Improved Ant Colony on Feature Selection and Weighted Ensemble to Neural Network Based Multimodal Disease Risk Prediction (WENN-MDRP) Classifier for Disease Prediction Over Big Data

Gakwaya Nkundimana Joel1*, S. Manju Priya2

1Research Scholar, Dept. Computer Science, Karpagam Academy of Higher Education.
2Associate Professor, Dept. of CS, CA & IT, Karpagam Academy of Higher Education.
E-mail: smanjupri@gmail.com
*Corresponding author E-mail: gajoel2000@gmail.com

Abstract

As the big data is growing in biomedical and healthcare communities, so are precise analyses of medical data aids, premature disease identification, patient care as well as community services. On the other hand, the accuracy of the analysis decreases, if the medical data quality is imperfect. As a result, the choice of features from the dataset turns out to be an extremely significant task. Feature selection has exposed its efficiency in numerous applications by means of constructing modest and more comprehensive models, enlightening learning performance and preparing clean and clear data. The proposed method analyzes the difficulties of feature selection for big data analytics. Improved Ant Colony Optimization based Feature Selection (IACO) algorithm is presented for resolving this issue. The reconstruction of missing data before the incomplete data available was performed with help of latent factor mode. Therefore, it was not easy to choose the best features from the structured and unstructured data. the unheard technique which is called Weighted Ensemble Based Neural Network for multimodal disease risk prediction (WENN-MDRP) algorithm is implemented in order to provide the best features selection among structured as well as unstructured data. The research method provides improved prediction accuracy when matched with conventional techniques. In the MATLAB environment, the presented classifiers are implemented. The outcomes are computed in regard to recall, precision, accuracy, f-measure and error rate.

Keywords: Disease prediction, machine learning, ensemble, neural network, prediction accuracy.

1. Introduction

50% of Americans contain one or more chronic diseases, 80% of American medical care fee is expended on chronic disease treatment in keeping with a report by McKinsey [1]. The occurrence of chronic disease is rising with the enhancement of living standards. On chronic disease treatment, the US has expended an average of 2.7 trillion USD yearly. This amount encompasses 18% of the complete annual GDP of the US. The healthcare issue of chronic diseases is as well extremely significant in numerous countries. Chronic diseases are the foremost reason of death in China.86.6% of deaths are occurred by chronic diseases in keeping with a Chinese report on nutrition as well as chronic diseases in 2015. So, it is vital to carry out risk assessments for chronic diseases. Gathering Electronic Health Records (EHR) [3] is progressively expedient with the development in medical data [2].

More consideration is paid to disease prediction with the growth of big data analytics technology. With the aim of enhancing the accurateness of risk classification, numerous researches were carried out by means of choosing the characteristics robotically from a huge amount of data [4-5], instead of the formerly chosen characteristics. At the present time, more than a few machine learning techniques [6] are implemented for getting the precious knowledge from enormously big data in disease prediction. On the other hand, they are unsuccessful while faces varied nature of feature values and that previous work typically took structured data. As data is of massive as well as complexity naturally, taking out precious knowledge from big data is not simple with conventional data was processing as well as data handling methods. Henceforth, for dealing with big data to get the valuable model, the growth of progressive technologies is needed. Deep as well as machine learning are the finest solution fortaking out beneficial knowledge as well as make suitable decisions from big data, [7]. The multiple or ensemble classifier is the mixtures of a diverse classifiers called base classifiers [8]. Typically, ensemble classifier gives improved classification accurateness compared to the single classifiers. Present years, use of deep learning quickly rising in numerous areas. In deep learning models, the neural network acts as a big role. This work presents the mixture of neural network as well as ensemble learning techniques in the big data environment. Previous data processing techniques are not suitable to deal with, examine and process. It is because of the complexity, unlabeled data and Un categorized. Therefore, in examining such enormous volumes of unsupervised data, the deep learning techniques acts as a vital role, which are suitable for exploiting huge amount of data and for examining raw data from numerous sources as well as indifferent styles [9]. Feature selection has shown its effectiveness in numerous applications by means of constructing modesties well as more
inclusive models, enlightening learning performance, and making clean, comprehensible data. In this research work, know about the challenges of feature selection for big data analytics. In order to choose features Improved Ant Colony Optimization based Feature Selection (IACO) technique is presented for resolving this issue. Previous to that firstly incomplete data, utilize a latent factor model to rebuild the missing data. IACO based feature selection for best selection of features from healthcare data. As well a novel Weighted Ensemble Based Neural Network based multimodal disease risk prediction (WENN-MDRP) technique is presented with the help of structured as well as unstructured data from hospital. No other previous method concentrated on both data types in the field of medical big data analytics. The remaining paper is organized in this manner. Section II provides the related work, and Section III elucidates the proposed method. Section IV offers the experimentation design and analysis. Section V summarizes the conclusion part.

2. Literature Review

The machine learning as well as data medical field have developed early detection systems. Subhapiya et al [8] presented the improved data mining method for healthcare application. The system encompasses predominantly three steps those are as follows anomaly detection, clustering, and classification. Classification algorithm utilizes the ensemble technique known as a random forest. This system precisely foresees the patient outcome from a noteworthy volume of data. Kathleen et al [10] presented an ensemble machine learning technology, which uses an adaptive Boosting algorithm implemented for precise heart disease prediction results. The research techniques used to 4 diverse data sets for heart disease diagnosis. Datasets comprise heart disease from Hungarian Institute of Cardiology (HIC), Cleveland Clinic Foundation (CCF), Switzerland University Hospital (SUH) and Long Beach Medical Center (LBMC). The outcomes of the novel algorithm proved that the performance of novel ensemble techniques is superior to previous techniques.

In [11], the disease risk prediction is expressed into a multilabel classification problem by Li et al. A new Ensemble Label Power-set Pruned datasets Joint Decomposition (ELPPJD) technique is presented in this research work. Initially, the multi label classification is transferred into a multiclass classification problem. As well recommended the pruned datasets as well as joint decomposition approaches to handle the imbalance learning problem. ELPPJD technique offers improved outcomes in the experimentation outcomes.

Saxena and Sharma [12] illustrated a structure, which guesses the risk level of patients. In order to make the accurate decision regarding the heart disease risk level, they concentrated on non-specialized Doctors. The research method focuses on rule generation parameters. The rules comprise Pruned Rules, Original Rules, Rules without duplicates, Sorted Rules, Classified Rules, and Polish.

In [13], Jaseena and Kovoor proposed a general idea of diverse deep learning methods for big data in biometrics and converses some problems and solutions. As well complete survey of all the associated work in deep learning for biometric big data is presented. In [14], Farid et al presented an ensemble learning in which a novel method that imparted boosting as well as decision tree classifier together. AdaBoost algorithm is utilized as the algorithm for boosting ensemble. Separate decision tree technique is used for each sample, and those outcomes are brought up to date. The attained cases were again categorized with the help of the voting technique.

In [15], Chen et al introduced a novel machine learning technique for valuable prediction of disease. With the help of structured and unstructured data, anovel Convolutional Neural Network (CNN) based multimodal disease risk prediction technique is explained. The uses for instance computer vision, speech recognition, and Natural language processing, its challenges, open research problems in addition to the future works are provided.

In [16], Chen et al presented design details, significant technologies and practical implementation techniques of smart clothing system. Distinctive applications powered by smart clothing as well as big data clouds are provided, for instance emotion care, medical emergency response, disease diagnosis, as well as real-time tactile interaction. Particularly, electrocardiograph signals gathered by smart clothing are utilized for mood monitoring as well as emotion identification. Lastly, as well highpoint certain design challenges as well as open problems, which must be stated to create smart clothing ubiquitous for an extensive range of applications.

In [17], a novel knowledge-based system was presented by Bates et al for diseases prediction with the help of clustering, noise removal, and prediction methods. As well, in order to produce the fuzzy rules to be utilized in the knowledge-based system, Classification and Regression Trees (CART) is utilized. The research technique was developed on numerous public medical datasets. Outcomes on Mesothelioma, Pima Indian Diabets, StatLog, WDBC, Cleveland and Parkinson’s telemonitoring datasets prove that there search technique enhances the diseases prediction accurateness. The knowledge-based system could aid medical practitioners in the healthcare practice as a medical analytical technique.

In [18], Qiu et al concentrated on the issue of data sharing problems in cloud computing as well as apresumptmethod, which utilizes dynamic programming with the aim of producing the best possible solutions to data sharing techniques. The research method is known as Optimal Telehealth Data Sharing Model (OTDSM) that takes transmission probabilities, increasing network capacities, and timing restraints. The experimentation outcomes have shown the flexibility as well as adoptability of the research technique.

In [19], Zhang et al presented a Cyber-Physical System (CPS) for patient-centric healthcare applications and services, known as Health-CPS, constructed on cloud and big data analytics technologies. This system encompasses a data collection layer with a unified standard, a data management layer for distributed storage as well as parallel computing, in addition to a data-oriented service layer. The outcomes of this research prove that the technologies of cloud as well as big data are utilizing for improving the performance of the healthcare system with the intention that humans could then enjoy numerous smart health care applications as well as services. In [20], Gakwaya N. Joel et reviewed on role of big data in healthcare as the pivot key to ameliorate the health of nations, and the use of big data in order to discover the medication for fighting certain diseases. The suggested technology such as hadoop freemarwok showed the advantage in analyzing huge data and storing those numerous data.

In [21], a new method was presented by Nori et al. Efficiently incorporated medical domain knowledge associating with the similarity amongst diseases as well as the similarity amongst Electronic Health Records (EHRs) into a data-driven method by means of integrating graph Laplacians into the regularization term to encrypt these similarities. The experimentation outcomes on a real dataset from a hospital validated the efficiency of the research technique. In addition to some results for disease predictive features which were highly reliable as per medical techniques additional analysis in the medical domain is still needed. In [22], introduced a machine learning technique by Bandypadhyay et al dependent upon Bayesian networks trained on EHD to foresee the likelihood of containing a CV event within 5 years. In these data, event status might be unknown for certain individuals, since the event time is right-censored because of disenrollment as well as imperfect follow-up. Outcomes prove that the research technique could result in improved predictive performance compared to the Cox proportional hazards model (that is to say a regression-based
method generally utilized for censored, time-to-event data) or a Bayesian network with ad hoc techniques to right-censoring. The research methods are inspired by as well as exemplified on data from a huge US Midwestern health care system.

3. Proposed System Methodology

Research technique focuses on feature selection as well as improved classification accurateness for healthcare data. In order to choose the important features, Improved Ant Colony Optimization based feature selection (IACO) method is presented in this section. After that a novel weighted ensemble based neural network disease prediction technique is presented for the structured as well as unstructured data from the hospital. Uniting ensemble as well as Convolutional Neural Networks (CNN) is probable by means of utilizing CNN as weak learners in the ensemble algorithm. Weights are attuned dependent upon the misclassification.

Improved Ant Colony Optimization Based Feature Selection (IACO) Algorithm

In any machine learning related task, Feature selection is an important process that identifies a subset of features, from the original feature set. ACO technique finds the best possible feature subset utilizing certain iterations. This research work uses an Improved Ant Colony Optimization Based Feature Selection (IACO) Algorithm for decreasing the redundancy.

Ensemble Algorithm

Ensemble algorithms uniting various collection of individual models together to enhance the predictive accurateness.

Weighted Ensemble Based Neural Network in Big Data

Combination ensemble algorithm with CNN is utilized in the research technique. Furthermore, this is developed in the big data environment using Hadoop – MapReduce.
4. Experiment Design and Analysis

In this part, define the hospital datasets utilize in this research. Also, offer disease risk prediction model as well as evaluation techniques. The data is gathered from UCI machine learning repository. It encompasses 100000 instances as well as 55 attributes. The data set denotes 10 years of clinical care at 130 US hospitals. It comprises over 50 features denoting patient as well as hospital results. Information was taken out from the database for encounters, which fulfilled the subsequent condition.

1. It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, that is, one for the period of which any type of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day as well as at most 14 days.
4. Laboratory tests were carried out for the period of the encounter.
5. Medications were administered in the course of the encounter.

The data encompasses attributes for instance patient number, age, gender, race, admission type, medical specialty of admitting physician, time in hospital, number of lab test carried out, diabetic medications, diagnosis, HbA1c test result, number of medication, and number of outpatient, inpatient, and emergency visits in the year beforehand the hospitalization, and so on.

For dataset assessment these traditional machine learning techniques, that is to say Naive Bayesian (NB), Convolutional Neural Network based Multimodal Disease Risk Prediction (CNN-MDRP), K-nearest Neighbour (KNN), and Decision Tree (DT) algorithm [23-24] are utilized to foresee the risk of cerebral infarction disease. This is for the reason that these three machine learning techniques are extensively utilized [25].

For the performance assessment, initially, represent TP, FP, TN and FN as true positive (the amount of instances appropriately predicted as needed), false positive (the amount of instances inaccurately predicted as needed), true negative (the amount of instances appropriately predicted as not needed) and false negative (the amount of instances imperfectly foreseen as not needed), correspondingly. After that, could get four measurements: precision, accuracy, recall and F1-measure in this manner:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \\
\text{Precision} = \frac{TP}{TP + FP} \\
\text{Recall} = \frac{TP}{TP + FN} \\
\text{F1-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

here the F1-Measure is known as the weighted harmonic mean of the precision as well as recall and denotes the complete performance. Along with the afore stated assessment condition, as well utilize Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC) for assessing the advantages and disadvantages of the classifier. The ROC curve proves the trade-off amid the True Positive Rate (TPR) and the False Positive Rate (FPR), in which the TPR and FPR are defined in this manner:

\[
\text{TPR} = \frac{TP}{TP + FN} \\
\text{FPR} = \frac{FP}{FP + FN}
\]

The model is good when the ROC curve is nearer to the upper left corner of the graph. The AUC is the area under the curve. The model is better while the area is closer to 1. In medical data, we concentrate on the recall instead of accuracy. The greater the recall rate, the lesser the probability that a patient who would contain the risk of disease is identified to have no disease risk. In Table 1, the performance of numerous familiar techniques on disease prediction is stated, and it proves that presented WEBNN algorithm contain improved accuracy compared to other provided classifiers.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>NB</th>
<th>DT</th>
<th>KNN</th>
<th>CNN-MDRP</th>
<th>WENN-MDRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>88.05</td>
<td>88.41</td>
<td>91.46</td>
<td>95.09</td>
<td>97.05</td>
</tr>
<tr>
<td>Recall</td>
<td>87.50</td>
<td>90.06</td>
<td>90.36</td>
<td>92.81</td>
<td>94.16</td>
</tr>
<tr>
<td>F-measure</td>
<td>87.77</td>
<td>89.23</td>
<td>90.91</td>
<td>93.94</td>
<td>95.58</td>
</tr>
<tr>
<td>Accuracy</td>
<td>80.50</td>
<td>82.50</td>
<td>85.00</td>
<td>90.00</td>
<td>92.70</td>
</tr>
<tr>
<td>Error rate</td>
<td>19.50</td>
<td>17.50</td>
<td>15.00</td>
<td>10.00</td>
<td>7.30</td>
</tr>
</tbody>
</table>

In the figure 1(a) the classifiers for instance NB, DT, KNN, CNN-MDRP and WENN-MDRP yields precision outcomes of 88.05%, 88.41%, 91.46%, 95.09% and 97.05% correspondingly. It yields recall outcomes of 87.50%, 90.06%, 90.36%, 92.81% and 94.16% correspondingly for classifiers. According to the figure 1(a) presented WENN-MDRP classifier yields greater precision outcomes of 9%, 8.64%, 5.59% and 1.96% while matched up with NB, DT, KNN, and CNN-MDRP classifiers correspondingly. According to figure 1(a) presented WENN-MDRP classifier yields
greater recall outcomes of 6.66%, 4.1%, 3.8% and 1.35% while matched up with NB, DT, KNN, and CNN-MDRP classifiers correspondingly.

![F-measure and accuracy under NB, DT, KNN, CNN-MDRP and WENN-MDRP](image1)

Figure 1 (a) – (c) exposes the performance comparison outcomes of five diverse classifiers under classification metrics. For example in figure 1(a) exposes the performance comparison outcomes of precision and recall metrics amid five diverse classifiers. In the figure 1(b) the classifiers for instance NB, DT, KNN, CNN-MDRP and WENN-MDRP yields f-measure outcomes of 87.7%, 89.2%, 90.9%, 93.9% and 95.8% correspondingly. It yields accuracy outcomes of 80.5%, 82.5%, 85.0%, 90.0% and 92.7% correspondingly for classifiers. According to the outcomes it proves that the presented classifier contains 7.8%, 6.3%, 4.6% and 1.6% greater f-measure outcomes while matched up with NB, DT, KNN, CNN-MDRP and WENN-MDRP classifiers. According to the outcomes it proves that the presented classifier contains 12.2%, 10.2%, 7.7% and 2.7% greater accuracy outcomes while matched up with NB, DT, KNN, CNN-MDRP and WENN-MDRP classifiers.

![Error rate comparison under NB, DT, KNN, CNN-MDRP and WENN-MDRP](image2)

Figure 1(c) exposes the error rate outcomes of all classifiers for instance 19.50%, 17.50%, 15.00%, 10.00% and 7.30% correspondingly. According to the outcomes it is clear that the presented WENN-MDRP classifier yields minimum error rate as well as greater accuracy outcomes while matched up with all classifiers.

![ROC curves under NB, DT, KNN, CNN-MDRP and WENN-MDRP classifiers](image3)

According to Figure 2, it is clear that the equivalent AUC of NB, DT, KNN, CNN-MDRP and WENN-MDRP are 0.51, 0.65, 0.71, 0.85 and 0.88, correspondingly. In summary, for data, the WENN-MDRP classification is the finest in experimentation[26]. On the other hand, it is as well noticed that presented classifier has precisely foreseen whether the patient is in a greater risk of diabetes in keeping with the patient’s gender, age, clinical laboratory and other structured data. [27]

5. Conclusion and Future Work

Presently, people are often distressed by numerous diseases. Day by day, the number of patients distressed by certain infection is rising. In order to choose features, Improved Ant Colony Optimization based Feature Selection (IACO) algorithm is presented. Previous to that in the beginning incomplete data, a latent factor model was utilized in order to rebuild the missing data. These two techniques namely IACO based feature selection for best selection of features from healthcare data and Weighted Ensemble Based Neural Network for multimodal disease risk prediction (WENN-MDRP) methods are combined in the big data medical field to improve the health. The experimentation carried out on structured as well as unstructured data in order to choose the important features for prediction. This method is mainly concentrated on feature selection parameter. Outcomes of the novel method prove that it outdoes previous techniques. The improved outcomes of this method aid the specialists to get organized for a greater determine as well as offer the patient to contain initial determination outcomes. According to the outcomes the presented classifiers yields greater accuracy outcomes of 92.7%, while other classifiers for instance NB, DT, KNN, and CNN-MDRP yields accuracy outcomes of 80.50%, 82.50%, 85.00%, 90.00% and 92.70% respectively. It shows that the research technique is superior matched with other classifiers. When matched with other distinctive classifiers, the classification accuracy of presented WENN-MDRP algorithm attains 92.70% with a convergence speed that is sooner compared to that of the CNN-MDRP algorithm. In future, work will concentrate on using the same algorithm to other disease prediction. Numerous optimization approaches are utilized to choose the best possible features also expanding the scope of future work.
References


