Journal of Computer Engineering & Information Technology

Perspective

A SCITECHNOL JOURNAL

Efficiency of Prediction Algorithms for Mining Biological Databases

Sandra Vahini*

Department of Computer Engineering, University of Busitema, Busitema, Uganda

***Corresponding Author:** Sandra Vahini, Department of Computer Engineering, University of Busitema, Busitema, Uganda. E-mail: vahinivik68@gmail.com

Received date: 23 August, 2022, Manuscript No. JCEIT-22-61687;

Editor assigned date: 25 August, 2022, Pre QC No. JCEIT-22-61687(PQ);

Reviewed date: 31 August, 2022, QC No. JCEIT-22-61687;

Revised date: 16 September, 2022, Manuscript No. JCEIT-22-61687 (R);

Published date: 29 September, 2022, DOI: 10.4172/jceit.1000250

Description

The inception of Information technology in all the spheres of human life, many new fields of study emerged. Bioinformatics is a field which has attracted many researchers because of its importance and influence in the modern world. The introduction of computer science in the medical science has brought remarkable changes in the old practices used in bioinformatics. It has introduced many new aspects in the medical sciences like Genomics, Proteomics, and Cell Biology. These concepts existences were difficult to believe in the past. Mining biological databases are one such option which transformed the medical science from its traditional reactive methods to modern proactive approach. The biological databases contain information about the life sciences and are enriched with the research based knowledge of proteomics, metabolomics, genomics, phylogenetic and microarray sciences. It also contains information about genes and protein sequence. The information present in the database is the real and true information though it can be current or old. In other words, the databases contain both historical data and the current facts. The information is saved in a predefined format. Mining methods have enabled the experts to extract information from the database in a way that it provides altogether a new set of information. This information is the one which cannot be easily observed by human beings in generalized reports. This is why, data mining is also known as Knowledge Discovery in Database.

Mining Biological Databases

The new set of information assists the decision makers and researchers in focusing the unexplored pattern or information of the field. Mining takes place with the help of artificial intelligence i.e. which explores information to extract new patterns. The extracted information is real but presents a perspective which is not considered so far. Mining biological databases is important for prediction of the likely biological patterns in the species. It helps the medical science identify the risk of potential diseases in the generation to come. It helps researchers to prepare the treatment and medication for the likely disease. It is a proactive approach which can save the generations from painful sufferings of diseases whose treatment is made possible only through the prediction practices. In case the predictions are not made, the medical practitioners will not be able to diagnose the problem in patient, the prescription may be the wrong one and the condition of patient is bound to worsen. Mining analyses information in a unique way that helps in prediction about future trends and likely patterns to occur. The major point of concern is about the validity if predicted information, which is largely determined by the efficiency of prediction algorithm. In Bioinformatics, prediction is made through a well-defined series of steps known as algorithm. The algorithm contains the set of rules and procedures which are followed in the information analysis process. In order to predict reliable information, it is mandatory that the underlying algorithm takes into consideration all the possible variables and their interactions which can cause changes in the predicted behavior.

Prediction is usually based on historical data; certain statistical methods are used in analyzing the data. The statistical methods are embedded in the algorithms and they increase the efficiency of prediction algorithm. For example, correlation throws light upon the type of association between variables and the strength of that relationship but it does not reveal the causality. In mining biological databases, where major health related developments are to be made based on the predictions, it is important to measure the causality as well. By using the linear regression model, the prediction is made easy but the accuracy of predicted value is highly questionable due to approximations and assumptions. The prediction on the basis of one variable tends to a misleading prediction. In reality, there are always multiple independent variables leading to change in the values of dependent variables. It is crucial to analyse the combined effect of the variables involved so that the most influential one can be controlled which warrants for a better statistical model like multiple correlation and multiple regression analysis to predict the current and future trends. The method usually employed to predict the value of an unknown variable is the regression analysis. The prediction is made on historical data only and it may be possible that the value predicted may be misleading. If a prediction model is applied to generate the future trends and the validity of the predicted values increase then the algorithm is considered as efficient. The prediction model includes the factors like environment, possible changes in the past data and the context in which the predicted values will be used. For better prediction, the context in which past data is gathered also plays a vital role.

Analysis of Testing Data

The ambient intelligence analyses the data accordingly and presents the trend which is workable for future decisions. The prediction model contains the built in mechanism to create a sequence of value observed over the period of time. The values are analyzed with respect to the various identified variables. The algorithm with ambient intelligence uses multiple dimensions of value analysis. The algorithm creates multiple sequences based on complex computational methodology and presents the report to the decision maker. The complexity of the algorithm also determines the result"s reliability. The analysis of sequential dataset is the most common feature of prediction algorithms applied in the present era. To support the claims made earlier we consider a case study in breast cancer. Thus the paper analyses the efficiency of predictive mining algorithm on the data set related to breast cancer. This data set has 286 instances described by 9 attributes + one class attribute. The set includes 201 instances of one class and 85 instances of another class. We consider two cases to make predictions –classification by considering data as categorical, classification by considering data as numeric and quantification is done by considering an identifier. The main aim of the first experiment

is to predict the class by finding the efficiency of different algorithms considering data as categorical. The experiment predicts whether the data being tested can be correctly predicted and classified as recurrence class or non-recurrence class. The experiment was done with five different algorithms namely Decision Tree, OneR, PART, JRip and ZeroR. Four different predictive methods – Cross validation, Percentage Split, Testing data and Training data are used in this analysis. For the cross validation the method used is 10 fold. While using percentage split the number of instances gets reduced to 97 since it uses the 2/3rd method. In the first part the mean absolute error for each of the algorithm was measured since it is one of the widely used statistics for regression.

Normally in medicine a false positive cause's unnecessary worry or treatment, while a false negative gives the patient the dangerous illusion of good health and the patient might not get an available treatment. In this case study True Positive denotes the number of no-recurrence events of cancer correctly classified as no recurrence events. False positive denotes the number of recurrence events of cancer incorrectly identified as no recurrence events. True negative denotes the number of recurrence events of cancer correctly identified as recurrence events of cancer. Using these values we calculate the different measures. The measures on which the prediction algorithm is analyzed are number of correctly classified and incorrectly classified instances, accuracy, sensitivity, specificity, positive predictive value (PPV), Negative predictive value (NPV). The meaning of each parameter is given below for immediate reference. Sensitivity measures the proportion of the actual positives that are correctly identified as true. Specificity measures the proportion of the negatives that are correctly identified. We note here that a perfect predictor should have 100% sensitivity and 100% specificity. A high PPV means that the predictor always classifies correctly but has to be taken with the value of NPV since predictive values are inherently dependent upon the prevalence.