ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Improvised K-Means for Improving Raw-Data in Diagnosis of Thyroid Disease

D.P.Gaikwad, Kunal Mahurkar

Asst. Professor, Department of Computer Engineering, All India Shree Shivaji Memorial Society's College of Engineering, India

Department of Computer Engineering, All India Shree Shivaji Memorial Society's College of Engineering, India

ABSTRACT: Classification and clustering algorithms were used separately in prediction of Thyroid Disease. Success and accuracy achieved by classification and clustering algorithm for thyroid prediction was less. We can take advantage of both classification and clustering algorithm for thyroid disease prediction by combining them in system. This paper focuses on potential benefit of using artificial neural networks (ANNs) as a classification algorithmand improvised K-means as clustering algorithm for the diagnosis of thyroid.We have examined the ANN architecture and assessed their robustness in the face of diagnostic noise. The experimental dataset of thyroid has previously studied by multivariate statistical methods and a variety of pattern-recognition techniques. For the experimentation, the thyroid dataset is divided into two subsets, one is training dataset for the networks and another for testing network performance. The test dataset has various proportions of cases with diagnostic noise to mimic real-life diagnostic situations.

KEYWORDS*:* Artificial Neural Networks, Improvised K-Means, Thyroid dataset, UCI Repository.

I. INTRODUCTION

Various classification and clustering methods are used by many experimenters to enhance the accuracy of predicting diseases before they actually occur like Diabetes Prediction [8]. Some methods like Genetic Programming [1], Semi-supervised Genetic Programming [3].Various authors have worked in the field of various data mining techniques. Multilayer Perceptron (MLP) for classification of thyroid diseases and achieved 96.74% ofaccuracy using WEKA tool.Disease Forecasting System Using Data Mining Methods is done using the K-means algorithm. This paper analyzes the heart disease predictions usingclassification algorithms. These hidden patterns can be used forhealth diagnosis in medicinal data. For classification the UCI machine learning repository have used various statistical and data mining method such as bayesand c4.5 algorithm they have developed the multilayer thyroid detection system to get higher efficiency.

The proposed model discusses about clustering algorithm improvised k-means in which the number of clusters i.e. value of K depends upon the min and max value of dataset, which also help in selecting feature . Further training and predicting is done using ANN algorithm.The data used for the prediction of thyroid

A.*OVERVIEW OF THYROID*

a)Thyroid Hormones

The thyroid hormones, triiodothyronine (T3) and its prohormone, thyroxine (T4), are tyrosine-based hormones produced by the thyroid gland that are primarily responsible for regulation of metabolism. T3 and T4 are partially composed of iodine (see molecular model). A deficiency of iodine leads to decreased production of T3 and T4, enlarges the thyroid tissue and will cause the disease known as simple goitre. The major form of thyroid hormone in the blood is thyroxine (T4), which has a longer half-life than T3. The thyroid hormones act on nearly every cell in the body. They act to increase the basal metabolic rate, affect protein synthesis, help regulate long bone growth (synergy with growth hormone) and neural maturation, and increase the body's sensitivity to catecholamines (such as adrenaline) by permissiveness. The thyroid hormones are essential to proper development and differentiation of all cells of the human

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

body. These hormones also regulate protein, fat, and carbohydrate metabolism, affecting how human cells use energetic compounds. They also stimulate vitamin metabolism. Numerous physiological and pathological stimuli influence thyroid hormone synthesis.
Thyroid hormone leads to heat generation in humans. However, the thyronamines function via some unknown mechanism to inhibit neuronal activity; this plays an important role in the hibernation cycles of mammals and the moultingbehaviour of birds. One effect of administering the thyronamines is a severe drop in body temperature.

b)Thyroid and Health Effects

i) Hyperthyroid
Hyperthyroidism is the condition that occurs due to excessive production of thyroid hormone by the thyroid gland.Hyperthyroidism is a condition of the thyroid. The thyroid is a small, butterfly-shaped gland located at the front of your neck. It produces tetraiodothyronine (T4) and triiodothyronine (T3), which are two primary hormones that control how your cells use energy. Your thyroid gland regulates your metabolism through the release of these hormones. Hyperthyroidism occurs when the thyroid makes too much T4, T3, or both. Diagnosis of overactive thyroid and treatment of the underlying cause can relieve symptoms and prevent complications. A variety of conditions can cause hyperthyroidism. Graves' disease, an autoimmune disorder, is the most common cause of hyperthyroidism. It causes antibodies to stimulate the thyroid to secrete too much hormone. Graves' disease occurs more often in women than in men. It tends to run in families, which suggests a genetic link.

ii)Hypothyroid.
Hypothyroidism, also called underactive thyroid or low thyroid, is a common disorder of the endocrine system in which the thyroid gland does not produce enough thyroid hormone. It can cause a number of symptoms, such as poor ability to tolerate cold, a feeling of tiredness, constipation, depression, and weight gain. Occasionally there may be swelling of the front part of the neck due to goitre. Untreated hypothyroidism during pregnancy can lead to delays in growth and intellectual development in the baby, which is called cretinism.

Worldwide, too little iodine in the diet is the most common cause of hypothyroidism. In countries with enough iodine in the diet, the most common cause of hypothyroidism is the autoimmune condition Hashimoto's thyroiditis. Less common causes include: previous treatment with radioactive iodine, injury to the hypothalamus or the anterior pituitary gland, certain medications, a lack of a functioning thyroid at birth, or previous thyroid surgery. The diagnosis of hypothyroidism, when suspected, can be confirmed with blood tests measuring thyroid-stimulating hormone (TSH) and thyroxine levels.

iii) Thyroid stimulating hormone (TSH):
A thyroid-stimulating hormone (TSH) blood test is used to check for thyroid gland camera.gif problems. TSH is produced when the hypothalamus releases a substance called thyrotropin-releasing hormone (TRH). TRH then triggers the pituitary gland camera.gif to release TSH.TSH causes the thyroid gland to make two hormones: triiodothyronine (T3) and thyroxine (T4). T3 and T4 help control your body's metabolism.Triiodothyronine (T3) and thyroxine (T4) are needed for normal growth of the brain, especially during the first 3 years of life. A baby whose thyroid gland does not make enough thyroid hormone (congenital hypothyroidism) may, in severe cases, be mentally retarded. Older children also need thyroid hormones to grow and develop normally.

Type of thyroid disease	T3	T4	TSH
Hypo	Increase	Increase	Reduction
Hyper	Reduction	Reduction	Increase

Table I. Classify thyroid disease based on parameters.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

II. CLASSIFICATION METHODS

A. The Artificial Neural Network (ANN):

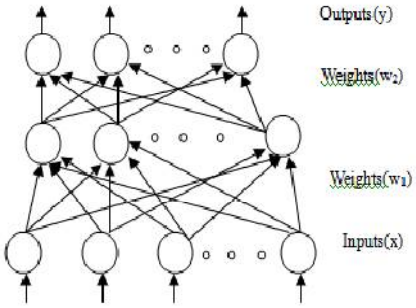

Fig.1. Feed Forward Neural Network

An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output. ANNs are considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found. ANN is also known as a neural network.An ANN has several advantages but one of the most recognized of these is the fact that it can actually learn from observing data sets. In this way, ANN is used as a random function approximation tool. These types of tools help estimate the most cost-effective and ideal methods for arriving at solutions while defining computing functions or distributions. ANN takes data samples rather than entire data sets to arrive at solutions, which saves both time and money. ANNs are considered fairly simple mathematical models to enhance existing data analysis technologies. ANNs have three layers that are interconnected. The first layer consists of input neurons. Those neurons send data on to the second layer, which in turn sends the output neurons to the third layer. Training an artificial neural network involves choosing from allowed models for which there are several associated algorithms.

Nowadays, one of the main issues to create challenges in medicine sciences by developing technology is the disease diagnosis with high accuracy. In the recent decades, Artificial Neural Networks (ANNs) are considered as the best solutions to achieve this goal and involve in widespread researches to diagnose the diseases. In this paper, we consider a Multi-layer Perception (MLP) ANN using back propagation learning algorithm to classify Thyroid disease. It consists of an input layer with 5 neurons, a hidden layer with 6 neurons and an output layer with just 1 neuron. The suitable selection of activation function and the number of neurons in the hidden layer and also the number of layers are achieved using test and error method. Our simulation results indicate that the performed optimization in MLP ANNs can achieve 98.6% accuracy.

B.Improvised K _means:

To extract useful information from huge data sets are considerable problem for researcher that is insufficient for conventional databases querying methods. K-means clustering algorithm is a one of the majorcluster analysis method that is commonly used in practical applications for extracting useful information interms of grouping data. But the standard K-means algorithm is computationally expensive by getting centroidsthat provide the quality of the clusters in results. The improved K-means divides the algorithm in two phases and uses different algorithms to make proposed algorithm more efficient and accurate. In the first phase the initial centroids are determined systematically to produce the clusters with betteraccuracy and in second phase it uses various methods described in the algorithm to assigned data points to the appropriate clusters.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

C.Improved K-means Clustering Algorithm
Related Concept

Definition 1:

The between data distance points and the cluster center. The distance formula of data point xi and cluster center kj defined as following:

$$d_{j,i} = \sqrt{(xi1 - kj1)2 + (xi2 - kj2)2 + \cdots + (xiw - kjw)2} \quad (2)$$

where w represents the number of attributes of the data points xi.

Definition 2:

The density parameter. The number of data points which is contained by a scope defined as density parameter. The scope is a round which takes space point of not statistics xi as the center, as the radius. The greater the density of xi, the greater the value of the density parameter are.

Algorithm 1: Improved K-means Algorithm

Input: data set x contains n data points; the number of cluster is k.
Output: k clusters of meet the criterion function convergence.

Program process:

Step 1. Initialize the cluster center.

Step 1.1. Select a data point xi from data set X, set the identified as statistics and compute the distance between xi and other data point in the data set X. If it meet the distance threshold, then identify the data points as statistics, the density value of the data point xi add 1.
Step 1.2. Select the data point which is not identified as statistics, set the identified as statis-tics and compute its density value. Repeat Step 1.2 until all the data points in the data set X have been identified as statistics.

Step 1.3. Select data point from data set which the density value is greater then the threshold and add it to the corresponding high-density area set D.

Step 1.4. Filter the data point from the corresponding high-density area set D that the density of data points relatively high, added it to the initial cluster center set. Followed to find the k-1 data points, making the distance among k initial cluster centers are the largest.

Step 2. Assigned the n data points from data set X to the closet cluster.

Step 3. Adjust each cluster center K by the formula (3).

Step 4. Calculate the distance of various data objects from each cluster center by formula (4), and redistribute the n data points to corresponding cluster.

Step 5. Adjust each cluster center K by the formula (3).

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Step 6. Calculate the criterion function E using formula (1), to determine whether the convergence, if convergence, then continue; otherwise, jump to Step 4.

III. UCI MACHINE LEARNING REPOSITORY

The UCI machine learning repository is a collection of databases that have 215 instances used by machine learning community for empirical analysis.

A. Classification Accuracy

In this study, the level of accuracy can be determined by equation (9) as follows [10]

$$Accuracy\ (T) = \frac{\sum_{i=1}^{|T|} assess\ (t_i)}{|T|},\ t_i \in T$$

$$assess\ (t) = \begin{cases} 1, & if\ classify\ (t) = t.c \\ 0, & otherwise \end{cases}$$

where*T* denotes dataset that will be classified, *tε T*, *t.c* denotes the class of item *t*, and *classify (t)* denotes the classification result of *t* resulted by each classification method. The java programming used for prediction thyroid diagnosis.

III.CONCLUSION

We have presented a combination two machine algorithms, Artificial Neural Network and K-means clustering algorithm to increase accuracy in prediction of thyroid disease. We have proposed improvised K-means algorithm with multilayer perceptron. The diagnosis of thyroid using artificial neural network is robust with respect to sampling variations. The cross validationtechnique provides decision makers with a method for examine predictive validity and hence the usefulness of theClassification method.

IV. ACKNOWLEDGMENT

I thanks to Mr. D.P.Gaikwad, Who have contributed towards development of this work. I, would like to thank the participants for the helpful discussion and constructive comments and suggestions.

REFERENCES

[1]Md. Dendi Maysanjaya, HanungAdiNugroho, Noor AkhmadSetiawan"A Comparison of Classification Methods onDiagnosis of ThyroidDiseases," International Seminar on Intelligent Technology and Its Applications, 2015, pp-978-1-4799-7711-6/15/

[2] F. Gorunescu, M. Gorunescu, E. El-Darzi, M. Ene, and S.
Gorunescu, "Statistical Comparison of a Probabilistic NeuralNetwork Approach in Hepatic Cancer Diagnosis," ProceedingsEurocon2005 -IEEE International Conference on "Computer as aTool", Belgrade, Serbia, November 21-24, pp 237-240, 2005.

[3] L. Ozyilmaz and T. Yildirim, "diagnosis of thyroid disease usingartificial neural networks methods," the 9th Inter. Conf. on neuralinformation processing, vol 4, pp2033-2036, 2002.

[4] Harrison' Principles of Internal Medicine, Endocrinology &Metabolism, 2001.

[5] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer, "Classification and Diagnostic Predictionof Cancers Using Gene Expression Profiling and Artificial Neural Networks," Nature Medicine, vol. 7, no. 6, pp. 673-679, June2001.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

[6] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M.Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio,W. Gerald, M. Loda, E.S. Lander, and T.R. Golub,"MulticlassCancer Diagnosis Using Tumor Gene Expression Signatures,"Proc. Nat'l Academy Sciences USA, vol. 98, no. 26, pp. 15149-15154,Dec. 2001.
[7] R.J.Schalkoff, *Artificial Neural Networks*, McGraw-Hill Inc, Singapore,1997.
[8] S.Haykin, *Neural Networks: A Comprehensive Foundation*,McMaster University, pp.278-300,1994.
[9] D. F .Specht, "A General Regression Neural Network", IEEE Trans.Neural Networks, pp.568- 576,1991.
[10]T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek,J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri,C.D. Bloomfield, and E.S. Lander, "Molecular Classification ofCancer: Class Discovery and Class Prediction by Gene ExpressionMonitoring," Science, vol. 286, no. 5439, pp. 531-537, Oct. 1999.