Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 3 ,July 2021 : 2671 - 2677

Research Article

Diagnosis of Diabetes using Modified PSO and FFNN Technique Their Application to WSD

Srishti Jain[1], Surya Prakash Sharma[2], Dr. CS Yadav[3]
1,2,3: Noida Institute of Engineering & Technology Gr. Noida (India)

Abstract— It is one of the most amazing and perplexing feats of human mind that we understand written and spoken communication in spite of enormous number of possibilities that exist because of multiple meanings of words that compose a sentence. Human beings can produce a correct sentence choosing words in their appropriate context with little effort. But the same problem becomes very hard and complex when it is sought to be automated. Therefore any system that proposes to implement Natural Language Processing (NLP) on a computer has to address very seriously the question of WSD. The most important application of WSD is diagnosis of diabetes. This paper is implementation of diagnosis of diabetes using PSO and FFNN. To classify the diabetic detection, we used a feed forward network (FFNN) in which the network can be trained to optimize the mean square error (MSE) by using a modified particle swarm optimization (PSO) algorithm. The effectiveness for this procedure was verified by an experimental analysis on a benchmark dataset which is publicly available on UCI learning repository. The result of our experimental analysis revealed that the improvement on the algorithm is significant with respect to FF trained by PSO. Using these attribute to Neural Network as input we have achieved the best known training classification accuracy of 76.0417%.
Keywords—diabetic dataset, particle swarm optimization, Feed forward neural network, classification

I. INTRODUCTION

WSD has been an area of research interest since the earliest days of understating of natural language by computer in the 1950's. WSD is an "intermediate task" [1] for many other NLP systems, including machine translation, information extraction, mono and multilingual information retrieval etc. It is essential for many Natural Language Understanding applications [2] such as text analytics, human computer interaction, etc. WSD finds several applications in many research areas within NLP, both as an explicit system as well as an implicit one [3].

The original problem that gave rise to WSD was MT. The first attempts to perform WSD were carried out in. All MT systems implement some sort of WSD as different senses of words often have completely different translations across languages. For example, the English word 'star' can be translated as a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior, an actor/celebrity/performer who plays a principal role, a plane figure with five or more points, star-shaped object, etc., depending on the context. The senses of a word in the source language are often represented directly as the translated words in the target language. A highly accurate WSD system is likely to enhance the performance of automatic translators to a great degree [4].

In modern technology, diagnosis of health is a very crucial task. Diabetes Mellitus is one of the serious challenging disease in both developed and developing countries [5]. In medical science diagnosis of disease data involves of a number of medical tests which are needed to diagnose a certain disease and the diagnosis are depend on the surgeon experience, if a less experience surgeon can diagnose a problem incorrectly. So, a surgeon needs to analyze a lot of issues and factors which makes the surgeon's job difficult for diagnosing the disease. In 2004, estimated that 3.4 million people are suffered from the high blood sugar [6] leading to diabetes mellitus. Diabetes mellitus is one of the fastest growing chronic infection, with set to raise more than double from 1.5 million to 3.5 million in the following 20 years and no more steps is done to address the increasing disease. Recently some research has been conducted on diseases which are frequently leads to misdiagnosis and death rate is increased due to this medical error as nearly 98,000 people per year. Early detection of diabetes mellitus would be great significant and the fact that 50% and 80% people with diabetes are unaware of their condition. A medical diagnosis is a classification process. The machine learning [7] and data mining techniques [8] have been considered to design automatic diagnosis system for diagnosing diabetes mellitus. In recent times many methods and algorithms are employed to mine the biomedical datasets for hidden information including

Neural networks (NNs), Decision Trees (DT), Fuzzy Logic Systems, Naive Bayes, Support Vector Machine (SVM), Ensemble model [9], Extreme Learning Machine (ELM) and logistic regression.

II. LITERATURE REVIEW

R. K.rMohanty et al. [1], in this article, the use of selection restrictions for automated sensory disambiguation in broad-covered settings is protected by the information method. The method incorporates methodological and knowledge-based techniques but as a starting point it takes the premise that sensory annotation 10 training text is not available as opposed to manycuerpus-based approaches to sensation disambiguation.

Muhammad Azeem Sarwar et al. [2], in this paper, belonging to the class of Knowledge based approaches is based on the assumption that words in a given “neighbourhood" (section of text) will tend to share a common topic. A simplified version of the Lesk algorithm is to compare the dictionary definition of an ambiguous word with the terms contained in its neighbourhood.

Munam Ali ShahS. Jha et al. [3], in this paper, use the conceptual distance between the senses of the context words and the sense of the target word as a measure for disambiguation. They proposed a formula for conceptual distance which is directly proportional to the length of the path between two synsets in the WordNet graph and inversely proportional to the depth of their common ancestor in the WordNet hierarchy.

Majid Ghonji Feshki et al. [4], the impressive developing of cardiovascular malady what’s more, its belongings and inconveniences and additionally the high expenses on society influences restorative group to look for answers for counteractive action, early recognizable proof and viable treatment with bring down expenses. In this manner, profitable information can be built up by utilizing counterfeit insight and information mining; the found learning makes enhance the nature of administration. As of not long ago, extraordinary investigates have been completed to foresee coronary illness in light of information mining strategies, for example, characterization and grouping strategies; be that as it may, what has been less seen is the correct conclusion of infection with the most reduced cost and time. In this paper, by utilizing highlight positioning on successful variables of malady identified with Cleveland facility database and by utilizing Particle Swarm Streamlining and Neural Network Feed Forward Back Propagation, powerful factors decreased to 8 enhanced highlights as far as cost and exactness. The evaluation of chose highlights of arranged strategies additionally demonstrated that PSO strategy alongside Neural Networks of Feed Forward Back.

Muhammad Waqar Aslam et al. [5], in this paper, thesaurus Based approach falling under Knowledge based approaches to WSD. The objective word is contained in the thesaurus group to which this meaning belongs and then the score for each dimension is determined by using the background terms. A reference word contributes more to the definition if the type of the expression thesaurus is the same as the concept.

Lan H. et al. [6], in this paper, introduced an adaptation of Lesk algorithm where WordNet and not standard dictionary was considered as the knowledge base for the glosses. The algorithm compares the glosses in a background window between each word pair. The longest series of consecutive terms in both glosses is an overlap. An overlap between the two glosses leads to a score equal to the square in the overlap and the candidate with the highest score is the winner.

Manjeevan et al. [7], diabetes happens when a body can't create or react legitimately to insulin which is expected to control glucose. Other than adding to coronary illness, diabetes likewise builds the dangers of creating kidney ailment, visual deficiency, nerve harm, and vein harm. Diabetes illness determination by means of legitimate elucidation of the diabetes information is an essential order issue. In this examination, a relative Pima diabetes infection conclusion was figured it out. For this reason, a multilayer neural system structure which was prepared by Levenberg– Marquardt (LM) calculation and a probabilistic neural system structure were utilized. The consequences of the examination were contrasted and the aftereffects of the past investigations announced concentrating on diabetes infection finding and utilizing the same UCI machine learning database. A similar report on Pima Indian diabetes sickness is symptomatic by utilizing multilayer neural system which was prepared by LM calculation and probabilistic neural system.

Sean N. Ghazavi et al. [8], measurement diminishment (DR) is critical in the handling of information in spaces, for example, interactive media or bioinformatics on the grounds that such information can be of high measurement. Measurement decrease in a directed learning setting is an all-around postured issue in that there is an unmistakable goal of finding a lessened portrayal of the information where the classes are very much isolated. By differentiate DR in an unsupervised setting is not well postured in that the general goal is less clear. All things considered fruitful unsupervised DR systems, for example, essential segment examination

(PCA) exist—PCA has the businesslike goal of changing the information into a lessened number of measurements that still catches the majority of the variety in the information.

III. METHODOLOGY

Particle Swarm Optimization:-

PSO algorithm was developed initially by Kennedy and Eberhart in 1995. The aforesaid Multiple Independent Augmentation Issues must concurrently augment the trajectory process and yield Pareto optimum resolutions. Pareto front is a set of Pareto optimum (non-conquered) resolutions, being contemplated to be optimum, if no goal can be enhanced devoid of foregoing at least one other goal. Instead, a resolution x* is discussed as conquered by another resolution x, if and only if, x is correspondingly good or improved than x * in respect of entire goals. For applying the element group augmentation strategy for solving multiple goal augmentation procedure, it is obvious that the original scheme has to be modified. In general, when solving a multi-objective problem, three main goals to achieve are:

- Make the most of the number of elements detected in the Pareto optimum set.
- Diminish the expanse of the Pareto front developed by the procedure with respect to the real (global) Pareto front (supposing it knows its location).
- Make the most of the detected spread of solutions; to enable it can have a dispersal of trajectories smoothly and uniformly to the extent possible.

$$x_k = \{x_{k1}, x_{k2}, x_{k3}.....x_{kn}\} \quad (1)$$

Where k=1,2,3,…,d and d is the number of particles in the swarm.
Each particle maintains its own velocity, let represented as given in equation 2.

$$v_k = \{v_{k1}, v_{k2}, v_{k3}.....v_{kn}\} \quad (2)$$

Also, in this algorithm each particle maintains its personal best position called as pbest and a best solution among all the particles called as gbest. In each iteration or generation, the particles move towards optimal solution by updating their velocity and position according to the formula given in Equation 3, 4

$$v_k(t+1) = w * v_k(t) + c1r1(p_k(t) - x_k(t)) + c2 * r2(g_k(t) - x_k(t)) \quad (3)$$

$$x_k(t+1) = x_k(t) + v_k(t+1) \quad (4)$$

In multiple goal issues, it can differentiate two essential attitudes for planning element group augmentation procedures. The first approach comprises procedures which contemplate every goal function individually. In such attitudes, every element gets assessed for only one unbiased process at a time, and the determining the finest locations is carried out in the same way as the single unbiased augmentation case. The chief test in such cases is the correct operation of the data being received from every unbiased operation for guiding the elements headed for Pareto optimum resolutions. The next tactic comprises procedures that assess the entire unbiased processes for every element and centred on the theory of Pareto optimum, they deliver finest non-conquered positions (often referred to as front-runners) that are applied to assist the elements.

2. **Feed Forward Neural Network (FFNN)**

Basically the Neural networks are typically composed of layers. The layers are made up of a large number of interconnected nodes which contain an activation function. The neural network consists of input layer, output layer and few hidden layers. The inputs such as patterns are applied to the network through input layer which further communicates to the one or more hidden layers. The hidden layers the place where the actual processing of information takes place and it is done in the system of weighted connections. The hidden layers then communicate with the output layer.

FFFNN is one of the simplest forms of Neural Network consisting of exactly three layers, namely input, hidden and output layer. The limitation of only three layers makes it simpler and show somehow the efficient neural network architecture (as shown in Fig. 1) .the idea of FFNN has been derived from function approximation. An FFNN Network positions one or more neurons in the space described by the predictor variables.

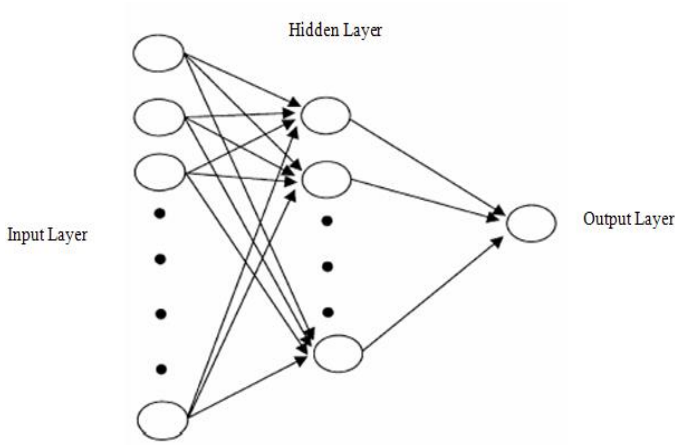

Figure 1: Feed Forward Neural Network Architecture

IV. PROPOSED METHODOLOGY

The new techniques we have implemented in our research are as under -

Algorithmic Description of the Proposed Model:-
The following algorithm /pseudo code describes the detailed structure of our proposed model:
Modified PSO-NN algorithm the diabetes dataset is given as input to the Modified PSO-Neural Network [7]. It consists of 768 samples and 9 attributes out of which the last attribute is 'class' that indicates rather positive or negative. In this rule looking out is additionally started from initializing a gaggle of random particles. During this manner, this modified rule is might notice Associate in Nursing optimum a lot of quickly. The procedure for this modified PSO-Neural Network is summarizing as follows:

Algorithm:
For each particle do
Initialize particle position and velocity
End
Calculate the inertia weight
For each particle do
Calculate fitness value (using MSE of FFNN)
If fitness value is better than best fitness value in particle history (pBest) then
Set current position as pBest
End If
End For
Choose the global best (gBest) as the particle with best fitness value among all the particles
For each particle do
Calculate particle velocity
Update particle position (Position & Weight vector)
End for
End while

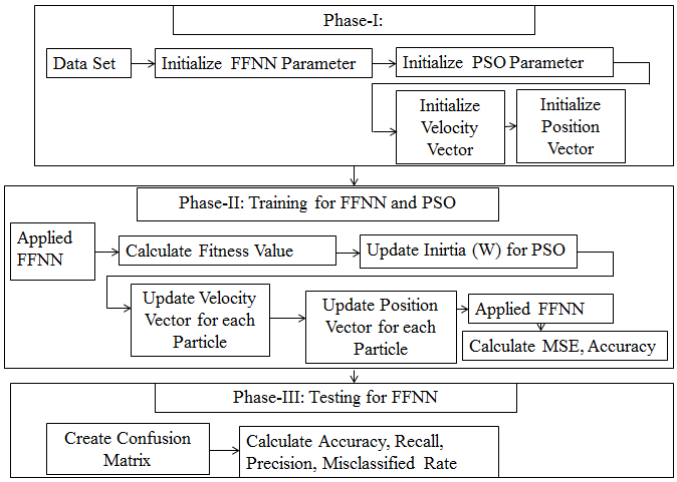

Figure 2: Flow Chart of Proposed Methodology

IV. EXPERIMENTAL RESULTS

The source of Pima Indians dataset diabetes dataset on which the experiment is performed is UCI machine learning repository with 768 data instances and 9 attributes. All patients in this dataset are Pima Indians women whose age is at least 21 years old and living near Phoenix , Arizona which denotes either “0” or “1”, where ‘0’ is tested as negative and ‘1’ is tested as positive for diabetes.

MATLAB (lattice lab) is numerical processing condition and fourth-age programming dialect. Made by Math Works, MATLAB grants framework controls, plotting of limits and data, use of computations, making of UIs, and interfacing with programs written in various lingos, including C, C++, Java, and Fortran.

Regardless of the way that MATLAB is proposed basically for numerical preparing, an optional toolbox uses the MuPAD delegate engine, empowering access to meaningful enlisting capacities. An additional package, Simulink, incorporates graphical multi-region proliferation and Model-Based Design for dynamic and embedded structures. MATLAB has structure data makes. Since all factors in MATLAB are exhibits, a more satisfactory name is "structure cluster", where every component of the exhibit has a similar field names.

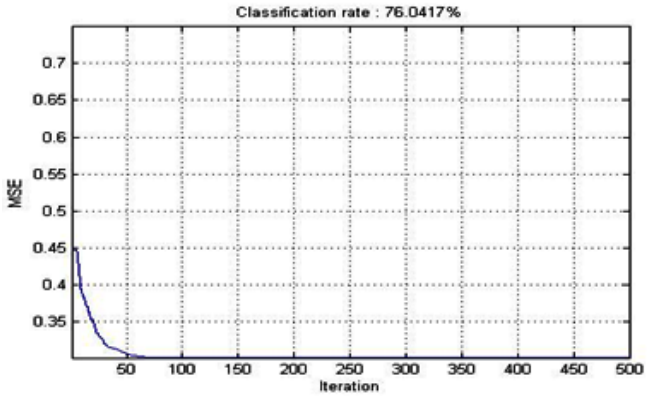

Figure 3: Graph for showing the classification Accuracy

Table 1: Comparative Study for tranning classification

Source	Techniques	Accuracy
Previous Technique	Hybrid PSO-NN	74.87%
Our Study	Modified Hybrid PSO-NN	76.04%

Evalution metrics: Generally, the evaluiton of a classification problem is based on a matrix calledd as a confusion matrix with the number of testing samples correctly classified and incorrectly classified represnted as so, the accuracy can be measured according to Eq. 5

$$Accurancy = \frac{TN + TP}{TN + TP + FN + FP} \qquad (1)$$

For a binary classification problem, the other measures include Precision, Sensitivity or Recall and Specificity. The formula to derive these measures is given in Eq. 6 and Eq. 7.

$$\Pr ecision = \frac{TP}{TP + FP} \qquad (2)$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \qquad (3)$$

In these relations ((1),(2) and (3) formula) TP means the number of samples that are healthy and properly diagnosed. FP indicates the number of samples that are healthy and have been diagnosed wrongly.

Table 2: Assessment of Different Classification Methods

	PLS	FMM	Hybrid PSO-NN	Modified hybrid PSO-NN
Accuracy	0.744	0.6928	0.917	0.947
Recall	0.732	0.6991	0.930	0.975
Precision	0.741	0.6928	0.919	0.955

V. CONCLUSION

Most of researchers have depended on Artificial Intelligence and Data Mining techniques and machine learning classifiers for constructing their classifier or forecaster models. In this research, several machine learning models have been implemented to predict and classify diabetes types. This classifier attempted to solve two problems such as categorizing patients in terms of diabetic types and to predict a diabetic and non-diabetic. The hybrid approaches yield better results than single classifiers. The objective of this work is to evaluate the Prima Indian Diabetic dataset based on machine learning algorithms and to classify the diabetic dataset. The comparative analysis shows that the promising results when compared with other techniques in order to classify the diabetes data with high classification accuracy with less execution time.

On comparison of these methods, we conclude that Modified PSO-FFNN has proved to be much better than the other two methods and is more efficient. In future, improvement in the execution time for large size data set could be treated as a research subject.

REFERENCES

[1] R. K.rMohanty et al., Synset Based Multilingual Dictionary:Insights, Applications and Challenges. GWC 2008: 4th Global WordNet conference, Szeged, Hungary, Januray 2018.
[2] Muhammad Azeem Sarwar, Nasir Kamal and Wajeeha Hamid, “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare”, International Conference on Automation & Computing, IEEE 2018.
[3] Munam Ali ShahS. Jha, D. Narayan, P. Pande, P. Bhattacharyya, A WordNet for Hindi. Workshop on Lexical Resources in Natural Language Processing, Hyderabad, India, January, 2017.
[4] Majid Ghonji Feshki and Omid Sojoodi Shijan, “Improving the Heart Disease Diagnosis by Evolutionary Algorithm of Feed Forward Neural Network”, 978-1-5090-2169-7/16/$31.00 ©2016 IEEE.
[5] Muhammad Waqar Aslam, Zhechen Zhu and Asoke Kumar Nandi, “Feature generation programming with comparative partner selection for diabetes classification”, “Expert Systems with Applications”,5402-5412 . 40, 2013.
[6] Lan H. Wittn, Eibe Frank, Mark A, Hall, “ Data Mining Practical Machine Learning Tools and Techinques” ,3rd Edition,Morgan Kaufmann Publishers is an imprint of Elsevier 30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
[7] Manjeevan SEERA AND Chee Peng Lim, “A modified intelligent system for medical data classification, “Expert Systems with Application”, 41(2014) 2239-2249

[8] Sean N. Ghazavi and Thunshun W. Liao, “Medical data mining by fuzzy modeling with selected features”, Ärtificial Intelligence in Medicine”, 2008 43,195-206.
[9] J. Saraswati, R. Shukla, S.Pathade, T. Solanki and P. Bhattacharyya, Challenges in Multilingual Domain-Specific Sense-Marking, Proceedings of the 5th Global WordNet Conference, Mumbai, Narosa Publishing House, India,Mumbai, Jan, 2017.
[10] P. Bhattacharyya, IndoWordnet. Proceedings of Lexical Resources Engineering Conference (LREC 2010), Malta, May, 2015.
[11] P. Vossen (ed.). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht.2016.
[12] D. Narayan, D. Chakrabarty, P. Pande and P. Bhattacharyya, An Experience in Building the Indo WordNet- a WordNet for Hindi. 1st International Conference on Global WordNet (GWC 02), Mysore, India, 2018.
[13] M. Sinha, M. K. Reddy, P. Bhattacharyya, P. Pandey and L. Kashyap, Hindi Word Sense Disambiguation. Proceedings of International Symposium on Machine Translation, Natural Language Processing and Translation Support System, Delhi India, 2017.
[14] P.Resnik. 2019, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the 5th annual international conference on Systems documentation, Toronto, Ontario, Canada, 2019.
[15] Lesk, The Use of Machine Readable Dictionaries in Sublanguage Analysis. In Analyzing Language in Restricted Domains, Grishman and Kittredge (eds), LEA Press, pp. 69-83, 2019.
[16] D. Yarowsky, Word sense disambiguation using statistical models of Roget's categories trained on large corpora. Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, 454-460, 2018.
[17] D. Yarowsky, Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL), Las Cruces, U.S.A., 88-95, 2018.