USE OF K-NEAREST NEIGHBOR IN THYROID DISEASE CLASSIFICATION

[1]Pratiksha Chalekar, [2]Shanu Shroff, [3]Siddhi Pise, [4]Suja Panicker
[1,2,3,4]Department of Computer Engineering,Maharashtra Institute of Technology, Pune, India.
[1]Pratiksha.chalekar@gmail.com, [2]shanushroff@gmail.com, [3]siddhi.pise92@gmail.com, [4]suja.panicker@mitpune.edu.in

***Abstract*—In medical disease classification, choice of classifier plays an important role. We have surveyed several methods like SVM, Neural Networks, k-Nearest Neighbor, Probabilistic Neural Network, Naive Bayes etc. In this paper we have selected two standard datasets from UCI Repository and have performed extensive experimentation using Nearest Neighbour Classifier. We performed Data Preprocessing and replaced the large number of missing values in Dataset2. kNN has yielded high classification accuracy of 96% and 95.5% respectively for Euclidean and Manhattan distance respectively on Dataset 1.Accuracy on Dataset 2 was 97% for both distance metrics. We did further experimentation by modifying certain distinguishing features of both datasets within allowable variance and noted results. Our results are quite promising and hint at the possible use of kNN in appropriate CAD systems to help doctors, researchers and medical students alike.**

Keywords—Thyroid Diagnosis, k-Nearest Neighbor, Classification.

I. INTRODUCTION

Thyroid is a small gland found at the base of neck, just below Adam's apple. It produces two main hormones - T3 (Triiodothyronine) and T4 (Thyroxine). An important gland, thyroid controls how quickly the body burns energy, makes proteins, and how sensitive the body should be to other hormones. The production of too little thyroid hormone causes Hypothyroidism (underactive thyroid) while production of too much thyroid causes Hyperthyroidism (overactive thyroid). These are the 2 most common thyroid problems [14].

KNN Classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. Each tuple represents a point in an n-dimensional space. When given an unknown tuple, a k-nearest neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. Based on these k training tuples are the k "nearest neighbors" of the unknown tuple. The unknown tuple is classified by a majority vote of its neighbors, and gets assigned to the class most common amongst its k-nearest neighbors. When given a training tuple k-Nearest Neighbor simply stores it and waits until it is given a test tuple. Hence it is a "lazy learner" as it stores the training tuples or the "instances", they are also known as "Instance-Based Learners".[20]

Thus, k is a positive integer and decides how many neighbours influence the classification. "Closeness" is defined in terms of a distance metric such as "Euclidean Distance" or "Manhattan Distance".

II. LITERATURE SURVEY

k-NN has been used in the past in medical disease classification. A brief survey of the same is as follows:

In [1] a small variation is added to k-NN where the selection of k neighbours depends on a parameter n taken as input from the user and depending on the size of the smallest class. Novel k-NN worked best when the value of k increased and some class sizes decreased.

Representation of outputs with various distances like-Euclidean, cosine, correlation, cityblock are used in the k-NN in [2].Helps to know the response of a classifier for the desired application. Comparative Study of the distances is done to find the efficient distance to be used in an algorithm.

[3] Focuses on computing the probability of occurrence of a particular ailment from the medical data by mining it using a unique algorithm which increases accuracy of such diagnosis by combining the key points of neural networks, Large Memory Storage and Retrieval, k-NN, and differential diagnosis all integrated into one single algorithm which include diagnosis of multiple diseases showing similar symptoms, diagnosis of a person suffering from multiple diseases.

[4] presents a CAD based technique for automatic classification of benign and malignant thyroid lesions in 3D contrast-enhanced ultrasound images. DWT and texture based features were extracted and the resulting feature vectors were used to train and test three different classifiers- k-NN, PNN and DT.

An automatic classification system is proposed for tumor classification of MRI images in [5]. Neural Network and k-NN were used to classify the tumor as normal or abnormal.

[6] diagnoses the erythemato squamous diseases by using the basic and weighted k-NN. It compares the performance between the basic and weighted k-NN; and also between Manhattan and Euclidean. The weighted k-NN method with the Manhattan distance measure gave a better accuracy rate of 96.36% and 95.53%respectively.

[7] checks the performance of fuzzy disease diagnosis by comparing its results with k-NN and NB to diagnose the diseases- chicken pox, dengue and flu.

[8] addresses breast cancer diagnosis as a pattern classification problem. The k-NN algorithm yields the best performance that is obtained on the breast cancer diagnosis problem.

To discriminate healthy people from those suffering from appendicitis; [9] presents a genetic based feature selection approach, Bayesian Classifier and k-NN were used. Results show that this feature selection in combination with NB performs much better than other techniques on appendicitis dataset.

An automatic system that classifies the thyroid images into Benign and Malignant (cancerous) Nodules. SVM, k NN and Bayesian are used. It is observed that the SVM gives much better accuracy than k-NN and Bayesian.

In [11] k-NN is used on heart dataset to investigate the diagnosis of heart disease by voting which is integrated with k-NN. Results yielded an accuracy of 97.4%, while also showing that applying voting could not enhance the k-NN accuracy in the diagnosis of heart disease.

In [12] an adaptive k-nearest neighbor (AdaNN) is developed to overcome the limitation of kNN. AdaNN algorithm finds out the suitable k for each test example. It finds the optimal value of k and selects the few number of the nearest neighbours to get the correct class label. Results show that AdaNN performs better than the traditional k-NN algorithm.

In [13] SVM and k-NN are used for CBIR using texture and shape feature. Feature optimization is done on the extracted features. Classification is done into three categories - normal, benign and malignant. The query image is classified by the classifier to a particular class and the relevant images are retrieved from the database.

Advantages of k-Nearest Neighbor are - very simple implementation, Robust with regard to the search space; for instance, classes don't have to be linearly separable, few parameters to tune: distance metric and k.

Disadvantages of k-Nearest Neighbor are - expensive testing of each instance, as we need to compute its distance to all known instances. This is problematic for datasets with a large number of attributes. Sensitiveness to noisy or irrelevant attributes, which can result in less meaningful distance numbers. large storage requirements.

2 datasets taken from UCI Learning Repository are used. Both the datasets are numerical. Each dataset has many attributes and a class label. The choice of assigning k is given to the user, depending on how many comparisons is desired.

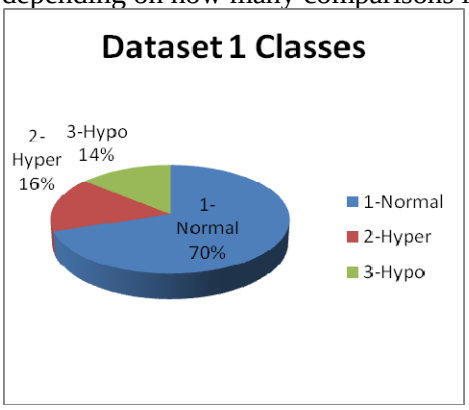

Fig.1. Class wise Distribution of DataSet1

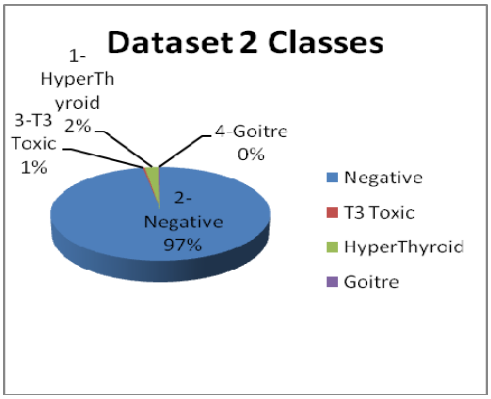

Fig.2. Class wise Distribution of DataSet1

III. DATA PREPROCESSING

A. *Attribute Selection*

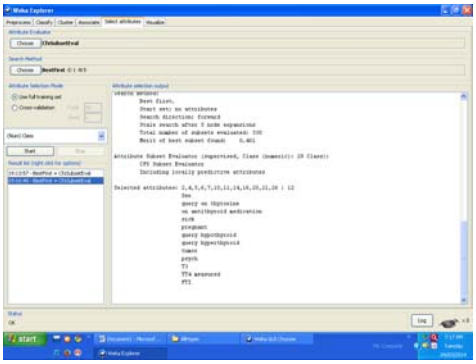

Fig.3. Feature Selection Using Weka Tool

Data Mining problems suffer from 'curse of dimensionality', hence we have performed Feature Selection using Weka tool.

The second dataset we are using has 30 attributes (including the class label) and 3772 total instances , of which 2800 are training tuples and 972 are test tuples. Feature Selecion was important here as we had to reduce the Feature Space from the present 28 dimensional space. Hence, with an aim to reduce the number of features used in Classification and maintain good classification accuracy, Search Method of Best Fit and Attribute Evaluator - CfsSubsetEval was applied on the dataset. As shown in Figure 3, only 12 relevant attributes were selected.

B. *Treatment of Missing Values*

Visual inspection of this Dataset revealed that there are several missing values. We wrote a program to compute missing values and came to a conclusion that there is large number of missing values in Dataset2. So there was a strong need to treat these, else accuracy would be affected adversely. As the data is skewed, we replaced each attribute having missing values by median of that attribute.

We wrote a program to normalize the data using Min Max Normalization. This normalized Dataset was used in all further process.

IV. EXPERIMENTAL RESULT

For Dataset 1 we performed Ten-fold stratified cross validation to evaluate k-NN. Here, the dataset was split into ten nearly equal parts, each part containing the same proportion of instances. In the first experiment, nine folds were used for training and the tenth fold was used for testing. This process was repeated nine more times with a different fold as test set.

Dataset 2 contains 2 separate files for training and testing. The training data was used for training the model in the first iteration and for testing the model in the second iteration. Similarly, the test data (in testing file) was used for training the model in the first iteration and for testing the model in the second iteration.

We have used both the distance measures – Euclidian and Manhattan in calculating closeness between tuples for the 2 Datasets.

TABLE I. k=1

Experiment No.	Test Dataset Used	% Classification	MAE	Classification Time(ms)
1	1	95	0.0952	17
2	2	95	0.0952	16
3	3	100	0	16
4	4	95	0.0952	17
5	5	100	0	15
6	6	95	0.0476	16
7	7	100	0	16
8	8	95	0.0476	16
9	9	100	0	16
10	10	85	0.2857	16

TABLE II. RESULT OF K-NN FOR EUCLIDEAN DISTANCE

Dataset Used	No. of Features in original Dataset	No. of Features considered	Optimum Value of k
Dataset 1	5	5	1
Dataset 2	29	12	5

TABLE III. RESULT OF K-NN FOR EUCLIDEAN DISTANCE

Dataset Used	% Classification	Classification Time
Dataset 1	96%	12 ms
Dataset 2	97%	3 s

TABLE IV. RESULT OF K-NN FORMANHATTAN DISTANCE

Dataset Used	No. of Features in original Dataset	No. of Features considered	Optimum Value of k
Dataset 1	5	5	3
Dataset 2	29	12	7

TABLE V. RESULT OF K-NN FORMANHATTAN DISTANCE

Dataset Used	% Classification	Classification Time
Dataset 1	95.5%	120ms
Dataset 2	97%	31s

TABLE VI. RESULT ANALYSIS FOR DISTANCE METRICS

Distance Metric used	Dataset 1		
	Value of k	% Classification	MAE
Euclidean	1	96%	0.0667
Manhattan	3	95.5	0.0666

TABLE VII. RESULT ANALYSIS FOR DISTANCE METRICS

Distance Metric used	Dataset 2		
	Value of k	% Classification	MAE
Euclidean	5	97%	0.0339
Manhattan	7	97%	0.0325

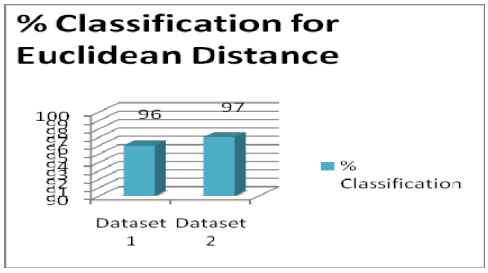

Fig.4. Classification Accuracy for k-NN with Euclidean Distance

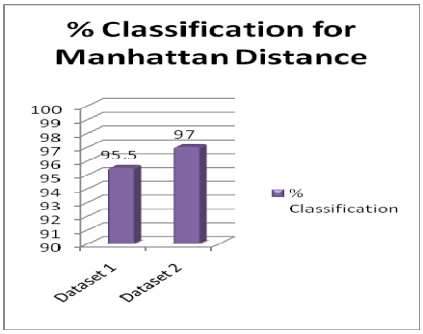

Fig.5. Classification Accuracy for k-NN with Manhattan Distance

Results of Classification accuracy on the two standard datasets using both the metrics- Euclidean and Manhattan are as shown in Fig 4 and 5. It is observed that Euclidian has given a better accuracy.

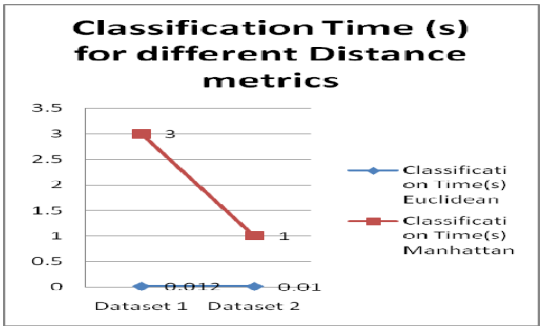

Fig. 6 Classification Time for Different Distance Metrics

As shown in Figure 6, the algorithm needed more time to classify the test tuples in the case of Manhattan metric.

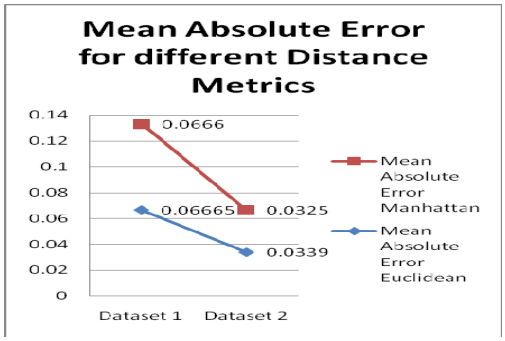

Fig.7. Mean Absolute Error for Different Distance Metrics

As shown in figure 7, it was observed that Dataset 2 yielded lower error rates of 0.0339 and 0.0325 respectively for Euclidian and Manhattan metrics respectively.

V. CONCLUSION

In this work, we have applied k-nearest neighbor classifier on two standard thyroid datasets. Our results show that the k-NN classifier presented a very high classification accuracy of 97% using Euclidean and Manhattan Distances respectively. The high accuracy encourages us to validate the system using a larger and a different medical dataset in order to establish its clinical applicability to assist doctors in thyroid classification and subsequent treatment regime.

REFERENCES

[1] Anjali Ganesh Jivani "The Novel k Nearest Neighbor Algorithm", International Conference on Computer Communication and Informatics (ICCCI -2013), IEEE 2013.

[2] Aman Kataria, M. D. Singh "A Review of Data Classification Using k-Nearest Neighbour Algorithm", International Journal of Emerging Technology and Advanced Engineering June 2013.

[3] Rahul Isola, Rebeck Carvalho, and Amiya Kumar Tripathy, "Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and k-NN", IEEE Transactions On Information Technology In Biomedicine, 2012.

[4] U.Rajendra Acharya, Vinitha Sree S, Filippo Molinari, Roberto Garberoglio, Agnieszka Witkowska, Jasjit S Suri, "Automated Benign & Malignant Thyroid Lesion Characterization and Classification in 3D Contrast-Enhanced Ultrasound", 34th Annual International Conference of the IEEE EMBS San Diego, California USA, 28 August - 1 September, 2012.

[5] Amer Al-Badarneh, Hassan Najadat, "A Classifier to Detect Tumor Disease in MRI Brain Images", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2012.

[6] Hatice Cataloluk, Metin Kesler, "A Diagnostic Software Tool for Skin Diseases with Basic and Weighted K-NN," IEEE 2012.

[7] Asaad Mahdi, Ahmad Razali, Ali AlWakil, "Comparison of Fuzzy Diagnosis with K-Nearest Neighbor and Naive Bayes Classifiers in Disease Diagnosis", Broad Research in Artificial Intelligence and Neuroscience Volume 2, Issue 2, May-June 2011.

[8] Manish Sarkar, Tze-Yun Leong, "Application of K-Nearest Neighbors Algorithm on Breast Cancer Diagnosis Problem."

[9] E.Sivasankar, R.S.Rajesh, "Design and Development of a Clinical Decision Support System for diagnosing appendicitis",IEEE 2012.

[10] Nikita Singh, Alka Jindal ,"Ultra sonogram Images for Thyroid Segmentation and Texture Classification in Diagnosis of Malignant (Cancerous) or Benign (Non-Cancerous) Nodules", IJEIT May 2012.

[11] Mai Shouman, Tim Turner, Rob Stocker "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients", International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012.

[12] Shiliang Sun, Rongqing Huang , " An Adaptive k-Nearest Neighbor Algorithm", Seventh International Conference on Fuzzy Systems and Knowledge Discovery 2010.

[13] Mohanapriya.S, Vadivel.M "Automatic Retrieval of MRI brain Image using multiqueries system".

[14] Jiawen Han, Micheline Kamber, "Data Mining Concepts and Techniques".

[15]http://archive.ics.uci.edu/UCIMachineLearningRepositoryThyroidDiseaseDataSet