Fuzzy Logic Classification Approach for Prediction of Patient Data in Health Care System using Data Mining Analysis

Gurinderjit Kaur[#1] , Sumeet Goyal[*2]

# Associate Professor, CGC College Of Engineering, Landran (Mohali)-Punjab
* Associate Professor, CGC College of Engineering, Landran (Mohali)-Punjab

ABSTRACT

Data Mining is the process of extracting data from huge information sets through various techniques drawn from the sector of Statistics, Machine Learning and information base Management Systems. Data processing, popularly known as data discovery in huge information, is carrying out of operations on data to retrieve, transform, or classify information. Authors propose a study of third-dimensional information deposit and mining approach to deal with the problems of organizing, reportage and documenting polygenic disease cases as well as causalities. Data processing procedures views representational process similarity and comparison of attributes extracted from the information gathered. Statistic statement takes the past values of a statistic and uses them to forecast the longer term values. Fuzzy regression strategies have been used to develop preferences models that correlate the engineering characteristics with shopper preferences relating to a replacement product; the patron preference models offer a platform, wherever by product developers will decide the engineering characteristics so as to satisfy shopper preferences before developing the merchandise. Recent analysis shows that these fuzzy regression strategies area units are normally used to model client preferences. We tend to propose a Testing the strength of Exponential Regression Model over regression Model to predict the patient data.

Keywords: *Health-care management systems, Fuzzy regression, Data mining, Data processing, Forecasting, Fuzzy membership function.*

INTRODUCTION

HEALTH CARE ANALYSIS:

In the health care domain, the major challenge is a way to offer higher health care services associated with nursing with the increasing range of individuals as the mistreatment restricts monetary and human resources. By providing attention and compliance, the medical technologies are executing innovations in patient care. Mining is a methodology of isolating gaining from gigantic content reports. [7].

A. Classification:

Along these lines information mining procedures, for example, grouping, combinations, relationship and clustering are for the most part used to remove the covered up, beforehand unobtrusive learning from ample databases. Information exploration arrangement is a method that directed machine learning system which makes predictions about future class cases by mapping cases of testing information to the predefined class marks which is gain from the supplied examples of classes with class names. There are a few models in characterizations, for example, probabilistic model, developmental algorithmic model and so forth.

Characterization includes of forecasting a definite lead to read of a given data. To anticipate the result, the calculation forms a preparation set containing a meeting of traits and also the separate result, usually known as objective or forecast characteristic. The calculation tries to search out connections between the traits that may create it conceivable to forecast the result. Next the calculation is given data settled not seen your time recently, known as forecast set, that contains identical arrangement of properties, other than the expectation property – not nonetheless illustrious. The method investigate the information and produces an expectation. The expectation characterizes exactitude, however "abundant" the calculation is. Definite procedure are provided for making an additional intensive smorgasbord of knowledge than to decline and is developing in eminence. Arrangement comprises of doling out a class name to an arrangement of unclassified cases. Administered Classification means the arrangement of conceivable classes is known ahead of time. Unsupervised Classification means set of conceivable classes are not known. After making arrangement we can attempt to dole out a name to that class. Unsupervised order is called grouping.

Arrangement is an alternate method of grouping. Bunching is an arrangement that the end-client/examiner know early how classes are characterized. It is vital that every record in the dataset used to manufacture the classifier as of now have a worth for the credit used to characterize classes. Since every record has a worth for the credit used to characterize the classes, and on the grounds that the end-client settles on the ascribe basis to utilize them. Arrangement is a great deal less exploratory than grouping. The target of a classifier is not to investigate the information to find intriguing fragments, but instead to choose how new records ought to be characterized. Characterization schedules in information mining utilizes a mixture of algorithms.

The information possessions of associations consistently stretch out to numerous terabytes of individual records that have collected over years of regularly impromptu or semi-arranged action. Organizations as often as possible don't have information arrangement systems and/or strategies set up that permit them to comprehend what information they hold and where it's found. This can have various pernicious impacts, for example in finding documents can turn into a troublesome errand, which can affect an association's capacity to pick up the full advantages of the learning amassed in its information. Legal implications can come about because of not having the capacity to rapidly discover or produce archives for a court hearing. Security may be traded off if information isn't effectively coordinated to staff access profiles.

B. Fuzzy based Classification

Fuzzy rule-based systems (FRBSs) square allow ways inside soft computing, supported fuzzy ideas to handle advanced real-world issues. They became a strong methodology to tackle numerous issues like uncertainty, impreciseness, and non-linearity [7]. They’re usually used for identification, classification, and regression tasks. FRBSs are deployed in a very variety of engineering and science areas.

FRBSs are referred to as fuzzy abstract thought systems or just fuzzy systems applied to specific tasks. They could collect specific names like fuzzy associative recollections or fuzzy controllers. They support the fuzzy pure mathematics that aims at representing the information of human specialists in an exceedingly set of fuzzy IF-THEN rules. Rather than victimization crisp sets as in classical rules, fuzzy rules use fuzzy sets. Rules were ab initio derived from human specialists through information engineering processes. However, this approach might not be possible once facing advanced tasks or once human specialists don't seem to be accessible. The basic need to adopt FRBS model with knowledge then victimization of learning ways.

REVIEW OF LITERATURE:

Shastri et.al [1] propose a strong metaphysics primarily based three-D knowledge deposit and mining approach to handle the problems of organizing, news and documenting polygenic disorder cases as well as causalities. data processing procedures, at intervals that map and data views depiction similarity and comparison of attributes extracted from warehouses, unit utilized during this studies, for understanding the ailments supported gender, age, geography, food habits and hereditary traits. Besides data image, data interpretation is planned for wealthy diagnosis, ensuing prescription and applicable medication. This approach provides a powerful back-end application for any web-based patient-doctor consultations and e-Health care management systems adopted by medical and social service suppliers.

Lan Yu et.al [13] proposes data mining on test data of physical health standard. In the mining experiment, scores of height/weight, vital capacity, grip strength, standing long jump and step test of a student are used for input attributes, and score level of the student is used for prediction attribute

Higuchi et al. [2] describes AN analysis technique supported fuzzy set for health scrutiny knowledge. This technique converts health info into fuzzy degree to manage a variable info analysis. The obtained fuzzy degree is taken under consideration as associate attribute price in interval [0, I]. The degree shows a normality of health condition. Throughout this study, fuzzy membership functions are created from commonplace divisions of reference interval of health medical info. As associate example, they calculated fuzzy degrees and ill health index from Japanese health medical info. Throughout this result, they confirmed that the obtained ill health index corresponded with medical established theories.

Kumar et al. [3] give the notion of intuitionistic fuzzy time series to handle the non-determinism in time series forecasting. An intuitionistic fuzzy time series forecasting model is also proposed. The proposed intuitionistic fuzzy time series forecasting method uses intuitionistic fuzzy logical relations on time series data. Performance of the proposed method is verified by applying it on two time series data. The effectiveness of the proposed intuitionistic fuzzy time series forecasting method is verified by comparing the forecasted output with others intuitionistic fuzzy sets based fuzzy time series forecasting methods using mean square error (RMSE) and average forecasting error (AFE).

Chowdary et al. [14] built up another system for choice tree for arrangement of information utilizing an information structure called Peano Count Tree (P-tree) which improves the productivity and adaptability. They apply Data Smoothing and Attribute Relevance methods alongside a classifier. Test results demonstrate that the P-tree strategy is altogether quicker than existing characterization systems and the favored technique for mining on information to be arranged.

Kishana et al. [15] concentrates on visual information digging applications for improving business choices. The product based framework is executed as a completely robotized and sufficiently shrewd to produce into results of every business exchange.

.

FUZZY BASED REGRESSION ANALYSIS:

Time Series Forecasting.

II. Time Series Forecasting

Profit analysis techniques embrace quantitative and qualitative foretelling ways. Quantitative prediction includes statistical analysis prediction and correlation analysis prediction. Statistical analysis is applied to the industries and firms that reveals the long run sales trends and history area unit consistent like food, energy, electricity, drugs and alternative basic industries. Stable and defensive corporations use moving averages, exponential smoothing and line analysis tools for future prediction. Correlation analysis forecast is being applied by the trade and firms, like building materials and business trade, in which external or internal factors have an effect on their sales, and that they believe the innovation-based industries and innovative corporations, like the knowledge trade. The sale factors is considered as independent variables and the profits as variable quantity to determine relationship between sale and profit by using regression model.

In ancient times, men have been obsessed with forecasting the future. There have been attempts to forecast the future by a wide range of means including paranormal and supernatural means. The ability to forecast the future is based on past history and is of utmost importance in various disciplines. To forecast time series accurately is essential in a wide range of domains such as weather forecasting, electric power demand forecasting, earthquake forecasting, and financial market forecasting. Because of the fact that these time series are affected by a multitude of interrelating macroscopic and microscopic variables, the underlying models that generate these time series are nonlinear and extremely complex.

FormallyTime series analysis is a statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals defined by wheret is the time index and n is the number of samples for observations. The aim of forecasting is to provide an algorithm that allows, with a certain level of confidence, the future values of the time series given by X_t+k where k$\in\mathbb{N}^+$ represents the prediction horizon of k steps ahead. According to chaos theory, forecasting accuracy exponentially grades with increase in the forecasting horizon. For instance, although we can forecast tomorrow’s temperature with a certain degree of accuracy, it is very difficult to forecast next year’s temperature with the same degree of accuracy. Therefore, long-range time series forecasting is challenging. Choice of time lags to represent time series is also a very important process in time series forecasting [1-2]. Real world time series are generated by nonlinear dynamical systems with an astronomical number of input variables. Such systems are extremely sensitive to initial conditions.

Time series occur in various domains in great number and heterogeneity. In general, a statistics may be represented as a sequence (x1, x2, x3 ..., xn) containing n information points xi. These information points will comprises real numbers, for instance of the river level or the voltage of an EEG derivation [8] measured at sure usually equal points in time; or additional advanced, they'll be extremely three-dimensional, e.g. the time of the transaction, a customer ID and bought items. Because of the fact that these time series are affected by a multitude of interrelating macroscopic and microscopic variables, the underlying models that generate these time series are nonlinear and extremely complex. Therefore, it is computationally infeasible to develop full-scale models with the present computing technology. Fully shaped applied mathematics models for random simulation functions, therefore on generate various versions of the statistic, representing what would possibly happen over non-specific time-periods within the future. Simple or totally shaped applied mathematics models to explain the probable outcome of the statistic within the immediate future, given data of the foremost recent outcomes (forecasting).

Modelling the Causal Time Series

With multiple regressions, we are able to use quite one predictor. It’s continuously best, however, that's to use as few variables as predictors as necessary to urge a fairly correct forecast. The forecast takes the form:
$Y = \beta_1 X_1 + \beta_{21} X_2 + \beta_3 X_3 + \_\_\_\_\_\_ \beta_n X_n 0$ Where β zero is that the intercept, β 1, β 2, . . . β n are coefficients representing the contribution of the freelance variables X1, X2,..., Xn. Statistical management limits are calculated in an exceedingly manner kind of like different internal control limit charts, however, the residual variance are used.

Modelling Seasonality and Trend
Seasonality could be a pattern that repeats for every amount. As an example annual seasonal pattern features a cycle that's twelve periods long, if the periods are months, or four periods long if the periods are quarters. We’d like to urge AN estimate of the seasonal index for every month, or different periods, like quarter, week, etc., reckoning on the information handiness.
The formula for computing seasonal factors is:
Si = Di/D, Where:
Si = the seasonal index for i^{th} amount
Di = the common values of i^{th} amount
D = grand average, i= i^{th} seasonal amount of the cycle.

SIMULATION METHODOLOGY:

a) Modelling and Simulation:

Here, we consider a medical dataset containing observations of patients in 12 months of the continuous dependent variable Y and independent variable X. Table 5 shows the medical dataset of patients over three years and forth year predicted value through Time Series Analysis. Let y_j denote the value of the variable y for observations, j (j = 1… N) Where N is number of months in a year, and let x_i be the observed value of the independent variable x for observation of number of patients. Suppose we have constants ‘a’ and ‘b’ for an exponential function
$y = a\, e^{(a+b)x}$ (1)

Where number of patients ‘y’ depends on the exponential function of ‘x’ with constants ‘a’ and ’b’. Now taking logarithmic analysis in both sides for the calculation of the two constants.

$\log_e y = \log_e a\, e^{(a+b)x}$ (2)
$\log_e y = \log_e a + (a+b)\, x$ (3)
$\log_e y = \log_e a + (a+b)x$ (4)

Now consider Y = loge y ; A = loge a, m = a+b and X = x – M , Where M is the mean of N observations.
We get the equation; Y = A + mX
To calculate the values of A and m we have the following equations:

$$A = \frac{\sum Y - b \sum X}{n} \qquad (5)$$

$$m = \frac{n \sum X\,Y - \sum X \sum Y}{n \sum X2 - (\sum X)2} \qquad (6)$$

Now X = x – M, X = x – 6.5 (6.5 is the mean value of 12 observations of x)
$\sum X = \sum x - \sum 6.5$ We get,
$\sum X = 0$
Putting $\sum X = 0$ in above values of constants A and m we get,

$$A = \frac{\sum Y - b \sum X}{n} \qquad (7)$$

$$M = \frac{n \sum X\,Y - \sum X \sum Y}{n \sum X2 - (\sum X)2} \qquad (8)$$

For the year 2014, after putting the values of X, Y and we get the values of the constants:
m = 30.223 and A = 562.08, Taking antilog of A we get, a = antilog A = 275.8 and b = -245.6
For the year 2015, after putting the values of X, Y and n we get the values of the constants:
m = 15.97 and A = 543.9 Taking antilog of A we get, a = antilog A = 228.14 and b =-212.14
For the year 2016, after putting the values of X, Y and n we get the values of the constants:
m = 20.50 and A = 606.25 Taking antilog of A we get, a = antilog A = 428.4 and b =-408
For the predicted year 2017, after putting the values of X, Y and n we get the values of the constants:
m = 39.2 and A = 661 Taking antilog of A we get, a = antilog A = 742.4 and b = -703.2
b) Result Analysis
Calculation of line equation for the month 2014:
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 = 30.223, a = 275.8, b = -245.6, A = 562.08. 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:

x	y1	y
1	275	373.6
2	300	505.4
3	370	683.7
4	475	925.0
5	550	1251.4
6	625	1693.0
7	800	2290.4
8	920	3098.7
9	880	4192.2
10	600	5671.5
11	500	7672.9
12	450	10380.5

Table 1: Calculation for the 12 months slope of the line after applying linear regression

Application of linear regression technique for the calculation of line equation for the month 2015:

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 = 15.97, a = 228.14, b = -212.14, A = 543.9
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:

x	y2	y
1	250	270.1
2	325	317.0
3	400	371.9
4	515	436.3
5	679	511.9
6	720	600.6
7	910	704.6
8	700	826.7
9	650	970.0
10	532	1138.0
11	456	1335.2
12	390	1566.6

Table 2: Calculation for the 12 months slope of the line after applying linear regression

Application of linear regression technique for the calculation of line equation for the month 2016:

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 = 20.50, a = 428.3, b = -407.8, A = 606.25
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:

x	y3	y
1	300	527.2
2	400	647.1
3	500	794.4
4	545	975.2
5	600	1197.1
6	750	1469.5
7	900	1804.0
8	850	2214.5
9	750	2718.4
10	700	3337.1
11	560	4096.5
12	400	5028.7

Table 3: Calculation for the 12 months slope of the line after applying linear regression

Application of linear regression technique for the calculation of line equation for the month 2017:

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): m = 39.2, a = 742.4, b =-703.2, A = 661. 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:

x	y4	y
1	145	109.9
2	240	162.7
3	371	240.8
4	489	356.5
5	642	527.7

6	919	781.2
7	1373	1156.4
8	1227	1711.8
9	1000	2534.0
10	750	3751.1
11	496	5552.6
12	281	8219.0

Table 3: Calculation for the 12 months slope of the line after applying linear regression

Comparison of slopes for four year:

X	y1	y2	y3	y4
1	373.6	270.1	527.2	109.9
2	505.4	317.0	647.1	162.7
3	683.7	371.9	794.4	240.8
4	925.0	436.3	975.2	356.5
5	1251.4	511.9	1197.1	527.7
6	1693.0	600.6	1469.5	781.2
7	2290.4	704.6	1804.0	1156.4
8	3098.7	826.7	2214.5	1711.8
9	4192.2	970.0	2718.4	2534.0
10	5671.5	1138.0	3337.1	3751.1
11	7672.9	1335.2	4096.5	5552.6
12	10380.5	1566.6	5028.7	8219.0

Table 4 Slope of four different years

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 1

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. 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. 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. By using time series analysis we can predict the data before the commencement of the particular period, which will enable us to make better arrangement for the treatment of patient so that no one may be mistreated.

Year	Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct	Nov	Dec
2014	275	300	370	475	550	625	800	920	880	600	500	450
2015	250	325	400	515	679	720	910	700	650	532	456	390
2016	300	400	500	545	600	750	900	850	750	700	560	400
Mean	275	341.7	423.3	511.7	609.7	698.3	870	823.3	760	610.7	505.3	413.3
Index	0.48	0.60	0.74	0.90	1.07	1.22	1.53	1.44	1.33	1.07	0.89	0.70
Expected 2017	145	240	371	489	642	919	1373	1227	1000	750	496	281

Table.5: Number of patients in different months, for three years and forecast of third year using linear regression.

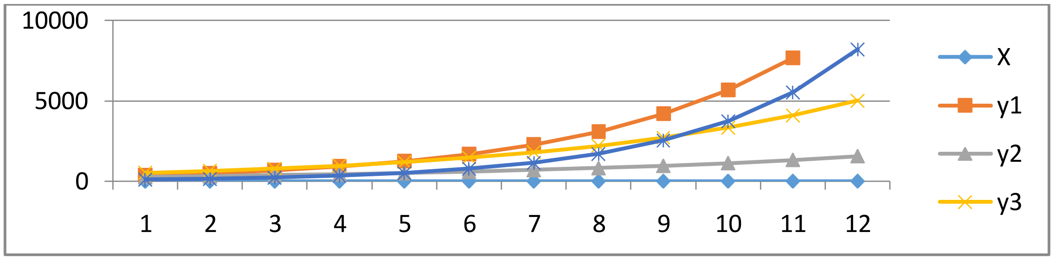

Figure 1: Comparison of slops for four years

References:

[1] Kit Yan Chan, Member, IEEE, Hak Keung Lam, Senior Member, IEEE, Tharam S. Dillon, Life Fellow, IEEE, and Sai Ho Ling, Senior Member, IEE "A Stepwise-Based Fuzzy Regression Procedure for Developing Customer Preference Models in New Product Developmen" IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 23, NO. 5, OCTOBER 2015

[2]H. Tanaka, S. Vejima, and K. Asai, "Linear regression analysis with fuzzy model," IEEE Trans. Syst., Man, Cybern., vol. SMC-12, pp. 903–907, 1982.

[3] Celikyilmaz A. and Turksen B., Fuzzy functions with support vector machines, Information Sciences, vol. 177, pp. 5163-5177, 2007.

[4] Chen S.P. and Dang J.F. A variable spread fuzzy linear regression model with higher explanatory power and forecasting accuracy, Information Sciences, vol. 178, pp. 3973-3988, 2008. [5] Chen X.B. and Ke H., Effect of fluid properties on dispensing processes for electronic packaging, IEEE Transactions on Electronic Packaging Manufacturing, vol. 29, no. 2, pp. 75-82, 2006.

[5] Kim H.K., Yoon J.H. and Li Y., Asymptotic properties of least squares estimation with fuzzy observations, Information Sciences, vol. 178, pp. 439-451, 2008.

[6] Takagi T. and Sugeno M., Fuzzy identification of systems and its application to modeling and control, IEEE Transactions on Systems, Man and Cybernetics, vol. 15, no. 1, pp. 116-132, 1985.

[7] Tanaka H. and Watada J., Possibilistic linear systems and their application to the linear regression model, Fuzzy Sets and Systems, vol. 272, pp. 275-289, 1988.

[8] 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.

[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] 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, Iași, Romania

[12] EndreIanosi, V. Vacarescu "Dialysis apparatus Technical and quality aspects (in Romanian)", Timisoara, Ed. OrizonturiUniversitare, 2002, ISBN 973-8391-26-1.

[13] Lan Yu "Data Mining on Test Data of Physical Health Standard", 978-1-4244-3894-5/09/$25.00 ©2009 IEEE.

[14] Lan Yu" Association Rules based Data Mining on Test Data of Physical Health Standard", 2009 International Joint Conference on Computational Sciences and Optimization.

[15] 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.