Sensitive Region Prediction using Data Mining Technique

Priyanka Pitale, Asha Ambhaikar

Abstract— Surveys provide huge amounts of healthcare data which, unfortunately, are not used to discover hidden information for effective decision making. Discovery of hidden patterns and relationships can provide a powerful prediction technique for predicting regions which are sensitive for several diseases. Advanced data mining techniques can help to predict future number of cases of a disease.

This research has developed a prototype, Sensitive Region Prediction System (SRPS), using data mining technique, called Linear Regression. Using historical data from various sources such as regional surveys and health reports, it can predict the number of cases of malaria disease. SRPS is user-friendly, platform independent, scalable, portable and expandable. It is implemented on the Java platform.

Index Terms—Decision Making, Hidden Patterns, Java, Linear Regression, SRPS

I. INTRODUCTION

Data Collected from various surveys can be used for prediction. Various sources of data collection provide huge number of live data. The current research developed a prototype; Sensitive Region Prediction System (SRPS) which uses the past data sets, collected from the various surveys and predict the number of cases of Malaria in future. A sample dataset taken from WORLD MALARIA REPORT 2010[1] is shown in table 1. The datasets are collected and saved into Attribute-Relation File Format (ARFF) [2]. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato [3] for use with the Weka machine learning software [4].After the dataset preparation the software create a model i.e. a function using linear Regression. At last model is used to predict the future values and analysis is done on predicted values for detecting the sensitive regions for a country. The overall working of SRPS is shown in Fig.2. The various phases and working of SRPS is described in further sections. The prototype is implemented in java Technology [5] and uses JFreeChart [6] library for generating plots and charts.

Table 1.Dataset Taken From World Malaria Report 2010

Year	Malaria Cases
2000	2031790
2001	2085484
2002	1841227
2003	1869403
2004	1915363
2005	1816569
2006	1785109
2007	1508927
2008	1532497
2009	1 563 344

Various data mining techniques are developed and are used for prediction [7-17]. Douglas c. Montgomery et.al.[18] performed simulation study on a prediction scheme using regression models with multicollinear predictor variables.

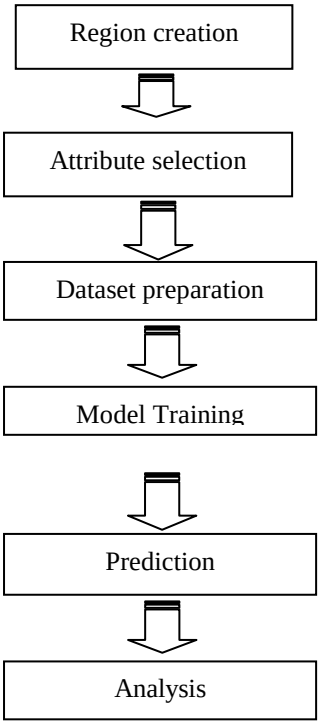

Fig. 2 overall working of SRPS

S Abdulsalam Sulaiman Olaniyi et.al [13] did Stock Trend Prediction Using Regression Analysis.

Manuscript Received June 22, 2011

Priyanka Pitale, Department of Computer Science, RCET,Bhilai, C.G, INDIA.

Asha Ambhaikar, Department of Computer Science, RCET,Bhilai, C.G, INDIA.

The tool also uses time series analysis for the prediction purpose.

II. METHODOLOGY

A. Region creation

The first step for prediction using SRPS is to create region for prediction. Fig.2 shows the main user interface and region creation window.

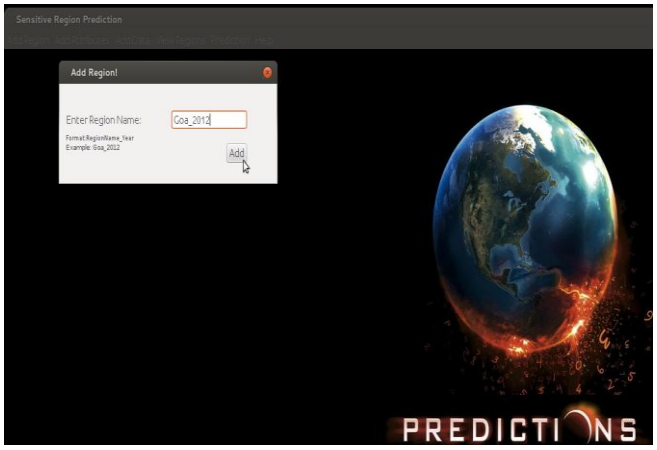

Fig2 Region creation using SRPS

SRPS takes the region name as Region_year format because it is assumed that the region is to be created to put data sets for a particular time period. After clicking Add button the system creates a file named same as region name, and to track the names of each region only region name is saved in a file named as Areaname.txt. the phase creates an .arff format file and put a line "@RELATION Region_Year " to it. This line indicates that a new relation is hcreated and name is "Region_Year". All these process are automated and runs in background.

B. Attribute selection

Attributes are the key part of prediction. The SRPS take two attributes namely Year and No. of cases from the user and put these attribute to selected regions arff file. The attribute selection window is shown in fig.3

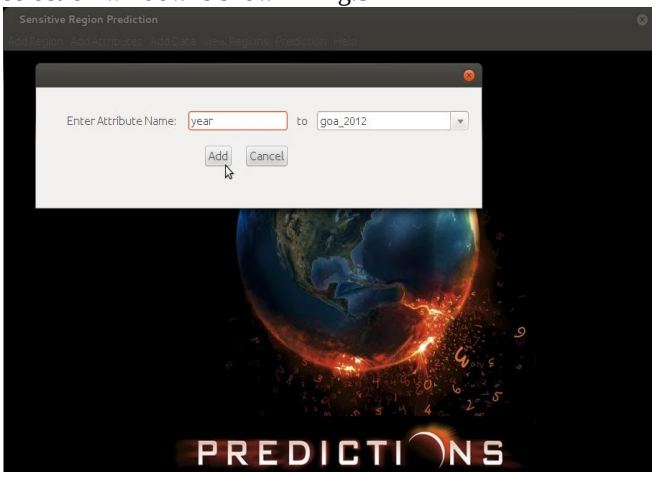

Fig. 3 Attribute selection using SRPS

The attribute names are appended to the region selected by the user with lines, "@ATTRIBUTE Year NUMERIC and @ATTRIBUTE Cases NUMERIC". These lines indicate that the two attributes i.e. Year and Cases are numeric and are part of a relation. These attribute plays the role of tuple and provides the interface to the datasets. The prediction is also based on these attributes.

C. Dataset preparation

Data set preparation is the task of providing the past data to the system. The data set performs the key role in predicting the future values. the data sets are first acquired from different sources like health reports , surveys, hospitals etc. The Sensitive Region Prediction System (SRPS) is trained and tested to detect the sensitive regions for Malaria disease.

The datasets on which the system is tested is taken from the WORLD MALARIA REPORT 2010, a report created by World Health Organization.

The system provides two ways for preparing data sets

1) *Manual Preparation*
2) *Importing from file*

Both of the approach works exactly the same thing, they append the data set to the selected regions file. The figure for creating datasets manually is shown in fig. 4.

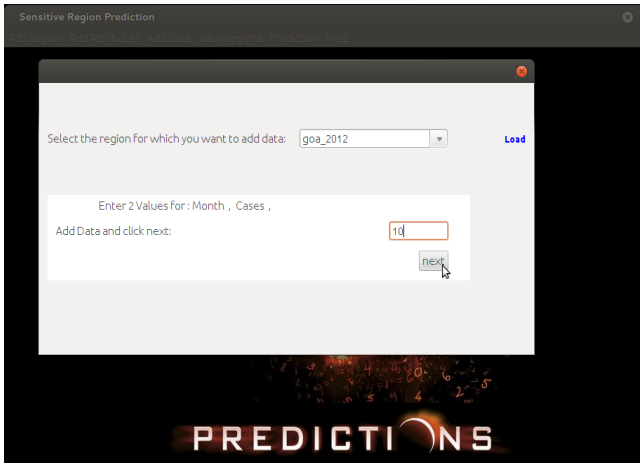

Fig. 4. Dataset preparation

The system put a line "@DATA" and append the data set line by line to the region file. The import option copies the dataset and appends it line by line to the region file. The coma separated datasets performs the key role for the prediction. Till this phase the region named file containing past data sets are ready for the prediction. The sample content of the goa_2000.arff file is shown below:

@RELATION goa_2000
@ATTRIBUTE Month NUMERIC
@ATTRIBUTE Cases NUMERIC

@DATA
1,100
2,200
3,90
4,250
5,300

Here the lines after @DATA are the datasets. The first part is the month value and second is no. of cases. For example "2,250" indicates 250 cases in the month of February 2000 in goa.

D. Model Training

The heart of the SRPS is the Model training module. A model in prediction system works as a core part of the system. In this technique data sets are used to generate a function that function is called model. Data set shown in table1 for example is used to create a model then first of all, all points for the data is first plotted in a 2D graph, and a function is then searched which will satisfy all the points. The simplest form of such a function is

$$y=b+wx \quad (1)$$

Where y is called response variable and x is predictor variable. Where the variance of y is assumed to be constant, and b and w are regression coefficients. The regression coefficients, w and b can also be thought of as weights so we can write

$$y=w0+w1*x \quad (2)$$

These coefficients can be solved by the method of least squares, which estimates the best fitted line as the one that minimizes the error between the actual data and estimate of the line. The values for the coefficient w1 is calculated from the formula-

$$W1= \{\sum (x - \bar{x}) *(y - \bar{y})\}/ \{\sum (x - \bar{x})*(x - \bar{x})\} \quad (3)$$

And then w0 can be calculated by the formula-

$$w0= \bar{y} - w1* \bar{x} \quad (4)$$

After putting the values of w0 and w1 calculated from the equation (3) and (4) into equation (2) we got a function of y in terms of x, this function is then called as model and the procedure is called model training.

Figure 5 shows the graph between Month and number of cases in Goa in the year of 2000. The data for the graph is shown previously. The method finds a line on graph which is closest to maximum points and returns a function for the line. This function is used for further process.

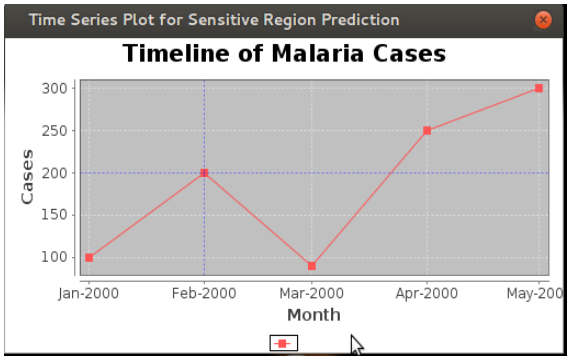

Fig. 5 Graph between Year and no. of cases

From the graph shown in figure one can easily observe that the malaria cases are increased in year of 2000 in Goa region.

E. Prediction

Prediction is the step in which, a query data set is passed to the trained model. The trained model uses function which is discussed in section D, and predicts the value according to the predictor variables and the regression coefficients b and w.

The SRP system provides three basic functionalities under prediction section i.e.

1) *Generate Prediction Formula:*

This option is used to generate the prediction formula. An extra ordinary feature of SRP System is that it is generalized, i.e. it can be utilized for other purpose like crime prediction and sales prediction. The current paper only focuses on the utilization on SRPS for the Malaria disease prediction. For the other purpose if one wants to see the relation of query data with the data sets then this option can show the details as a formula. The option is shown in fig. 6.

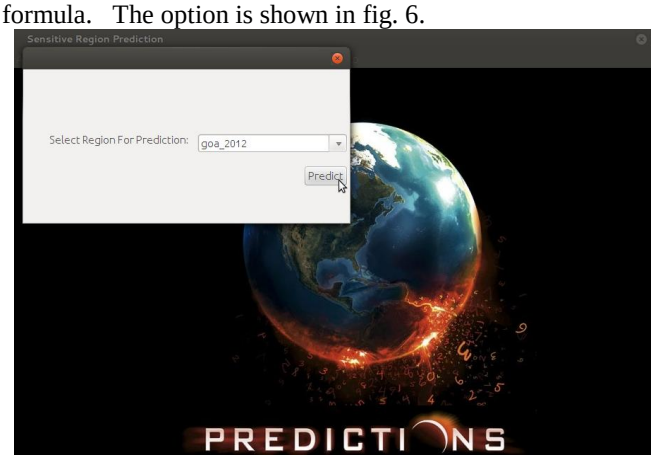

Fig. 6 Formula Generation using SRPS

User can select the region or the dataset file for which he wants to generate the formula, after clicking predict button the formula is shown in a text box. The formula generated for goa_2000.arff dataset file is –

Cases = 45 * Month + 53

Where 45 is w1 and 53 is w0 which is calculated by the equations (3) and (4).

2) *Predict future values:*

This option is used to predict the future values for the particular area and for the future date. For example if the data set for goa region from last 10 years is available, the system can predict the number of cases for the November month 2012. The system shows the predicted value in a text box and in the form of graph. The sample predicted value for the goa region for February month 2012 is shown figure.7. From the graph user can easily visualize the predicted result. For example whether the number of cases will increase or decrease? Figure- 7 shows the window for predicting the future value for a region.

The window requires region name, year and month for which prediction is to be done. After putting all the details the system searches all dataset files for the region name goa, it extracts all past datasets for the inputted month and put the data to a newly created dataset file. Now the newly created dataset file is then used for the prediction and a result is generated.

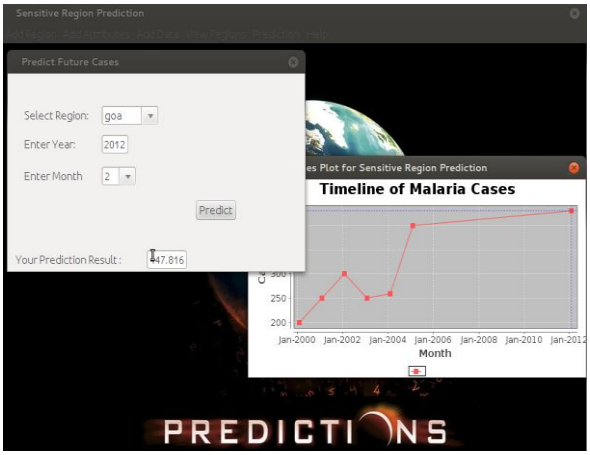

Fig. 7 Predict future values using SRPS

3) Predict sensitive regions:

This is the final and most important phase of the system. in this phase the system takes the month and year for which the sensitive region is to be predicted and shows the most sensitive region. Sensitive region is a region where the cases of malaria will be increased in maximum number. Window for prediction sensitive region is shown in figure 8.

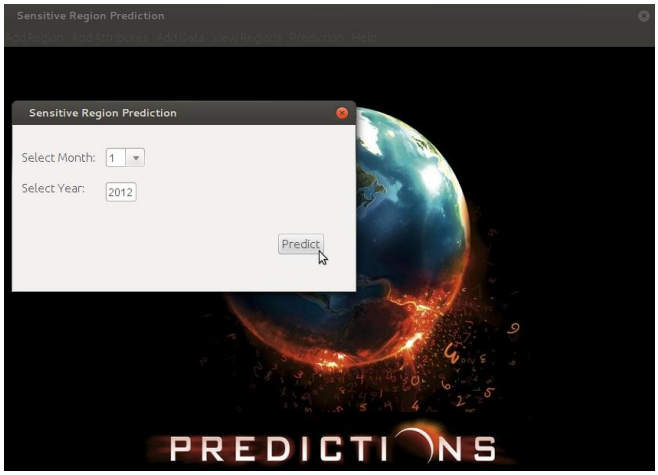

Fig. 8 Predict sensitive regions using SRPS

The system generates a pie chart to indicate the sensitive region. The sample chart generated by the system is shown in figure 9. The pie chart shows the sensitive regions in dark red color and also segments it from the other region. This type of chart can provide a good and easy interface to the forecasters.

F. Analysis

The forecasters can predicts the future number of cases and from the system generated pie chart they can easily analyze that which region is sensitive.

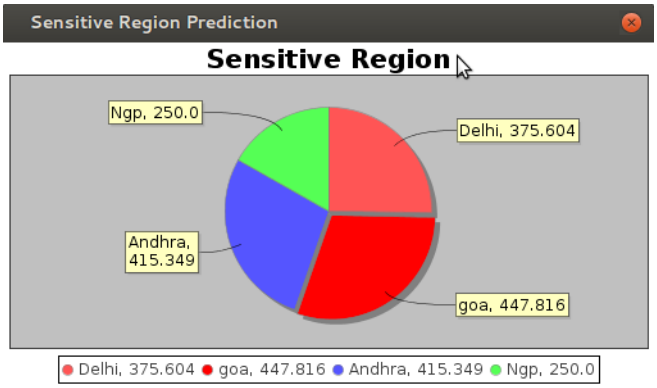

Fig. 9 Sensitive region prediction using SRPS

Hence the system can provide an analytical power to the predictors to take proper decision to control the diseases in the sensitive regions.

III. RESULTS AND DISCUSSION

We presented regression analysis data mining technique and proposed a tool for exploiting especially time series data in crime analysis. A prediction system is proposed that uses data mining technique to produce periodically forecasts about crimes. Our technique complement proven numeric forecasting method using regression analysis with technology taking as input the crime information. Several series of experiments were run using WEKA [15] frame work to achieve model training and prediction. In place of WEKA, a simple module based on java, is in development process so that necessary changes or up gradation can be done.

It is observed that the accuracy for the prediction is dependent on data sets. If data sets are of huge numbers high level of model training is achieved and in case of small data sets a low level of model training is achieved.

It is also observed that the accuracy of the prediction is depends upon the level of training. If model is highly trained then the prediction will be most accurate and if model is not highly trained the accuracy will be low.

ACKNOWLEDGMENT

The authors would like to express deep gratitude to Rungta College of Engineering and Technology for providing necessary support towards the project.

REFERENCES

1. http://www.who.int
2. www.cs.waikato.ac.nz/ml/weka/arff.html
3. http://www.cs.waikato.ac.nz/
4. http://www.cs.waikato.ac.nz/~ml/weka/
5. www.java.com/en/about/
6. www.jfree.org/jfreechart/
7. Takashi Kimoto, Kazuo Asakawa ,Morio Yoda and Masakazu Takeoka, “Stock Market Prediction System with Modular Neural Networks ”,1990, IJCNN International Joint Conference on Neural Networks, Vol-1, San Diego, CA , USA pp. 1-6.
8. Sayan Mukherjee ,Edgar Osuna,Federico Girosi, “Nonlinear Prediction of Chaotic Time Series Using Support Vector Machine ” , Proceeding of IEEE NNSP'97, Amelia Island, FL, 24-26 Sep., 1997

9. Jaehyun Sim , Seung-Yeon Kim and Julian Lee, "Prediction of protein solvent accessibility using fuzzy k -nearest neighbor method ", Oxford Journals ,Vol. 21 no. 12 2005, pages 2844–2849 .
10. Kari Laasonen, "Prediction of Mobile User Routes from Cellular Data", 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005. Proceedings, Springer Berlin / Heidelberg, Monday, November 07, 2005 ,pp-569-576.
11. Steven Rovnyak , Stein Kretsinger , James Thorp , Donald Brown , "Decision Trees For Real-Time Transient Stability Prediction " , IEEE Transactions on Power Systems, Vol. 9, No. 3. August 1994 , pp-1417-1426.
12. Kiryung Lee, Dong Sik Kim, Taejeong Kim, "Regression-based prediction for blocking artifact reduction in JPEG-compressed images ", IEEE Transactions on Image Processing, Electron. & Telecommun. Res. Inst., Daejeon, South Korea,14 Issue: 1, pp-36 - 48 .
13. Hongyu Sun, Henry X Liu, Heng Xiao, Bin Ran, "Short Term Traffic Forecasting Using the Local Linear Regression Model", Center for Traffic Simulation Studies, Institute of Transportation Studies, UC Irvine, 07-01-2002.
14. C C Toner, C J Broomhead, I H Littlejohn,G S Samra "Prediction of postoperative nausea and vomiting using a logistic regression model", British Journal of Anaesthesia. 1996, pp-347-351.
15. Panagiotis Sentas, Lefteris Angelis, "Categorical missing data imputation for software cost estimation by multinomial logistic regression" , Journal of Systems and Software, Volume 79, Issue 3, March 2006, Pages 404-414.
16. Xiaobo Zhou, Xiaodong Wang, Edward R. Dougherty, "Gene prediction using multinomial probit regression with Bayesian gene selection", EURASIP Journal on Applied Signal Processing, Volume 2004, 1 January 2004, pp- 115-124.
17. F.J. Nogales, J Contreras, A.J Conejo, R. Espinola, "Forecasting next-day electricity prices by time series models", IEEE Transactions on Power Systems, Volume 17Issue:2, pp- 342 – 348.
18. Douglas C. Montgomery , David J. Friedman, "prediction using regression models with multicollinear predictor variables", IIE Transactions, Volume 25, Issue 3, 1993, pp- 73-85.

AUTHORS PROFILE

Priyanka Pitale pursued her Bachelor of engineering in Information Technology from Pt. Ravishankar shukla University, Raipur. She is persuing her masters of technology in Computer Technology from C.S.V.T.U., Bhilai. Her Research area mainly focus on data mining and prediction field

Asha Ambhaikar was born in July 1965 in Nagpur district, MH, India. She is graduated from Nagpur University, Nagpur, India, in Electronics Engineering in the year 2000 and later did her post graduation in Information Technology from Allahabad Deemed University, Allahabad India. She has submitted her PhD on the topic "Design and Development of Manet Routing Protocol for Improving Scalability". Currently she is working as an Associate Professor in RCET Bhilai, and she has published more than 17 research papers in reputed national and international journal's and conferences. Her research area includes, networking, data warehousing and mining, Distributed system, signal processing, image processing, and information systems and security.