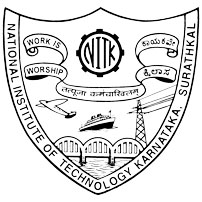
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA



Soft Computing Project Report

IT355

SUBMITTED TO: Dr. Nagamma Patil,  
 Assistant Professor,  
 Department of Information Technology,  
 National Institute of Technology Karnataka.

SUBMITTED BY: P Pavan [15IT224]  
 Lokaksha B M [15IT124]  
 Sachin K Rathod [15IT253]  
 Lakshmish M Gond [15IT215]

ACKNOWLEDGEMENT

# We wish to express our sincere gratitude to Dr. Nagamma Patil, Assistant Professor, Department of Information Technology, NITK for providing us an opportunity to work on a project on Soft Computing.

# We sincerely thank Mr. Sanjay Bankapur and Mr. Manjunath for their guidance and encouragement in carrying out this project work. We also wish to express our gratitude to all teaching officials who rendered help during the course of the project.

# DECLARATION

# We hereby declare that the project entitled “Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction” submitted by us in partial fulfillment for the completion of the course Soft Computing is a record of bonfied project work carried out by us under the guidance of Mr. Sanjay Bankapur. We further declare that the work done in this project will not be submitted either in full or part for the reward for degree or diploma in this institute or any other institute.

# 

# Signatures of the candidates:

# (P Pavan) (Lokaksha BM)

# (Sachin K Rathod) (Lakshmish M Gond)

# Place: Surathkal

# Date : 14/11/2017

TABLE OF CONTENTS

1. Abstract. [Page no: 6]
2. Introduction. [Page no: 7]
3. Sampling Method. [Page no: 8]
   1. Under Sampling. [Page no: 8]
   2. Over Sampling. [Page no: 9]
      1. Resample. [Page no: 9]
      2. Smote. [Page no: 9]
4. Multi-Layer Perceptron. [Page no: 10-12]
5. Data Description. [Page no: 12-13]
6. Result and Analysis. [Page no: 13-16]
7. Conclusion. [Page no: 16]
8. References. [Page no: 17]

LIST OF TABLES

Table 1: Data Description [Page no: 13]

Table 2: Multilayer perceptron with no sampling [Page no: 14]

Table 3: Multilayer perceptron with under sampling [Page no: 15]

Table 4: Multilayer perceptron with resample [Page no: 15]

Table 5: Multilayer perceptron with smote [Page no: 16]

# \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction

ABSTRACT:

Imbalanced data is an important problem for medical data analysis. Medical datasets are often not balanced in their class labels. The traditional classifiers can be seriously affected by the imbalanced class distribution in the data. This is because they aim to optimize the overall accuracy without considering the relative distribution of each class. This study searches the effect of class imbalance in training data when developing neural network classifier for medical decision making on chronic kidney disease. Neural networks are widely used in a number of applications including data mining and decision systems. Back propagation networks are a popular type of neural networks that can be trained to recognize different patterns. The importance of these networks was considered and a comparative study of some sampling algorithms was performed based on multilayer perceptron with different learning rate values for the prediction of chronic kidney disease.

1. Introduction

A well balanced dataset is very important for developing a good prediction model in classification research. Medical datasets are usually not balanced in their class distributions. An imbalanced data set contains a disproportionately high number of data in one or more classes than those for a class that is of interest. Traditional classification algorithms do not take into account the imbalance of class. They give the the same attention to the majority class and the minority class. The cost in mispredicting minority classes is higher than that of the majority class for many medical applications.  
Classification is an important task of data mining and knowledge discovery in databases. A range of classification modeling algorithms, such as decision tree, nearest neighbor and neural network have been developed and successfully applied to many domains. The assumptions built into most of these algorithms are: maximizing accuracy is the goal, and the classifier will operate on data drawn from the same distribution as the training data.  
Most original classification algorithms pursue to minimize the error rate: the percentage of the incorrect prediction of class labels. They ignore the difference between types of misclassification errors. In particular, they implicitly, assume that all misclassification errors cost equally. In many real world applications, the differences between different misclassification errors can be quite large.   
A comparison of some sampling algorithms was performed based on multilayer perceptron with different learning rate values for the prediction of chronic kidney disease. Chronic kidney disease or chronic renal failure, as it was historically a term that encompasses all degrees of decreased renal function, from damaged at risk through mild, moderate, and severe chronic kidney failure. Chronic kidney disease is a worldwide public health problem.  
The main aim is to help identify the challenges in imbalanced data problems in medicine and highlight the effects of learning rate parameter on multilayer perceptron model using back propagation algorithm.

1. Sampling Methods

The class imbalance problem is an important issue in many data mining applications. Standard machine learning algorithms fail to classify imbalanced data. Due to the goal of optimizing overall accuracy, these learning algorithms prefer training instances from the majority class, and therefore reducing the predictive accuracy on the minority class. Therefore, many researchers in machine learning area focused their attention on the class imbalance problem. There are different algorithms and techniques that handle the imbalanced data sets. At the data level, the objective is to rebalance the class distribution by re-sampling the data space.

Sampling is one of the important technique manage imbalanced data. The most direct ways for dealing with class imbalance is to alter the class distributions toward a more balanced distribution. These solutions include many different forms of re-sampling such as random over-sampling, random under-sampling, improved sampling approaches and combinations of these techniques. Random over-sampling is a non-heuristic method replicate examples of the minority class in order to achieve a more balanced distribution. Random under sampling is also a non-heuristic method aim to balance the data set by eliminating examples of the majority class.

1. Under Sampling.

Under sampling method removes examples from the majority class to make the data set balanced. This method tries to balance the distribution of class by randomly removing majority class samples. The drawback of under sampling method is that it can discard potentially useful information that could be important for classifiers. Under sampling methods are divided into random and informative. Random under sampling randomly eliminates examples from the majority class till the data set gets balanced. Informative under sampling method selects only the required majority class examples based on a pre-specifies selection criterion to make the data set balanced.

1. Over Sampling

Over sampling is a sampling approach which balances the data set by replicating the examples of minority class. The advantage of this method is that there is no loss of data in under sampling technique. The disadvantage of this technique is it may lead to over fitting and can introduce an additional computational cost if the data set is already fairly large but imbalanced. Like under sampling, oversampling is also divided into two types: random oversampling and informative oversampling. Random oversampling is the method which balances the class distribution by replicating the randomly chosen minority class examples. Random over sampling generates duplicated data without creating any new information and this method is the simplest approach to over sampling where members from the minority class are chosen at random; these randomly chosen members are then duplicated and added to the new training set. This technique duplicates instances in the minority class. Informative oversampling method synthetically generates minority class examples based on a pre-specifies criterion. In summary, over sampling may cause longer training time of over-fitting. The alternative to over sampling is under sampling. This approach is better than over sampling in terms of time and memory complexity. In this study, following algorithms are used for sampling:

1. Resample

This algorithm produces a random subsample of a dataset, sampling with replacement.

1. Smote

This algorithm creates artificial data based on the feature space similarities between existing minority examples. Chawla proposed Synthetic Minority Over-Sampling Technique (SMOTE) an over-sampling approach in which the minority class is over-sampled by creating synthetic examples rather than by over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.

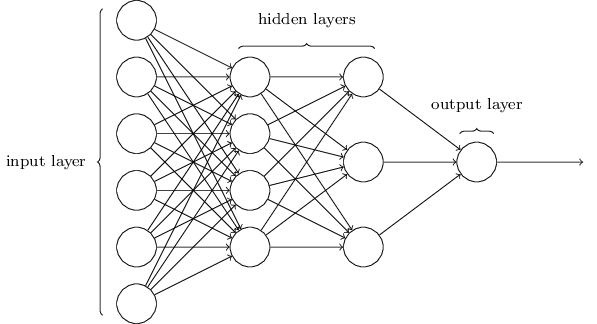
1. Multi-Layer Perceptron

Neural networks have emerged as a result of simulation of biological nervous system, such as the brain on a computer. Neural networks are represented as a set of nodes called neurons and connections between them. The connections have weights associated with them, representing the “strength” of those connections. Nowadays neural networks can be applied to problems that do not have algorithmic solutions or problems for which algorithmic solutions are too complex to be found. In other words, the kind of problems in which inputs and outputs variables does not have a clear relationship between them, a neural network models due to its clear architecture and comparably simple algorithm. Especially, back propagation networks are a popular type of neural network is an efficient approach in such problems. Most neural network architecture has three layers in its structure. First layer is input layer which provides an interface with the environment, second layer is hidden layer where computation is done and last layer is output layer where output is stored. Data is propagated through successive layers, with the final result available at the output layer. Many different types of neural networks are available and multilayer neural networks are the most popular. Multilayer neural networks popularity is due to more then one hidden layer in its structure which help sometimes in solving complex problems which a single hidden layer neural network cannot solve.

A node in multilayer perceptron can be modeled as an neuron which computes the weighted sum of the inputs at the presence of the bias and passes this sum through the activation function. The whole process is defined as follows:

Where vj is the linear combination of inputs Xl,X2,…,Xp, θJ is the bias, Wji is the connection weight between the input Xi and the neuron j, and fj(vj) is the activation function of the jth neuron and yJ is the output.

The sigmoid function is a common choice of the activation function, as defined:



[fig 1]

A multilayer perceptron is a feed forward neural network model that maps sets of input data onto a set of appropriate output. Back propagation is the most widely applied learning algorithm for multilayer perceptron in neural networks. Back propagation employs gradient descent to minimize the squared error between the network output values and desired values for those outputs. These error signals are used to calculate the weight updates which represent knowledge learnt in the networks. The performance of back propagation algorithm can be improved by adding a momentum term.

The basic back propagation algorithm works as follows:

* Initialize all the connection weights W with small random values from a pseudorandom sequence generator.
* Repeat until convergence (either when the error E is below a preset value or until the gradient ∂B(t)/∂W is Smaller than a preseny value).
  + Compute the update using
* Compute the error E(t+1)

Where t is the iteration number, w is the connection weight, and η is the learning rate. The error E can be chosen as the mean square error (MSE) function between the actual output y and the desired output dj;

The back propagation algorithm has some shortcomings. If the learning rate is set small enough to minimize the total error, the learning process will be slowed down. On the other hand, a larger learning rate may speed up learning process at the risk of potential oscillation. Another problem is that, partial minimal points or stable stages on error surface are often encountered during the learning process. Using a momentum term is the simplest method to avoid oscillation problems during the search for the minimum values on the error surface.

1. Data Description

The data in this study was taken from UCI Machine Learning Repository which is publicly available. It includes 400 patients with 25 attributes collected from each of these patients; 250 of them have Chronic Kidney Disease (CKD). The ages of these patients vary from 2 to 90 with mean of 51:483 and a standard deviation of 17:17. The characteristics of data set is shown in Table I. As seen the Table 1, the class distribution of the data set is imbalanced. It contains CKD (%62.5) and Not CKD(%37.5).

Table 1: Data Description

|  |  |  |
| --- | --- | --- |
| Number of Instances | 400 | |
| Number of attributes | 25 | |
| Class distribution | CKD (62.5%) | Not CKD (27.5%) |

1. Results and Analysis

Chronic Kidney Disease data set was used to compare different sampling methods for the prediction of disease. Multilayer perceptron was selected to evaluate classification accuracy. Respectively, Resample, SMOTE and Spread Sub Sample algorithms were used for sampling. Same experiment was run in incremental steps by varying the learning rate between 0.1 and 0.8. WEKA 3.7.3 software was used. WEKA is a collection of machine learning algorithms for data mining tasks and is open source software. The software contains tools for data pre-processing, feature selection, classification, clustering, association rules and visualization.

There are many performance measures for the evaluation of the classification results, where *TP/TN* is the number of True Positives/Negatives instances, *FP/FN* is the number of False Positives/Negatives instances but some of them are used in this study. Precision is a proportion of predicted positives which are actual positive:

Recall is a proportion of actual positives which are predicted positive:

Precision and recall measures are utilized to find the best method, but it is not easy to make decision. Thus, F-measure was used to get a single measure to evaluate results. The F-measure is the harmonic mean of precision and recall:

The comparison analysis by root mean squared error was also performed and described in Table III–VI where n is the number of data patterns, ypm indicates the predicted, tmm is the measured value of one data point m and t¯mom is the mean value of all measure data points. Root Mean Squared Error (RMSE) can be written as follows:

Table 2: Multilayer perceptron with no sampling

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Learning Rate** | **Precision** | **Recall** | **F – Measure** | **RMSE** |
| 0.1 | 0.993 | 1.0 | 0.996 | 0.0577 |
| 0.3 | 0.882 | 1.0 | 0.937 | 0.2581 |
| 0.5 | 0.655 | 1.0 | 0.791 | 0.5131 |
| 0.8 | 0.375 | 1.0 | 0.545 | 0.9128 |

Table 3: Multilayer perceptron with under sampling

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Learning Rate** | **Precision** | **Recall** | **F – Measure** | **RMSE** |
| 0.1 | 0.993 | 1.0 | 0.996 | 0.0577 |
| 0.3 | 0.986 | 1.0 | 0.993 | 0.0816 |
| 0.5 | 0.684 | 1.0 | 0.813 | 0.4795 |
| 0.8 | 0.5 | 1.0 | 0.666 | 0.7071 |

Table 4: Multilayer perceptron with resample

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Learning Rate** | **Precision** | **Recall** | **F – Measure** | **RMSE** |
| 0.1 | 0.996 | 1.0 | 0.998 | 0.0577 |
| 0.3 | 0.988 | 1.0 | 0.994 | 0.1 |
| 0.5 | 0.5 | 1.0 | 0.666 | 0.9128 |
| 0.8 | 0.5 | 1.0 | 0.666 | 0.9128 |

Table 5: Multilayer perceptron with smote

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Learning Rate** | **Precision** | **Recall** | **F – Measure** | **RMSE** |
| 0.1 | 0.993 | 0.994 | 0.932 | 0.1511 |
| 0.3 | 0.853 | 0.947 | 0.898 | 0.3696 |
| 0.5 | 0.735 | 0.961 | 0.833 | 0.4831 |
| 0.8 | 0.364 | 1.0 | 0.533 | 0.9295 |

1. Conclusion

Imbalanced data set is common problem with most medical data sets. Most of machine learning algorithms cannot handle imbalanced class distribution and sampling algorithms play important role in classification accuracy. In this study, a comparative experiment was carried out on sampling algorithms to predict chronic kidney disease. The effect of class imbalance in training data on performance was evaluated for multilayer perceptron. Three sampling algorithms were used to analyze the data set and their performance was evaluated by multilayer perceptron by varying learning rate. The results were evaluated on accuracy metrics and execution time. Among the sampling algorithms, Resample method has better accuracy results on the data set than the others. In terms of execution time. The results highlight that sampling can significantly effect on the performance of classification studies and the learning rate parameter has to be carefully selected for any multilayer perceptron classification problem to obtain higher accuracy. In conclusion, this study makes contributions in medical data mining studies and reveals that sampling algorithms can improve the performance of multilayer perceptron with optimum learning rate parameter for learning process.

1. References

**Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction**[*Pinar Yildirim*](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Pinar%20Yildirim.QT.&newsearch=true) *Fac. of Eng., Dept. of Comput. Eng., Okan Univ., Istanbul, Turkey*

<http://ieeexplore.ieee.org/document/8029917/>

<https://en.wikipedia.org/wiki/Multilayer_perceptron>