Abnormality Detection from Blood Report using Hierarchical Clustering Algorithm

T. Rajasekaran[1], V.Sowmeya[2], S.Suha[3], S.V.Vinodhini[4]

[1]*Assistant Professor, Department of Computer Science & Engineering, KPR Institute of Engineering and Technology, Arasur, Coimbatore.*

[2,3,4] *UG Scholars, Department of Computer Science & Engineering, KPR Institute of Engineering and Technology, Arasur, Coimbatore*

***Abstract*: In modern medical applications data mining techniques are very popular and produce accurate results, diagnosing a blood test report is a complicated process that largely depends on the doctor's knowledge, experience, ability to evaluate the patient's current test results and analyze risk factors that might be causation of illness. Therefore, a need for system to assist physician in making accurate and fast decision has arisen. The main focus of the present paper is to analyze the performance of "Hierarchical clustering algorithm" for blood reports. The results are compared with the normal values given in the medical books and shown that the hierarchical clustering technique was sufficiently effective to diagnose medical dataset especially, blood test reports and suggested that these results may be used for developing automatic abnormality detection Expert Systems.**

***Keywords*: Orange tool, Hierarchical clustering, Map distance, Data selection.**

I. INTRODUCTION

Before modern science began to take shape most doctors depended on Independent knowledge, skill and slight luck when diagnosing people[4].Little knowledge was available on most common sicknesses, Doctor's knew the symptoms and possibly the name of the disease, but other than that, most physicians could do little to treat the infection, and often times, the cure appeared far more drastic than the condition itself[6]. Not until healthcare analytics come into play did the medical field begin to comprehend what caused such problems in local citizens.If health care analytics existed in the middle Ages, avoiding the black plague may have saved millions of lives[11].

Here the blood test report of the patient in structured format is taken as input for the orange tool. In orange tool the blood report is analyzed and it is compared with the actual content of the blood and checked for abnormality and final report is generated which detect the abnormality which makes the work simple for doctor. And it can able to analyze the type of the abnormality for the person.These reports are stored in a database and it can be retrieved when needed.

II. DATA ANALYTICS

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making [3]. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics,Computer programming and operations research to quantify performance. Analytics often favors data visualization to communicate insight.

III. HEALTHCARE

Data analytics is an essential resource for any Profession. The collection of data and information is capable of forecasting the future. From understanding what services customers deem necessary to the cost effectiveness of a recently implemented technology, data analytics is a vital part of any corporation, business or organization [11]. Analytics plays a more pivotal role for health care than it might in financial and business markets. Understanding data points to ever-changing trends, including new research findings, emergency situations and outbreaks of disease. Thus, effective use of analytics in the healthcare industry can improve current care but more importantly can facilitate preventive care [10].

IV. MODERN ANALYTICS

Before the advent of modern analytics, researches and analysts were forced to pour over thousands of pages of data, resulting in thousands of hours of labor to make a simple conclusion based on the combined data [10]. Often times, information was missed during this transition period, in which data was analyzed, but not to a great extent, simply because no methodology allowed large spectrum of data to be studied relationally [11]. Sample groups could provide insights as to what the larger, general public might like or enjoy, but outside of these sample groups, looking over the needs for hundreds of thousands of individuals proved costly and far too time consuming. With the growth of information, big data is growing every larger, necessitating a system to create understanding from multiple data sets [12]. A flexible platform is essential to accept multiple data sources and should have these important features:

- The ability to search information by relationships between entities.
- Index information from any source type including social media, feeds, databases and file shares
- Ability to customize the environment based on individual needs and data sets.
- Tunable search algorithms for significance, relevance, temporal decay and geo-spatial decay.
- Simple third party integration using JSON, XML, RSS and/or KML
- Visual interface that is easy to use
- Natural language processing
- Ability to combine structured and unstructured data

V. ABOUT ORANGE

Orange is a machine learning and data mining suite for data analysis through Python scripting and visual programming. Orange library is a hierarchically-organized toolbox of data mining components [7]. The low-level procedures at the bottom of the hierarchy, like data filtering, probability assessment and feature scoring, are assembled into higher-level algorithms, such as classification tree learning. [8]. This allows developers to easily add new functionality at any level and fuse it with the existing code.

Data management and preprocessing for data input and output, data filtering and sampling, imputation, feature manipulation and feature selection, classification with implementations of various supervised machine learning algorithms (trees, forests, instance-based and Bayesian approaches, rule induction), borrowing from some well-known external libraries such as LIBSVM , Regression including linear and lasso regression, partial least square regression, regression trees and forests, and multivariate regression splines, Association for association rules and frequent item sets mining, Ensembles implemented as wrappers for bagging, boosting, forest trees, and stacking, Clustering, which includes k-means and hierarchical clustering approaches, Evaluation with cross-validation and other sampling-based procedures, functions for scoring the quality of prediction methods, and procedures for reliability estimation, Projections with implementations of principal component analysis, multi-dimensional scaling and self-organizing maps.

VI. HIERARCHICAL CLUSTERING

We defined several different ways of measuring distance between the rows or between the columns of the data matrix, depending on the measurement scale of the observations [8]. As we remarked before, this process often generates tables of distances with even more numbers than the original data, but we will show now how this in fact simplifies our understanding of the data [9].Distances between objects can be visualized in many simple and evocative ways. In this we shall consider a graphical representation of a matrix of distances which is perhaps the easiest to understand – a dendrogram, or tree – where the objects are joined together in a hierarchical fashion from the closest, that is most similar, to the furthest apart, that is the most different. The method of hierarchical cluster analysis is best explained by describing the algorithm, or set of instructions which creates the dendrogram results [12].

There are two approaches in the hierarchical clustering algorithm.

In Hierarchical Clustering – Agglomerative, Data objects are represented in a bottom-up fashion with data objects are initially in its own cluster and then combines these tiny clusters into larger clusters, until all of the data objects are in a single cluster or until certain termination condition specified by the user is satisfied.

Where as in Hierarchical Clustering - Divisivedata objects are represented in a top down fashion with all objects are in one cluster initially and then the cluster is subdivided into smaller pieces, until waiting each data object forms a own cluster or certain termination condition specified by the user is satisfied. Here distance between objects in two clusters may be Single link, Average link and complete link based on the distance between clusters is small, average and large respectively. In this paper Hierarchical Clustering is considered because, Tree representation of the cluster is more informative compared to all the remaining clustering algorithms.

VII. ALGORITHM

Cluster analysis is a unsupervised learning technique used for classification of data.Data elements are partitioned into groups called clusters that represent proximate collection of data elements based on distance dissimilarity function.

D={t1,t2,.....tn} be the set of elements.

E be the Euclidean distance.

E=sim(ti,tsj) U sim(tk,tn)

Clustering: After merging t_i and t_j, the similarity of the resulting cluster to another cluster, t_k, is:

sim((ti U tj),tk) = max(sim(ti,tk),sim(tj,tk))

Cluster the E as single Linkage.

Predefine the normality value to the parameters using Select data component.

Then find the abnormal patients from the list.

VIII. METHODOLOGY

The dataset contains the blood report of the patient and it was in the excel sheet. The data set can be converted into the tab delimiter format. We are using orange data mining tool for the process. That tool only accept the data set in the format of the tab.

Then the file can be converted into the table format by using the data table component in the tool. After that Euclidean distance can be calculate for separating the male and female patient because the normal values are differ from each other. Then Euclidean distance can be calculated for the parameter gender. Then the similarity of gender can be calculated and it can be clustered as the single linkage cluster. Single linkage is the clustering of nearest similar data. Gender can have only two values that are male and female. So Single linkage clustering is used. Then the normal values for the male and female are included by using the select data component in the tool. For that different parameters for the different diseaseare included separately and get the result from the Data table. Here different data tables are used to indicate the different disease from the report.

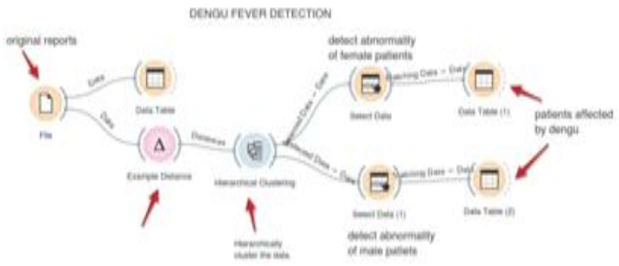
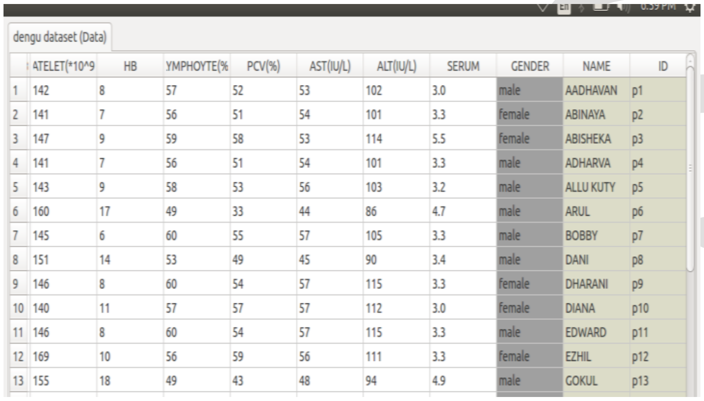


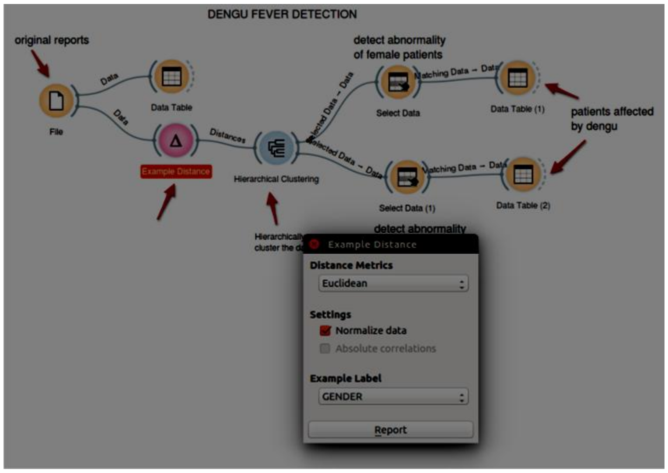
Figure 8.1 Process flow



dengu dataset (Data)

	ATELET(*10^9	HB	YMPHOYTE(%	PCV(%)	AST(IU/L)	ALT(IU/L)	SERUM	GENDER	NAME	ID
1	142	8	57	52	53	102	3.0	male	AADHAVAN	p1
2	141	7	56	51	54	101	3.3	female	ABINAYA	p2
3	147	9	59	58	53	114	5.5	female	ABISHEKA	p3
4	141	7	56	51	54	101	3.3	male	ADHARVA	p4
5	143	9	58	53	56	103	3.2	male	ALLU KUTY	p5
6	160	17	49	33	44	86	4.7	male	ARUL	p6
7	145	6	60	55	57	105	3.3	male	BOBBY	p7
8	151	14	53	49	45	90	3.4	male	DANI	p8
9	146	8	60	54	57	115	3.3	female	DHARANI	p9
10	140	11	57	57	57	112	3.0	female	DIANA	p10
11	146	8	60	54	57	115	3.3	male	EDWARD	p11
12	169	10	56	59	56	111	3.3	female	EZHIL	p12
13	155	18	49	43	48	94	4.9	male	GOKUL	p13

Figure 8.2 Input Dataset



Example Distance

Distance Metrics
Euclidean

Settings
☑ Normalize data
☐ Absolute correlations

Example Label
GENDER

Report

Figure 8.3 Euclidean Distance for Gender

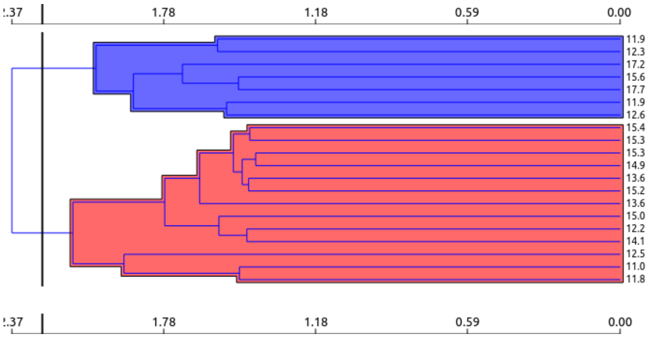
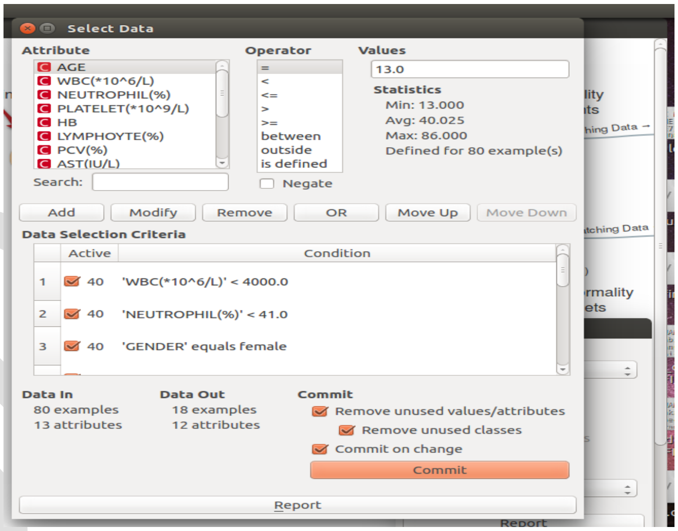


Fig 8.4 Clustering of Data by Gender



Select Data

Attribute: AGE, WBC(*10^6/L), NEUTROPHIL(%), PLATELET(*10^9/L), HB, LYMPHOYTE(%), PCV(%), AST(IU/L)

Search:

Operator: =, <, <=, >, >=, between, outside, is defined

☐ Negate

Values: 13.0

Statistics
Min: 13.000
Avg: 40.025
Max: 86.000
Defined for 80 example(s)

Add | Modify | Remove | OR | Move Up | Move Down

Data Selection Criteria

	Active	Condition
1	☑ 40	'WBC(*10^6/L)' < 4000.0
2	☑ 40	'NEUTROPHIL(%)' < 41.0
3	☑ 40	'GENDER' equals female

Data In
80 examples
13 attributes

Data Out
18 examples
12 attributes

Commit
☑ Remove unused values/attributes
☑ Remove unused classes
☑ Commit on change

Commit

Report

Fig 8.5 Define a normal value

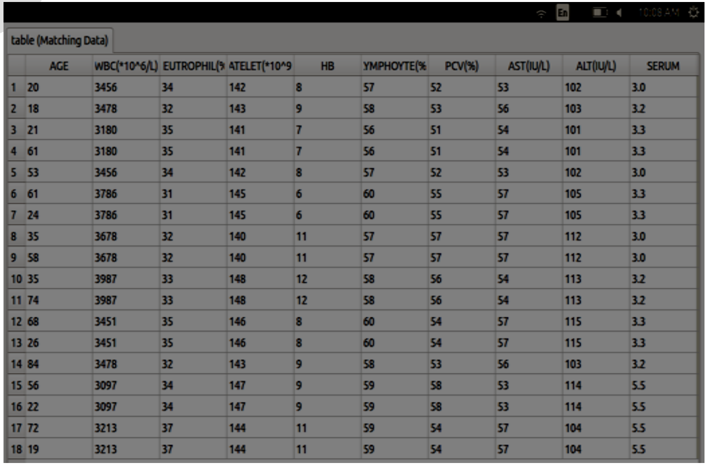


table (Matching Data)

	AGE	WBC(*10^6/L)	EUTROPHIL(%	ATELET(*10^9	HB	YMPHOYTE(%	PCV(%)	AST(IU/L)	ALT(IU/L)	SERUM
1	20	3456	34	142	8	57	52	53	102	3.0
2	18	3478	32	143	9	58	53	56	103	3.2
3	21	3180	35	141	7	56	51	54	101	3.3
4	61	3180	35	141	7	56	51	54	101	3.3
5	53	3456	34	142	8	57	52	53	102	3.0
6	61	3786	31	145	6	60	55	57	105	3.3
7	24	3786	31	145	6	60	55	57	105	3.3
8	35	3678	32	140	11	57	57	57	112	3.0
9	58	3678	32	140	11	57	57	57	112	3.0
10	35	3987	33	148	12	58	56	54	113	3.2
11	74	3987	33	148	12	58	56	54	113	3.2
12	68	3451	35	146	8	60	54	57	115	3.3
13	26	3451	35	146	8	60	54	57	115	3.3
14	84	3478	32	143	9	58	53	56	103	3.2
15	56	3097	34	147	9	59	58	53	114	5.5
16	22	3097	34	147	9	59	58	53	114	5.5
17	72	3213	37	144	11	59	54	57	104	5.5
18	19	3213	37	144	11	59	54	57	104	5.5

Fig 8.6 Abnormal patient List

IX. BENEFITS

This system makes the prediction of blood easily and reduce the time for the analyzing the report. Here there is no knowledge of doctor. Educated people can easy to understand their report easily. It can able to analyze many of blood test reports at the same time. It can able to analyze the disease and able to predict the possibility of the disease.

X. RESULT

Thus the abnormality of the patient was detected by separating the male and female dataset by using the hierarchical clustering algorithm and the possibility of the particular disease was predicted.

XI. FUTURE WORK AND CONCLUSION

The enhancement of the project is to predict the possibility of occurring the disease in the particular season. It was used to forecast the spreading of the disease by analyzing the previous year disease occurrence. It will be used to improve the medical facilities and to analyze the supply of the medicines and vaccinations.The performance evaluation is conducted with respect to the performance parameter: Accuracy and found that the Proposed Hierarchical Clustering Algorithm applied on data set exhibits more accurate using classification. These results are used in developing the blood report automation Expert system for decision making in diagnosing the both patients and doctors. The details of the proposed expert system are included in this paper.

REFERENCES

[1]. "Survey of Clustering Algorithms"Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.

[2]. .Data Analysis, Wikipedia, https://en.wikipedia.org/wiki/Data_analysis

[3]. Analytics, Wikipedia,https://en.wikipedia.org/wiki/Analytics.

[4]. Data analytics for healthcare, www.iknow.com

[5]. Introduction to predictive analysis tool, www.uky.edu.

[6]. Practical Predictive Analytics for Healthcare", Steven S. Eisenberg, MD.

[7]. Orange open source tool, Wikipedia.

[8]. "An Implementation of Hierarchical Clustering on Indian Liver Patient Dataset" Prof. M.S. Prasad Babu,2014.

[9]. "Hierarchical clustering",David M. Blei

[10]. "A survey on Data Mining approaches for Healthcare", Divya Tomar and Sonali Agarwal Indian Institute of Information Technology, Allahabad, India.

[11]. ."The value of analytics in healthcare", By James W. Cortada, Dan Gordon and Bill Lenihan.

[12]. ."Privacy and security for analytics on healthcare data",Albana Gaba, Yeb Havinga

[13]. "Health care Analytics and managing population health", Victoria Tiase, MS, RN, Director Informatics Strategy, New York-Presbyterian Hospital.

[14]. "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu ,International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011.

[15]. "A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis", Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu,IJCSI International Journal of Computer Science Issues,vol. 9,Issue 3, No 2,May 2012 .

[16]. "Healthcare2015 and care delivery: Delivery models refined, competencies defined."Adams, Jim, Richard Bakalar, MD, Michael Boroch, Karen Knecht, Edgar L. Mounib and Neil Stuart. IBM Institute for Business Value. June 2008.

[17]. "Data Mining Application in Healthcare", H. C. Koh and G. Tan ,Journal of Healthcare Information Management, 2005.

[18]. D. Hand, H. Mannila and P. Smyth, "Principles of data mining", MIT, 2001.

[19]. J. Han and M. Kamber, "Data mining: concepts and techniques". The Morgan Kaufmann Series, 2006.

[20]. L. Duan, W. N. Street & E. Xu, Healthcare information systems: data mining methods in the creation of a clinicalrecommender system, Enterprise Information Systems, 2011.