PREDICTIVE ANALYSIS OF BLOOD TEST REPORT

Mr.T.Rajasekaran (Assistant Professor)1, V.Sowmeya2, S.Suha3, S.V.Vinodhini4

1,2,3,4 (UG Scholars) DEPARTMENT OF COMPUTER SCIENCE &ENGINEERING
KPR INSTITUTE OF ENGINEERING AND TECHNOLOGY
ARASUR, COIMBATORE.

rajasekaran30@gmail.com ,sowmeya95be@gmail.com, suhasanmugam@gmail.com , vinovimal@gmail.com

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 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.

I. INTRODUCTION

Before modern science began to take shape in the 19th century, most doctors depended on Independent knowledge, skill and slight luck when diagnosing people. Little knowledge was Available on most common sicknesses, let alonewhat caused it? Doctor's knew the symptoms andpossibly the name of the disease, but other thanthat, most physicians could do little to treat theinfection, and often times, the cure appearedfar more drastic than the condition itself. Notuntil healthcare analytics come into play did themedical field begin to comprehend what causedsuch problems in local citizens.

If healthcare analytics existed in the Middle Ages,avoiding the black plague may have saved millionsof lives. With proper data analysis, individualsin the medical field could have pinpointed onesimilarity in all cases: the water. In Germany forexample, many of the citizens came from differentareas of the country, most of which ate differentfood. However, all drank the same water. Yet,those people seemingly immune to the blackplague were the monks, isolated from the rest ofthe community in abbeys, drinking not water butbeer. Modern healthcare analytics quickly wouldhave pointed to water as the death-causing culprit,While the boiling process eliminated bacteria in thewater, keeping monks healthy.Sadly, many in the medical community remain inthe dark ages by not properly using healthcareanalytics to determine the needs of patients andtheir community. This is a limitation in the current

Medical facility, resulting in poor treatment optionsfor local patients. Data analysis obtained directlyfrom a community is capable of identifying keyneeds local citizens have, what their most commonailments are, services they require and just aboutany other variety of information needed. It isessential for a medical provider to tailor its servicestowards the community and effective use ofanalytics turns big data into actionable informationfor local care providers.

Here the blood test report of the patient in structured format is given as input to 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 then the final report is generated the report includes the detected abnormality which makes the work simple for doctor. These reports are stored in a database and it can be retrieved when needed. Updating the record can also be done here.

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. 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. DATA ANALYTICS IN HEALTHCARE

Data analytics is an essential resource for any Profession. This collection of data and informationis capable of forecasting the future. Fromunderstanding what services customers deemnecessary to the cost effectiveness of a recentlyimplemented technology, data analytics is a vitalpart of any corporation, business or organization. Analytics plays a more pivotal role for healthcarethan it might in financial and business markets. Understanding data points to ever-changingtrends, including new research findings, emergency situations and outbreaks of disease. Thus, effective use of analytics in the healthcareindustry can improve current care but moreimportantly can facilitate preventive care.

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. 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 spectrums Of data to be studied relationally. 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. 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. 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. This allows developers to easily add new functionality at any level and fuse it with the existing code. The main branches of the component hierarchy are:

Data management and preprocessingfor data input and output, data filtering and sampling, imputation, feature manipulation (discretization, continuation, normalization, scaling and scoring), 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 (Chang and Lin, 2011), **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.

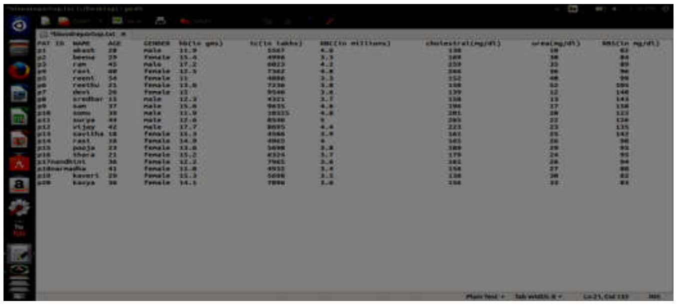
VI. HIERARCHIAL CLUSTERING

We defined several different ways of measuring distance (ordissimilarity as the case may be) between the rowsor between the columns of the datamatrix, depending on the measurement scale of the observations. As we remarked before,this process often generates tables of distances with even more numbers than the originaldata, but we will show now how this in fact simplifies our understanding of the data.

Distances between objects can be visualized in manysimple and evocative ways. In this we shall consider a graphical representation of a matrix of distances which isperhaps the easiest to understand – a dendrogram, or tree – where the objects are joinedtogether 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 bydescribing the algorithm, or set of instructions, which creates the dendrogram results. Inthis chapter we demonstrate hierarchical clusteringon a small example and then list thedifferent variants of the method that are possible.

VII. METHODOLOGY

The dataset contains 20 patients' blood report with 9 attributes. Which important attributes that are used to find abnormality are hemoglobin(hg),Total cell count(tc), RedBloodCells (rbc), cholesterol(ch), and rbs.

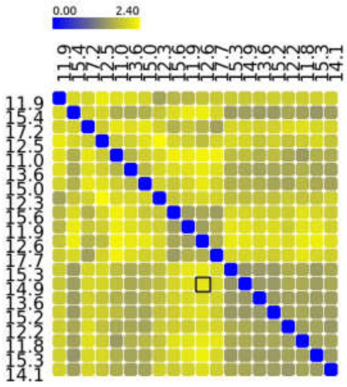


```
PAT ID   NAME      AGE   GENDER  hb(in gms)   tc(in lakhs)   RBC(in millions)   cholestral(mg/dl)   urea(mg/dl)   RBS(in mg/dl)
p1       akash     28    male    11.9         5567           4.0                138                 18            62
p2       beena     25    female  15.4         4996           3.3                169                 38            84
p3       ram       45    male    17.2         6023           4.2                259                 35            89
p4       ravi      40    female  12.5         7362           4.8                366                 36            90
p5       reena     54    female  11           4086           3.3                152                 40            99
p6       reethu    21    female  13.6         7230           3.8                158                 52            101
p7       devi      26    female  13           9546           3.6                139                 12            140
p8       sreedhar  11    male    12.3         4321           3.7                138                 13            143
p9       sam       37    male    15.6         9635           4.6                196                 27            158
p10      somu      38    male    13.9         10335          4.8                281                 28            123
p11      surya     44    male    17.6         8540           5                  265                 22            134
p12      vijay     42    male    17.7         8695           4.4                223                 23            135
p13      kavitha   18    female  13.3         4566           3.9                167                 25            142
p14      ravi      18    female  14.9         4065           4                  165                 26            98
p15      pooja     23    female  13.6         5698           3.8                189                 29            93
p16      thara     23    female  15.2         4324           3.7                179                 24            95
p17nandhini        38    female  12.2         7965           3.6                182                 28            94
p18narmadha        41    female  11.8         4933           3.4                154                 27            88
p19      kaveri    29    female  15.3         5698           3.5                138                 30            82
p20      kavya     36    female  14.1         7896           3.0                156                 33            83
```

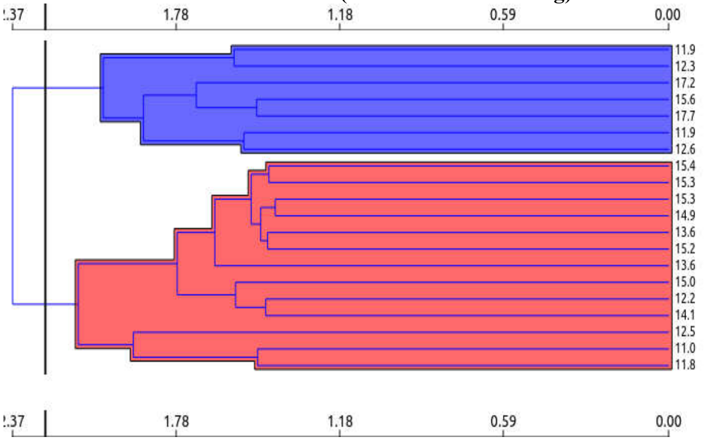
Fig 7.1 Input for the process

The given data set is converted into table structure and the map distance is founded for that data's by comparing each data for the single attribute. If the map distance was found then the data will be visualized and the distance matrix for the particular attribute data will be obtained. The data's will be clustered in the form of hierarchical.The similar set of data will be combined together and the dissimilar set of data will be combined together.Then again the clustered data will be converted in the form of table. In that it separate the abnormality with normality values.

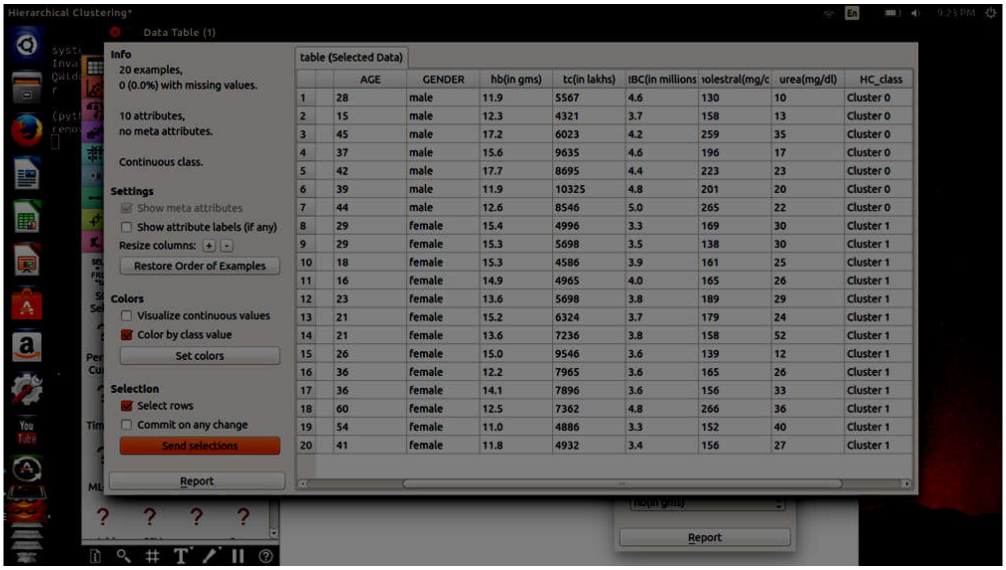
(distance map visualization)



(hierarchical clustering)



VIII. RESULT



Info
20 examples,
0 (0.0%) with missing values.

10 attributes,
no meta attributes.

Continuous class.

Settings
☑ Show meta attributes
☐ Show attribute labels (if any)
Resize columns: + -
Restore Order of Examples

Colors
☐ Visualize continuous values
☑ Color by class value
Set colors

Selection
☑ Select rows
☐ Commit on any change
Send selections

table (Selected Data)

		AGE	GENDER	hb(in gms)	tc(in lakhs)	:BC(in millions	iolestral(mg/c	urea(mg/dl)	HC_class
1		28	male	11.9	5567	4.6	130	10	Cluster 0
2		15	male	12.3	4321	3.7	158	13	Cluster 0
3		45	male	17.2	6023	4.2	259	35	Cluster 0
4		37	male	15.6	9635	4.6	196	17	Cluster 0
5		42	male	17.7	8695	4.4	223	23	Cluster 0
6		39	male	11.9	10325	4.8	201	20	Cluster 0
7		44	male	12.6	8546	5.0	265	22	Cluster 0
8		29	female	15.4	4996	3.3	169	30	Cluster 1
9		29	female	15.3	5698	3.5	138	30	Cluster 1
10		18	female	15.3	4586	3.9	161	25	Cluster 1
11		16	female	14.9	4965	4.0	165	26	Cluster 1
12		23	female	13.6	5698	3.8	189	29	Cluster 1
13		21	female	15.2	6324	3.7	179	24	Cluster 1
14		21	female	13.6	7236	3.8	158	52	Cluster 1
15		26	female	15.0	9546	3.6	139	12	Cluster 1
16		36	female	12.2	7965	3.6	165	26	Cluster 1
17		36	female	14.1	7896	3.6	156	33	Cluster 1
18		60	female	12.5	7362	4.8	266	36	Cluster 1
19		54	female	11.0	4886	3.3	152	40	Cluster 1
20		41	female	11.8	4932	3.4	156	27	Cluster 1

The blood report will be analyzedAnd using hierarchical clustering algorithm the data will be clustered and it define a similar and dissimilar value separated and the resulting table had same value.There is an additional information called cluster.If the row had a cluster 0 value there is an abnormality in the blood.If the row had cluster 1 value there is no abnormality in the particular person's bloo

XI. BENEFITS

1. Finding the abnormality of patients from many test records will be easy.
2. There is no need of doctor's knowledge to understand the report.
3. Work load for doctor's must be reduced instead of analyzing the report.
4. Reports of patient will be maintained and updated for various testing purposes.

X. CONCLUSION

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 by 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
[3].Analytics, Wikipedia
[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.
[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, NewYork-Presbyterian Hospital*