Fuzzy Logic Classification based Approach for Linear Time Series Analysis in Medical Data Set

Dr. Manish Pandey*, Meenu Talwar**, Sachin Chauhan*** and Gurinderjit Kaur****
* Principal, CGC-College of Engineering, Mohali, Punjab
principal.cgccoe@gmail.com
** Assistant Professor, CGC-College of Engineering, Mohali, Punjab
cgccoe.cse.meenu@gmail.com
***Assistant Professor, Chandigarh Engineering College, Mohali Punjab
Sachindiet@gmail.com
****Associate Professor, CGC-College of Engineering, Mohali, Punjab
cgccoe.appsc.gn@gmail.com

Abstract: Health-care management systems are of great relevance now days due to provision of an easy and quick management in all aspects of a patient, not necessarily medical. Furthermore, there are more and more cases of pathologies in which diagnosis and treatment can be only carried out by using medical imaging techniques. With an ever-increasing prevalence, medical images are directly acquired in or converted into digital form, for their storage as well as subsequent retrieval and processing. Data Mining is the process of extracting information from large data sets through using algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems. Traditional data analysis methods often involve manual work and interpretation of data which is slow, expensive and highly subjective. Authors propose a robust ontology based multidimensional data warehousing and mining approach to address the issues of organizing, reporting and documenting diabetes cases including causalities. Data mining procedures, in which map and data views depicting similarity and comparison of attributes extracted from warehouses, are used in the present studies, for understanding the ailments based on gender, age, geography, food habits and hereditary traits. Time series forecasting takes the past values of a time series and uses them to forecast the future values. In this paper, we have proposed a new algorithm for multistep ahead time series forecasting. The original time series and differenced series are classified using Competitive Learning Neural Network.

Keywords: Neural Network, Time Serious Analysis, Regression Technique, Fuzzy Logic.

Introduction

Data Mining and Fuzzy Logic

Health care can be considered as the prevention, treatment, and management of illness and the preservation of mental and physical well being through the services offered by the medical, nursing, and allied health professions. Health care embraces all the goods and services designed to promote health, including preventive, curative and palliative interventions, whether directed to individuals or to populations. The organized provision of such services may constitute a health care system [1]. To implement the conceptions about the quality of the applied medical treatment the quality management system of the organism (medical center) must realize some characteristic activities, which are classified as following points:

- Identification of the basic elements (compartments) that necessary defines and realizes patient's global activity (treatment of the patients); Coordination of the activities of these elements;
- Definition and adaptation of criteria's and methods to make the control of the activity of the composing elements and the entire medical process;
- Make a permanent supervision, measurement and analyze of the process (based on the defined criteria's);
- Implement the proposed activity's to obtain the planed results, and realize a continuous amelioration of the process.

Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [9] and "the science of extracting useful information from large data sets or databases" [10]. It is the core principle of the knowledge discovery process, which also includes data selection, pre processing and cleaning, transformation and reduction, evaluation, and visualization. In the context of healthcare and biomedicine, data mining is often viewed as a potential mean to identify various biological, drug discoveries, and patient care knowledge embedded in the extensive data collected. Furthermore, data mining provides results that possibly highlight vaguely understood

doctrine and provide useful insights to help in decision making processes. Fuzzy Logic is usually known as an appropriate method for sorting and handling large amounts of data, but has also proven in recent years to be an excellent choice for many control system applications since it mimics human control logic. Fuzzy logic handles the concept of partial true that is true values between "completely true" and "completely false. It is very robust and forgiving of operator and data input and often works when first implemented with little or no tuning [7].

Linear Regression in Fuzzy Logic

A linear regression is a major factor evolved as an independent variable to explain the dependent variable change. In the real studies, the dependent variable changes often by several important factors. When more than one independent variable and dependent variables is in linear relationship, regression analysis is carried out by multiple linear regression Setting up y as the dependent variable, x1, x2, ... xk as independent variables, There is a linear relationship between the k independent variables and n the dependent variable, then multiple linear regression function is:

$y = b_0 + b_1x_1 + b_2x_2 \ldots + b_kx_k + e$

Where b_0 is a constant, b1, b2, ... bk for the regression coefficients. Multiple regression model parameter estimation is the same as linear regression equation.

Related Work

Castiglione et al. [1] has presented that in the clinical evaluation and treatment on a patient, there are mostly two relevant factors: the former is the training and experience of the physicians, the latter is the amount of information on which they may rely on. Such information may include the diagnosis of previous identical or similar cases, estimates on the possible healing times and any other information or comments considered clinically relevant. Ianosi [2] has presented in their paper regarding health management, life style modification which is the important solution against not just healthy life expectancy but medical crisis. Obviously, medical field plays a quite important role for prevention and treatment of disease. However, almost all patients consume very short time such as 3 to 5 minutes in medical treatment. On the other side, our ordinary life is at home, here the words of home is used as expanded meaning including our time for working and other activities. Lan Yo [3] has presented regarding Data Mining, which is the process of extracting information from large data sets through algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems. Lan Yo [4] has presented in the paper that in practice, many data mining exercises using data drawn from patients with particular conditions are performed to provide medical researchers with some insight into the disease that could lead to a greater understanding of the condition and suggest possible interesting directions for research. Frawley [5] has presented that Healthcare organizations practicing evidence-based medicine strive to unite their data assets in order to achieve a wider knowledge base for more sophisticated research as well as to provide a matured decision support service for the care givers..

Time Series Forecasting

Time series occur in various domains in great number and heterogeneity. In general, a time series s can be described as a sequence (x1, x2, x3 ..., xn) containing n data points xi. These data points can consist of real numbers, for instance of the river level or the voltage of an EEG derivation [8] measured at certain typically equidistant points in time; or more complex, they can be highly multidimensional, for example in market basket analysis [9], where xi corresponds to a customer's transaction containing e.g. the time of the transaction, a customer ID and bought items. In recent work on model-free analyses, wavelet transform based methods (for example locally stationary wavelets and wavelet decomposed neural networks) have gained favor. Multi scale (often referred to as multi resolution) techniques decompose a given time series, attempting to illustrate time dependence at multiple scales.A number of different notations are in use for time-series analysis: $X = \{X_1, X_2, \ldots\}$

is a common notation which specifies a time series X which is indexed by the natural numbers. Another common notation is: $Y = \{Y_t : t \in T\}$.

Conditions

There are two sets of conditions under which much of the theory is built as: Stationary process and Ergodicity. In addition, time-series analysis can be applied where the series are seasonally stationary or non-stationary. Situations where the amplitudes of frequency components change with time can be dealt with in time-frequency analysis which makes use of a time–frequency representation of a time-series or signal.[4].

Autoregressive Models

The general representation of an autoregressive model, well-known as AR(p), is

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t$$

Where the term ε_t is the source of randomness and is called white noise. It is assumed to have the following characteristics:

1. $E[\varepsilon_t] = 0$
2. $E[\varepsilon_t^2] = \sigma^2$
3. $E[\varepsilon_t \varepsilon_s] = 0 \quad \forall t \neq s$

Forecasting Methods

One of the main goals of time series analysis is to forecast future values of the series. A trend is a regular, slowly evolving change in the series level. The use of intuitive methods usually precludes any quantitative measure of confidence in the resulting forecast. In the Single-Equation Regression Models the variable under study is explained by a single function (linear or nonlinear) of a number of explanatory variables. The equation will often be time-dependent (i.e., the time index will appear explicitly in the model), so that one can predict the response over time of the variable under study to changes in one or more of the explanatory variables.

Modelling the Causal Time Series

With multiple regressions, we can use more than one predictor. It is always best, however, to be parsimonious, that is to use as few variables as predictors as necessary to get a reasonably accurate forecast. The forecast takes the form:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$,

where β_0 is the intercept, $\beta_1, \beta_2, \ldots \beta_n$ are coefficients representing the contribution of the independent variables X_1, X_2,..., X_n.

Statistical control limits are calculated in a manner similar to other quality control limit charts, however, the residual standard deviation are used.

Modelling Seasonality and Trend

Seasonality is a pattern that repeats for each period. For example annual seasonal pattern has a cycle that is 12 periods long, if the periods are months, or 4 periods long if the periods are quarters. Seasonal index represents the extent of seasonal influence for a particular segment of the year. The calculation involves a comparison of the expected values of that period to the grand mean. A seasonal index is how much the average for that particular period tends to be above (or below) the grand average. Therefore, to get an accurate estimate for the seasonal index, we compute the average of the first period of the cycle, and the second period, etc, and divide each by the overall average. The formula for computing seasonal factors is:

$S_i = D_i/D$,

where:

S_i = the seasonal index for i^{th} period — D_i = the average values of i^{th} period

D = grand average — i = the i^{th} seasonal period of the cycle.

Result Analysis

Numerical Application

The table 6.1 provides number of patients on monthly basis at the hospital. The number of patient shows a seasonal pattern. The final step is represented in figure 6.1, shows the forecast is to use the seasonal index to adjust the trend projection. One simple way to forecast using a seasonal adjustment is to use a seasonal factor in combination with an appropriate underlying trend of total value of cycles.

Linear Regression Equations

If we expect a set of data to have a linear correlation, it is not necessary for us to plot the data in order to determine the constants *m* (slope) and *b* (y-intercept) of the equation y = mx + c. Instead, we can apply a statistical treatment known as linear regression to the data and determine these constants. Given a set of data (x_i, y_i) with *n* data points, the slope, y-intercept and correlation coefficient, *r*, can be determined using the following:

$$m = \frac{n\sum(xy) - \sum x \sum y}{n\sum(x^2) - (\sum x)^2}$$

Table 6.1: Number of patients in different months, for three years and forecast of third year using linear regression

	Jan	Feb	March	April	May	June	July	Aug	Sept	Oct	Nov	Dec	Sum
2010	202	456	512	1167	1209	13	402	387	893	329	516	787	6873
2011	219	28	425	530	1398	1059	1544	747	654	266	523	769	8162
2012	442	659	629	363	1021	179	806	1637	2070	1313	1027	1110	11256
Mean	287.667	381	522	686.667	1209.33	417	917.333	923.667	1205.6	636	688.67	888.67	8763.67
Index	0.41	0.54	0.74	0.97	1.71	0.59	1.30	1.31	1.71	0.90	0.98	1.26	12.42
Expected 2013	180	356	465	353	1750	106	1048	2143	3537	1183	1002	1398	13521

$$b = \frac{\sum y - m\sum x}{n}$$

$$r = \frac{n\sum(xy) - \sum x \sum y}{\sqrt{\left[n\sum(x^2) - (\sum x)^2\right]\left[n\sum(y^2) - (\sum y)^2\right]}}$$

(Note that the limits of the summation, which are *i* to *n*, and the summation indices on *x* and *y* have been omitted.), implicitly applying regression to the sample data.

Application of linear regression technique for the calculation of line equation for the month 2010

Table 6.2: Calculation for the Summation of the Table Content for the Calculation

∑x	78
∑y1	6873
∑x*y1	45798
∑x*x	650
∑y1*y1	5437311
(∑x)²	6084
(∑y)²	47238129

After applying the following linear regression formula of slope and constant for the line equation, we will get the value of m and b. Where n = 12 (for 12 months in a year), after applying the formula for the slope of the line and the constant b for the line equation y = mx + b, we get: m = 7.856643 and b = 521.6818
Now for the plot of the graph for x = month of the year, y = variable from line equation after applying the value of slope (m) and constant (b), and y1 = number of patients in the month. We get:

Table 6.3: Calculation for the 12 months slope of the line after applying linear regression

x	y	y1
1	529.538	202
2	537.395	456
3	545.252	512
4	553.108	1167
5	560.965	1209
6	568.822	13
7	576.678	402
8	584.535	387
9	592.392	893
10	600.248	329
11	608.105	516
12	615.962	787

Application of linear regression technique for the calculation of line equation for the month 2011

Table 6.4: calculation for the summation of table content in year 2011

∑x	78
∑y2	8162
∑x*y2	57325
∑x*x	650
∑y2*y2	7891462
$(\sum x)^2$	6084
$(\sum y)^2$	66618244

After applying the following linear regression formula of slope and constant for the line equation, we will get the value of m and b: m = 29.87413 and b = 485.9848, Where n = 12 (for 12 months in a year) Now for the plot of the graph for x = month of the year, y = variable from line equation after applying the value of slope (m) and constant (b), and y2 = number of patients in the month. We get:

Table 6.5: Calculation for the 12 months slope of the line after applying linear regression in 2011

x	Y	y2
1	515.859	219
2	545.733	28
3	575.607	425
4	605.481	530
5	635.355	1398
6	665.23	1059
7	695.104	1544
8	724.978	747
9	754.852	654
10	784.726	266
11	814.6	523
12	844.474	769

Application of linear regression technique for the calculation of line equation for the month 2012

Table 6.6: Calculation for the Summation of the Table Content for the Calculation in 2012

∑x	78
∑y3	11256
∑x*y3	86393
∑x*x	650
∑y3*y3	13856640
$(\sum x)^2$	6084
$(\sum y)^2$	126697536

After applying the following linear regression formula of slope and constant for the line equation, we will get the value of m and b: m = 92.51049 and b = 336.6818, Where n = 12 (for 12 months in a year)
Now for the plot of the graph for x = month of the year, y = variable from line equation after applying the value of slope (m) and constant (b), and y3 = number of patients in the month. We get:

Table 6.7: Calculation for the 12 months slope of the line after applying linear regression in 2012

x	y	y3
1	429.192	442
2	521.703	659
3	614.213	629
4	706.724	363
5	799.234	1021
6	891.745	179
7	984.255	806
8	1076.77	1637
9	1169.28	2070
10	1261.79	1313
11	1354.3	1027
12	1446.81	1110

Application of linear regression technique for the calculation of line equation for the month 2013

Table 6.8: Calculation for the Summation of the Table Content for the Calculation

∑x	78
∑y4	13520.611
∑x*y4	109022.57
∑x*x	650
∑y4*y4	26129000
$(\sum x)^2$	6084
$(\sum y)^2$	182806926

After applying the following linear regression formula of slope and constant for the line equation, we will get the value of m and b. m = 147.8224 and b = 165.8723, Where n = 12 (for 12 months in a year):
Now for the plot of the graph for x = month of the year, y = variable from line equation after applying the value of slope (m) and constant (b), and y1 = number of patients in the month. We get:

Table 6.9: Calculation for the 12 months slope of the line after applying linear regression

x	y	y4
1	313.695	180
2	461.517	356
3	609.339	465
4	757.162	353
5	904.984	1750
6	1052.81	106
7	1200.63	1048
8	1348.45	2143
9	1496.27	3537
10	1644.1	1183
11	1791.92	1002
12	1939.74	1398

Comparison of slopes for four year

Table 6.10: Comparison of slops for four years

x	y1	y2	y3	y4
1	514	496	421	418
2	521	528	520	466
3	529	560	618	514
4	537	592	716	562
5	544	624	815	610
6	552	656	913	658
7	559	689	1012	706
8	567	721	1110	754
9	574	753	1208	802
10	582	785	1307	850
11	590	817	1405	898
12	597	849	1503	946

We get the different slope for different years

Table 6.11 Slope of four different years

m (2010)	7.85664
m (2011)	29.8741
m (2012)	92.5105
m (2013)	147.822

On comparing the four different slopes of four years we got that the predicted slope of the fourth year is about the average of the three year slopes hence we can say that the fourth year prediction is the good quality prediction for the patient data. The statistical reports in the following pages shows the various reports and analysis of patient data that can be represented using figure 6.6.

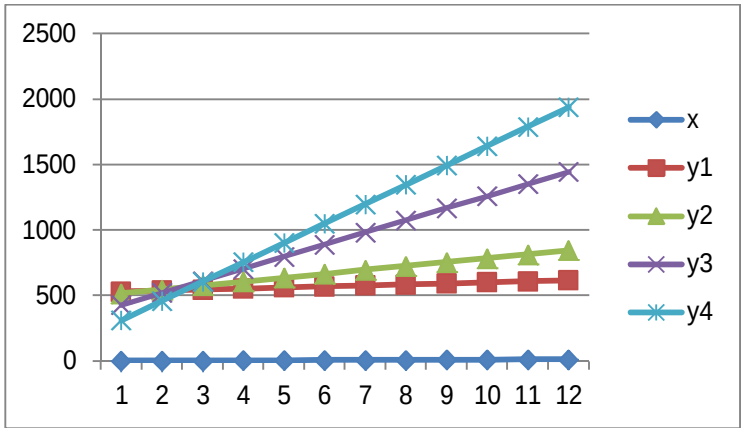

Figure 6.6: Comparison of slops for actual data of three years and the predicted value for fourth year

Conclusion

Inherent in the collection of data taken over time is some form of random variation. There exist methods for reducing of cancelling the effect due to random variation. Widely used techniques are smoothing. This technique, when properly applied, reveals more clearly the underlying trends. However, the data is not properly managed. As a result of this, majority of out-patients do not have full medical record. With this situation, the physician's time is wasted since they have to collect this information again and in addition, it becomes very difficult for them to keep track of the patients. This reduces the ability to carry out high quality clinical research in the hospitals, and compromises the continuity of healthcare as well as the quality of

healthcare delivery in the hospital. A Data Mart has been designed to collect, store, organize and retrieve the medical information of patients. A simple way of detecting trend in seasonal data is to take averages over a certain period. If these averages change with time we can say that there is evidence of a trend in the series.

References

[1] IANOSI ENDRE "Considerations about efficient health care management systems", Proceedings of the 3rd International Conference on E-Health and Bioengineering - EHB 2011, 24th-26th November, 2011, Iaşi, Romania
[2] Endre Ianosi, V. Vacarescu "Dialysis apparatus Technical and quality aspects (in Romanian)", Timisoara, Ed. Orizonturi Universitare, 2002, ISBN 973-8391-26-1.
[3] Lan Yu "Data Mining on Test Data of Physical Health Standard", 978-1-4244-3894-5/09/$25.00 ©2009 IEEE.
[4] Lan Yu" Association Rules based Data Mining on Test Data of Physical Health Standard", 2009 International Joint Conference on Computational Sciences and Optimization.
[5] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, "Knowledge Discovery in Databases: An Overview", in G. Piatetsky-Shapiro and W. J. Frawley (eds.), Knowledge Discovery in Database. AAAI/MIT Press, pp.127, 1991.
[6] Sashi K Nimmagadda "Multidimensional Data Warehousing & Mining of Diabetes & Food-domain ontologies for e-Health", 978-1-4577-0434-5/11/$26.00 ©2011 IEEE.
[7] Stefan Kleinmann, Ralf Stetter, Praveen Kumar Kubendra Prasad" Optimization of a Pump Health Monitoring System using Fuzzy Logic", 2013 Conference on Control and Fault-Tolerant Systems (SysTol) October 9-11,2013. Nice, France.
[8] T. Schluter and S. Conrad, "An approach for automatic sleep stage ¨ scoring and apnea-hypopnea detection," in Proc. of the 10th IEEE Int. Conf. on Data Mining (ICDM), 2010, pp. 1007–1012
[9] T. Schluter and S. Conrad, "TEMPUS: A Prototype System for Time ¨ Series Analysis and Prediction," in IADIS European Conf. on Data Mining 2010. IADIS Press, 2010, pp. 11–1.
[10] A.S. Chen, M.T. Leung and H. Daouk, "Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index," Computers and Operations Research 30, 2003, 901-923.
[11] W. Kreesuradej, D. Wunsch, and M. Lane, "Time-delay neural network for small time series data sets," in World Cong. Neural Networks, San Diego, CA, June 1994.
[12] H. Tan, D. Prokhorov, and D. Wunsch, "Probabilistic and timedelay neural network techniques for conservative short-term stock trend prediction," in Proc. World Cong. Neural Networks, Washington, D.C., July 1995.