Asian Journal of Information Technology 13 (1): 29-37, 2014
ISSN: 1682-3915

Preprocessing and Generation of Association Rules for Bone Marrow Analysis Data of Haematology for Acute Myeloid Leukemia

[1]D. Minnie and [2]S. Srinivasan
[1]Department of Computer Science, Madras Christian College, Chennai, India
[2]Department of Computer Science and Engineering, Anna University Regional Office, Madurai, India

Abstract: Clinical pathology uses laboratory tests on body fluids such as blood and urine to diagnose diseases. Haematology is the study of blood and blood forming organs such as bone marrow. In this study, researchers analyze the components of the bone marrow and the structure of the bone marrow analysis database. The Knowledge Discovery in Databases (KDD) steps are briefly explained. The 18,000 bone marrow analysis records are collected from a reputed hospital and this raw data is transformed into a preprocessed data using the pre-processing phases of KDD such as data cleaning, data selection and data transformation. Eliminate the tuple technique is used to clean the data. The attributes related to the bone marrow components are selected. The ranges of low, high and normal values for the individual attributes are used to transform the data. The data mining techniques are studied and the Apriori algorithm is selected for finding frequent itemsets that are used for the generation of association rules.

Key words: Association rule mining, bone marrow analysis, haematology, knowledge discovery in databases, blood

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 that can improve the quality of service in the medical fields.

Clinical pathology is a study that is concerned with conducting laboratory experiments on body fluids such as blood and urine to diagnose diseases. Hematology is the study of blood, diseases related to blood and blood forming organs such as bone marrow. Hematology department of clinical pathology performs various tests on blood.

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) as discussed by Han and Kamber (2006) and Dunham (2007) is used to extract useful knowledge from the raw bone marrow data. The raw data is preprocessed using the KDD preprocessing steps such as data cleaning, data transformation and the resultant data is used to generate knowledge using the data mining techniques.

Data mining is the process of extracting implicit, useful, previously unknown, non-trivial information from data. The data mining techniques are broadly grouped as classification, clustering, association rule mining and prediction. Various algorithms are developed for each technique. The methods developed for the association rule mining are given by Agrawal *et al*. (1993a, b). The Apriori algorithm as discussed by Srikant and Agrawal (1995) is used in this project on the preprocessed data to generate associations between the attributes of the bone marrow analysis database.

Medical data is taken most of the times from medical records as discussed by Cerrito and Cerrito (2006) and the data is found to be heterogeneous in nature. The bone marrow analysis data consists of both numerical data as well as the detailed description about the data and the final impression about the patient. The privacy issues are to be finalized before handling medical data and is described by Cios and Moore (2002). The bone marrow analysis data is de-identified so that the privacy of the patient is protected.

Corresponding Author: D. Minnie, Department of Computer Science, Madras Christian College, Chennai, India

A major source of error in clinical pathology is specimen mislabeling that can be reduced by collecting and trending the data on mislabeled samples with timely feedback to patient care as shown by Quillen and Murphy (2006). Auto verification of results as discussed by Duca (2002) in a laboratory information system is used to verify the correctness of a result.

Aslandogan and Mahajani (2004) discusses the various combinations of data mining classification algorithms used on medical data for efficient classification of the data. Toussi *et al.* (2009) presents some of the ways of using sequences of clustering algorithms to mine temporal data.

Dogan and Turkoglu (2008) uses association rule mining to diagnose hyperlipidemia disease. Goh and Ang (2007) shows how association rule mining can be used in the application of counseling and help seeking behavior of adolescents. Li *et al.* (2005) describes how risk patterns can be identified from medical data.

The automated blood cell counter data is a hematology data that contains the Complete Blood Count (CBC) details such as RBC, WBC, MCH, etc. and they are subjected to delta checks to ensure the quality of the tests as presented by Minnie and Srinivasan (2011, 2012). Minnie and Srinivasan (2012) presents how association rules are generated among the attributes of the automated blood cell counter data.

The literature studied for this project is on medical data, issues in handling medical data, KDD, data mining techniques, association rule mining and the application of various data mining techniques on various medical data and are discussed above. This bone marrow analysis data was not found to be used for mining knowledge and hence the association rule generation for bone marrow analysis data is selected for this project.

MATERIALS AND METHODS

System architecture: The system architecture of this research is shown in Fig. 1 and it consists of studying the bone marrow components, analyse bone marrow data structure, bone marrow analysis data collection, KDD preprocessing steps data cleaning, data selection and data transformation and data mining using association rule generation.

Bone marrow components: 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.

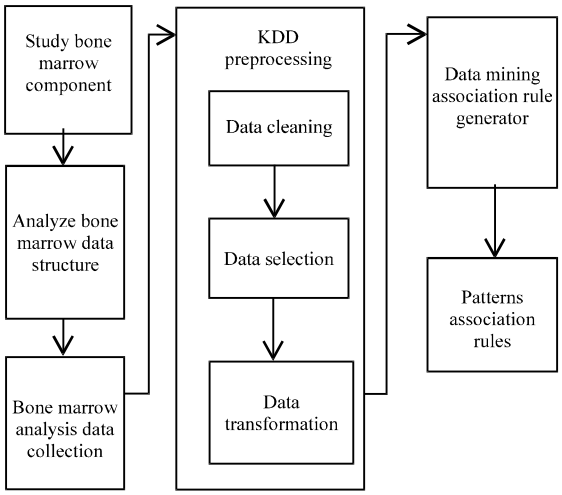


Fig. 1: System architecture

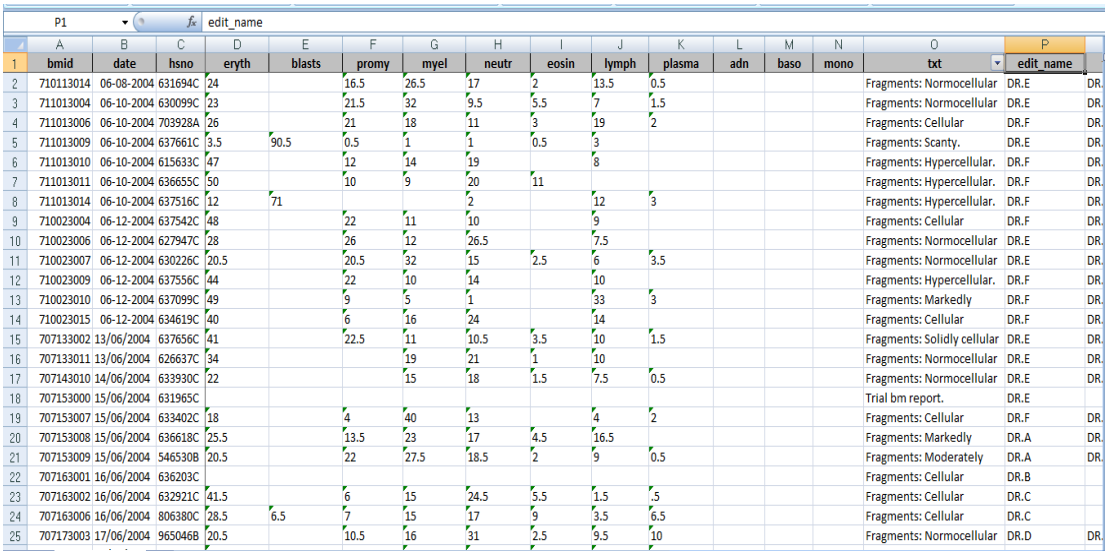
Data collection: Eighteen thousand bone marrow data are collected from a Clinical Pathology department of a reputed hospital. The data is present as an excel file. The bone marrow analysis data is de-identified to preserve the privacy of the patient, doctor and the hospital. A sample of the data file is shown in Fig. 2.

Bone marrow data format: The bone marrow analysis 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 of the attributes is shown in Table 1.

Knowledge Discovery in Databases (KDD): 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 data from 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



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	bmid	date	hsno	eryth	blasts	promy	myel	neutr	eosin	lymph	plasma	adn	baso	mono	txt	edit_name
2	710113014	06-08-2004	631694C	24		16.5	26.5	17	2	13.5	0.5				Fragments: Normocellular	DR.E
3	711013004	06-10-2004	630099C	23		21.5	32	9.5	5.5	7	1.5				Fragments: Normocellular	DR.E
4	711013006	06-10-2004	703928A	26		21	18	11	3	19	2				Fragments: Cellular	DR.F
5	711013009	06-10-2004	637661C	3.5	90.5	0.5	1	1	0.5	3					Fragments: Scanty.	DR.E
6	711013010	06-10-2004	615633C	47		12	14	19		8					Fragments: Hypercellular.	DR.F
7	711013011	06-10-2004	636655C	50		10	9	20	11						Fragments: Hypercellular.	DR.F
8	711013014	06-10-2004	637516C	12	71			2		12	3				Fragments: Hypercellular.	DR.F
9	710023004	06-12-2004	637542C	48		22	11	10		9					Fragments: Cellular	DR.F
10	710023006	06-12-2004	627947C	28		26	12	26.5		7.5					Fragments: Normocellular	DR.E
11	710023007	06-12-2004	630226C	20.5		20.5	32	15	2.5	6	3.5				Fragments: Normocellular	DR.E
12	710023009	06-12-2004	637556C	44		22	10	14		10					Fragments: Hypercellular.	DR.F
13	710023010	06-12-2004	637099C	49		9	5	1		33	3				Fragments: Markedly	DR.F
14	710023015	06-12-2004	634619C	40		6	16	24		14					Fragments: Cellular	DR.F
15	707133002	13/06/2004	637656C	41		22.5	11	10.5	3.5	10	1.5				Fragments: Solidly cellular	DR.E
16	707133011	13/06/2004	626637C	34			19	21	1	10					Fragments: Normocellular	DR.E
17	707143010	14/06/2004	633930C	22			15	18	1.5	7.5	0.5				Fragments: Normocellular	DR.E
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
20	707153008	15/06/2004	636618C	25.5		13.5	23	17	4.5	16.5					Fragments: Markedly	DR.A
21	707153009	15/06/2004	546530B	20.5		22	27.5	18.5	2	9	0.5				Fragments: Moderately	DR.A
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
24	707163006	16/06/2004	806380C	28.5	6.5	7	15	17	9	3.5	6.5				Fragments: Cellular	DR.C
25	707173003	17/06/2004	965046B	20.5		10.5	16	31	2.5	9.5	10				Fragments: Normocellular	DR.D

Fig. 2: Sample bone marrow analysis data

Table 1: Attributes of the bone marrow analysis data

Attribute names	Attribute full name	Attribute description
bmid	Patient ID	The ID of a patient
date	Test date	The date in which the test is taken
hsno	Hospital number	The hospital ID
eryth	Erythrocyte count	Erythrocytes are the red blood cells formed in the bone marrow
blasts	Blast cell count	Blast cells are immature cells that do not do any function
promy	Promyelocyte count	Promyelocytes are immature cells developed from myeloblasts
myel	Myeloblast count	Myeloblast is a unipotent stem cell
plasma	Plasma cell count	Plasma cells secrete large quantities of antibodies
neutr	Neutrophil count	Neutrophil is a type of White Blood Cell (WBC) that fights harmful foreign particles and bacteria
eosin	Eosinophil count	Eosinophil is a type of WBC that fights multicellular parasites and controls allergy and asthma
lymph	Lymphocyte count	Lymphocyte is a type of WBC that produces antibodies
baso	Basophil count	Basophil is a type of WBC that fights infections
mono	Monocyte	Monocyte is a type of WBC that fights viruses
txt	Test results	Detailed description of test results
edit_name	Doctor's name	Doctor who entered the txt
fin_name	Doctor's name	Doctor who gave final impression
imp	Impression	The final impression of the tests

identified from the processed data. The discovered knowledge is visually presented to the user in the knowledge representation process.

Data mining: Data mining is the knowledge discovery stage of KDD. 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 attributes in a database.

Association Rule Mining (ARM): ARM is used to find frequent patterns, associations and correlations among sets of items in databases and any other information repositories. Association rule correlates the presence of one set of items with that of another set of items in the same transaction. The quality of an association rule is measured using its support and confidence values and several efficient methods are developed to generate association rules.

Let N be the number of records in a database, N(I) be the number of records with item set I, X be an item set with k elements $X_1, X_2, \ldots, X_k$ and Y be an item set with h elements $Y_1, Y_2, \ldots, Y_h$.

An association rule $X \Rightarrow Y$ can be generated if the support of X and that of Y is above the minimum support value and also the confidence of the rule $X \Rightarrow Y$ is above the minimum confidence specified.

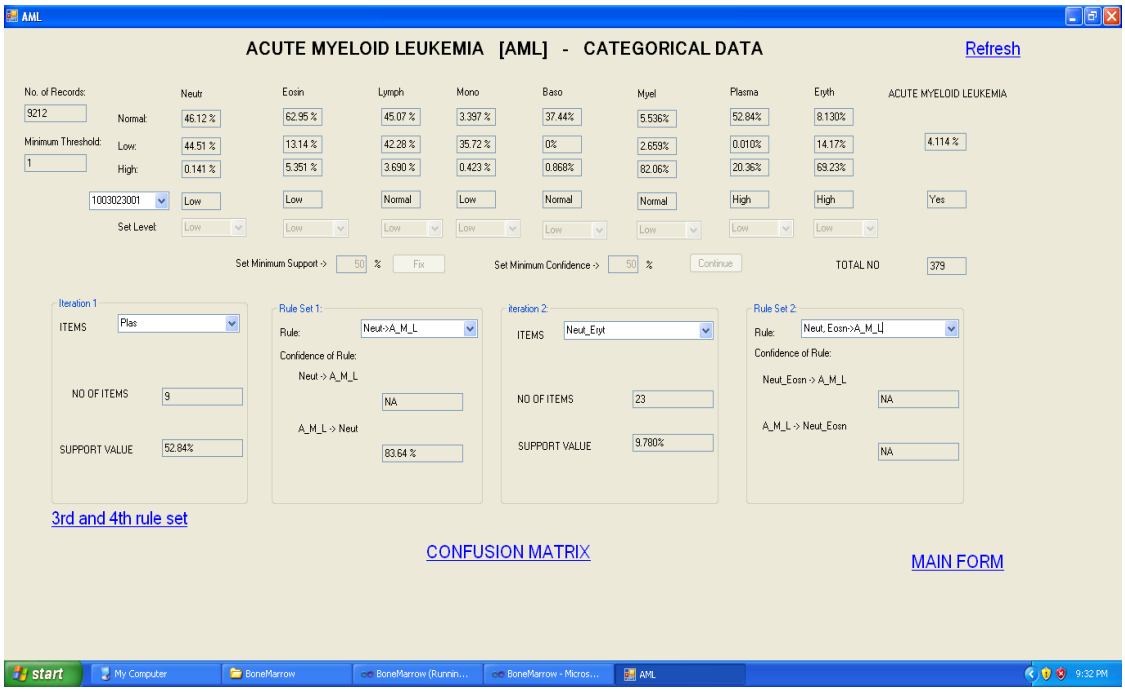


Fig. 3: Application interface for association rule generation for AML

Support S, of X is a probability that a transaction contains X and is given in Eq. 1 and 2:

$$S(X) = P(X) \quad (1)$$

$$S(X) = \frac{N(X)}{N} \quad (2)$$

Also, support of X⇒Y is a probability that a transaction contains both X and Y as given in Eq. 3 and 4:

$$S(X \Rightarrow Y) = P(X \bigcup Y) \quad (3)$$

$$S(X \Rightarrow Y) = \frac{N(X \bigcup Y)}{N} \quad (4)$$

Confidence C, of X⇒Y is a conditional probability that a transaction that contains X contains Y also. Confidence can be calculated as in Eq. 5-7.

$$C(X \Rightarrow Y) = P(X|Y) \quad (5)$$

$$C(X \Rightarrow Y) = \frac{N(X \bigcup Y)}{N(X)} \quad (6)$$

$$C(X \Rightarrow Y) = \frac{S(X \bigcup Y)}{S(X)} \quad (7)$$

Association rule mining consists of the two steps finding frequent item sets and generating association rules.

The set of all item sets are called as candidate item sets. The item sets which occur frequently in a database are called frequent item sets. The frequent item sets satisfy the minimum support, min_sup specified by the user or designer. If the min_sup is 10% then the item sets whose support % are ≥10 are considered as frequent item sets.

The association rules are generated using the frequent item sets and they should satisfy the minimum confidence min_conf of 50%. Apriori algorithm is used to find frequent item sets in a database and to generate association rules from the frequent item sets. Apriori algorithm uses the apriori principle "All subsets of an infrequent item set are infrequent" and hence those subsets need not be considered for further processing.

Application interface: The interface for this application to generate association rules between the attributes of the bone marrow analysis data are given in Fig. 3.

RESULTS AND DISCUSSION

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 or invalid values in the bone marrow data cannot be replaced by any other value and hence those records were eliminated from further process. 18,429 records out of 18,449 records were selected for further process as they have valid values.

Data selection: The cleaned bone marrow data was taken as input for data selection. The numerical attributes representing the bone marrow components such as bid, neutr, eosin, lymph, baso, mono, myel, plasma, eryth and impression are identified for the generation of association rules and hence they are selected for further process.

Training and testing data: The cleaned and transformed data is divided into two data sets training data and testing data. The training data consist of 9212 records and testing data consist of 9217 records. The training data is used to generate association rules and the testing data is used to measure the quality of the association rules generated.

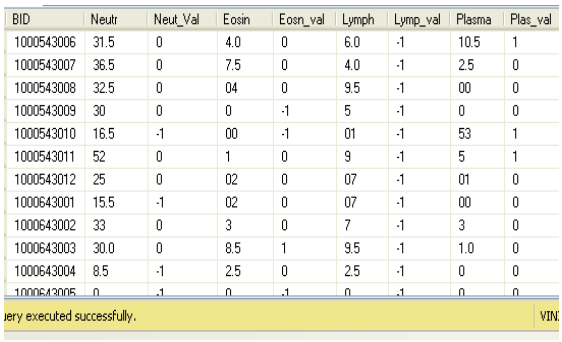
Data transformation: In the data transformation stage the data are transformed or consolidated in to forms appropriate for mining. The ranges of values for the bone marrow component attributes are used to find out whether the value is low, normal or high. New attributes low, normal and high are generated for each of the attributes. Hence, the individual attribute values are used to generate the status of the attributes as low, normal or high.

A value 0 is stored for the normal values, 1 is stored for the high values and -1 is stored for the low values for the newly generated attributes depending on the values of the corresponding attributes.

The disease AML data is collected from the impression attribute and a new attribute is inserted for AML that can hold Boolean values 0 for the absence of the disease AML and 1 for the presence of the disease AML for that record. The flattened data is shown in Fig. 4 and 5. Also, the excel data was converted into a SQL Data base.

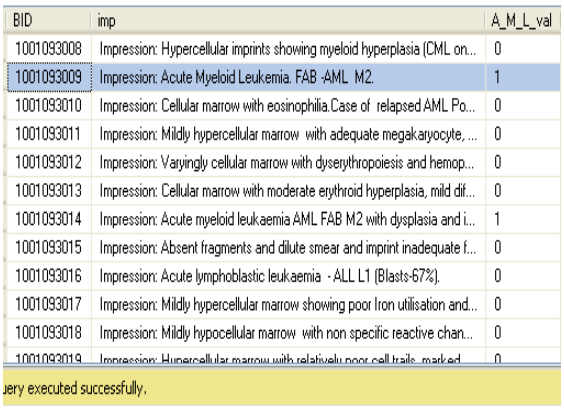
Attribute combinations: The attributes have a normal, low and high support associated with them. Various combinations of these are considered for generating the association rules. The combination selected to produce association rules is shown in Table 2.

Frequent item sets: The minimum support value is set as 0.1% to include all the item sets including outliers for the association rule generation process.



BID	Neutr	Neut_Val	Eosin	Eosn_val	Lymph	Lymp_val	Plasma	Plas_val
1000543006	31.5	0	4.0	0	6.0	-1	10.5	1
1000543007	36.5	0	7.5	0	4.0	-1	2.5	0
1000543008	32.5	0	04	0	9.5	-1	00	0
1000543009	30	0	0	-1	5	-1	0	0
1000543010	16.5	-1	00	-1	01	-1	53	1
1000543011	52	0	1	0	9	-1	5	1
1000543012	25	0	02	0	07	-1	01	0
1000643001	15.5	-1	02	0	07	-1	00	0
1000643002	33	0	3	0	7	-1	3	0
1000643003	30.0	0	8.5	1	9.5	-1	1.0	0
1000643004	8.5	-1	2.5	0	2.5	-1	0	0

Fig. 4: Flattened bone marrow numerical data



BID	imp	A_M_L_val
1001093008	Impression: Hypercellular imprints showing myeloid hyperplasia (CML on...	0
1001093009	Impression: Acute Myeloid Leukemia. FAB -AML M2.	1
1001093010	Impression: Cellular marrow with eosinophilia.Case of relapsed AML Po...	0
1001093011	Impression: Mildly hypercellular marrow with adequate megakaryocyte, ...	0
1001093012	Impression: Varyingly cellular marrow with dyserythropoiesis and hemop...	0
1001093013	Impression: Cellular marrow with moderate erythroid hyperplasia, mild dif...	0
1001093014	Impression: Acute myeloid leukaemia AML FAB M2 with dysplasia and i...	1
1001093015	Impression: Absent fragments and dilute smear and imprint inadequate f...	0
1001093016	Impression: Acute lymphoblastic leukaemia - ALL L1 (Blasts-67%).	0
1001093017	Impression: Mildly hypercellular marrow showing poor Iron utilisation and...	0
1001093018	Impression: Mildly hypocellular marrow with non specific reactive chan...	0

Fig. 5: Flattened bone marrow categorical disease data

Table 2: Combinations of levels selected for attributes

Attributes	Levels	Support value	Support (%)
Neutrophils	Low	4101	44.51802
Eosinophils	Low	1211	13.14590
Lymphocytes	Normal	4152	45.07165
Monocytes	Low	3291	35.72514
Blasts	High	1104	11.98437
Myeloblast	High	7560	82.06687
Plasma Cells	Normal	4868	52.84412
Erythrocyte	High	6378	69.23578
A_M_L	Present	1052	11.41989

Table 3: Frequent item set generation count

Item sets	Count
Frequent 1	9
Frequent 2	36
Frequent 3	84
Frequent 4	121

The frequent 1 item sets for the combinations of levels selected for the attributes are given in Table 2. The number of frequent 1, 2, 3 and 4 item sets are given in Table 3.

Association rules generated without disease detail: The minimum confidence is set as 50% to select strong

Table 4: Association rules generated with 2 attributes and without disease details

Association rules	Confidence
Neut_Eosn->Mono	73.43
Neut_Eosn->Myel	53.96
Neut_Eosn->Plas	62.50
Neut_Lymp->Myel	91.23
Neut_Lymp->Plas	55.80
Neut_Lymp->Eryt	81.53
Neut_Mono->Myel	75.09
Neut_Mono->Plas	70.49
Neut_Mono->Eryt	64.68
Neut_Blas->Myel	51.97
Neut_Myel->Plas	55.24
Neut_Myel->Eryt	82.79
Neut_Plas->Eryt	74.78
Eosn_Lymp->Mono	63.74
Eosn_Lymp->Myel	83.94
Eosn_Lymp->Plas	63.74
Eosn_Lymp->Eryt	66.18
Eosn_Mono->Myel	62.42
Eosn_Mono->Plas	70.44
Eosn_Blas->Plas	64.48
Eosn_Myel->Plas	64.32
Eosn_Myel->Eryt	75.52
Eosn_Plas->Eryt	54.92
Lymp_Mono->Myel	94.15
Lymp_Mono->Plas	74.43
Lymp_Mono->Eryt	80.55
Lymp_Blas->Myel	63.35
Lymp_Myel->Plas	60.60
Lymp_Myel->Eryt	83.89
Eryt->Lymp_Myel	51.61
Lymp_Plas->Eryt	84.49
Mono_Blas->Myel	52.36
Mono_Blas->Plas	71.57
Mono_Myel->Plas	73.84
Mono_Myel->Eryt	82.46
Mono_Plas->Eryt	76.29
Blas_Myel->Plas	52.92
Myel_Plas->Eryt	84.94
Eryt->Myel_Plas	59.72

association rules that are formed using the numerical attributes of the bone marrow data and the categorical disease data is not considered.

The associations generated with 2 attributes are given in Table 4 with 3 attributes are given in Table 5 and are shown graphically in Fig. 6. The 4 attribute associations and the 5 attribute association rules have the confidence level lesser than the minimum confidence of the confidence level lesser than the minimum confidence of 50% and hence they are not discussed.

Association rules generated with disease AML: The minimum confidence is set as 50% to select strong association rules. The association rules are generated with the disease AML and other attributes of the bone marrow analysis data. The associations generated with AML and 1 attribute is given in Table 6, 2 attributes are given in Table 7 with 3 attributes are given Table 8. The 4 attribute ones are shown in Table 9 and all associations are shown in Fig. 7.

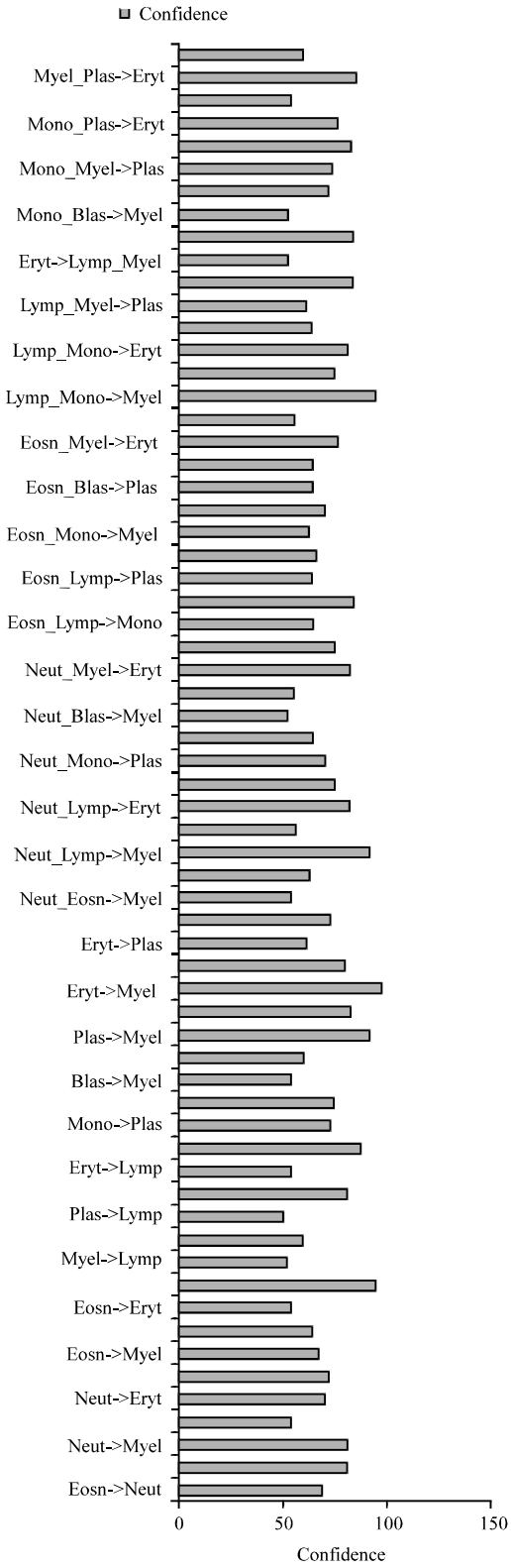


Fig. 6: Confidence of association rules generated without disease details

Table 5: Association rules generated with 3 attributes and without disease details

Association rules	Confidence
Neut_Eosn->Mono	73.43
Neut_Eosn->Myel	53.96
Neut_Eosn->Plas	62.50
Neut_Lymp->Myel	91.23
Neut_Lymp->Plas	55.80
Neut_Lymp->Eryt	81.53
Neut_Mono->Myel	75.09
Neut_Mono->Plas	70.49
Neut_Mono->Eryt	64.68
Neut_Blas->Myel	51.97
Neut_Myel->Plas	55.24
Neut_Myel->Eryt	82.79
Neut_Plas->Eryt	74.78
Eosn_Lymp->Mono	63.74
Eosn_Lymp->Myel	83.94
Eosn_Lymp->Plas	63.74
Eosn_Lymp->Eryt	66.18
Eosn_Mono->Myel	62.42
Eosn_Mono->Plas	70.44
Eosn_Blas->Plas	64.48
Eosn_Myel->Plas	64.32
Eosn_Myel->Eryt	75.52
Eosn_Plas->Eryt	54.92
Lymp_Mono->Myel	94.15
Lymp_Mono->Plas	74.43
Lymp_Mono->Eryt	80.55
Lymp_Blas->Myel	63.35
Lymp_Myel->Plas	60.60
Lymp_Myel->Eryt	83.89
Eryt->Lymp_Myel	51.61
Lymp_Plas->Eryt	84.49
Mono_Blas->Myel	52.36
Mono_Blas->Plas	71.57
Mono_Myel->Plas	73.84
Mono_Myel->Eryt	82.46
Mono_Plas->Eryt	76.29
Blas_Myel->Plas	52.92
Myel_Plas->Eryt	84.94
Eryt->Myel_Plas	59.72

Table 6: Association rules generated with AML and 1 attribute

Rules	Confidence
A_M_L->Neut	69.29
Blas->A_M_L	62.40
A_M_L->Blas	65.49
A_M_L->Myel	52.09

Table 7: Association rules generated with AML and 2 attributes

Association rules	Confidence
Neut_Blas->A_M_L	66.44
A_M_L->Neut_Blas	55.89
Eosn_Blas->A_M_L	66.97
Lymp_Blas->A_M_L	64.88
Mono_Blas->A_M_L	62.59
Blas_Myel->A_M_L	60.48
Blas_Plas->A_M_L	63.08
Blas_Eryt->A_M_L	52.23

Table 8: Association rules generated with AML and 3 attributes

Association rules	Confidence
Neut_Eosn_Blas->A_M_L	69.10
Neut_Lymp_Blas->A_M_L	69.78
Neut_Mono_Blas->A_M_L	66.29
Neut_Blas_Myel->A_M_L	66.52

Table 8: Continue

Association rules	Confidence
Neut_Blas_Plas->A_M_L	67.85
Neut_Blas_Eryt->A_M_L	55.87
Eosn_Lymp_Blas->A_M_L	73.83
Eosn_Mono_Blas->A_M_L	67.07
Eosn_Blas_Myel->A_M_L	72.95
Eosn_Blas_Plas->A_M_L	68.11
Eosn_Blas_Eryt->A_M_L	66.66
Lymp_Mono_Blas->A_M_L	65.58
Lymp_Blas_Myel->A_M_L	59.43
Lymp_Blas_Plas->A_M_L	64.61
Mono_Blas_Myel->A_M_L	60.47
Mono_Blas_Plas->A_M_L	60.97
Mono_Blas_Eryt->A_M_L	51.64
Blas_Myel_Plas->A_M_L	60.06
Blas_Plas_Eryt->A_M_L	51.35

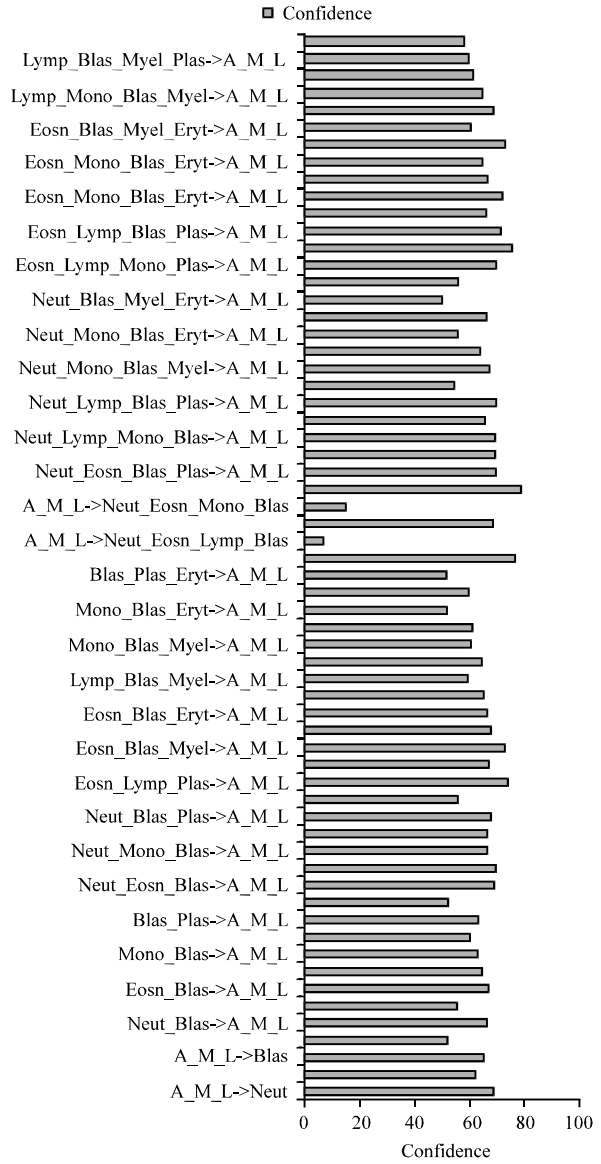


Fig. 7: Confidence of association rules generated with disease AML

Table 9: Association rules generated with AML and 4 attributes

Association rules	Confidence
Neut_Eosn_Lymp_Blas->A_M_L	77.000
A_M_L->Neut_Eosn_Lymp_Blas	7.319
Neut_Eosn_Mono_Blas->A_M_L	68.660
A_M_L->Neut_Eosn_Mono_Blas	15.200
Neut_Eosn_Blas_Myel->A_M_L	79.040
Neut_Eosn_Blas_Plas->A_M_L	70.100
Neut_Eosn_Blas_Eryt->A_M_L	69.810
Neut_Lymp_Mono_Blas->A_M_L	69.850
Neut_Lymp_Blas_Myel->A_M_L	66.010
Neut_Lymp_Blas_Plas->A_M_L	70.300
Neut_Lymp_Blas_Eryt->A_M_L	54.380
Neut_Mono_Blas_Myel->A_M_L	67.450
Neut_Mono_Blas_Plas->A_M_L	64.340
Neut_Mono_Blas_Eryt->A_M_L	56.090
Neut_Blas_Myel_Plas->A_M_L	66.660
Neut_Blas_Myel_Eryt->A_M_L	50.230
Neut_Blas_Plas_Eryt->A_M_L	55.900
Eosn_Lymp_Mono_Blas->A_M_L	70.230
Eosn_Lymp_Blas_Myel->A_M_L	75.860
Eosn_Lymp_Blas_Plas->A_M_L	72.220
Eosn_Lymp_Blas_Eryt->A_M_L	66.660
Eosn_Mono_Blas_Myel->A_M_L	72.410
Eosn_Mono_Blas_Plas->A_M_L	67.030
Eosn_Mono_Blas_Eryt->A_M_L	65.000
Eosn_Blas_Myel_Plas->A_M_L	73.490
Eosn_Blas_Myel_Eryt->A_M_L	60.860
Eosn_Blas_Plas_Eryt->A_M_L	68.420
Lymp_Mono_Blas_Myel->A_M_L	64.640
Lymp_Mono_Blas_Plas->A_M_L	61.260
Lymp_Blas_Myel_Plas->A_M_L	59.550
Mono_Blas_Myel_Plas->A_M_L	58.550

CONCLUSION

A brief study of clinical pathology, hematology, bone marrow components and bone marrow analysis data are presented in the study. The format of the bone marrow analysis database was described and few of the attributes were selected for generating association rules, based on the knowledge given by the clinical pathologist. The KDD steps were explained and were applied on the bone marrow analysis data to convert the raw data into a transformed data that was used for generating more knowledge from the system. Frequent item sets of the medical data are generated using Apriori algorithm and association rules are generated using the records that satisfy the selected levels of range for the attributes. The association rules could not be generated for the other combinations as the confidence of the association rules generated are lesser than the minimum confidence level (50%) set for this application. The association rules can be generated for the records with other diseases such as malaria, MMM and ALL.

ACKNOWLEDGEMENT

Researchers 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 components of the bone marrow and also for providing the de-identified bone marrow analysis data.

REFERENCES

Agrawal, R., T. Imielinski and A. Swami, 1993a. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 25-28, 1993, Washington, DC., USA., pp: 207-216.

Agrawal, R., T. Imielinski and A. Swami, 1993b. Database Mining: A performance perspective. IEEE Trans. Knowledge Data Eng., 5: 914-925.

Aslandogan Y.A. and G.A. Mahajani, 2004. Evidence combination in medical data mining. Proceedings of the International Conference on Information Technology: Coding and Computing, Volume 2, April 5-7, 2004, pp: 465-469. 10.1109/ITCC.2004. 1286697.

Cerrito, P. and J.C. Cerrito, 2006. Data and text mining the electronic medical record to improve care and to lower costs. Proceedings of the 31st SAS Users Group International Conference on Data Mining and Predictive Modeling, March 26-29, 2006, San Francisco, CA., USA., pp: 1-20.

Cios, K.J. and G.W. Moore, 2002. Uniqueness of medical data mining. Artif. Intell. Med., 26: 1-24.

Dogan, S. and I. Turkoglu, 2008. Diagnosing hyperlipidemia using association rules. Math. Comput. Appl., 13: 193-202.

Duca, D.J., 2002. Autoverification in a laboratory information system. Lab. Med., 33: 21-25.

Dunham, M.H., 2007. Data Mining: Introductory and Advanced Topics. Prentice Hall, USA.

Goh, D.H. and R.P. Ang, 2007. An introduction to association rule mining: an application in counseling and help-seeking behavior of adolescents. Behaviour Res. Methods 39: 259-266.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publisher, San Fransisco, USA., ISBN: 1-55860-901-6.

Li, J., A. Wai-Chee Fu, H. He, J. Chen and H. Jin, 2005. Mining risk patterns in medical data. Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 21-24, Chicago, Illinois, USA., pp: 770-775.

Minnie, D. and S. Srinivasan, 2011. Application of knowledge discovery in database to blood cell counter data to improve quality control in clinical pathology. Proceedings of the 6th International Conference on Bio-Inspired Computing: Theories and Applications, September 27-29, 2011, Penang, Malaysian, pp: 338-342.

Minnie, D. and S. Srinivasan, 2012. Preprocessing and generation of association rules for automated blood cell counter data in haematology. Proceedings of the International Conference on Recent Advances in Computing and Software Systems, April 25-27, 2012, Chennai, India, pp: 27-32.

Quillen, K. and K. Murphy, 2006. Quality improvement to decrease specimen mislabeling in transfusion medicine. Arch. Pathol. Lab. Med., 130: 1196 -1198.

Srikant, R. and R. Agrawal, 1995. Mining generalized association rules. Proceedings of the 21st International Conference on Very Large Databases, September 11-15, 1995, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA., pp: 407-419.

Toussi, M., J.B. Lamy, P.L. Toumelin and A. 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 Med. Inform. Decis. Making, Vol. 9 10.1186/1472-6947-9-28.