International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 22 Issue 1 – MAY 2016.

LOGISTIC REGRESSION BASED APPROACH FOR CLASSIFYING DERMATOLOGY DISEASES

Neha Sharma[1], R. K. Gupta[2]
[1,2]*Department of CSE & IT,* [1,2]*Madhav Institute of Technology & Science, Gwalior (India)*

***Abstract*- Text mining is an adaptable technology due to its applicability for numerous different tasks in the biomedical domain. It is seen that during the past years there is a significant development in the area of biomedical text mining because of the increasing number of electronic medical records in the database. We present a system based on logistic regression combined with one versus all strategy for classifying the dermatology diseases into various classes based on thirty three attributes. For the performance of this task, a dataset was divided into two sets 178 medical records were taken as a training set and 179 records were taken as a testing set. We will show that our system has relatively performed well with 95.34% recall, 94.82% precision and 98.36% accuracy. The dermatology database is accessible through a web interface at http://archive.ics.uci.edu, where all datasets are uninhibitedly accessible for download.**

Index Terms- Text Mining, Logistic Regression, One versus All Strategy, XLMiner

I. INTRODUCTION

Traditionally, documents were stored in paper files, within folders and filing cabinets and a major disadvantage of this type of document file organization was the time it takes to access the document file particularly when documents are in large quantity. Due with the advancement of technology documents are being stored digitally, but at the same time it is difficult to retrieve relevant documents from the stored document if they are not categorized properly. Text categorization has become the subject of considerable research interest in the area of information retrieval, information extraction, text classification & clustering and topic detection [1]. The process of text categorization is the task of arranging a set of documents in a preordained set of categories [2]. Prior this process had been performed manually, which requires huge cost and time, therefore there is a requirement for an automatic text categorization system. In the field of medicine, the concepts and techniques of text categorization are being used nowadays. As we know that documents can be classified using two classes or more, the system can be designed to perform binary classification in which classification is done only for two classes or it can be designed to perform multiclass classification in which classification is done for more than two classes so later technique is more beneficial than previous one because it can be used to solve the real world problems [3]. Our present work concentrates on the multiclass classification of the dermatology diseases based on thirty three attributes. We have used logistic regression with one versus all strategy for multiclass text categorization. The structure of the paper is as per the following: Section 2 talks about the foundation work done in this field. Segment 3 exhibits the proposed approach. Section 4 clarifies the examination setup. Close and devise future work.

II. BACKGROUND WORK

A significant measure of work has been done in the field of text categorization of medical records. The Authors have proposed distinctive strategies for categorizing the documents. In [4] KNN was applied as a white-box multiclass classifier to classify epileptic diagnoses into a standard code. In [5] dictionary based tagger was combined with a scoring strategy for named entity recognition of human genes and diseases. In [6] lexical KNN algorithm was used in which lexemes (tokens) were selected from abstracts of medical documents and utilized for classification by matching them with the standard list of keywords specified as (MESH) medical subject heading. In our work we have combined logistic regression with the one versus all strategy for the multiclass classification of the dermatology diseases.

III. PROPOSED METHODOLOGY

We have proposed a logistic regression combined with one versus all strategy for performing the task of classifying dermatology diseases into several classes.

A. One versus All Strategy

The simplest approach is to reduce the problem of classifying among K classes into K twofold problems, where each problem discriminates a given class from the other K −1 classes. For this methodology, we require N = K twofold classifiers, where the k^{th} classifier is trained with positive cases belonging to class k and negative cases belonging to the other K − 1 classes [3]. When testing an unknown example, the classifier delivering the maximum yield is considered the winner, and this class label is allocated to that case.
This method has an advantage that the number of binary classifiers is equal to the number of classes. However, there are some limitations. Firstly, during the training phase memory requirement is high.

B. Logistic Regression

From numerous points of view logistic regression resemble ordinary regression. It requires a dependent variable, y, and one or more independent variables [5]. In multiple regression analysis, the mean or expected estimation of y is alluded to as the multiple regression equation.

$$E(y) = \beta_1+\beta_1x_1+\beta_2x_2+.....+\beta_px_p \quad (1)$$

In logistic regression, statistical theory as well as practice has shown that the relationship between *E* (y) and x1, x2,……, xp is better described by the following nonlinear equation.

$$E(y) = e^{\beta1+\beta1x1+2x2+.....+\beta pxp} / (1+e^{\beta1+\beta1x1+\beta2x2+.....+\beta pxp}) \quad (2)$$

In the above equation y is the dependent variable and x is the independent variable, in our proposed methodology y predicts the class and x1, x2,……, x3 takes the attribute values which were taken for the disease diagnosis. For example If the two values of the dependent variable *y* are coded as 0 or 1, the value of E(*y*) in equation provides the probability that *y*=1 given a particular set of values for the independent variables *x*1, *x*2, . . . , *xp*. Because of the interpretation of *E*(*y*) as a probability, the logistic regression equation is often written as follows.

$$E(y) = P(y=1|x1, x2,..., xp) \quad (3)$$

C. Multiclass Classification

When Classification is done with more than two classes, it is known as multiclass classification. For example, classify a set of images of vegetables which may be cabbage, cauliflower, or peas. Multiclass classification makes the assumption that each sample is assigned to one and only one label: a vegetable can be either a cauliflower or a pea but not both at the same time. A multiclass classification is more complex than the binary classification problem since the generated classifier must be able to separate the data into larger number of categories which increases the chance of conferring errors.

D. Concept

Suppose we need to diagnosis medical symptoms such as cold, fever, flu. Each class is assigned a value. Class cold is assigned with a value 1, fever with value 2, and flu with value 3.

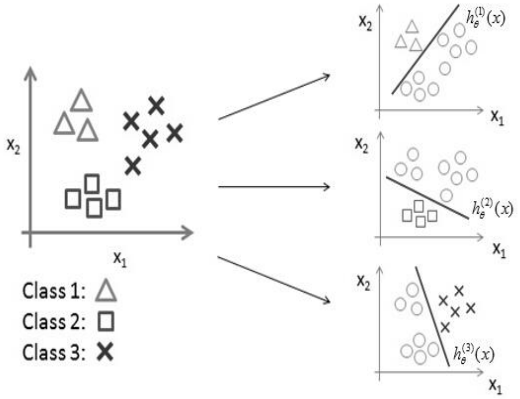

Fig. 1. Illustration of multiclass Classification Problem Using One versus All Scheme

In above figure training set is converted into three binary training set, now let us first consider the triangle i.e. y=1 (cold), and take it as a positive class and rest as negative classes, similarly for y=2 (fever), take it as a positive class and rest as negative class and for y=3 (flu), take it as a positive class and rest as negative class. Now for each binary training train the logistic regression classifier P (y=i|x) for each class to predict the probability that y=i, on a new input x, to make a prediction, pick the class i that maximizes max P (y=i|x).

Equation computed for the first two class training set.

$$E(y1) = P(y=1|x) \quad (4)$$

Equation computed for the first two class training set.

$$E(y2) = P(y=2|x) \quad (5)$$

Equation computed for the first two class training set.

$$E(y3) = P(y=3|x) \quad (6)$$

Find the maximum value among three i.e. max (E (y1), E (y2), E (y3)), class is picked with maximum value.

IV. ARCHITECTURE OF PROPOSED METHODOLOGY

In the proposed methodology architecture, the following modules are taken. These modules will be explained further.

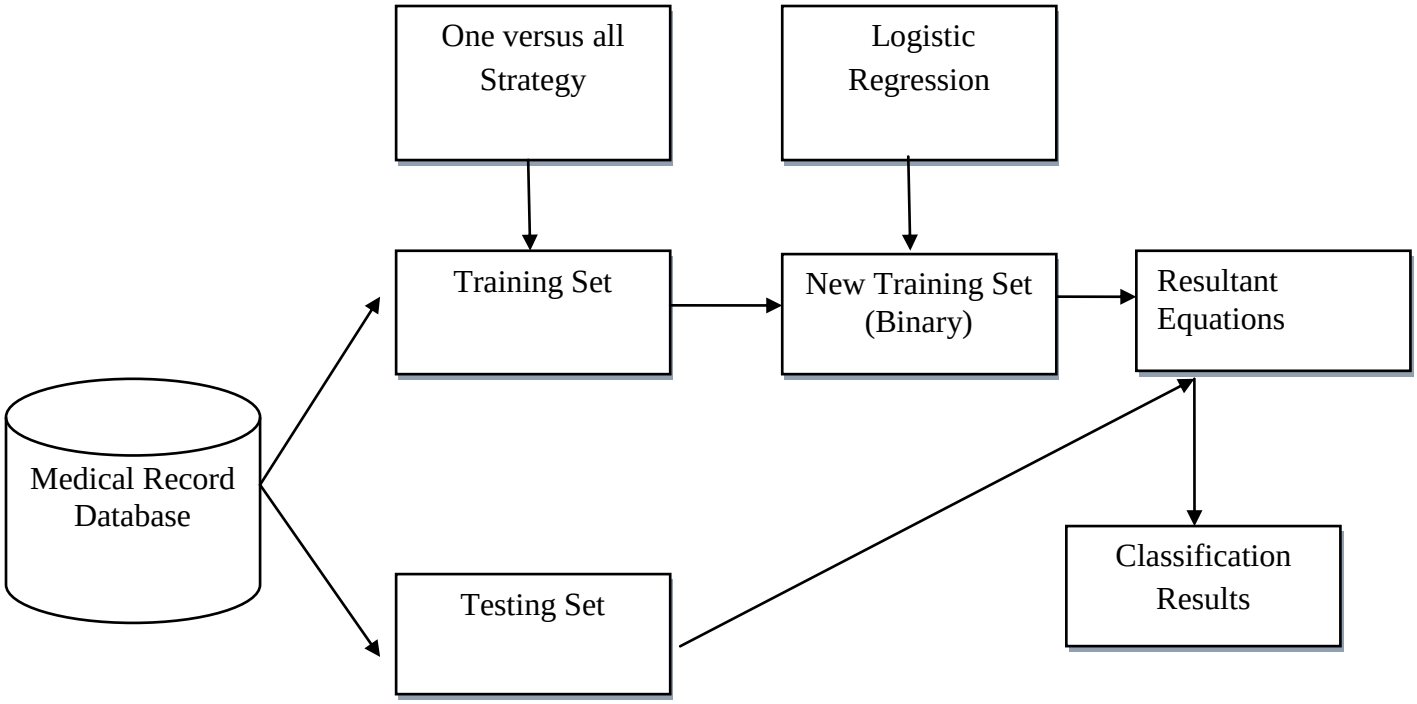

Fig. 2. Architecture of Proposed Methodology

The following architecture can be explained as follows:
Step 1: Medical Record Database is divided into two sets, training set consisting of 178 records and testing set consisting of 179 records.
Step 2: One versus All (one versus rest) strategy is used where training set is converted into k binary problems for the k classes such that each binary problem is considered as a new training set where k class is treated as a positive class and k-1 classes are treated as a negative classes.
Step 3: Logistic regression classifier is fitted is on the recently produced training set.
Step 4: Equations are yielded.
Step 5: Test the new record taken from the test set and fit them in the resultant equations.
Step 6: Computation is being carried out and the set which has the greatest worth, that specific class is picked.

V. EXPERIMENTAL SETUP AND RESULTS

This includes various sections which are discussed below:

A. Frameworks

XLMiner is one of the best tools for mining data available in Excel Worksheets. It includes capabilities that allow a miner to work with partitioning, neural systems, classification and regression, association rules, nearest neighbor etc. [7]. XLMiner can work with large data sets which may exceed the limits in Excel. A standard methodology is to test data from a greater database, pass on it into Excel to fit a model and, in the case
of supervised learning routines, score yield back to the database. In the standard release of XLMiner, this feature is upheld for SQL server, Oracle and Access databases. There are several features of XLMiner which includes partition data, data utilities (sampling), classification, association rules, prediction, time series analysis, data reduction and exploration and charting. In our work it is used to perform logistic regression on the training set due to ease of use and learning. Secondly proposed methodology was implemented in Java swings. Swing library is an official Java GUI toolkit released by Sun Microsystems; it is utilized to make graphical client interfaces with Java, there are many features in Java swings, including platform independent, customizable, adaptable, configurable and lightweight which has made it successful in creating full featured desktop applications [8].

B. Datasets

The process was tested with a genuine dataset, developed with genuine anonymous patient records, found on the electronic medical record [9] provided by a hospital. The records contain dermatology disease diagnoses. For each record, there are thirty three attributes which were taken into consideration for the diagnosis purpose. Each attribute was assigned with 0, 1, 2, 3 values. Here 0 indicates that the absence of feature, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values. Based on the above information dermatology diseases were needed to be classified among six classes which were psoriasis, Seboreic dermatitis, lichen planus, Pityriasis rosea, chronic dermatitis, and Pityriasis rubra Pilaris and In addition, classification codes were used to describe a particular class.

I. Class distribution

Class code	Class
1	psoriasis

2	Seboreic dermatitis
3	lichen planus
4	Pityriasis rosea
5	chronic dermatitis
6	Pityriasis rubra Pilaris

C. Evaluation Metrics

To handle a multiclass issue as the one, present different approaches can be taken after. We will partition text mining issue into several two-class problems, utilizing a one-versus all methodology. To evaluate the decision task, we first define conceivable results of the classification:

True positive: A true positive test outcome is one that identifies the condition when the condition is present.
False positive: A true negative test outcome is one that does not identify the condition when the condition is absent.
False negative: A false positive test outcome is one that identifies the condition when the condition is absent.
True negative: A false negative test outcome is one that does not distinguish when the condition is present. Several measures have been defined based on these values.

Recall (R): It is the fraction of retrieved instances that are relevant. Recall is calculated by (TP / TP + FN)) formula and **Precision (P):** It is the fraction of relevant instances that are retrieved. Precision is calculated by (TP / (TP + FP)) [10].
F1 measure: It is computed by combining recall and precision in a single score. F1 is calculated by (2*P*R/ (P+R)). F1 is one of the most appropriate measures for text classification,, since it deals well with unbalanced scenarios, common in text classification.
Finally **accuracy** is calculated which refers to the ability of the model effectively anticipate the class name of new or beforehand inconspicuous information. It can be computed by ((TP+TN)/ (TP+TN+FP+FN)).

D. Learning and Results

Machine Learning has distinctive procedures of deducting models (capacities) from data, which can be used to map new documents. A multiclass classification using one-versus-all with logistic regression algorithm was chosen. Each new document was tested against each logistic regression equation which was computed for every two class new training set and whosoever had the maximum value was chosen and considered as the class of the new document. After performing the task of assigning categories to each new document results are analyzed.

II. Four Categories to Express the Quality of Classification

class	TP	TN	FP	FN	F1
Psoriasis	53	126	2	2	96.36%
Seboreic dermatitis	27	152	3	3	90%
Lichen planus	37	142	0	0	100%
Pityriasis rosea	24	155	0	4	92.30%
Cronic dermatitis	19	160	4	0	90.47%
Pityriasis rubra pilaris	10	169	0	0	100%

In table 3 results of the performance measures are shown

III. Performance Measures

class	Recall	Precision	Accuracy
Psoriasis	96.36%	96.36%	97.81%
Seboreic dermatitis	90%	90%	96.75%
Lichen planus	100%	100%	100%
Pityriasis rosea	85.71%	100%	97.81%
Cronic dermatitis	100%	82.60%	97.81%
Pityriasis rubra pilaris	100%	100%	100%

Above tabulated results can be shown on graph illustrated below:

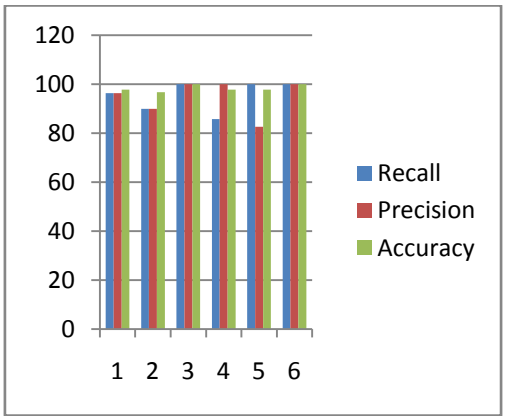

Fig. 3. Graph of Performance Measures

As shown in graph logistic regression has performed well than the other classifiers.

VI. CONCLUSIONS AND FUTURE WORK

This work means to build up a logistic regression model to classify dermatology diseases into various classes. This specific work is important considering the expanding number of electronic medical records in the databases. On the assessment of the model performance on the basis of the 178 records of the training data set and the 179 records of the testing data set, the model is able to detect classes with 98.36% accuracy including 94.85%. Future work will expand the real dataset and deal with dynamic issues, as new patient records appear every day.

REFERENCES

[1] Zhou, Xuezhong, Yonghong Peng, and Baoyan Liu. "Text mining for traditional Chinese medical knowledge discovery: a survey." *Journal of biomedical informatics* 43, no. 4 (2010): 650-660.
[2] Zong, Wei, Feng Wu, Lap-Keung Chu, and Domenic Sculli. "A discriminative and semantic feature selection method for text categorization." *International Journal of Production Economics* 165 (2015): 215-222.
[3] Mehra, Neha, and Surendra Gupta. "Survey on multiclass classification methods." (2013).
[4] Pereira, Luis, Rui Rijo, Catarina Silva, and Margarida Agostinho. "ICD9-based text mining approach to children epilepsy classification." *Procedia Technology* 9 (2013): 1351-1360.
[5] Pranoto, Hady, Fergyanto E. Gunawan, and Benfano Soewito. "Logistic Models for Classifying Online Grooming Conversation." *Procedia Computer Science* 59 (2015): 357-365.
[6] Jindal, Rajni, and Shweta Taneja. "A Lexical Approach for Text Categorization of Medical Documents." *Procedia Computer Science* 46 (2015): 314-320.
[7] "Introduction to XL-Miner" slideshare. 03 Feb. 2010.dataminingtools.<http://www.slideshare.net/dataminingtools/introduction-to-xlminer>.
[8] Babak danyal. "Swing and Graphical User Interface in Java" slideshare. 11 Apr. 2014. <http://www.slideshare.net/lineking/graphical-user-interface-33422176>.
[9] Kushima, Muneo, Kenji Araki, Muneou Suzuki, Sanae Araki, and Terue Nikama. "Text data mining of the electronic medical record of the chronic hepatitis patient." In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1. 2012.
[10] Guru, D. S., and Mahamad Suhil. "A Novel Term_Class Relevance Measure for Text Categorization." *Procedia Computer Science* 45 (2015):13-22.

AUTHORS

Neha Sharma, she is a research scholar in Madhav Institute of Technology and Science, Gwalior (M.P.), India under the supervision of Dr.R.K.Gupta. She has completed her bachelor degree in information technology from Institute of Technology and Management, Gwalior (M.P.), India and currently pursuing master of technology (M.Tech) degree in information technology. Research area s of her interest includes data mining, text mining and classification techniques etc.

Dr. R. K. Gupta, he is working as head of the department of computer science and information technology department in Madhav Institute of Technology and Science, Gwalior (M.P.) India .He has received PhD degree from ABV-IIITM Gwalior (M.P.) India .He has completed his post graduation (M.Tech) from IIT Delhi, India and bachelor degree (B.E.) from Madhav Institute of Technology and Science Gwalior (M.P) India. He has many years of teaching experience and guided many Ph.D. students as well as M.Tech students. Numbers of research paper has been published by him in data mining .His areas of interest are data mining, web mining etc.