general methodology for the mining of Temporal Association Rules on sequences of hybrid events is proposed. The experimental results show that the method can be a practically used for the evaluation of the care delivery flow for specific pathologies. In [11] the work done in[10] has been extended to focus on the care delivery flow of Diabetes Mellitus, and an algorithm is proposed for the extraction of temporal association rules on sequences of hybrid events. This work has been extended in [12] to show how the method can be used to highlight cases and conditions which lead to the highest pharmaceutical costs. Considering the perspective of a regional healthcare agency, this method could be properly exploited to assess the overall standards and quality of care, while lowering costs.

In [13] an efficient technique to match and retrieve the sequence of different lengths has been proposed. A number of the research works proposed earlier were concentrated on similarity matching and retrieval of sequences of the same length using Euclidean distance metric. In the matching process a mapping among non-matching elements is created to check for the unacceptable deviations among them. An indexing scheme is proposed for efficient retrieval of matching sequences, which is totally based on lengths and relative distances between sequences.

In [14] the analysis of sequential data streams to unearth any hidden regularities is discussed and also the applications of it in various field ranging from finance to manufacturing processes to bioinformatics is explained. The notions of sequential patterns or frequent episodes represent only the currently popular structures for patterns. The field of temporal data mining is relatively young hence new developments in the near future is yet to come. The paper discuss such several issues and others like what constitutes an interesting pattern in data, problem of defining data structures for interesting patterns, linking pattern discovery methods etc.

b) Association Rule Based Classifier

Given a set of cases with class labels as a training set, classification is to build a model (called classifier) to predict class label of future data objects. Associative classification is an integrated framework of association rule mining and classification. A special subset of association rules whose right-hand-side is restricted to the classification class attribute is used for classification. This subset of rules is referred as the class association rules (CAR). Extensive performance studies show that association based classification may have better accuracy in general [15], [16], [17]. The major advantages of new Predictive Model over the other models are-

- Fast training mechanism regardless of the size of the training set.
- Training sets with high dimensionality can be handled easy.
- Classification can be very fast with a compact set of rules.
- The classification model is easily understandable to humans (interpretability) well-organized, and easier to use model.
- Provides better accuracy than traditional decision tree classification algorithms.
- In medicine we are interested in creating understandable to human descriptions of medical concepts, or models. Associative classifiers are used for achieving this goal, since they can create a model in terms of intuitively transparent rule of the form $X \rightarrow Y$. On the other hand, unintuitive black box methods, like artificial neural networks, may be of less interest.

In section III we have discussed some basic definition for sequence mining. In section IV the different steps of sequence mining is discussed. In section V the algorithm weighted associative classifiers is discussed. In section VI conclusion and future work has been discussed.

III. Problem Definition

Definition1: Sequence Database: A sequence database D is a set of records D[0], D[1],...,D[n] where record D[i] represents the record of ith patient consists of ordered sequences, S(i,1), S(i,2), S(i,3), …S(i,j),…, where each sequence S(i,,j) is observed at time stamp tj , $1 \leq j \leq n$, n is positive integer. Si represents a sequence observed at time stamp ti. In database D the size of record may be varying because the number of visits for the complete treatment of one patient may be different from other patient.

Example: For patient 3 the number of sequence is i, whereas the number of sequence for patient 1, 2 and 4 is m.

Definition 2: Sequence: An ordered sequence si is set of elements ek, where $1 \leq k \leq l$,

i.e. Si = (e1, e2, e3…el)

Each element ek belongs to some domain representing quantitative or categorical or binary value corresponding to any preliminary test results like blood pressure, blood sugar, body mass index, or other pathological test results or medication recommendation based on the test result at time stamp t_i.

Definition 3: Sequence length: Length of sequence is defined as number of elements in the sequence.
length (S_i) = number of elements in S_i.

Definition 4: Sequence Structure: Structure of sequence is defined as the length of sequence and the elements

and their order in the sequence. The exact sequence length and structure of sequence will be based on the disease for which the training data is collected. A typical example of structure of sequence and sequence in case of heart patient may be-

Example: Structure of the sequence is (Blood_ pressure_upper, Blood_pressure_lower, Fasting_Blood _Sugar, BMI, test1, test2, Medicine1, Medicine2, Medicine3) and corresponding sequence is (190,50, 150, result_test1, result_test2, med1, med2, med3).

Definition 5: The sequence for one patient at different time stamp may be same or varying, also the sequence at same time stamp for the different patient may be same or varying. i.e.

1. Let Si is a sequence at ti and Sj is sequence at tj and S_i, $S_j \in D[i]$ then $S_i = S_j$ is possible. Let at time stamp ti , Si ∈ D[0] and Sj ∈ D[1]

then Si≠Sj or Si= Sj is possible.

The operator = and ≠ are discussed in Definition 6.

2. *Example:* patient 2 and 4 have given same treatment at same time stamp, also patient 2 has been given same treatment from time stamp t_1 to t_i.

Table 1 : Relational database D with set of temporal records

	Time Dimension→			t_1	…	t_i	…	t_m	
Patient Record→	P_Id	age	gender	S_1	…	S_i	…	S_m	class_label
	1	45	f	(190,50,150,result_test1, result_test2, med1)	…	(200,90, 150, result_test1, result_test2, med1, med2)	…	(200,90,150, result_test1, result_test2, med1, med2	Disease_ cured
	2	30	f	(200,90,150, result_test1, result_test2, med1, med2)	…	(200,90, 150, result_test1, result_test2, med1, med2)	…	(190,50,150,result_test 1, result_test2, med1, med2)	Disease_ Notcured
	3	55	m	(190,50,150,result_test1, result_test2, med1)	…	(200,90, 150, result_test1, result_test2, med1, med2)	NA	NA	Disease_ cured
	4	35	m	(200,90,150, result_test1, result_test2, med1, med2)	…	(200,90, 150, result_test1, result_test2, med1, med2)	…	(190,50,150,result_test 1, result_test2, med1, med2)	Disease_ Notcured

IV. Sequence Mining using Weighted Association Rule

a) Data Preparation

Data preparation process includes preparation of the data of interest to be used for mining and convert this data to the format suitable to perform apriori type algorithm. The database of the form shown in Table I have to be converted into the form as shown in Table IV.

i. Discretisation/ Normalisation

In the database firstly we perform Discretisation/ Normalisation for the non temporal attributes like age, gender etc. Discretisation is the process of converting the range of possible values associated with a continuous data item (e.g. a double precision number) into a number of sub-ranges each identified by a unique integer label; and converting all the values associated with instances of this data item to the corresponding integer labels. For example for attribute age the sub-range can be as shown in Table II

Table 2 : Discretisation Of Numeric Attribute

age	categorical value
20-30	1
31-40	2
41-50	3
51-60	4

Normalisation is the process of converting values associated with nominal data items so that they correspond to unique integer labels. TableIII shows normalization for attribute gender.

Table 3 : Normalization Of Attribute Gender

Gender	integer label
male	5
female	6

We use *DN (discretization/ normalisation) software Version 2* available at site http://cgi. Csc.liv.ac.uk/~frans/ KDD/ Software/ LUCS -KDD - DN/exmpleDnnotes. html to perform Discreti sation/ Normalisation process.

b) Similarity Search for Sequences using Euclidean Distance

This is an important step in this framework. As once the database has been converted from database of sequences to the database of items, the apriori algorithm can be applied to find the association among the items, and ultimately the CAR rules can be generated for prediction.

To assign a unique numeric label to every unique sequences corresponding to each patient, sequence comparison method is required. There can be number of methods to compare the similarity of sequence.

Many time series representations and distance measure techniques have been proposed for more than one decade. Some of these approaches work well for short time series data, but they fail to produce satisfactory results for long sequences. There are two kinds of similarities: shape-based similarity and structure-based similarity. Shape based similarity is suitable for short sequences only. For the two sequences Si and Sj, shape-based determines the similarity based on local comparisons.

The well-known distance measure in data mining is Euclidean distance, in which sequences are aligned in the point-to-point fashion, i.e. the ith point in sequence Si is matched with the ith point in sequence Sj. Euclidean distance works well in many cases. Dynamic Time Warping (DTW) is another distance measure technique that overcomes the limitation by determining the best alignment to produce the optimal distance. Euclidean distance is a special case of DTW, where no warping is allowed, the dips and peaks in the sequences are miss-aligned and therefore not matched. In DTW, the dips and peaks of sequences are aligned and it provides more robust distance measure than Euclidean distance, compensation to that DTW a lot more computationally intensive as discussed in [13].

To determine the similarity for long sequence a more appropriate is to measure their similarity based on higher-level structures. Several methods for structure or model-based similarities have been proposed.

In this paper we use Euclidean distance measure for similarity search, For matching sequences we would like to address the following points.

- The relative times that the corresponding samples are collected are almost same in both sequences. This means that the lengths of sequences should be close to each other to be matched.
- The elements of both sequences are taken from the lifetime of the experiment in a rather uniformly manner.
- In numeric sequences from medical domains, since the elements are real numbers obtained from various pathological tests with a limited precision, elements from different sequences should be matched based on proximity.
- In non-numeric sequences, matching is done based on equality of their domain.

Definition 6: Sequence Similarity: Consider two sequences S_1 and S_2 having length x and y respectively, and $e_1, e_2 \ldots e_x$ are matching $q_1, q_2, \ldots q_y$.

1. The sequences S1 and S2 matches each other if-
   i. Their length is same as
   ie $length(S_i) = length(S_j)$
   ii. Distance $(e_k, q_k) = 0$,for all values of k.

Also the distance between two elements ek and qk can be defined as follows.

- For numeric elements,
  distance $(e_k, q_k) = |e_k - q_k|$.
- Non-numeric sequences can be matched based on equality. In that case, the distance between any two elements is defined to be

$$\text{distance}(e_k, q_k) = \begin{cases} 0 & , \text{if } \text{dom}(e_k) = \text{dom}(q_k) \\ \text{positive number}, & \text{if } \text{dom}(e_k) \neq \text{dom}(q_k) \end{cases}$$

2. and $S_i \neq S_j$ if either condition i or ii is false.

c) Representation of Temporal Sequences

In order to perform the apriori like operation in the above dataset, we transform the original dataset consisting of sequences into the relational database consisting of numeric labels like 1, 2, 3…..etc, where each numeric label represents unique sequence. Sequences in one record are compared for their similarity and unique symbol is assigned to unique sequence.

In the database D consisting of m columns and n rows, we precede row wise from top to bottom and in each row we will precede from left to right. A unique numeric label num is assigned to the first sequence S(0,0) of first patient and maintains the processed sequence and num assigned to that sequence in arr as shown in Table V. Then we pick the next sequence S(I,j)and compare (using Euclidean distance) it with already processed sequences stored in arr. If the sequence matching is found then assign the same numeral to new sequence and no need to assign new numeric label. If the sequence is not present in the List

arr then assign a num++ to the sequence and store the sequence and num to arr. Comparing the sequence in the list will always starts from first entry in arr but it will not be a time consuming process as there will be finite number of sequence in the original database D. This way entire database D is preprocessed to convert the database D' as shown in TableIV. The algorithm is discussed in figure 3.

Table 4 : Transformed Database D'

Patient Record→	Time Dimension→			t_1	…	t_i	.. .	t_m	
	P_Id	age	gender	s_1	.. .	s_i	…	S_m	Class_ label
	101	2	5	7	.. .	8	…	8	10
	102	1	5	8	.. .	8	…	9	11
	103	4	6	7	.. .	8	0	0	10
	104	3	6	8	.. .	8	…	9	11

Table 5 : List Consisting of Sequences and Numeric Labels

Sequence	Numeric labels
$S_{(0,0)}$	7
$S_{(0,1)}$	8
$S_{(0,5)}$	9
-	-
-	-
$S_{(n,m)}$	30

```
Initialize num with available numeric value
Initialize k=0;
for each i=0 to n-1
{for  each j=0 to m-1
{ exists=false
for each l= 0 to k
{ if arr[l,0]=D[i,j]
exists=true
D[i,j]=arr[l,1] } }
If not exist then
{arr[k,0]=D[l,j]
arr[k,1]=num;
D[i,j]=num
k++ and num++} }
```

Figure 2 : Algorithm to convert D to D'

d) Assigning Weight to the Sequences using Maximum Likelihood Estimation.

The weighted concept is used to improve the performance in terms of accuracy and number of rules generating as mentioned in [18]. In this paper the weighed concept have been utilized to assign more weights to the sequence (pathological test and medication to the patient at particular time period) having much impact on treatment of patient. Attribute is assigned weight based on the domain. For example item in supermarket can be assigned weight based on the profit on per unit sale of an item. In medical domain, symptoms can be assigned weight based on their significance on prediction capability. Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. By using iterative technique, the maximum likelihood estimator is measured upon varying probability values of items in the training dataset.

e) Frequent Sequence Mining

The problem of sequence mining has now been converted to frequent itemset mining in the database D' where items are nothing but sequence represented by numeric labels. Hence the following section contains terms and basic concepts to define sequence weight, sequencesetweight, recordweight, weighted support and weighted confidence for weighted associative classifiers.

The transformed training dataset D' consists of n distinct set of records i.e. D'= $\{r_1, r_2, r_3 \ldots r_n\}$. Where each record is collection of varying number of labels (representing temporal sequence) and value of class label. Each record has unique identifier called PID.

Definition 7: Sequence weight It is same as Item weight in WARM[19]. In this work we have extended the definition for the sequences. Each sequence S_i is assigned weight wi, denoted by $w(S_i)$, where $0<w_i<=1$. Weight is used to illustrate the significance of the sequence. Attribute is assigned weight based on the domain. For example item in supermarket can be assigned weight based on the profit on per unit sale of an item. In medical domain, symptoms can be assigned weight based on their significance on prediction capability. Weight is calculated from training data using maximum likelihood estimation and denoted by wi. Table V shows the synthetic weight assigned to the sequences.

Table 5 : Synthetic Weight Assigned To The Sequences

S.No	Numeric labels	*Sequence*	Weight
1	*7*	*(190,50,150,result_test1, result_test2, med1)*	0.5
2	*8*	*(200,90, 150, result_test1, result_test2, med1, med2)*	0.6
3	*9*	*(190,50,150,result_test1, result_test2, med1, med2)*	0.8

Definition 8 : Sequence Set Weight: It is same as Itemset weight in WARM[19]. In this work we have extended the definition for the sequences set weight. Weight of sequence set X is denoted by W(X) and is calculated as the average of weights of enclosing attribute. And is given by

$$W(X) = \frac{\sum_{i=1}^{|X|} W(S_i)}{\text{Number of sequences in } X}$$

Definition 9 : Record Weight: The tuple weight or record weight can be defined as type of sequence set weight. It is average weight of sequences in the patient record. If the transactional table is having m number of sequence then Record Weight is denoted by W(rk) and given by

$$W(r_k) = \frac{\sum_{i=1}^{|r_k|} W(S_i)}{m}$$

Definition 10 : Weighted Support: In associative classification rule mining, the classification association rules R: X→Y is special case of association rule where Y is the class label. Weighted support (WSP) of rule X→Class_label, where X is non empty set of sequences, is fraction of weight of the record that contain above sequence set relative to the weight of all transactions. This can be given as

Here *m* is the total number of records.

Example: Let sequence Si= (190,50, 150,result_test1, result_test2, med1) and

Sj= (200,90, 150, result_test1, result_test2, med1,med2)

Consider a rule R: ((Si, Sj) → Class_label) then Weighted Support of R is calculated as:

$$WSP(X \rightarrow Class\_label\ ) = \frac{\sum_{i=1}^{|X|} W(r_i)}{\sum_{k=1}^{|n|} W(r_k)}$$

Definition 11 : A frequent sequence is a set of sequences whose support is greater or equal than a user-specified threshold called minimum weighted support (WMin_sup). Given a dataset and WMin_sup, the goal of sequence mining is to determine in the dataset all the frequent sequences set whose support are greater than or equal to WMin_sup.

Definition 12 : Weighted Confidence: Weighted Confidence (WC) of a rule X→Y where Y represents the Class label and can be defined as the ratio of Weighted Support of (X→Y) and the Weighted Support of (X).

$$WC(X \rightarrow Y) = \frac{WSP\ (X \rightarrow Y)}{WSP\ (X)}$$

f) Classification Association Rule Generation

After generating all frequent item sets CAR rules are filtered using frequent item set having one of the class labels. Frequent item sets that does not contain any of the class label has to be removed. To generate significant CAR rules the weighted confidence threshold is used. Using WAC algorithm the set of CAR rules are generated as shown in figure 4. Finally the numeric labels are replaced by corresponding sequence using arr database and frequent sequences are generated as shown in figure 5.

(5, 8, 10, 12) → Class_label1
(9, 13, 15, 17) → Class_label2
(12, 19, 22, 25) → Class_label2

Figure 3 : CAR Rules Consisting Of Numeric Labels

$(S_{(0,2}, S_{(1,8)}, S_{(3,9)}, S_{(5,6)}) \rightarrow Class\_label1$
$(S_{(91)}, S_{(8,2)}, S_{(91,5)}, S_{(61,7)}) \rightarrow Class\_label2$
$(S_{(81,2)}, S_{(121,9)}, S_{(32,2)}, S_{(202,5)}) \rightarrow Class\_label2$

Figure 4 : CAR rules consisting of sequences

ALGORITHM 1:**WAC**(D_{tr}, *WMS, WMin_conf, W,* D_{ts})
Input: 1 D_{tr}: Training data, 2 D_{ts}: Test Data, 3 W_s: Weighted support, *WC*: Weighted confidence,
Output: *Weighted Class Association Rule Base(WCARB)*
WCARB← **WeightedAssociationRuleMiner**(D_{tr}, W_s, *W, WMin_conf)*
WeightedAssociationRuleMiner (D_{tr}, *WMS, W, WMin_conf*)
Apply Apriori type algo. using weighted support to find frequent attribute sets *a* where $a=a_1,a_2...a \in D_{tr}$
for (all frequent itemset $a_i \in a$)
if a_i does not contain $c_i \in C$ (Set of Class labels)
remove a_i from *a*
else
generate Rule $R_i=(a_i - c_i) \rightarrow c_i$
if(Weightedconfidence(R_i)> *WMin_conf*)
Store R_i to *WCARB*
BuildPredictiveModel(*WCADB*, D_{ts})
Sort Rules R_i of *WCARB* w.r.t. their *WMin_conf* and store CAR rule in Rule_Base
For each record $r_i \in D_{ts}$ predict class label using Rule-Base
Find Accuracy of the system
If (accuracy> minimum threshold)
The Model is suitable for prediction

Figure 5 : The WAC Algorithm

V. Experiments and Results

We present the Temporal WAC results on real data collections of blood cancer disease.

a) Representation and Modeling

The data has been collected for 30 patient. The database consists of maximum 5 time stamp as shown in figure 7, two Class label with the values –cure and not cure.
Structure of sequence is-

{Cancerous cell% , Therapy, Medicine(s) }

Example of few sequences available in the dataset are-

- { 30% , Chemotherapy, Zofran, Busulphan, Kadian}
- {40%, Raditiontherapy, Aclarbicin, Azacitidine}
- {30%, Immunotherapy, Adriamycin IV, Elspar inj}
- {20%, Targeted terapy, Nilotinib, GastroMARK}
- {92%, StemCelltransplantation, Aclarbicin, Photofrin}

In the above sequence, first element is the percentage of cancerous cell, second element is therapy given to the patient and rest elements are medicines.

The Experiments have been performed step by step following the framework shown in figure1.

P_Id	Age	Gender	Time_slot1	Time_slot2	Time_slot3	Time_slot4	Time_slot5	Class_Label

Figure 6 : Schema of cancer dataset

b) Similarity Measure (Pre-Processing)

Euclidian distance measure is used to convert the database consisting of above sequence to database consisting of numeric labels.

Total 26 unique sequences have been identified and 27, 28 are assigned as the numeric labels for "cure" and "not cure" class labels respectively.

c) Mining Operation

The WAC algorithm shown in figure 4 is used to mine the database and CAR rule are generated for the different support value. With the CAR rules the accuracy is calculated using same training data and the result is shown in Table 5.

Table 6 : Car Rules Consisting Of Numeric Labels

S. No	Support Value	CAR rules	Confidence	Accuracy
1	0.15	16,20,21 →28	100%	66%
2	0.20	19,22→27 8,23→27 7, 16→28	100% 100% 80% 100%	80%
		16,20→28		
3	0.25	16,20→28	100%	66%
4	0.30	22→27 17→27 20→28	100%	90%

With the result shown in table 5 we conclude that accuracy is better in case of having CAR rules for all the class labels. The reason of less accuracy in case of single CAR rule may be the default class label we are

assigning during accuracy calculations. The Efficient CAR rules can be generated using enough training record. The purpose of this experiment is to show that Framework shown in figure1 is possible to implement and can generate useful result in medical domain can be used for the purpose stated in introduction section. The authors are confident enough that improved result will be obtained if the experiment were to be performed on real data with little or no modifications.

VI. Conclusion and Future Work

This work presents a new foundational approach to mine frequent sequence using weighted associative classifiers whose core idea is to assign weights to the attributes depending upon their importance in predicting the class labels. The proposed model can be used as an alternative, computerized decision aid to assist physicians to find the sequence of treatment that can be given to the patient. The author feels confident in declaring that the framework is feasible one in the medical domain.

References Références Referencias

1. Riccardo Bellazzi,Fulvia Ferrazzi, and Lucia Sacchi, Predictive data mining in clinical medicine: a focus on selected methods and applications, @2011 John Willey & Sons , Inc . Volume 00, Januar y / Februar y 2011.
2. Cláudia M. Antunes , and Arlindo L. Oliveira, Temporal Data Mining: an overview, Lecture Notes in Computer Science pp 1-15.
3. Stefano concaro, temporal emporal data mining for the analysis of healthcare data,.ph.d. Thesis (2006-2009).
4. Tu Bao Ho, Trong Dung Nguyen,Saori Kawasaki, Si Quang Le, Dung Duc Nguyen, Hideto Yokoi, Katsuhiko Takabayashi, Mining Hepatitis Data with Temporal Abstraction.
5. Luca Chittaro, Carlo Combi, Giampaolo Trapasso, Data Mining on Temporal Data:a Visual Approach and its Clinical Application to Hemodialysis, Journal of Visual Languages and Computing, vol. 14, no. 6, December 2003, pp. 591-620.
6. Weiqiang Lin, Mehmet A. Orgun, and Graham J. Williams, Mining Temporal Patterns from Health Care Data.
7. Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R. (2005) Temporal data mining for the quality assessment of hemodialysis services. Artificial Intelligence in Medicine 34:25–39.
8. Sacchi, L., Larizza, C., Combi, C., Bellazzi, R. (2007) Data mining with temporal abstractions: Learning rules from time series. Data Mining Knowledge Discovery 15, 217–247.
9. Concaro, S., Sacchi, L., Cerra, C., Fratino, P., Bellazzi, R. (2008) Temporal Data Mining for the Analysis of Administrative Healthcare Data. In Proceedings of IDAMAP 2008 Workshop, Washington, 75–80, http://labmedinfo.org/download/lmi503.pdf.
10. Stefano CONCARO , Lucia SACCHI , Carlo CERRA , Riccardo BELLAZZI, Mining Administrative and Clinical Diabetes Data with Temporal Association Rules, Medical Informatics in a United and Healthy Europe, IOS Press, 2009 European Federation for Medical Informatics pp574-578.
11. Stefano Concaro, MS, Lucia Sacchi, Carlo Cerra, Mario Stefanelli, Temporal Data Mining for the Assessment of the Costs Related to Diabetes Mellitus Pharmacological Treatment, AMIA 2009 Symposium Proceedings Page – 119-123.
12. Stefano Concaro, Lucia Sacchi, Carlo Cerra, Pietro Fratino, and Riccardo Bellazzi, Mining Healthcare Data with Temporal Association Rules: Improvements and Assessment for a Practical Use, AIME 2009, LNAI 5651, pp. 16–25, Springer-Verlag Berlin Heidelberg 2009.
13. Matching and Indexing Sequences of Different Lengths, Tolga Bozkaya Nasser Yazdani Meral ¨Ozsoyo ˇglu.
14. SRIVATSAN LAXMAN and P S SASTRY A survey of temporal data mining , Vol. 31, April 2006, pp. 173–198.
15. B. Liu, W. Hsu, and Y. Ma. “Integrating classification and association rule mining”, In KDD’98, New York, NY, Aug.1998.
16. W. Li, J. Han, and J. Pei. “CMAR: Accurate and efficient classification based on multiple class-association rules” In ICDM'01, pp. 369-376, San Jose, CA, Nov.2001.
17. Yin, X. & Han, J. “CPAR: Classification based on predictive association rule”, In Proceedings of the SIAM International Conference on Data Mining. San Francisco, CA: SIAM Press, pp. 369–376, 2003.
18. Sunita Soni , O.P.Vyas, Performance Evaluation of Weighted Associative Classifier in Health Care Data Mining and Building Fuzzy Weighted Associative Classifier, D. Nagamalai, E. Renault, and M. Dhanushkodi (Eds.): PDCTA 2011, CCIS 203, pp. 224–237, 2011.© Springer-Verlag Berlin Heidelberg 2011.
19. Feng Tao, Fionn Murtagh and Mohsen Farid. “Weighted Association Rule Mining using Weighted Support and Significance Framework”, In proceedings of the ninth ACM SIGKDD International conference on Knowledge Discovery and Data Mining 2003, pp:661-666.

This page is intentionally left blank