Hierarchical Clustering For Cancer Discovery Using Range Check And Delta Check

Breetha S*, Kavinila R**

*** Department of Information Technology, Sri Krishna College of Technology, Coimbatore**

**** Department of Information Technology, Sri Krishna College of Technology, Coimbatore**

Abstract: Class discovery is one of the most important tasks in cancer classification using biomolecular data. To perform this, a multiple clustering approach called Hierarchical clustering is used. It uses one of the metrics called Manhattan Distance which measures the distance between the values of the data set and builds a hierarchy of clusters after analysing it. The clustering result enables to classify the cancer types and it is further evaluated by Range check and Delta check. The various test results are compared with the known initial range of values using Range check. Delta check is performed on the current test result and the immediate previous test result for better results. These techniques are used to improve the diagnosis of cancer.

Index Terms- Class Discovery, Hierarchical clustering, Range check, Delta check

I. INTRODUCTION

Cancer diagnosis and treatment involves discovering and classifying cancer types. Most of the previous works involve the single clustering algorithms. In Golub'swork [3], the self-organizing feature map and neighbourhood analysis were adopted to discover two types of human acute leukemia, which are acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). In Wigle's work [4], clustering approaches and statistical analysis were adopted to identify non-small cell lung cancer (NSCLC) from the normal cases. These works include certain limitations such as lack of robustness, stability and accuracy. But in our paper, we have adopted the concept called Hierarchical clustering which is one of the multiple clustering algorithms. This is the powerful method for improving both the robustness as well as the stability of unsupervised classification solutions.

Cancer classification using biomolecular data poses a major challenge in monitoring the levels of thousands of genes[1] [2], and this can be overcome by using the machine learning technique. For this purpose we use the WEKA tool which is the popular machine learning workbench. WEKA contains simple implementations of algorithms for classification, clustering, and association rule mining, along with graphical user interfaces and visualization utilities for data exploration and algorithm evaluation.

II. HIERARCHICAL CLUSTERING

Hierarchical methods for unsupervised and supervised data mining give multi-level description of data. It is relevant for many applications related to information extraction, retrieval navigation and organization [5]. Hierarchical clustering algorithm is of two types:

i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and

ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

We focus on agglomerative probabilistic clustering. This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. This way we go on grouping the data until one cluster is formed. Now on the basis of dendogram graph we can calculate how many numbers of clusters should be actually present.

Algorithmic steps for Agglomerative Hierarchical clustering

Let X = {x1, x2, x3, ..., xn} be the set of data points.

1) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.

3) Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].

4) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The

distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).

5) If all the data points are in one cluster then stop, else repeat from step 2).

Divisive Hierarchical clustering - It is just the reverse of Agglomerative Hierarchical approach.

The probabilistic scheme enables automatic detection of the final hierarchy level.

III. INTRODUCTION TO WEKA

WEKA may prove useful to others involved in the development of open-source machine learning software. It contains implementations of algorithms for classification, clustering, and association rule mining, along with graphical user interfaces and visualization utilities for data exploration and algorithm evaluation, which are described as the main features. WEKA uses the java language and therefore satisfies the promise of platform independence. Weka's standard file format is ARFF, any file which is used to be used in Weka should end with .arff format.

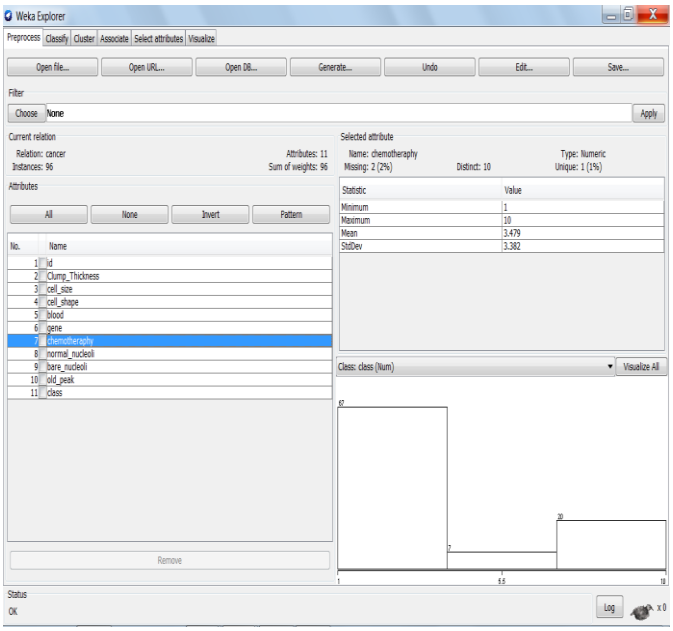
A. Sample cancer Data set

	id no	clump thic	cell size	cell shape	Blood cont	gene	chemother	Normal Nu	Bare Nuck	Old peak	Class
2	1000025	5	1	1	1	2	1	3	1	1	2
3	1002945	5	4	4	5	7	10	3	2	1	2
4	1015425	3	1	1	1	2	2	3	1	1	2
5	1016277	6	8	8	1	3	4	3	7	1	2
6	1017023	4	1	1	3	2	1	3	1	1	2
7	1017122	8	10	10	8	7	10	9	7	1	4
8	1018099	1	1	1	1	2	10	3	1	1	2
9	1050718	6	1	1	1	2	1	3	1	1	2
10	1054590	7	3	2	10	5	10	5	4	4	4
11	1054593	10	5	5	3	6	7	7	10	1	4
12	1056784	3	1	1	1	2	1	2	1	1	2
13	1057013	8	4	5	1	2	?	7	3	1	4
14	1059552	1	1	1	1	2	1	3	1	1	2
15	1065726	5	2	3	4	2	7	3	6	1	4
16	1066373	3	2	1	1	1	1	2	1	1	2
17	1066979	5	1	1	1	2	1	2	1	1	2
18	1067444	2	1	1	1	2	1	2	1	1	2
19	1070935	3	1	1	1	1	1	2	1	1	2
20	1071760	2	1	1	1	2	1	3	1	1	2
21	1072179	10	7	7	3	8	5	7	4	3	4
22	1074610	2	1	1	2	2	1	3	1	1	2

Cancer data set with missing values

The sample Cancer data set contains some of the missing values and these can be replaced by choosing the unsupervised filter in WEKA and select Replace Missing Values and click apply. After this all the missing values are replaced with an appropriate value either by taking mean or any nearer value. This will reduce the effort and the risk of replacing the values manually.

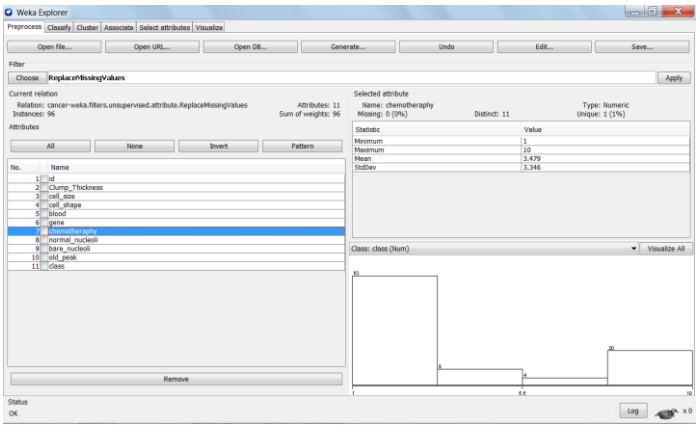
B. Output before replacing the missing values



Here in this data set there is 2% of missing value in the data chemotherapy and these missing values can be replaced by using filters. The output can be visualized at the bottom corner. As well as all of the data outputs can also be visualized by selecting visualize all.

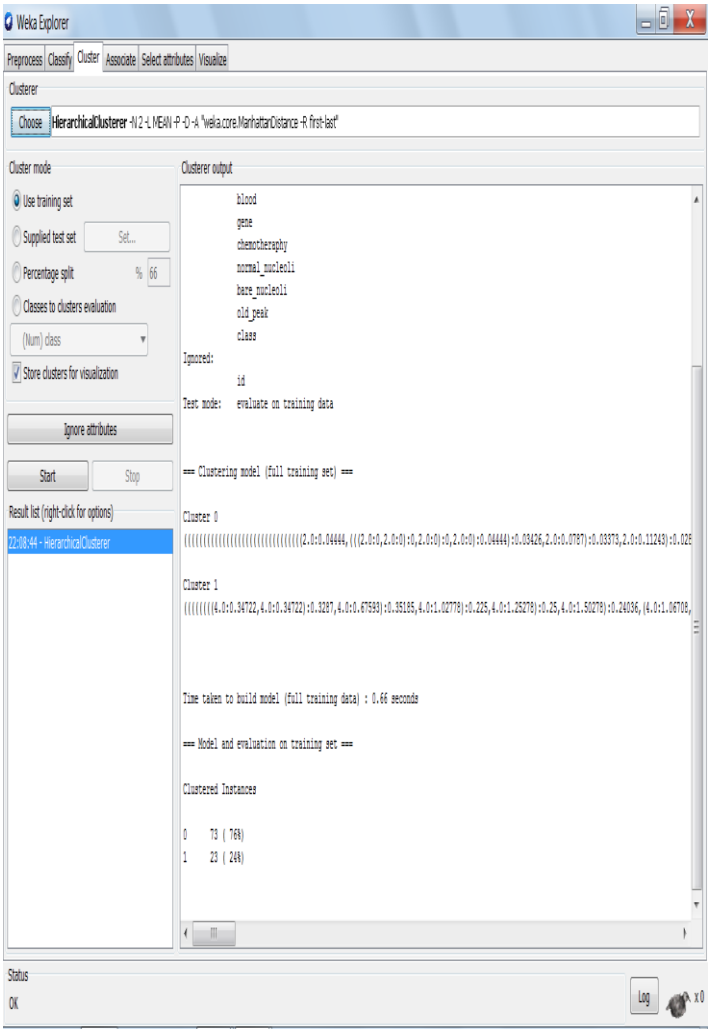
C. Output after replacing the missing values

Here the missing values are replaced by selecting the unsupervised filter. And since the Weka tool is an automated machine learning the missing values will be replaced automatically by selecting the filter.



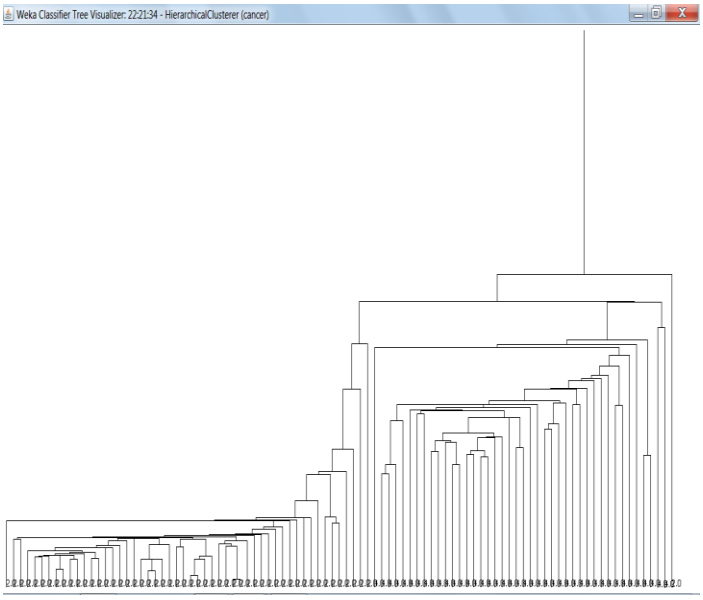
IV. CLUSTER OUTPUT

After choosing the Hierarchical clustering method in the weka tool select one of the class and number of clusters. If any attributes has to be ignored it can also be selected and ignored. Then click on start button below, the output window will show the mean, standard deviation, instances, attributes, and percentage split of clusters [6] [7].



V. HIERARCHICAL TREE VIEW

The cluster is formed by grouping the data and the dendogram grath (i.e., the hierarchical tree view for the cancer data set is shown). This also shows the number of clusters has been partitioned and how they are grouped.



VI. RANGE CHECK AND DELTA CHECK

Range check and Delta check are the techniques used to validate the results generated by the Hierarchical clustering approach.

The automated system normally uses a range check technique as a quality control measure. In **range check**, the results of the various tests are compared with known normal range of values for the tests and the normal results are accepted and stored. The results that are not within the normal range of values are identified as panic values and a smear of the blood is prepared from those samples and it is sent for an equivalent manual procedure to ensure quality of the system

Delta check is performed on the current and the immediate previous test result for better results and in order reduce the number of manual reviews.

VII. CONCLUSION

In this paper we have used one of the multiple clustering algorithms named Hierarchical clustering algorithm which replaces the existing use of the single clustering algorithm where certain limitations such as lack of robustness, stability, and accuracy follows. To avoid these limitations multiple clustering algorithm is used in this paper. The implementation of the algorithm was carried out by WEKA tool where the unsupervised data can be managed by using filters and the hierarchical tree view of the data set can be evaluated easily.

And further the techniques such as Range check and Delta check are used to validate the results generated by the Hierarchical clustering approach, in order to maintain accuracy of the result.

REFERENCES

[1] ZhiwenYu, Hau-SanWongb, JaneYou, QinminYang, and Hongying Liao "Knowledge Based Cluster Ensemble for Cancer Discovery from Biomolecular Data" IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL.10, NO.2, JUNE2011.

[2] Jung-HsienChiang*, Senior Member, IEEE, and Shing-HuaHo "A Combination of Rough-Based Feature Selection and RBF Neural Network for Classification Using Gene Expression Data" IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL.7, NO.1, MARCH 2008.

[3] T.R.Golub, D.K.Slonim, P.Tamayo, C.Huard, M.Gaasenbeek, J.P.Mesirov, H.Coller, M.Loh, J.Downing, M.Caligiuri, C.Bloomfield, and E.Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression," Science, vol.286, no.5439, pp. 531–537, 1999.

[4] D.A.Wigle, I.Jurisica, N.Radulovich, M.Pintilie, J.Rossant, N.Liu, C.Lu, J.Woodgett, I.Seiden, M.Johnston, S.Keshavjee, G.Darling, T.Winton, B.J.Breitkreutz, P.Jorgenson, M.Tyers, F.A.Shepherd, and M.S.Tsao, "Molecular profiling of non-small cell lung cancer and correlation with disease-free survival," Cancer Res.,vol.62, pp. 3005–3008, 2002.

[5] Anna Szymkowiak, Jan Larsen, Lars KaiHansen Informatics and Mathematical Modeling Richard Petersens Plads, Build.321,Technical University of Denmark, DK-2800KongensLyngby, Denmark "Hierarchical Clustering for Datamining"

[6] Zdravko Markov Central Connecticut State University, Ingrid Russell University of Hartford "Introduction to WEKA".

[7] RemcoR.Bouckaert, EibeFrank, MarkA.Hall, Geoffrey Holmes, BernhardPfahringer, PeterReutemann, IanH.Witten, Department of Computer Science, University of Waikatoi Hamilton, New Zealand, "WEKA—Experiences with a Java Open-Source Project".

AUTHORS

First Author – Breetha S, Department of Information Technology, Sri Krishna College of Technology, breetha92@gmail.com

Second Author – Kavinila R, Department of Information Technology, Sri Krishna College of Technology, kavinilabtechit@gmail.com