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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

A Novel Approach to Cancer Prediction Using Weka Tool

Ms. Vennila Santhanam

Asst. Professor, Department of Computer Science, Auxilium College (Autonomous), Vellore, Tamil Nadu, India

ABSTRACT: Cancer is a major root cause of disease among human deaths in many developed countries. Cancer classification in medical practice trusted on clinical and histopathological facts may produce incomplete or misleading results. Cancer is a dreadful disease. Millions of people died every year because of this disease. It is very essential for medical practitioners to opt a proper treatment for cancer patients. Therefore cancer cells should be identified correctly. The method predicted will help in medical diagnosis and guide researchers to develop most cost effective and user friendly systems, processes and approaches for clinicians. Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields. We recommend the weka tool for applying different clustering algorithms and predict a useful result that will be very helpful for the prediction of cancer. This will help the medical practitioners to treat the cancer patients in the right way.

KEYWORDS: Data clustering, K-Means Clustering, Hierarchical Clustering, DBScan Clustering, EM Algorithm Data mining algorithms, Weka tools, K-means algorithms, Clustering methods etc.

I. INTRODUCTION

History of Cancer

The genetic basis of cancer was recognised in 1902 by the German zoologist Theodor Boveri, professor of zoology at Munich and later in Würzburg. He discovered a method to generate cells with multiple copies of the centrosome, a structure he discovered and named. He postulated that chromosomes were distinct and transmitted different inheritance factors. He suggested that mutations of the chromosomes could generate a cell with unlimited growth potential which could be passed on to its descendants. He proposed the existence of cell cycle check points, tumour suppressor genes and oncogenes. He speculated that cancers might be caused or promoted by radiation, physical or chemical insults or by pathogenic microorganisms.

The political 'war' on cancer began with the National Cancer Act of 1971, a United States federal law. The act was intended "to amend the Public Health Service Act so as to strengthen the National Cancer Institute in order to more effectively carry out the national effort against cancer". It was signed into law by then U.S. President Richard Nixon on December 23, 1971.In 1973, cancer research led to a cold war incident, where co-operative samples of reported oncoviruses were discovered to be contaminated by HeLa. Since then the war against cancer continued.

Despite this substantial investment, the country has seen just a five percent decrease in the cancer death rate (adjusting for size and age of the population) between 1950 and 2005. Longer life expectancy may be a contributing factor to this, as cancer rates and mortality rates increases significantly with age, more than three out of five cancers are diagnosed in people aged 65 and over.

Over the past decades, a continuous evolution related to cancer research has been performed. Scientists applied different methods, such as screening in early stage, in order to find types of cancer before they cause symptoms. Moreover, they have developed new strategies for the early prediction of cancer treatment outcome. With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians

Cancer is a major root cause of disease among human deaths in many developed countries. Cancer classification in medical practice trusted on clinical and histopathological facts may produce incomplete or misleading results.

Methods to diagnose cancer

Various machine learning techniques are used for cancer classification. Neural network techniques are very useful for detection and monitoring of cancer. Artificial neural network is a robust tool recently used as either clustering or classification of gene expression data. Supervised models are used for classification while unsupervised models are used for clustering. Neural network model has been successfully implemented in various classification problems.

Classification is crucially important for cancer diagnosis and treatment. A classification problem occurs when an object needs to be allotted into a predefined group or class based on a number of observed attributes related to that object.

Clustering can be considered as unsupervised learning problem, it deals with finding a structure in collection of unlabeled data. Artificial neural network had been proven a very effective method for pattern recognition. This made them very useful for diagnosis of cancer disease at very early stages.

As a result, Machine Learning(ML) methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type. Given the significance of personalized medicine and the growing trend on the application of ML techniques, we here present a review of studies that make use of these methods regarding the cancer prediction and prognosis. In these studies prognostic and predictive features are considered which may be independent of a certain treatment or are integrated in order to guide therapy for cancer patients, respectively In addition, we discuss the types of ML methods being used, the types of data they integrate, the overall performance of each proposed scheme while we also discuss their pros and cons.

Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the likeness (or homogeneity) within a group, and the greater the disparity between groups, the ―better‖ or more distinct the clustering.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Basic types of Clustering

Broadly speaking, clustering can be divided into two subgroups :

- **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each customer is put into one group out of the 10 groups.
- **Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each costumer is assigned a probability to be in either of 10 clusters of the retail store.

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

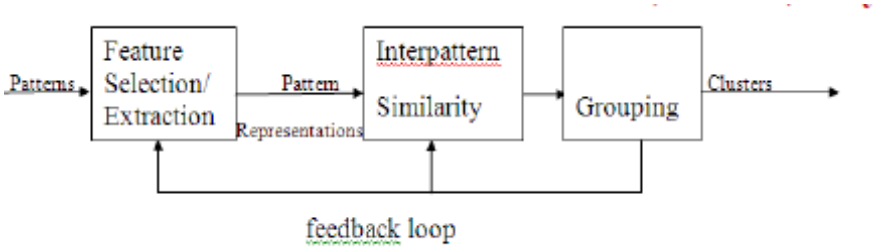

Fig. 1 Steps in Clustering

Clustering Models

Since the task of clustering is subjective, the means that can be used for achieving this goal are plenty. Every methodology follows a different set of rules for defining the '*similarity*' among data points. In fact, there are more than 100 clustering algorithms known. But few of the algorithms are used popularly, let's look at them in detail:

- **Connectivity models:** These models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches.These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.
- **Centroid models:** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.
- **Distribution models:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.
- **Density Models:** These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.

Types of Clustering

I. K Means Clustering

K means is an iterative clustering algorithm that aims to find local maxima in each iteration. k-means clustering is a partitioning based clustering technique of classifying/grouping items into k groups (where k is user specified number of clusters). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid. A centroid (also called mean vector) is "the center of mass of a geometric object of uniform density".

Advantages

- Easy to implement
- With a large number of variables, K-Means may be computaHonally faster than hierarchical clustering (if K is small).
- k-Means may produce Hghter clusters than hierarchical clustering
- An instance can change cluster (move to another cluster) when the centroids are recomputed.

Disadvantages

- Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome).

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

- Fixed number of clusters can make it difficult to predict what K should be.
- Does not work well with non-globular clusters.

II. HIERARCHICAL CLUSTERING

Hierarchical clustering, as the name suggests is an algorithm that builds hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.

Advantages

- Hierarchical clustering outputs a hierarchy, ie a structure that is more informaHve than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrogram .
- Easy to implement

Disadvantages

- It is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around.
- Time complexity: not suitable for large datasets
- The order of the data has an impact on the final results
- Very sensitive to outliers

III. MEAN-SHIFT CLUSTERING

Mean shift clustering is a sliding-window-based algorithm that attempts to find dense areas of data points. It is a centroid-based algorithm meaning that the goal is to locate the center points of each group/class, which works by updating candidates for center points to be the mean of the points within the sliding-window. These candidate windows are then filtered in a post-processing stage to eliminate near-duplicates, forming the final set of center points and their corresponding groups

IV. DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE (DBSCAN)

DBSCAN is a density based clustered algorithm similar to mean-shift. Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature

Advantages

- It does not require a pe-set number of clusters at all.
- It also identifies outliers as noises unlike mean-shift which simply throws them into a cluster even if the data point is very different.
- It is able to find arbitrarily sized and arbitrarily shaped clusters quite well.

Disadvantages

- It doesn't perform as well as others when the clusters are of varying density. This is because the setting of the distance threshold ε and minPoints for identifying the neighborhood points will vary from cluster to cluster when the density varies.

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

- The above limitation also occurs with very high-dimensional data since again the distance threshold ε becomes challenging to estimate.

V. EXPECTATION–MAXIMIZATION (EM) CLUSTERING USING GAUSSIAN MIXTURE MODELS (GMM)

One of the major drawbacks of K-Means is its naive use of the mean value for the cluster center. We can see why this isn't the best way of doing things by looking at the image below. On the left hand side it looks quite obvious to the human eye that there are two circular clusters with different radius' centered at the same mean. K-Means can't handle this because the mean values of the clusters are a very close together. K-Means also fails in cases where the clusters are not circular, again as a result of using the mean as cluster center.

Advantages

- GMMs are a lot more **flexible** in terms of **cluster covariance** than K-Means; due to the standard deviation parameter, the clusters can take on any ellipse shape, rather than being restricted to circles. K-Means is actually a special case of GMM in which each cluster's covariance along all dimensions approaches 0.
- Since GMMs use probabilities, they can have multiple clusters per data point. So if a data point is in the middle of two overlapping clusters, we can simply define its class by saying it belongs X-percent to class 1 and Y-percent to class 2. I.e GMMs support **mixed membership**.

VI. COBWEB

It is an incremental system for hierarchical conceptual clustering. COBWEB was invented by Professor Douglas H. Fisher, currently at Vanderbilt University.COBWEB incrementally organizes observations into a classification tree. Each node in a classification tree represents a class (concept) and is labeled by a probabilistic concept that summarizes the attribute-value distributions of objects classified under the node. This classification tree can be used to predict missing attributes or the class of a new object.[3]

There are four basic operations COBWEB employs in building the classification tree. The operations are:

1. **Merging Two Nodes**
   Merging two nodes means replacing them by a node whose children is the union of the original nodes' sets of children and which summarizes the attribute-value distributions of all objects classified under them.
2. **Splitting a node**
   A node is split by replacing it with its children.
3. **Inserting a new node**
   A node is created corresponding to the object being inserted into the tree.
4. **Passing an object down the hierarchy**
   Effectively calling the COBWEB algorithm on the object and the subtree rooted in the node.

VII. CLUSTERING IN WEKA

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering. Here we are working only with the clustering because it is most important process, if we have a very large database. Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Using Weka tools for clustering is a very effective method. It provides a better interface to the user than compared the other data mining tools.

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

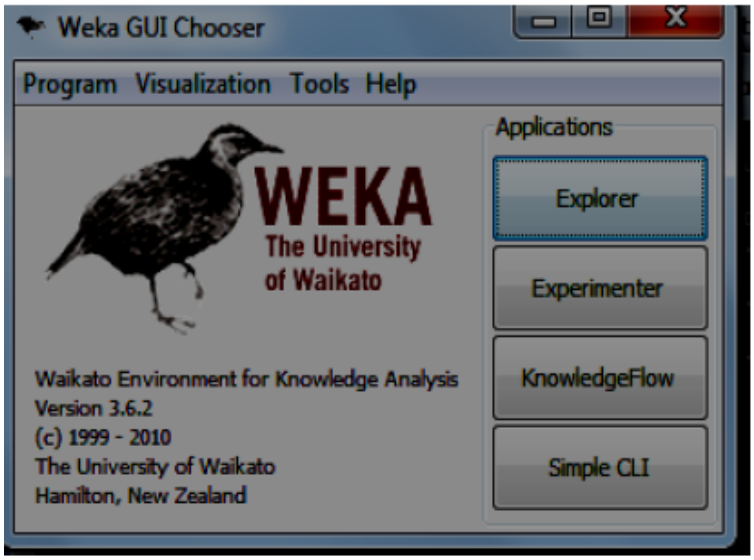

Fig.2 Front Page Of Weka Tool

Weka

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License

Methodology

The methodology is very simple. Take the past project data from the repositories and apply it on the weka. Apply different clustering algorithms and predict a useful result.

Performing Clustering In Weka

For performing cluster analysis in weka. We have to load the data set in weka. For the weka the data set should have in the format of CSV or .ARFF file format. If the data set is not in arff format we need to be converting it. To perform clustering, we have to click the cluster button. After that we need to choose which algorithm is applied on the data. And then click ok button.

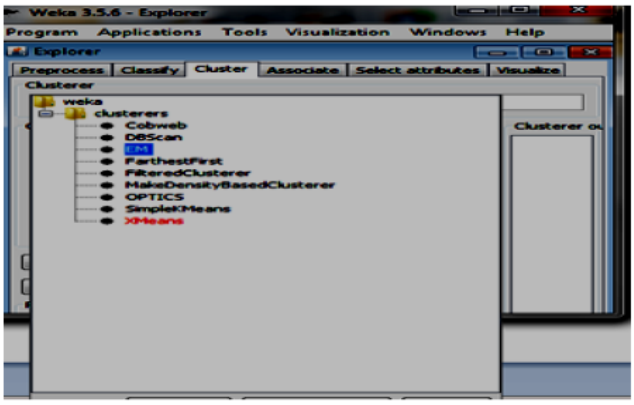

Fig. 3 Various clustering algorithms in Weka

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Cobweb Clustering Algorithm

The COBWEB algorithm yields a clustering dendrogram called classification tree that characterizes each cluster with a probabilistic description. Cobweb generates hierarchical clustering [2], where clusters are described probabilistically.

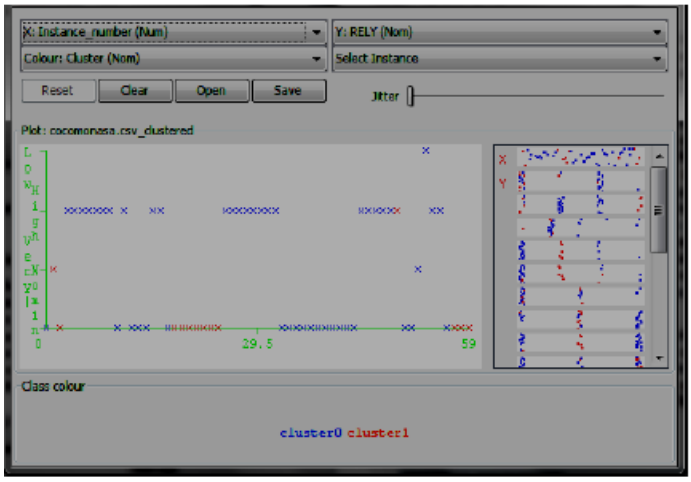

Fig.4 Cobweb clustering algorithm

DBSCAN Clustering Algorithm

DBSCAN (for density-based spatial clustering of applications with noise) is a data clustering algorithm proposed by Martin. DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. DBSCAN does not require you to know the number of clusters in the data a priori, as opposed to k-means.

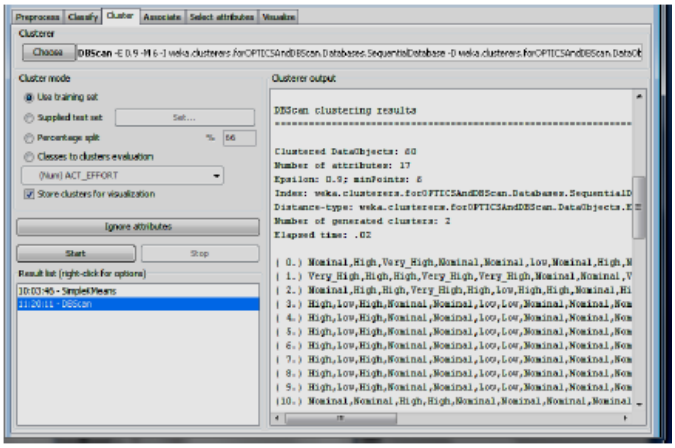

Fig.5 DBSCAN Algorithm

EM algorithm

This algorithm gives extremely useful result for the real world data set. This algorithm can be used when you want to perform a cluster analysis for a small scene or region-of- interest and are not satisfied with the results obtained from the k-means algorithm.

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

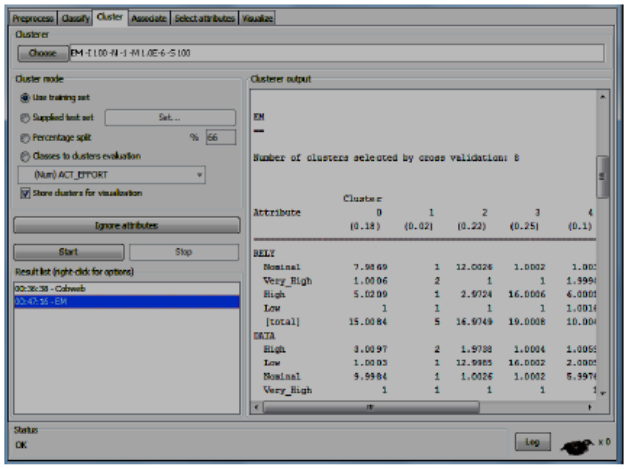

Fig. 6 EM algorithm

K-Means Clustering Algorithms

This algorithm can be used to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results into a partitioning of the data space into Verona cells. With a large number of variables, K-Means may be computationally faster than hierarchical clustering.

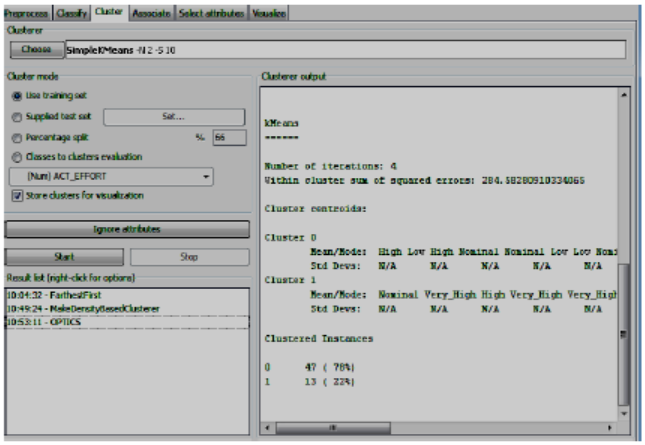

Fig. 7 K- Means Clustering Algorithms

K-Means Clustering In Weka Interface

K-means only allow numerical values for attributes. In that case, it may be necessary to convert the data set into the standard spreadsheet format and convert categorical attributes to binary. WEKA provides filters to accomplish all preprocessing tasks. But these are not necessary for clustering in WEKA. This is because WEKA SimpleKMeans algorithm automatically handles a mixture of categorical and numerical attributes. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations. The WEKA SimpleKMeans algorithm uses Euclidean distance measure to compute distances between instances and clusters.

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Fig. 8 Weka Interface

To perform clustering, select the "Cluster" tab in the Explorer and click on the "Choose" button. This results in a drop down list of available clustering algorithms. In this case we select "SimpleKMeans". Next, click on the text box to the right of the "Choose" button to get the pop-up window.

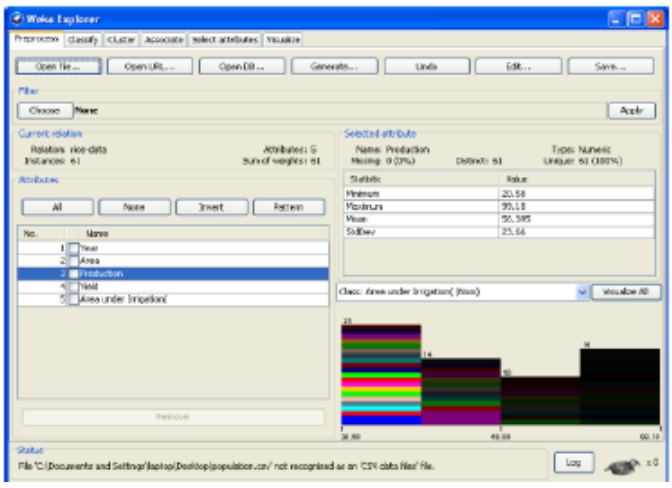

Fig. 9 WEKA Explorer interface

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

VIII. COMPARISON & RESULTS USING WEKA CLUSTERING TOOL

Name	Number of clusters	Cluster Instances	Number of Iteration	Within clusters sum of squared errors	Time taken to build model	Log likelihood	Unclustered Instances
DBSCAN	3	0:10 (40%) 1:6 (24%) 2:9 (36%)			1.03 Seconds		575
EM Algorithm	6	0:31 (5%) 1:97 (16%) 2:65 (11%) 3:184(31%) 4:92(15%) 5:131(22%)			76.94 seconds	-21.09024	0
Hierarchical Clustering	2	0:599(100%) 1:1 (0%)			1.16 Seconds		0
Density based Clusters	2	0:239(40%) 1:361(60%)	4	2016.6752520938053	0.06 Seconds	-22.04211	0

IX. CONCLUSION

Data mining techniques cover all the area in our life. We are using data mining techniques in mainly in the medical, banking, insurances, education etc. before start working in the with the data mining models, it is very necessary to knowledge of available algorithms. Weka is the data mining tools. It is the simplest tool for classify the data various types.. It is provide the past project data for analysis. We have shown the working of various algorithms used in weka. We have also shown the advantages and disadvantages of each algorithm. Every algorithm has their own importance and we use them on the behavior of the data, but on the basis of this research we found that k-means clustering algorithm is simplest algorithm as compared to other algorithms. Here we show only the clustering operations in the weka, Weka interface also plays an important role in the field of data mining.

.

REFERENCES

[1] https://en.wikipedia.org/wiki/History_of_cancer
[2] https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
[3] Privacy-Preserving K-Means clustering over vertically Partitioned Data- By Jaideep Vaidya and Chris Clifton, Deptt. Of Computer Sciences, Purdue University, 250 N University St, West Lafayette, IN 47907-2066
[4] K-means clustering via Principal component analysis – By Chris Ding and Xinofeng He, Computational research division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, Preceeding of 21st International conference on Machine Learning,Banff, Canada, 2004. [3] K-means clustering Tutorial- By Kardi Teknomo, Ph.D
[5] Khalid Hammouda, Prof. Fakhreddine Karray,‖A Comparative Study of Data Clustering Techniques‖ University of Waterloo, Ontario, Canada N2L 3G1
[6] Pavel Berkhin,―Survey of Clustering Data Mining Techniques‖, Accrue Software, Inc.
[7] Tapan kanungo, senior member IEEE David M. Mount, member IEEE ―An Efficient –means algorithm: analysis and implementation‖