A Rule Discovery Algorithm Appropriate for Electrocardiograph Signals

S Konias, N Maglaveras

Laboratory of Medical Informatics, Medical School, Aristotle University of Thessaloniki, Greece

Abstract

In this paper the problem of discovering rules, which are associations among patterns in the same time series, is considered. A novel algorithm, namely Rule Discovery Algorithm (RDA), appropriate for periodic time series data, like Electrocardiograms (ECGs) can be considered, is proposed. The first phase of the algorithm aims to break the sequence (i.e. an ECG) into overlapping, reconfigured length subsequences according to the sampling frequency and the types of ECG abnormalities to be studied. The Pearson Correlation Coefficient was chosen as the categorization metric, which is independent of the base line shifts and the amplitude scales. At the following phase the categorized sequence was scanned, so the most efficient rules would be mined. The format of those rules is "IF A occurs THEN B occurs WITHIN time T", where A and B are categorized subsequences and T the time duration between A and B. RDA was evaluated on 60 congestive heart failure patients' ECGs from a home care monitoring database. The mined rules are complementary to the ECGs' plots allowing the physician to test various hypotheses and discover hidden knowledge.

1. Introduction

Data Mining (DM) is a new and interdisciplinary field merging ideas from statistics, artificial intelligence, and databases. Its main motivation is the phenomenal growth of data in all domains of human knowledge including medicine and the need of discovering valid and useful information from the collected data.

Mining association rules in demographic dataset is one of the most popular data mining technique. To our knowledge association rules have firstly proposed in [1] aiming at capturing significant dependences between items in transactional datasets, like market basket analysis and financial analysis. For instance, in market basket analysis, where the items are the things bought and itemsets are several items together an association rule could be the "Beer→chips (83%)", which states that 83% of the times someone buys Beer he or she also buys chips. Generally, an association rule is an expression of the form "X→Y", where X and Y are sets of items from a given transaction database.

The problem of mining association rules can be decomposed into two subproblems, firstly finding all frequent itemsets and secondly generating all rules with support (the most frequent) and confidence (the most efficient) higher than specified thresholds. However, many algorithms have suggested for extracting rules from transaction databases [2,3].

During the last years there has been a great research investment in data mining information seeking knowledge building in medicine [4-6]. In [4] an uncertainty rule algorithm, named Uncertainty Rule Generation (URG-2), is proposed appropriate for discovering rules with confidence higher that a reconfigured threshold. URG-2 mines rules from dynamic databases containing missing values, without the need of recovering the already existing itemsets from the beginning. In [5] an improvement of the aforementioned algorithm called Adaptive Uncertainty Rule Generation (AURG) is illustrated.

In this paper a novel algorithm, namely Rule Discovery Algorithm (RDA), appropriate for mining rules from periodic time series, like Electrocardiographs (ECGs), is proposed. The format of the extracted rules is "IF A occurs THEN B occurs WITHIN time T", where A and B are categorized subsequences and T the time duration between A and B.

The remaining part of the paper is organized as follows. In Section 2 the problem of time series categorization and its solutions are presented. Section 3 briefly reviews the rule format and the required metrics. In Section 4 the RDA algorithm is described, while in Section 5 the results of an experimental evaluation of our approach on ECGs coming from a home care database are provided. Finally Section 6 concludes with the possibility of using the mined rules further in order to fill missing parts in an ECG.

2. Time series categorization

To cluster a given time series, which can be considered as sequence, into subsequences distance notions are required. Suppose a sequence s={x_1, x_2, …, x_n} and a window's length w are given. Then all subsequences s_1, s_2, …, s_{n-w+1} length's w are obtained,

0276−6547/04 $20.00 © 2004 IEEE

where s_i ={xi, x_{i+1}, …, x_{i+w-1}} and be denoted as W(s)={s_i /i=1, 2, …,n-w+1}.

As a following task is to cluster the set of all subsequences W(s). The latter can be realized by trying the W(s) of length w as elements of R^w and use various of distance metrics. Euclidian distance is one of the most used in time series [7] and is defined as:

$$L_2(\overline{x}, \overline{y}) = \sqrt{\sum_i (x_i - y_i)^2} \qquad (1)$$

where $\overline{x} = (x_1, ..., x_w)$ and $\overline{y} = (y_1, ..., y_w)$.

Other alternative distance measures between time series can provide interesting results, as well. For instance, the general L_p metrics and the L_∞ metric defined by :

$$L_p(\overline{x}, \overline{y}) = \sqrt[p]{\sum_i (x_i - y_i)^2} \text{ for } p\geq 1 \qquad (2)$$

$$L_\infty = \max_i | x_i - y_i | \qquad (3)$$

However, for many applications one would like to see the shape of the subsequences as the main factor in distance determination [8]. It is frequent for two subsequences to have essentially the same shape although they differ in their amplitudes and baselines. One way of achieving this is by normalizing firstly the subsequences and then using one of the above metric on the normalized subsequence.

A simple and well known way to normalize a sequence is by shifting the time series by its mean and then scaling by its standard deviation.

$$\overline{x}' = \frac{\overline{x} - \mathrm{avg}(\overline{x})}{\mathrm{std}(\overline{x})} \qquad (4)$$

where $\mathrm{avg}(\overline{x})$ is the average of $\overline{x}$ and $\mathrm{std}(\overline{x})$ is the standard deviation for $\overline{x}$.

An alternative way to deal with the problem of the different baselines or amplitude scales among time series is the Pearson Correlation Coefficient similarity measure. In this paper the above measure has selected, which is widely used in statistical analysis, pattern recognition and image processing [9].

$$\mathrm{corr}(\overline{x}, \overline{y}) = \frac{\mathrm{avg}(\overline{x} * \overline{y}) - \mathrm{avg}(\overline{x})\mathrm{avg}(\overline{y})}{\mathrm{std}(\overline{x})\mathrm{std}(\overline{y})} \qquad (5)$$

where $\overline{x} * \overline{y}$ is the inner product between $\overline{x}$ and $\overline{y}$.

The Pearson Correlation Coefficient ranges from –1 to 1. A negative coefficient indicates a negative relation while a positive coefficient indicates a positive relation. Finally, a coefficient equal to zero indicates that the compared time series do not comprise any relation. In this paper the Pearson Correlation Coefficient is chosen in order to find similar parts among an ECG.

3. Rule format and metrics

In this section the rules format, which are mined from a set of categorized sequences, and the required metrics are exemplified. The rule format of those mined rules from an ECG is “IF A occurs THEN B occurs WITHIN time T”, where A and B are basic shapes appeared in the ECG, i.e. a value of the categorized sequence, and T the time duration between A and B. The previous rule can be written as “A=>B within T”.

Given a categorized sequence the frequency f(A) of A is the number of occurrences of A in the sequence, while the relative frequency of A, called support, is f(A)/n, where n is the whole number of the values in the categorized sequence. The confidence c(A=>B within T) of the rule “A=>B within T” is the quotient of occurrences of A that are followed by B within T, i.e.:

$$\text{c(A=>B within T)} = \frac{f(A, B, T)}{f(A)} \qquad (6)$$

where f(A,B,T) is the number of occurrences of A that followed by a B within T.

The above metrics are correspondingly to the well known notion of support and confidence in association rules [1]. Support is used in order to mine the most frequent itemsets in a sequence while confidence is responsible for discovering the most efficient rules.

4. The RDA algorithm

As it is already mentioned RDA is an algorithm for discovering the most efficient association rules amongst subsequences in a sequence. In this study sequences corresponds to ECGs when each subsequence corresponds to a waveform of the ECG.

The first phase in RDA algorithm is to define some parameters. The initial sequence is categorized according to the windows length parameter and the correlation threshold. If the correlation between two subsequences is higher than the specified threshold, then they are categorized in the same category. Moreover, a time step is required scanning the sequence. Optionally a starting point and a search range can be selected, otherwise the entire sequence will be scanned. At last, a minimum support and a minimum confidence are needed.

At the following phase the categorized sequence is scanned, so that rules with support and confidence higher than the reconfigured thresholds to be mined. For that reason for each possible value for T frequent of each f(A,B,T) is counting. The number of possible rules is $a \times b^2$, where "a" is the number of the different categories in the sequence and "b" is the number of different possibilities for T. In Figure 1 the main idea of the RDA algorithm is demonstrated in pseudo code.

```
**Rule_Discovery_Algorithm** (MinCorrelation,
                MinSupport, MinConfidence, TimeStep)
**for** all subsequences $s_j$
  **for** all already defined categories $c_i$
    corr=correlation($c_i$,$s_j$)
    **if** (corr>=MinCorrelation) **then** define $s_j$ as $c_i$
  **endfor**
  **if** (no category math) **then** define $s_j$ as a new category $c_k$
**endfor**
**for** (t=0;t<T;t=t+TimeStep)
  **for** each categorized subsequence $c_i$
    **if** (itemset {$c_i$ ,$c_{i+t}$ } is new one) **then** add this itemset
                        into to the itemset list and count
                        itemsets {$c_i$ ,$c_{i+t}$ } frequent
**endfor**
**for** each counted itemset {$c_i$ ,$c_j$}
  **if** (|{$c_i$ ,$c_{i+t}$ }|>=MinSupport &&
                |{$c_i$ ,$c_j$ }|/|{$c_i$}|>=MinConfidence)
    **then** add  the rule $c_i$ =>$c_j$ within t into to the rule list
**endfor**
```

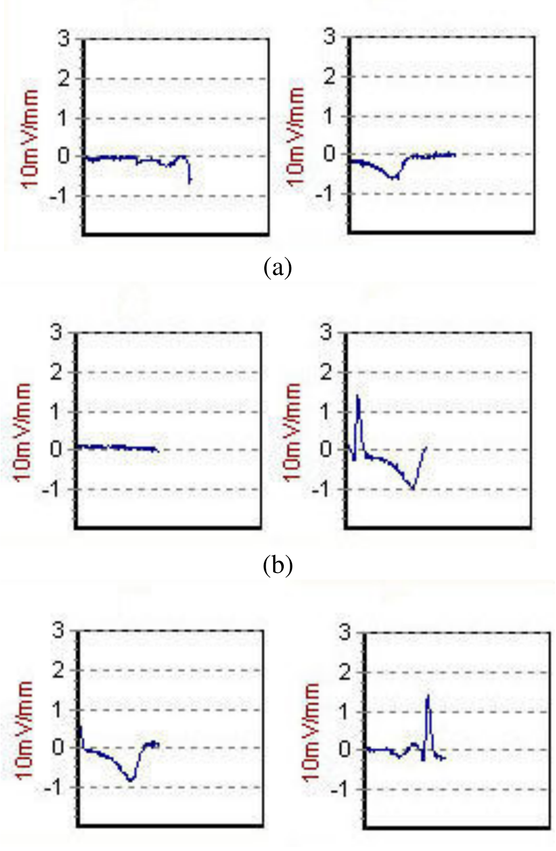
Figure 1. RDA Algorithm

5. Experimental results

In this study, the used ECGs' database was created at the Laboratory of Medical Informatics in the Aristotle University of Thessaloniki, Greece. It consists of ECGs of 11 Congestive Heart Failure (CHF) patients who participated in an EC project named CHS (Citizen Health System) between September 2001 and January 2003 [10]. The main goal of CHS was to develop a generic contact center which in its pilot stage could be used in the monitoring, treatment and management of chronically ill patients at home in Greece, Spain and Germany. Patients were able to communicate with the contact center via a variety of interfaces, like public telephone, Internet or a mobile device. During the period of this project, the heart failure patients were monitored for a period of 8-13 months and they were sending their ECG (record with the help of electronic microdevices) once a week.

We ran the RDA algorithm for each ECG separately (over than five ECGs for each patient). The initial ECGs (32 seconds duration) were categorized into 100 values length subsequenses. The standard metrics of 50% minimum confidence and a range of 75-90% minimum correlation were applied. The correlation threshold was depended on the quality of the corresponding ECG, as better quality a specific ECG had as higher threshold correlation was selected. Applying those metrics a mean of 1643.7 rules/ECG were mined (ranging from 1405 to 1912).

A significant number of the extracted rules could be considered as diagnostic criteria [11] aiding the physician to establish various hypothesis. For example, in patient's number 1 (#1) ECG 66.7% of the times his PR interval is following by T complex within 0,124 (±0.008) seconds. For patient's #2 ECG his TP segment is followed by QRS complex with in 0.266 (±0.008) seconds with a confidence of 100% while in an other ECG of the same patient 85,7% of the times the duration of QRS complex and of PR interval is less than 0.3 seconds. Figures 2 shows the graphs of the above examples.



(a)

(b)

(c)

Figure 2. In (a) PR interval is following by T complex within 0,124 with confidence of 66.7%. In (b) TP segment is followed by QRS complex with in 0.266 seconds with a confidence of 100%. In (c) the duration of QRS complex and of PR interval is less than 0.3 seconds in 85.7% of the times.

From the previous examples, the Initial ECGs was categorized into 100 values length subsequences using 90% for minimum correlation and 2 values length for the time step.

6. Conclusion

On the contrary from the previous algorithms that have mentioned before, the RDA is an algorithm appropriate for discovering associations rules in periodic time series. Subsequences are considered as items in such rules. RDA was evaluated on 60 ECGs coming from 11 congestive heart failure patients. The results have expressed that RDA can be an complementary way to ECGs' plots allowing the physician to test various hypotheses and discovering hidden knowledge.

In the future, an improved algorithm for RDA could be developed for filling missing parts in a periodic time series. First of all, the rules be obtained by the use of the RDA algorithm could be continuously combined in order to fill missing parts internally according to the rest of the time series. Particularly, for each missing part only the corresponding rules have to be mined so that significant CPU time will be saved.

Acknowledgements

This work was supported in part by the EC projects IST-1999-13352-CHS, IST-2001-33369-PANACEIA-ITV, as well as, by PROMESIP and IRAKLITOS both funded by the Greek Ministry of Education (EPEAEK I & II).

References

[1] Agrawal R, Imielinski T, Swami A: Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD International Conference on the Management of Data, 1993; 207-216.

[2] Zaki M. J, Parthasarathy S, Ogihara M, Li W: New algorithms for fast discovery of association rules. In Proc. of the 3rd KDD Conference, 1997; 283-286.

[3] Shen L, Shen H, Cheng L: New Algorithms for Efficient Mining of Association Rules. Information Sciences, 1999; 118, 251-268.

[4] Konias S, Giaglis G.D, Gogou G, Bamidis P.D, Maglaveras N: Uncertainty Rule Generation on a Home Care Database of Heart Failure Patients. In Proc of Computers in Cardiology, IEEE Comp. Soc. Press, 2003; Vol. 30: 765-768.

[5] Konias S, Bamidis P.D, Maglaveras N: Efficient Mining of Uncertainty Rules using Adaptive Thresholds in Medical Data. In Proc. of 3rd Hellenic Conference on Artificial Intelligence, 2004; 32-41.

[6] Lavrac N: Machine Learning for Data Mining in Medicine. AIMDM'99, LNAI 1620, 1999; 47-62.

[7] Das G, Lin K, Mannila, H, Renganathan G, Smyth P: Rule Discovery from Time Series. In Proc. of Int Conference on Knowledge Discovery and Data Mining, 1998; 16-22.

[8] Agrawal, R, Faloutsos C, Swami A.N: Efficient Similarity Search in Sequence Databases. In Proc. of 4th International Conference of Foundations of Data Organization and Algorithms, Chicago, 1993; 69-84.

[9] Rodgers J.L, Nicewander J.L, Nicewander W.A: Thirteen Ways to Look at the Correlation Coefficient. American Statistician 42, 1995; 59-66.

[10] Maglaveras N, Koutkias V, Chouvarda I, Goulis D.G, Avramides A, Adamidis D, Louridas G, Balas E.A: Home Care Delivery through the Mobile Telecommunications Platform: The Citizen Health System (CHS) Perspective, International Journal of Medical Informatics 68, 2002; 99-111.

[11] Macfarlane P.W, and Lawrie V: Comprehensive Electrocardiology. Volume 3, Pergamon Press, 1989.

Address for correspondence

Sokratis Konias
Aristotle University of Thessaloniki
Lab of Medical Informatics, POB 323
Thessaloniki, 54124, Greece
E-mail address: sokratis@med.auth.gr