CLUSTERING THE PREPROCESSED BONE MARROW DATA USING MODIFIED K-MEANS ALGORITHM

D.Minnie*

Department of Computer Science, Madras Christian College, Tambaram,
Chennai, Tamilnadu, India†
minniearul@yahoo.com

S.Srinivasan

Department of Computer Science and Engineering, Anna University Regional Office Madurai,
Madurai, Tamilnadu, India
sriniss@yahoo.com

Abstract

In this paper we analyse the components of the bone marrow and the bone marrow analysis data. Eighteen thousand bone marrow records are collected from a reputed Clinical Pathological Laboratory and this raw data is transformed into a pre-processed and flattened data using the pre-processing phases of the Knowledge Discovery in Databases. The transformed bone marrow data is used to create clusters of the database using various fields of the bone marrow. The K-Means clustering algorithm is slightly modified and is applied on the Bone Marrow Database to form various clusters.

***Keywords*:** Hematology, Bone Marrow, Knowledge Discovery in Databases, Data Mining, Clustering, K-Means Clustering.

1. Introduction

A huge volume of automated medical data are currently available in various forms such as text, numbers, combination of text and numbers, images, scan reports, video and audio reports. This data are used along with various analysis techniques to generate results that can be used by the health care professionals in efficient decision making.

Hematology is the study of blood, diseases related to blood and blood forming organs such as bone marrow. Clinical Pathology is a study that is concerned with conducting laboratory experiments on body fluids such as blood and urine to diagnose diseases. Hematology department of Clinical Pathology performs various tests on blood. Some of the common tests on blood are the Complete Blood Count (CBC) to diagnose diseases such as anemia and some types of blood cancers, Erythrocyte Sedimentation Rate (ESR) to diagnose inflammation and Prothrombin Time (PT) to diagnose coagulation disorders.

Bone Marrow is the flexible tissue found in the interior of bones. It produces the cellular elements of the blood such as red blood cells, white blood cells and platelets. Bone Marrow analysis refers to the pathologic analysis of bone marrow samples obtained by bone marrow aspiration that yields semi-liquid bone marrow or a bone marrow biopsy that yields a cylindrical shaped solid bone marrow. The bone marrow sample is used to diagnose diseases such as leukemia, anemia, neimen pick disease and so on.

The methodology used for studying the bone marrow data is Knowledge Discovery in Databases (KDD). KDD is used to extract useful knowledge from the raw bone marrow data. The raw data is first transformed into a form that is appropriate for the Data Mining process using the KDD pre-processing steps. The data mining technique clustering is then applied on the preprocessed data to generate knowledge.

* Department of Computer Science, Madras Christian College, Tambaram, Chennai, Tamilnadu, India.

2. Methods

2.1. *Bone Marrow Data*

The bone marrow contains various cells such as erythrocytes (red blood cells), blasts, promyelocyte, myeloblasts, plasma cells, white blood cells such as neutrophils, eosinophils, basophils and monocytes and the components are shown in figure 1. The bone marrow analysis gives the details and quantities of those cells in the bone marrow, the patient id, date in which the test is taken, hospital id, detailed description of the results and the final impression as identified by a clinical pathologist.

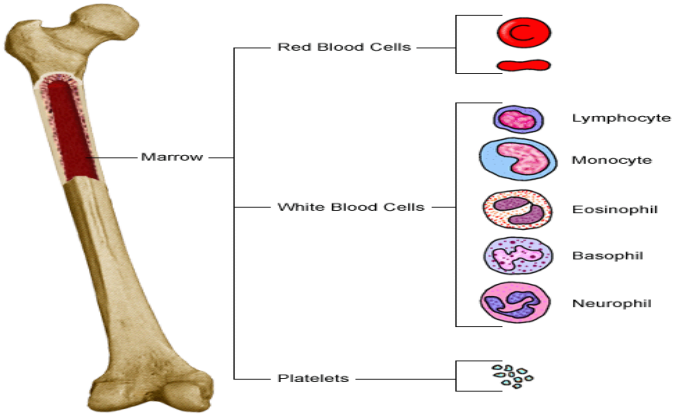


Figure 1. Bone Marrow Components

2.2. *Data collection*

Eighteen thousand bone marrow data are collected from a Clinical Pathology department of a reputed hospital. The data is present as an excel file. A sample of the data file is shown in figure 2.

	bmid	date	hsno	erythblast	promy	myel	neutr	eosin	lymph	lasm	adn	baso	monc	txt	dit_nam	in_name	diffcell	txt1	imp
2	710113014	06/08/2004	631694C	24		16.5	26.5	17	2	13.5	0.5				Fragments: Normocellular	DR.E	DR.A1	200 Cells	Fragments: Normocellular	Impression: Cellular
3	711013004	06/10/2004	630099C	23		21.5	32	9.5	5.5	7	1.5				Fragments: Normocellular	DR.E	DR.B1	200 Cells	Fragments: Normocellular	Impression: Cellular
4	711013006	06/10/2004	703928A	26		21	18	11	3	19	2				Fragments: Cellular	DR.F	DR.B1	200 Cells	Fragments: Cellular	Impression: Cellular
5	711013009	06/10/2004	637661C	3.5	90.5	0.5	1	1	0.5	3					Fragments: Scanty.	DR.E	DR.B	200 Cells	Fragments: Scanty.	Impression: Acute
6	711013010	06/10/2004	615633C	47		12	14	19		8					Fragments: Hypercellular.	DR.F	DR.B	200 Cells	Fragments: Hypercellular.	Impression: Hypercellular
7	711013011	06/10/2004	636655C	50		10	9	20	11						Fragments: Hypercellular.	DR.F	DR.B	200 Cells	Fragments: Hypercellular.	Impression: Iron
8	711013014	06/10/2004	637516C	12	71			2		12	3				Fragments: Hypercellular.	DR.F	DR.B	200 Cells	Fragments: Hypercellular.	Impression: Acute
9	710023004	06/12/2004	637542C	48		22	11	10		9					Fragments: Cellular	DR.F	DR.B1	200 Cells	Fragments: Cellular	Impression: Cellular
10	710023006	06/12/2004	627947C	28		26	12	26.5		7.5					Fragments: Normocellular	DR.E	DR.B1	200 Cells	Fragments: Normocellular	Impression:
11	710023007	06/12/2004	630226C	20.5		20.5	32	15	2.5	6	3.5				Fragments: Normocellular	DR.E	DR.B1	200 Cells	Fragments: Normocellular	Impression: Cellular
12	710023009	06/12/2004	637556C	44		22	10	14		10					Fragments: Hypercellular.	DR.F	DR.B	200 Cells	Fragments: Hypercellular.	Impression: Hyprecellular
13	710023010	06/12/2004	637099C	49		9	5	1		33	3				Fragments: Markedly	DR.F	DR.B	200 Cells	Fragments: Markedly	Impression: Aplastic
14	710023015	06/12/2004	634619C	40		6	16	24		14					Fragments: Cellular	DR.F	DR.B	200 Cells	Fragments: Cellular	Impression: Reactive
15	707133002	13/06/2004	637656C	41		22.5	11	10.5	3.5	10	1.5				Fragments: Solidly cellular	DR.E	DR.C1	200 Cells	Fragments: Solidly cellular	Impression: Severe Iron
16	707133011	13/06/2004	626637C	34			19	21	1	10					Fragments: Normocellular	DR.E	DR.B1	200 Cells	Fragments: Normocellular	Impression: Reactive
17	707143010	14/06/2004	633930C	22			15	18	1.5	7.5	0.5				Fragments: Normocellular	DR.E	DR.B1	200 Cells	Fragments: Normocellular	Impression: Cellular
18	707153000	15/06/2004	631965C												Trial bm report.	DR.E				
19	707153007	15/06/2004	633402C	18		4	40	13		4	2				Fragments: Cellular	DR.F	DR.B1	200 Cells	Fragments: Cellular	Impression: Cellular
20	707153008	15/06/2004	636618C	25.5		13.5	23	17	4.5	16.5					Fragments: Markedly	DR.A	DR.B1	200 Cells	Fragments: Markedly	Impression: Cellular
21	707153009	15/06/2004	546530B	20.5		22	27.5	18.5	2	9	0.5				Fragments: Moderately	DR.A	DR.B1	200 Cells	Fragments: Moderately	Impression: Cellular
22	707163001	16/06/2004	636203C												Fragments: Cellular	DR.B				
23	707163002	16/06/2004	632921C	41.5		6	15	24.5	5.5	1.5	.5				Fragments: Cellular	DR.C			Fragments: Cellular	Impression: Cellular
24	707163006	16/06/2004	806380C	28.5	6.5	7	15	17	9	3.5	6.5				Fragments: Cellular	DR.C		200 Cells	Fragments: Cellular	Impression: Cellular
25	707173003	17/06/2004	965046B	20.5		10.5	16	31	2.5	9.5	10				Fragments: Normocellular	DR.D	DR.B	200 Cells	Fragments: Normocellular	Impression: Cellular
26	707173011	17/06/2004	642024C	22.5			14	15.5	9	17.5	2.5				Fragments: Scanty.	DR.E	DR.B1	200 Cells	Fragments: Scanty.	Impression: Marrow
27	707193008	19/06/2004	640665C	35		7.5	15	20	3.5	6	3				Fragments: Hypocellular	DR.C		200 Cells	Fragments: Hypocellular	Impression: Cellular
28	707193009	19/06/2004	639178C	28		8	18.5	25	.5	6.5	3.5				Erythroid Maturation:	DR.C		200 Cells	Erythroid Maturation:	Impression: Cellular
29	707203002	20/06/2004	641746C	43	1	10	17	10	7	5					Fragments: Cellular	DR.C		200 Cells	Fragments: Cellular	Impression: Cellular
30	707203008	20/06/2004	636642C	17	0.5	5	15.5	12.5	6	20	3.5				Fragments: Normocellular	DR.E	DR.C1		Fragments: Normocellular	Impression:

Figure 2 Sample Bone Marrow Analysis Data

2.3. *Knowledge Discovery in Databases*

The data is subjected to the KDD processes to generate knowledge from it. The processes include Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Generation of Patterns and Knowledge Interpretation.

In Data Cleaning the irrelevant data are removed from the collected data. In Data Integration multiple sources are combined into a data warehouse. The Data Selection process is involved with the selection of data relevant to the analysis and extracting them from the integrated data. The selected data is transformed to the appropriate form for the mining procedure.

The process of extracting useful and implicit information from the transformed data is referred to as Data Mining. In Pattern Evaluation interesting patterns are identified from the processed data. The discovered knowledge is visually presented to the user in the Knowledge Representation process.

2.4. *Data Mining*

Data Mining is the Knowledge Discovery stage of KDD and it is the process of extracting implicit, useful, previously unknown, non-trivial information from data. The methods or techniques involved in Data Mining are grouped as Classification, Clustering, Association Rules and Sequences that represent the knowledge generated from the data.

Classification is a supervised learning process and it maps data into known classes using Decision Trees, Neural Networks and Genetic Algorithms. Clustering is an unsupervised learning and it groups similar data into unknown clusters using K-Means, Nearest Neighbour and various other algorithms. Association Rule Mining (ARM) uncovers relationships among data in a database.

2.5. *Data Mining*

Data Mining is the Knowledge Discovery stage of KDD and it is the process of extracting implicit, useful, previously unknown, non-trivial information from data. The methods or techniques involved in Data Mining are grouped as Classification, Clustering, Association Rules and Sequences that represent the knowledge generated from the data.

2.6. *Clustering*

Clustering is the task of assigning a set of objects into groups so that objects in the same group are more similar to each other than the objects in other groups. Clustering is an unsupervised algorithm and it does not use class labels. The class labels are needed for the Classification algorithms and are not required for the clustering algorithms.

Some of the major clustering models are Hierarchical Clustering, Centroid Based Clustering and Density Based Clustering. The Hierarchical clustering creates a hierarchical decomposition of the data objects and it doesn't require the number of clusters k to be given as the input whereas the other clustering techniques require the k value. But it requires a terminating condition to stop the clustering process. The technique is further divided into agglomerative or divisive algorithms. Centroid Based Clustering uses partitioning methods. It constructs k partitions of the given k elements where $k \leq n$ and each portion are referred as a cluster. The cluster should have at least one element and one element should be present in only one cluster. Density Based Clustering is used to form clusters that have arbitrary shape instead of the traditional spherical shaped clusters formed using the partitioning methods. It is used to identify noise and outliers and to exclude them in grouping the clusters.

2.7. *K-Means Clustering*

The K-Means Clustering is a Centroid based clustering model in which the database is partitioned into K clusters in which each record belongs to the cluster with the nearest mean value. The algorithm starts with given initial set of mean values and allocates each object to a cluster with nearest mean value. The mean values for each cluster are calculated then using the elements in each cluster.

Let D be a data set containing n elements or objects to be clustered and k the number of clusters to be formed where $k \leq n$. The clusters formed are represented as C_i where i = 1 to k. The mean values of each cluster C_i is represented as m_i. When the clustering process is started initial mean values m_1, m_2 …… m_k are identified from

the given set of elements or objects. In the next step, the elements or objects are selected one by one and the distance between the element and each of the cluster means m_i is found. The mean squared error E for each element p from the various clusters Ci is calculated for finding the strength of the clustering technique as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} | p - m_i |^2 \quad (1)$$

The elements are added to the clusters for which the mean squared error is lesser. The mean values of the clusters are recalculated and the process is repeated until there is no change in the cluster means..

2.8. *Bone Marrow Analysis Data Format*

The Bone Marrow Data is given as an excel file. A sample of it is given in table 2. The Bone Marrow data consists of values for each sample of bone marrow for the various attributes such as erythrocytes, blasts, promyelocyte, myeloblasts, neutrophils, eosinophils, plasma cells, basophils and monocytes, the patient id, date in which the test is taken, hospital id, detailed description of the results and the final impression.
The list of attributes along with a detailed description is shown in table 1.

Table 1. Bone Marrow Data Attributes

Attribute Name	Attribute Description		Attribute Name	Attribute Description
bmid	Patient Id		lymph	Lymphocyte count
date	Test Date		plasma	Plasma cell count
hsno	Hospital Number		baso	Basophil count
eryth	Erythrocyte count		mono	Monocyte
blasts	Blast cell count		txt	Detailed description of the test results
promy	Promyelocyte count		edit_name	Doctor's name
myel	Myelocyte count		fin_name	Doctor's name
neutr	Neutrophil count		imp	Impression
eosin	Eosinophil count			

The algorithm starts with given initial set of mean values and allocates each object to a cluster with nearest mean value. The mean values for each cluster are calculated then using the elements in each cluster.

3. Previous Research

Rakesh Agrawal, T. Imielinski and A. Swami (1993) present some of the ways of using sequences of clustering algorithms to mine temporal data.

Michael Goebel and Le Gruenwald (1999) present a survey of various Data Mining Tools wherein each of the tools is designed to handle a specific type of data and to perform a specific type of task.

Georg Berks, Diedrich Graf v. Keyserlingk, Jan Jantzen, Mariagrazia Dotoli and Hubertus Axer (2000) used fuzzy c-means clustering algorithms to assign symptoms to various types of aphasia disease categories.
K.Y.Yeung, D.R.Haynor and W.L.Ruzzo (2001) analyses the performances of K-Means clustering algorithm with other clustering algorithms on the gene expression data.

Cios KJ and Moore GW (2002) discusses on the uniqueness of medical data mining with respect to the heterogeneity of the data, ethical, legal and social issues, privacy and security issues related to handling medical data.

Alp Aslandogan Y. and Gauri A.Mahajani (2004) discuss how various combinations of Data Mining classification algorithms are used on medical data for efficient classification of the data.

Derek Greene, Alexey Tsymbal, Nadia Bolshakova, P´adraig Cunningham (2004) presented several ensemble generation and integration strategies, and evaluated each approach on a number of synthetic and real-world datasets from the UCI repository.

Patricia Cerrito and John C. Cerrito (2006) discusses on how the electronic medical records are created from manual records and how data and text mining are used to improve the quality and reduce costs.

Massoud Toussi, Jean-Baptiste Lamy, Philippe Le Toumelin, and Alain Venot (2009), use data mining techniques on type 2 diabetes medical data to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations.

Anil K.Jain (2010) presents a detailed study on well-known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and point out some of the emerging and useful research directions, including semi-supervised clustering, ensemble clustering, simultaneous feature selection during data clustering, and large scale data clustering.

Minnie D and Srinivasan S (2011) discuss the steps in Knowledge Discovery in Databases and how they are applied on the Automated Blood Cell Counter Data to ensure the quality of results.

Minnie D and Srinivasan S (2012) generate association rules from the Automated Blood Cell Counter Data.

Minnie D and Srinivasan S (2012) have applied various initial means for the K-Means clustering algorithm for the Automated Blood Cell Counter Data.

4. Results and Discussion

The bone Marrow Data was taken as a raw data and the preprocessing phase of the KDD process was applied on the data to generate transformed data that was used to extract knowledge from the data.

4.1 *Data Cleaning*

The process of detecting and correcting or removing corrupt or inaccurate records from a record set, table, or database is Data Cleaning. The missing values in the Bone Marrow data cannot be replaced by any other value and hence those missing entries are replaced by the text “NA” referring “Not Applicable”.

The attributes myel, neutr, eosin, plasma, adn, eryth, blasts, promy, mono, baso, were required for analyzing the bone marrow data and hence the records without these fields were removed. The resultant excel file contained the records with bmid, myel, neutr, eosin, plasma, adn, eryth, blasts, baso, mono, promy were selected for further processing.

4.2 *Data Selection*

The cleaned bone marrow data was taken as the data source for data selection process. The patient records are to be grouped using various attributes and those attributes are selected for the clustering process. The numerical attributes representing the bone marrow components are selected for further process. A sample of the preprocessed data is shown in table 2.

4.3 *Data Transformation*

In the Data Transformation stage the data are transformed or consolidated in to forms appropriate for mining. The excel data is converted into a SQL Data base. The values 0 and NA are not included in the clustering process and they are shown as two independent clusters.

4.4 *Clustering*

The initial mean for clustering is selected as the first distinct k values from the dataset. The selected and transformed attributes are clustered for k values ranging from 2 to 5. A sample result for the attribute myel and k value 4 is shown in figure 3.

Table 2. Sample Preprocessed Data

bmid	eryth	blasts	promy	myel	neutr	eosin	lymph	plasma	adn	baso	mono
710113014	24	NA	16.5	26.5	17	2	13.5	0.5	NA	NA	NA
711013004	23	NA	21.5	32	9.5	5.5	7	1.5	NA	NA	NA
711013006	26	NA	21	18	11	3	19	2	NA	NA	NA
711013009	3.5	90.5	0.5	1	1	0.5	3	NA	NA	NA	NA
711013010	47	NA	12	14	19	NA	8	NA	NA	NA	NA
711013011	50	NA	10	9	20	11	NA	NA	NA	NA	NA
711013014	12	71	NA	NA	2	NA	12	3	NA	NA	NA
710023004	48	NA	22	11	10	NA	9	NA	NA	NA	NA
710023006	28	NA	26	12	26.5	NA	7.5	NA	NA	NA	NA
710023007	20.5	NA	20.5	32	15	2.5	6	3.5	NA	NA	NA
710023016	2.2	0.0	0.0	0.6	2.4	0.2	0.0	0.0	94.6	0.0	0.0
710163013	59.4	0.4	1.4	8.0	20.0	2.0	7.0	1.4	0.0	0.2	0.2

The records are sorted on the myel values and the first 4 distinct elements are taken as the starting mean values m1, m2, m3 and m4. All the elements are compared with the mean values and the records are placed in the cluster in which the element value and the mean value are closer. If there is a tie the element is placed in the first cluster among the set of equal clusters. The final cluster mean values are also generated.

The number of items placed, minimum value, maximum value and average value in each of the clusters when the value of k is 4 is given in figure 4.

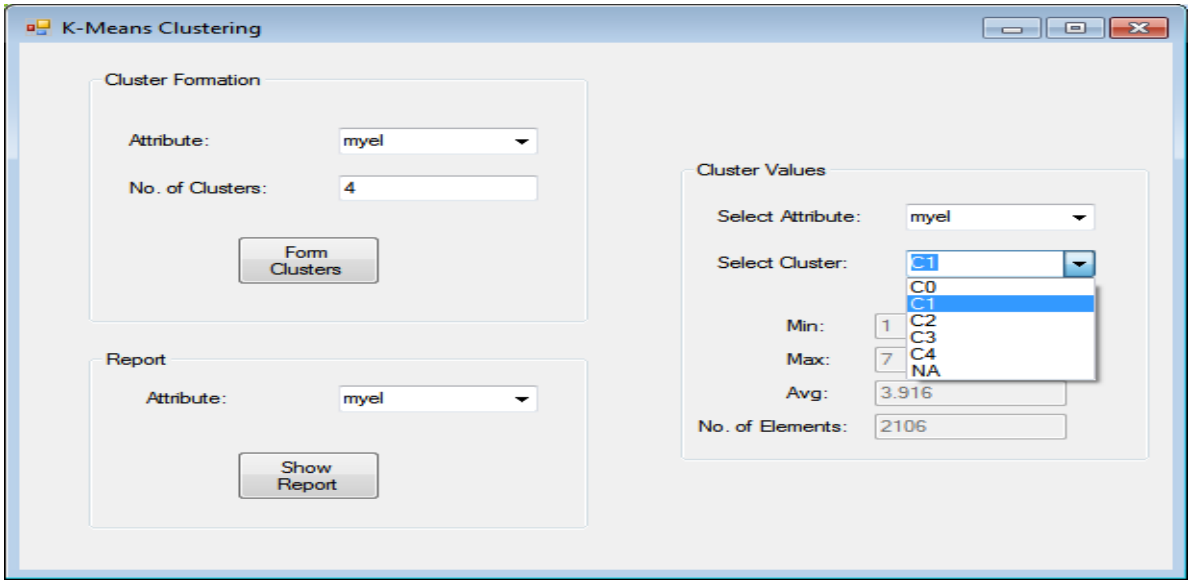
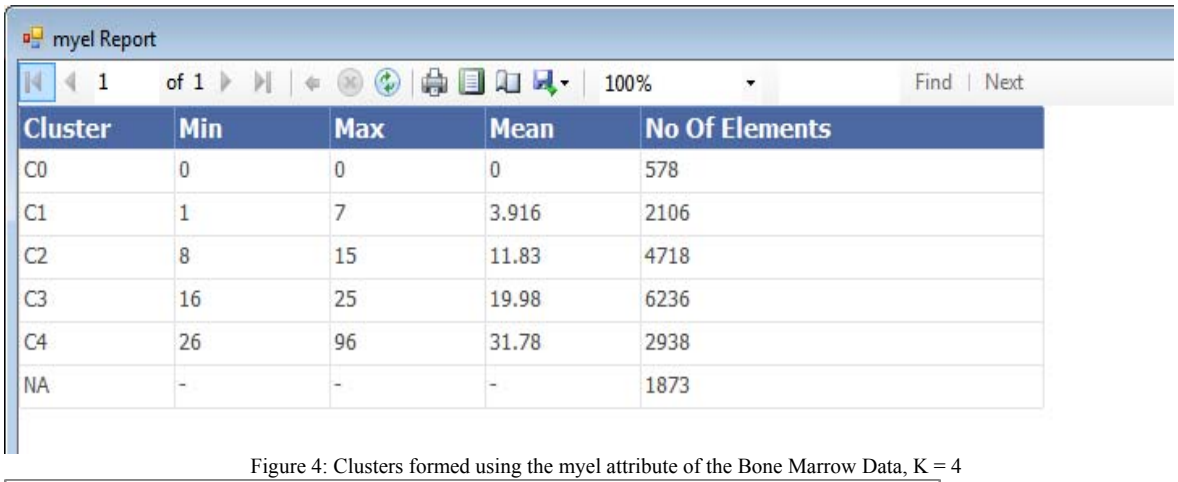


Figure 3: Clustering using myel attribute with k =4



Cluster	Min	Max	Mean	No Of Elements
C0	0	0	0	578
C1	1	7	3.916	2106
C2	8	15	11.83	4718
C3	16	25	19.98	6236
C4	26	96	31.78	2938
NA	-	-	-	1873

Figure 4: Clusters formed using the myel attribute of the Bone Marrow Data, K = 4

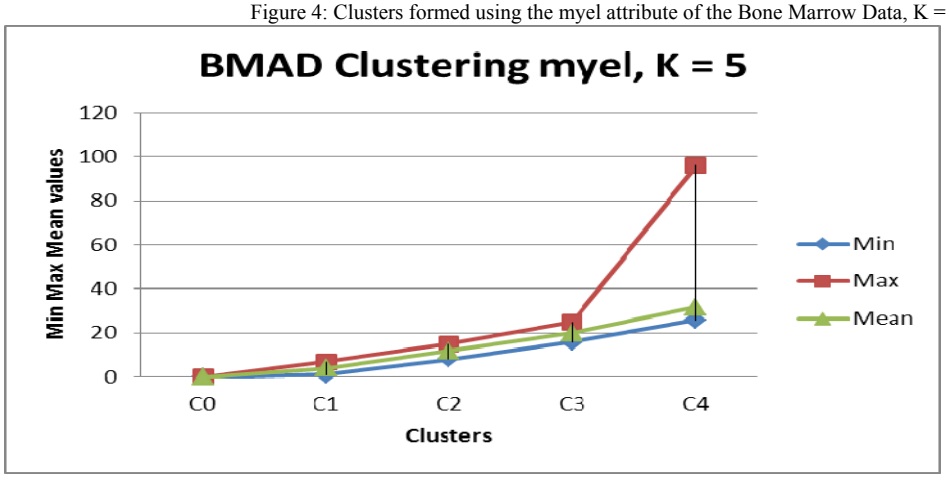


Figure 5: Min, Max and Mean values of Clusters for Bone Marrow Analysis Data attribute myel, K=4

Table 3: Min, Max, Mean values of Clusters and no. of elements per cluster for Bone Marrow Analysis Data attribute myel, K = 3

Cluster	Min	Max	Mean	NoOfElements
C0	0	0	0	578
C1	1	11	6.647	4092
C2	12	23	17.23	8007
C3	24	96	29.96	3899
NA	-	-	-	1873

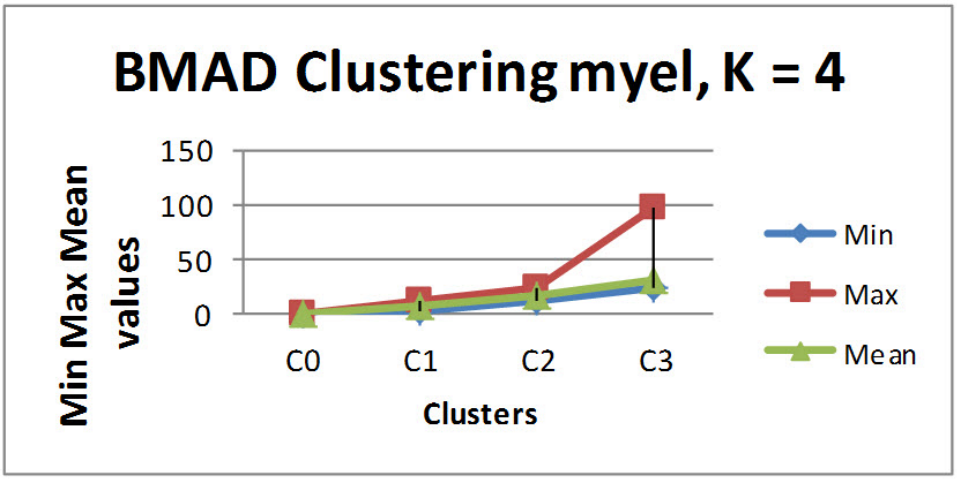


Figure 6: Min, Max and Mean values of Clusters for Bone Marrow Analysis Data attribute myel, K=3

Table 4. Clusters formed using various attributes of the Bone Marrow Data when K = 5

Attribute Name	Number of elements in Cluster i					
	i = 1	i = 2	i = 3	i = 4	i = 5	NA
myel	578	2106	4718	6236	2938	1873
neutr	502	2924	4246	6129	2869	1779
eosin	2129	5153	4510	2740	124	3793
plasma	3786	1	5	9596	353	5008
adn	6033	164	76	61	93	12022
baso	6747	91	99	35	6	11471
mono	6214	780	198	35	15	11207
promy	5425	1677	4813	1095	127	5312

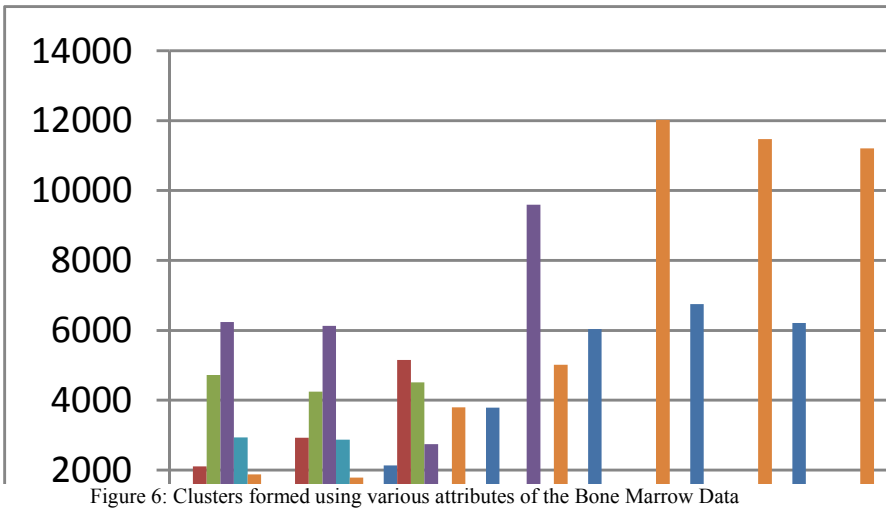


Figure 6: Clusters formed using various attributes of the Bone Marrow Data

5. Summary and Concluding Remarks

A brief study of Clinical Pathology, Hematology, Bone Marrow examination, Bone Marrow data is presented in the paper. The format of the bone marrow analysis result was described and few of the attributes were selected for processing, based on the knowledge given by the Clinical Pathologist. The KDD steps were explained and were applied on the Bone Marrow Data to convert the raw data into a transformed data that was used for generating more knowledge from the system. Various clusters are formed based on the various numerical attributes of the bone marrow data.

5. Running Heads

Clustering preprocessed bone marrow data using k-means algorithm.

Acknowledgments

The authors wish to thank Dr. Joy John Mammen, MD, Department of Transfusion Medicine and Immunohematology, Christian Medical College, Vellore, Tamilnadu, India for sharing his knowledge in Hematology, specially the functionalities of the bone marrow pathological data and also for providing the De-identified bone marrow data.

References

[1] Alp Aslandogan Y. and Gauri A.Mahajani, 2004. “Evidence Combination in Medical Data Mining”, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC’04), Volume 2, pp. 465 – 469
[2] Anil K Jain, “Data clustering: 50 years beyond K-means, Pattern Recognition Letters”, Volume 31, Issue 8, 1 June 2010, Pages 651-666
[3] Berks, Georg, Diedrich Graf V. Keyserlingk, Jan Jantzen, Mariagrazia Dotoli, and Hubertus Axer, 2000. "Fuzzy clustering-a versatile mean to explore medical database." ESIT2000, Aachen, Germany.
[4] Cios KJ, Moore GW, 2002. “Uniqueness of Medical Data Mining”, Artificial Intelligence in Medicine, 2002, Sep-Oct; 26(1-2), pp. 1-24.
[5] Derek Greene, Alexey Tsymbal, Nadia Bolshakova, P´adraig Cunningham, 2004. “Ensemble clustering in medical diagnostics”, Proceedings of 17th IEEE Symposium on Computer-Based Medical Systems, pp. 576 – 581
[6] Jaiwei Han, Michelinne Kamber, Data Mining : Concepts and Techniques, Morgan Kaufmann Publishers, Second Edition, 2006
[7] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, 2001. “Validating clustering for gene expression data”, Bioinformatics 2001 17(4), pp.309-318
[8] Margaret H.Dunham, Data Mining: Introductory and Advanced Topics, Pearson Education, 2007.
[9] Massoud Toussi, Jean-Baptiste Lamy, Philippe Le Toumelin, and Alain Venot, 2009. “Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes”, BMC Medical Informatics and Decision Making 2009; pp. 9:28
[10] Michael Goebel, Le Gruenwald, 1999. “A Survey of Data Mining and Knowledge Discovery Software Tools”, SIGKDD Explorations, ACM SIGKDD.
[11] Minnie D, Srinivasan S, 2011. “Application of Knowledge Discovery in Database to Blood Cell Counter Data to Improve Quality Control in Clinical Pathology”, Proceedings of 6th International Conference on Bio Inspired Computing – Theory and Applications 2011, pp 338 – 342.
[12] Minnie D, Srinivasan S, 2012. “Preprocessing and Genearation of Association Rules for Automated Blood Cell Counter Data in Haematology”, Proceedings of International Conference on Recent Advances in Computing and Software Systems 2012, pp 27 – 32.
[13] Minnie D, Srinivasan S, 2012. “Clustering the Preprocessed Automated Blood Cell Counter Data using modified K-Means algorithms and Generation of Association Rules”, International Journal of Computer Applications 2012, pp 38 – 42.
[14] Patricia Cerrito, John C. Cerrito, 2006. “Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs”, Proceedings of SUGI 31, March 26 – 29, paper 077-31.
[15] Rakesh Agrawal, T. Imielinski, A. Swami, 1993. “Database Mining: A Performance Perspective”, IEEE Transactions on Knowledge and Data Engineering, Volume 5 Issue 6, pp. 914 – 925.