IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. IV (May-Jun. 2016), PP 87-91
www.iosrjournals.org

Heart Disease Detection using EKSTRAP Clustering with Statistical and Distance based Classifiers

Terence Johnson[1], Dr. S. K. Singh[2], Vaishnavi Kamat[3], Aishwarya Joshi[4], Lester D'Souza[4], Poohar Amonkar[4], Devyani Joshi[4], Anirudha Kulkarni[4]

[1]*(Ph.D Scholar, Information Technology Department, AMET University, India)*
[1]*(Asst. Prof., Computer Engineering Department, AITD, Goa University, India)*
[2]*(Head, Information Technology Department, TCSC, Mumbai University, India)*
[3]*(Asst. Prof., Computer Engineering Department, AITD, Goa University, India)*
[4]*(Students, Computer Engineering Department, AITD, Goa University, India)*

Abstract : *The heart is the most important organ in the human body which pumps blood to various parts of the body. If there is inefficient circulation of blood in body organs like brain will suffer. If heart stops pumping blood it results in death. An individual's life is very much dependent on how efficiently the heart works. Using data mining technique proposed in this paper we are trying to detect if a patient has heart disease or not. The system uses 13 attributes like age, gender, blood pressure, cholesterol etc to detect the same. The system uses a hybrid technique which uses Enhanced K STRAnge Points(EKSTRAP) clustering algorithm , output of which is given to different classifiers like statistical –Naïve Bayes classifier and Distance Based – MSDC (Modified Simple Distance Classifier).*

Keywords: *Enhanced K Strange, clustering, Heart Disease, Naïve Bayes Classifier*

I. Introduction

When we have large pre-existing dataset, we can examine it and generate new information previously not observed by processing the data. For processing we can use different data mining techniques, in this paper clustering and classification techniques are used [1]. Clustering is an unsupervised process of grouping similar objects from a given dataset. The similarity is determined using many techniques like Euclidian Distance [2]. Clustering is achieved by the Enhanced k strange points clustering algorithm [3]. Classification on the contrary is a supervised technique to check, to which group the point belongs to given the groups. Statistical classification [4] and Distance based classification [5] techniques are used. Naïve Bayes [6], a simple probabilistic classifier is the statistical classifier used and the Modified Simple Distance classifier is the Distance based classifier used in this paper. Heart Disease Detection [7] tells whether the new data object has heart disease or not based on the training data given to the classifier. The training data is obtained by clustering. The dataset used is cleveland dataset which is taken from the UCI Repository. The block diagram below explains the flow of the implemented system. Entire dataset is grouped into classes using clustering algorithm. The output of the clustering algorithm along with the new tuple is given to the classifier which then detects to which class the new tuple belongs.

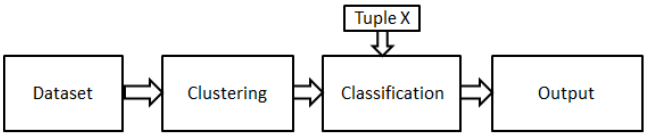

Fig.1.Block Diagram

The clustering and classification is done based on the following attributes:-

1. Age-Age in years.
2. Gender-
a. 1 is male
b. 0 is female.
3. Cp –Chest Pain type
a. value 1-Typical angina
b. value 2-Atyical angina
c. value 3-Non Anginal pain
d. value 4 -Asymptomatic

4. trestbps-resting blood pressure(in mm Hg on admission to the hospital)
5. chol-(Cholestrol) serum cholesterol in mg/dl.
6. fbs- (fasting blood sugar)
a. if >120 mg/dl then 1
b. else 0
7. restecg-resting electrocardiographic results.
a. Value 0 –normal
b. Value 1-having ST-T wave abnormality
c. Value2 –showing probable or definite left ventricular hypertrophy
8. Thalach-maximum heart rate achieved
9. Exang- exercise induced angina
a. 1-yes
b. 0-no
10. Oldpeak-ST depression induced by exercise relative to rest
11. Slope-the slope of the peak exercise ST segment.
a. Value1-Upsloping
b. Value2-Flat
c. Value3-downsloping
12. Ca-number of major vessels(0-3) colored by flouroscopy
13. Thal-
a. 3-Normal
b. 6-Fixed Detect
c. 7-reverable detect
14. Num-diagnosis of heart disease

II. Motivation

If the clustering accuracy is improved then the classification result will improve. The existing system uses k means for clustering. K Means has many limitations. When the number of attributes increase the number of iterations increases to converge and sometimes it may not converge, so we may have to forcefully stop the computation using stopping criteria‘t’ (maximum number of iterations). This may lead to wrong clusters. So to improve the efficiency of the classifier we replaced k means with enhanced k strange points clustering algorithm. The benefit of using EKSTRAP (Enhanced K STRAnge Points clustering algorithm) is that there is no need of iterations so it converges faster. And even when the number of attributes is more, the performance of EKSTRAP (Enhanced K STRAnge Points clustering algorithm) is better.

III. Proposed Methodology

Here we have replaced K-Means with EKSTRAP (Enhanced K STRAnge Points clustering algorithm).
Input:-Heart Disease Dataset.
Output: - Tells if the new tuple falls in cluster 0 or cluster 1.

Step 1:-Dataset is given as input to Enhanced k strange points clustering algorithm, which will form two clusters, class 0 and class 1.
Cluster 0- No Heart disease.
Cluster 1-Has Heart disease.
This will act as training data.

Step 2 :- Using this training data, whenever a new tuple is given using Naïve Bayes classifier we try to detect to which class the newly entered tuple belongs.
Enhanced K strange points clustering algorithm (EKSTRAP)
Input:-Number of clusters required (N) and dataset.
Output:-Set of N clusters.

Step 1:- Find the point which will be at the minimum distance from origin “min” using Euclidian Distance.

Step 2:-Find the point which will be at the maximum distance from” min” call it “max”.

Step 3:-Now find the strange point “strange” which will at maximum distance from “min” and “max”.

Step 4:-This Strange point may not be equidistant from “min” and “max” so we make it equidistant by applying one of the following formulae.

If point is near to “min”:- Str=Str(prev)+Xm[max-str(prev)]/[N-1]
If point is near to “max”:- Str=min+Xm[max-min]/[N-1]
Xm ranges from 1,2,3,…..to N-2.

Step 5:- Repeat step 4 until all N strange points are located.

Step 6:- Assign remaining points from dataset to these N different clusters.

Step 7:-Output the N clusters.

IV. Experimental Results

Clustering: The entire dataset contains 297 tuples. For this dataset after clustering using K-Means, Class 0 contains 186 tuples and Class 1 contains 111. This result was compared to the actual dataset, in Class 0, 110 tuples were clustered correctly and in Class 1, 61. The same dataset when given to EKSTRAP Clustering Algorithm, Class 0 contains 193 tuples and Class 1 contains 104 .This result when compared to the actual dataset 144 tuples in Class 0 and 88 in Class 1 were correctly clustered. The actual dataset contains 160 in Class 0 and 137 tuples in Class 1.

Table.1. Clustering Output

Algorithm Used	Class 0	Class 1
Actual	160	137
K-Means	186	111
EKSTRAP	193	104

Tabel.2. Clustering Accuracy

Algorithm Used	Class 0	Class 1
K-Means	68.75%	44.53%
EKSTRAP	90.00%	64.23%

Statistical Classifier:

Naïve Bayes Classifier: The Naive Bayes is a conditional probability system. Given a unlabeled tuple to be classified, represented by a vector x=(x_1,…..,x_n) representing some n independent dimension, it assigns to this the tuple probabilities

$$p(C_k|x_1,\dots,x_n)$$

for each of *K* possible categories

The issue with the above design is that if the number of dimensions *n* is large or if a dimension can take on a large number of values, then building such a model on probability tables is not viable. We therefore reframe the model to make it more compliant. Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

In simple terms, using Bayesian probability jargon, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

For classification 10 tuples were removed, these 10 and the remaining 287 tuples were clustered separately using K-Means and EKSTRAP Clustering Algorithm. The clusters formed by the 287 tuples were given to the Naive Bayes Classification as training data. Then these tuples were given to the classification algorithm. When the 287 tuples were clustered, Class 0 had 180 tuples and Class1 had 107 tuples using k-means, whereas using EKSTRAP, class 0 had 185 and class 1 had 102.

Table.3. Classification Output (K-Means)

k-Means: Clustering o/p	Classification o/p using k-means	Comment
Class 0	Class 1	Wrong
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct

Class 1	Class 1	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct
Class 1	Class 0	Wrong

Table.4. Classification Output(EKSTRAP)

EKSTRAP: Clustering o/p	Classification o/p using EKSTRAP	Comment
Class 1	Class 1	Correct
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 0	Class 1	Wrong
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct

Table.5. Classification Accuracy

Training Data	Accuracy
o/p of K-Means	80%
o/p of EKSTRAP	90%

Distance based Classifier:

Modified Simple Distance Classifier: In the simple distance classifier the means of each of the classes are computed and the distance of the unlabelled tuple from each of these means is calculated. The unlabelled tuple is assigned to the class whose mean is closest to the unlabelled tuple.

In the modified simple distance classifier the minimum of one class and the maximum distance from the minimum of another class is computed. If there are three classes a point in the third class is found which is farthest from the minimum and the maximum of the earlier two classes. The distance of the unlabelled tuple from each of these farthest points is calculated. The unlabelled tuple is assigned to the class having the minimum if distance from the unlabelled tuple to the minimum is lesser that its distance to the maximum and third farthest point.

If however the distance of the unlabelled tuple is lesser to the maximum than to the minimum and 3^{rd} farthest point, it is assigned to the class having the maximum. Similarly, if the distance of the unlabelled tuple is lesser to the 3^{rd} farthest point than to the minimum and maximum, it is assigned to the class having the 3^{rd} farthest point.

For classification 25 tuples were extracted from the dataset and these 25 and the remaining 272 tuples were clustered separately using K-Means and EKSTRAP Clustering Algorithm. The clusters formed by the 272 tuples were given to the Simple Distance Classifier as training data. Then these tuples were given one by one to the classification algorithm.

When the 272 tuples were clustered, Class 0 had 110 tuples and Class1 had 177 tuples using k-means, whereas using EKSTRAP, class 0 had 178 and class 1 had 94.

Table.6. Classification Output (EKSTRAP)

EKSTRAP: Clustering o/p	Classification o/p using EKSTRAP	Comment
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct
Class 1	Class 1	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct

Class 1	Class 1	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct
Class 0	Class 0	Correct
Class 1	Class 1	Correct
Class 0	Class 0	Correct

Table.7. Modified Simple Distance Classification Accuracy

Training Data	Accuracy
o/p of EKSTRAP	100%

V. Conclusion

In this paper it was proved that improving the clustering algorithm, the output of classification can be improved. The existing system used k-means clustering algorithm which was replaced by EKSTRAP (Enhanced K STRAnge Points) clustering algorithm. The accuracy of the algorithms are 68.75% for class 0, 44.53% for class 1 for k means and 90.00% for class 0 and 64.23% for class 1 using EKSTRAP. EKSTRAP clustering algorithm eliminates the randomness used while selecting the centroid for k means. The accuracy of the Naïve Bayes classifier also improves, as it classifies 8 out of the 10 tuples correctly using k-means output as training data whereas while using EKSTRAP clustering algorithm it classifies 9 out of the 10 tuples correctly. The accuracy with the modified simple distance classifier is also 100% as it correctly classifies 15 out of the 15 class 0 tuples and 10 out of the 10 class 1 tuples accurately. Hence the proposed system accuracy is good. In our future work, this can further enhanced and expanded. For detecting heart disease the number of attributes can be increased.

References

[1]. David Hand, Heikki Mannila and Padhraic Smyth, *principles of data mining*, Cambridge, MA: MIT Press, (2001).

[2]. Jiawei Han and Micheline Kamber, *data mining - concepts and techniques*, Elsevier, Second Edition, Original ISBN: 978-1-55860-901-3, Indian Reprint ISBN: 978-81-3120535-8

[3]. Johnson Terence,Dr.Santosh Kumar Singh," Enhanced K-Strange Points Clustering Algorithm" International Conference on Emerging Information Technology and Engineering Solutions(EITES 2015),78-1-4799-1838-6/15,IEEE, DOI:10.1109/ EITES.2015.14,pp 32-37.

[4]. Kuncheva L, "On the equivalence between fuzzy and statistical classifiers" Int. J. of Uncertainity, Fuzziness and Knowledge based Systems. Vol. 4, No. 3, 1996, pp 245-253

[5]. Margaret Dunham , *data mining – introductory and advanced topics*, Pearson Education 2006.ISBN: 8177587854, 9788177587852

[6]. N. Friedman, D. Geiger, and Goldszmidt M. Bayesian network classifiers. Machine Learning, 29:131–163, 1997

[7]. Rucha Shinde , Sandhya Arjun ,Priyanka Patil and Prof Jayashree Waghmare "An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naive Bayes Algorithm" ,International Journal of computer science and Information Technology Vol 6 (1) 2015 637-639