Grasshopper Optimization with Stacked Autoencoder based Data Mining Approach for Healthcare Sector

A. Subashini
Assistant Professor, Department of Computer Application, Government Arts College, C. Mutlur, Chidambaram
E-Mail: subanandh31@gmail.com

Abstract

Data mining (DM) is the procedure of pattern discovery and abstraction where massive quantity of data is involved. It finds its applicability in several areas, particularly healthcare. The use of DM in the healthcare industry enables to design of effective early disease diagnosis models based on the available medical data. The recently developed machine learning (ML) models pave the way to design effective DM models for healthcare sector. This article introduces a grasshopper optimization with stacked autoencoder based data mining (GOSAE-DM) technique for healthcare sector. The proposed GOSAE-DM technique primarily aims to handle the class imbalance problem and then performance classification process. The GOSAE-DM technique involves ADASYN technique for handling class imbalance issues. Besides, SAE model is applied for the classification of balanced data, which determines the existence of the disease. In order to boost the classifier results of the SAE model, the GOA is applied and thereby enhances the classifier results. A wide range of experiments was carried out using benchmark medical dataset and the results are inspected under several dimensions. The comparative study reported the betterment of the GOSAE-DM technique over the recent techniques in terms of different measures.

Keywords: Data mining, Healthcare, Class imbalance problem, Stacked autoencoder, Grasshopper optimization, Machine learning

1. Introduction

In every field, information is accumulated and collected at a vivid pace. There is an urgency for a new generation of computation tools and theories to help individuals in extracting beneficial data (knowledge) from the quickly expanding amount of digital information. The core process is the application of certain data mining (DM) techniques for pattern extraction and discovery [1]. Amongst the DM methods proposed recently, the DM method includes characterization, generalization, clustering, classification, evolution, association, meta-rule guided mining, data visualization, and pattern matching. [2]. There is considerable scope for
DM application in medical field. In general, it is classified into management of healthcare; the calculation of treatment effectiveness; detection of abuse and fraud; and customer relationship management. The DM application is designed for evaluating the efficiency of medicinal treatment. By contrasting and comparing symptoms, courses, and causes of treatment, DM method could transport an analysis where course of action proves efficacious. To help health care organizations, DM application is designed for effectivity tracking and identifying high-risk patients, and chronic disease states, designing a suitable intervention, and reducing amount of hospital claims and admissions [3].

Information technology has been implemented gradually in health care organizations for addressing the requirements of doctors in operational decision-making activity. The Machine learning (ML) technique assist in primary care and making decisions in emergency situations. Furthermore, ML methods are utilized for helping physician diagnoses patient, particularly in case when results are barely to choose and predict optimal process technique [4]. Nowadays, ML method in health care is inevitable. Optimist predicts that artificial intelligence and ML diagnoses disease earlier and better, engage patients more efficiently, and treat illness more precisely in future health care [5]. Current advancements in ML method have established that ML could construct algorithm which implements on par with physicians. Recently, health care instances namely predicting healthcare operational decisions, detection and management of prostate cancer, dosage trials for intravenous tumor treatment, medical image processing, and analyzing [6]. The ML method emphasizes the expansion of computation algorithm that could access data for looking patterns in data and makes good calculations regarding upcoming instances on the basis of instance. The ML method is named supervised if known label is provided with instance in the training stage, where instance is unlabelled in unsupervised ML method [7].

Kanakaraddi et al. [8] highlight the usage of DM methods, such as DT, NB, RF, and LR, to detect brain tumor and cancer, through ML capability on the benchmark database from Kaggle. The efficacy measure of recognition with DM approaches is completely acceptable. Dahiwade et al. [9] presented disease diagnosis-based symptoms of the person. For predicting the disease, the study employs Convolutional neural network (CNN) and K-Nearest Neighbor (KNN) method for predicting diseases accurately. In order to predict disease, disease symptom datasets are required. With that regard, disease diagnosis is the living habit of individual, and check-up data is considered for the precise diagnosis. Mohan et al. [10] developed an approach which focuses on detecting considerable features by employing ML methods results in enhancing the
precision in the diagnosis of heart diseases. The predictive method is presented by using distinct integrations of features and many known classification methods. Si et al. [11] implement research that investigates the metaheuristic performances as learning algorithm for training the ANN for medicinal data classification tasks. The experiment is conducted on fifteen familiar medicinal data sets. Khan et al. [12] designed a higher-order Functional Link Neural Network (FLNN). The presented method is combined with a meta-heuristic-based searching method named Accelerated Particle Swarm Optimization (APSO). The efficiency of the presented ASPO Functional Link Neural Network (APSOFLNN) method is validated by several medicinal data sets.

This article introduces a grasshopper optimization with stacked autoencoder based data mining (GOSAE-DM) technique for healthcare sector. The proposed GOSAE-DM technique primarily aims to handle the class imbalance problem and then performance classification process. The GOSAE-DM technique involves ADASYN technique for handling class imbalance issues. Besides, SAE model is applied for the classification of balanced data, which determines the existence of the disease. In order to boost the classifier results of the SAE model, the GOA is applied and thereby enhances the classifier results. A wide range of experiments was carried out using benchmark medical dataset and the results are inspected under several dimensions.

![Fig. 1. Overall process of GOSAE-DM technique](image-url)
2. The Proposed Model

In this study, a novel GOSAE-DM technique has been developed to handle the class imbalance problem and then performance classification process in the healthcare sector. The GOSAE-DM technique involves different stages of operations such as ADASYN based class imbalance data handling, SAE based classification, and GOA based parameter optimization. Fig. 1 demonstrates the overall process of GOSAE-DM technique.

2.1. ADASYN based Class Imbalance Data Handling

The fundamental model of ADASYN oversampling technologies are to define the weighted distributions of minority sample with considering the degree of learning difficulty of minority sample. In order to minority samples with learning difficulty, further novel instances are created along with them [13]. To binary classifier issue, the dataset \(D_{tr}\) of \(m\) instances are written as \({x_i, y_i}\), where \(i = 1, ..., m\), \(x_i\) refers the instance of \(n\) dimensions feature space \(X\), and \(y_i\) denotes the label of instance \(x_i\), \(y_i \in Y \{ -1, 1 \}\). The amount of majority sample was \(m_1\), and the amount of minority sample was \(m_s\). At this point, \(d \in (0,1]\) refers the degree of imbalances from the dataset. The lesser value of \(d\) is, the further imbalance data is, and further, the value of \(d\) attends 1, the additional balance data is. \(d_{th}\) denotes the set maximal tolerated imbalance thresholds. If \(d_{th}\) is 1, it refers that it is only be recognized when the amount of instances from distinct types are equivalent. \(\beta \in [0,1]\) is the parameter utilized for setting the chosen balanced degree of synthetic data set than creating sample. When \(\beta = 1\), the dataset to create a novel instance is entirely balanced, i.e., the amount of samples from distinct sets is similar. \(K\) signifies the parameter to define KNN. To create instance set \(S\) return by technique, it is combined with novel dataset \(D_{tr}\) as to a novel dataset as novel trained set of techniques. This technique creates further novel instances from the regions in which learning was complex to minority samples that are effectual strengthen the model learning of minority samples, so enhancing the model rate of detection of the minority samples forecast.

2.2. SAE based Classification

Once the dataset is balanced, the next process is to classify the data into appropriate class labels. AE is a form of unsupervised learning framework which retains 3 layers: output, input, and hidden layers [13]. The procedure of AE trained involves 2 parts: encoding and decoding. The encoding was utilized to map input data as to hidden illustration, and decoding was demonstrated for recreating input data in the hidden demonstration. To provide the unlabeled
input data set \( \{x_n\}_{n=1}^{N} \), whereas \( x_n \in \mathbb{R}^{m \times 1} \), \( h_n \) signifies the hidden encoding vector computed in \( \hat{x}_n \), and \( \hat{x} \) signifies the decoding vector of output layer. So the encoder method is as follows:

\[
h_n = f(W_1 x_n + b_1)
\]  

(1)

where \( f \) stands for the encoder functions, \( W_1 \) denotes the weight matrix of encoding, and \( b_1 \) implies the bias vectors. The decoded method was determined as:

\[
\hat{x}_n = g(W_2 h_n + b_2)
\]  

(2)

where \( g \) denotes the decoder functions, \( W_2 \) refers the weight matrix of decoding, and \( b_2 \) denotes the bias vectors. The parameter set of AE was optimized for minimizing the reconstruction error:

\[
\phi(\theta) = \arg\min_{\theta'} \frac{1}{n} \sum_{i=1}^{n} L(x^i, \hat{x}^i)
\]  

(3)

where \( L \) denotes the loss functions \( L(x, \hat{x}) = \|x - \hat{x}\|^2 \). The infrastructure of SAE was stacking \( n \) AEs as to \( n \) hidden layers with unsupervised layer-wise learning technique and then fine-tuning by supervised technique. Thus, the SAE based technique was separated as to 3 stages:

1) Training the primary AE by input data and attain the learn feature vectors;

2) The feature vector of prior layer was utilized as input to next layer, and this process was repeating still the trained ends.

3) Afterward, each hidden layer is trained, BP technique was utilized for minimizing the cost function and upgrading the weight with labeled trained set for achieving fine-tuned.

2.3. GOA based Parameter Tuning

In order to modify the weight and bias values of the SAE model, the GOA is applied and thereby the classifier results get improved. The GOA method is an evolutionary model proposed by the simulation behavior of swarm of grasshoppers while searching for food. Typically, they are insects of, destructive nature; cause harm to agricultural produce and harvest production [14]. The growth of a full-grown grasshopper drives as egg, nymph, and adults. It can be mathematically modeled by the following equation for resolving different optimization issues.
\[
Y_i^d = cx \left\{ \sum_{j=1}^{n} c x(u l_d - l l_d / 2)s f(|Y_j^d - Y_i^d|)^{Y_j^d - Y_i^d / D_{ij}} \right\}
\]

Here, \(Y_j, Y_i\) represents the location of jth and ith grasshopper. The jth and ith locations of the grasshopper in \(D\)th dimension are represented by \(Y_j^d\) and \(Y_i^d\), correspondingly. The distance, grasshopper number, and social interaction among jth and ith grasshoppers are symbolized as, \(s f\), and \(D_{ij}\) respectively. \(T_d\) indicates the value of the target in the \(D\)th dimension, while \(u l_d\) and \(l l_d\) denotes the upper and lower limits in Dth dimension. The adoptive variable \(c z\) is utilized for reducing the comfort zone. To balance exploitation and exploration of the grasshopper swarm near the optimal global solution, the initial \(c x\) value is used. Moreover, attraction, repulsion zone, and comfort zone amongst the grasshoppers are minimized by utilizing second \(c x\) value. The coefficient \(c x\) minimizes the comfort zone proportionate to the iteration count as follows

\[
cz = cz_{\text{max}} - t(t_{\text{max}})
\]

In which \(cz_{\text{max}}\) means the maximal value, \(cz_{\text{min}}\) shows the minimal value, \(t\) represents the existing iteration, and \(t_{\text{max}}\) indicates the maximal amount of iterations.

3. Results and Discussion

The performance validation of the proposed model takes place utilizing 2 benchmark datasets such as heart disease and diabetes datasets. Fig. 2 shows the correlation matrix of the attributes that exist in the heart disease dataset. Similarly, the correlation matrix of the attributes involved in the diabetes dataset is depicted in Fig. 3.
Fig. 2. Confusion matrix of GOSAE-DM technique under heart disease dataset

Fig. 3. Confusion matrix of GOSAE-DM technique under diabetes dataset
Fig. 4 inspects the $sens_y$ analysis of the GOSAE-DM technique with recent methods under heart disease dataset. The figure reported that the GOSAE-DM technique has accomplished enhanced classification outcomes. For instance, with 2000 instances, the GOSAE-DM technique has obtained higher $sens_y$ of 96.51% whereas the NB approach, SVM method, DT system, and OSSDA techniques have attained lower $sens_y$ of 88.24%, 83.45%, 93.66%, and 94.59% respectively. Moreover, with 10000 instances, the GOSAE-DM technique has obtained increased $sens_y$ of 99.61% whereas the NB approach, SVM method, DT system, and OSSDA techniques have reached reduced $sens_y$ of 89.09%, 84.54%, 96.67%, and 98% respectively.

Fig. 5 examines the $spec_y$ analysis of the GOSAE-DM system with existing approaches under heart disease dataset. The figure stated that the GOSAE-DM technique has accomplished enhanced classification outcomes. For instance, with 2000 instances, the GOSAE-DM method has obtained maximum $spec_y$ of 94.73% whereas the NB approach, SVM method, DT system, and OSSDA methodologies have attained lower $spec_y$ of 83.78%, 80.55%, 92.75%, and 94.59% correspondingly. In addition, with 10000 instances, the GOSAE-DM approach has
gained maximum $spec_y$ of 95.65% whereas the NB approach, SVM method, DT system, and OSSDA techniques have reached lower $spec_y$ of 87.36%, 84.51%, 90.18%, and 93.22% correspondingly.

Fig. 6 scrutinizes the $acc_y$ analysis of the GOSAE-DM technique with recent algorithms under heart disease dataset. The figure described that the GOSAE-DM system has accomplished enhanced classification outcomes. For sample, with 2000 instances, the GOSAE-DM technique has reached superior $acc_y$ of 95.74% whereas the NB approach, SVM method, DT system, and OSSDA techniques have obtained minimal $acc_y$ of 76.06%, 73.75%, 91.09%, and 94.02% correspondingly. Besides, with 10000 instances, the GOSAE-DM approach has obtained increased $acc_y$ of 97.27% whereas the NB approach, SVM method, DT system, and OSSDA systems have achieved decreased $acc_y$ of 82.36%, 81.32%, 92.29%, and 94.86% correspondingly.

![Graph showing specificity analysis of GOSAE-DM technique](image)

**Fig. 5.** $spec_y$ analysis of GOSAE-DM technique under heart disease dataset
Fig. 6. Acc analysis of GOSAE-DM technique under heart disease dataset

Fig. 7 examines the sens analysis of the GOSAE-DM technique with existing techniques under diabetes dataset [16-18]. The figure exhibited that the GOSAE-DM technique has resulted in improved classifier results. For instance, with 2000 instances, the GOSAE-DM technique has attained maximum sens of 97.67% but the NB system, SVM algorithm, DT method, FNC approach, and OSSDA techniques have reached minimum sens of 86.67%, 83.34%, 92.66%, 93.83%, and 96.11% respectively. Likewise, with 8000 instances, the GOSAE-DM technique has gained better sens of 99.02% while the NB system, SVM algorithm, DT method, FNC approach, and OSSDA techniques have accomplished decreased sens of 87.72%, 82.73%, 97.68%, 98.09%, and 98.98% respectively.
Fig. 7. Sens$_y$ analysis of GOSAE-DM technique under diabetes dataset

Fig. 8 demonstrates the spec$_y$ analysis of the GOSAE-DM approach with existing systems under diabetes dataset. The figure outperformed that the GOSAE-DM algorithm has resulted in enhanced classifier results. For instance, with 2000 instances, the GOSAE-DM technique has reached maximal spec$_y$ of 98.29% while the NB system, SVM algorithm, DT method, FNC approach, and OSSDA techniques have obtained lower spec$_y$ of 83.29%, 81.94%, 92.29%, 94.94%, and 97.08% correspondingly. Also, with 8000 instances, the GOSAE-DM approach has reached better spec$_y$ of 95.04% but the NB system, SVM algorithm, DT method, FNC approach, and OSSDA methods have accomplished decreased spec$_y$ of 87.04%, 84.36%, 90.23%, 91.63%, and 93.73% respectively.
Fig. 8. Spec_y analysis of GOSAE-DM technique under diabetes dataset

Fig. 9 scrutinizes the acc_y analysis of the GOSAE-DM technique with existing techniques under diabetes dataset. The figure exhibited that the GOSAE-DM technique has resulted in enhanced classifier results. For instance, with 2000 instances, the GOSAE-DM methodology has attained increased acc_y of 96.05% whereas the NB system, SVM algorithm, DT method, FNC approach, and OSSDA systems have reached lower acc_y of 77.05%, 73.34%, 92.72%, 92.93%, and 95.19% respectively. Likewise, with 8000 instances, the GOSAE-DM algorithm has gained optimum acc_y of 98.83% while the NB system, SVM algorithm, DT method, FNC approach, and OSSDA methodologies have accomplished lesser acc_y of 82.83%, 79.15%, 93.06%, 94.42%, and 96.75% correspondingly.
Fig. 9. $\text{Acc}_c$ analysis of GOSAE-DM technique under diabetes dataset

4. Conclusion

In this study, a novel GOSAE-DM technique has been developed to handle the class imbalance problem and then performance classification process in the healthcare sector. The GOSAE-DM technique involves different stages of operations such as ADASYN based class imbalance data handling, SAE based classification, and GOA based parameter optimization. The SAE model is applied for the classification of balanced data, which determines the existence of the disease. For improving the classifier outcomes of the SAE technique, the GOA is applied and thereby enhances the classifier results. A wide range of experiments was carried out using benchmark medical datasets and the results are inspected under several dimensions. The comparative study reported the betterment of the GOSAE-DM technique over the recent techniques in terms of different measures. In future, data clustering techniques can be included in the GOSAE-DM technique to raise the classifier results.
References


