Clustering the Preprocessed Automated Blood Cell Counter Data using modified K-Means Algorithms and Generation of Association Rules

D. Minnie
Madras Christian College
Tambaram East, Chennai
Tamil Nadu, India

S. Srinivasan
Anna University of Technology Madurai
Madurai
Tamil Nadu, India

ABSTRACT
The raw data from an Automated Blood Cell Counter is transformed in to a Preprocessed and Flattened data using the preprocessing phases of the Knowledge Discovery in Databases and the transformed data is used to create clusters of the database in this paper. The K-Means algorithm is applied on the database to form various clusters. Twelve thousand records are taken from a clinical laboratory for processing. Associations among the various attributes of the database are generated.

General Terms
Algorithms.

Keywords
Hematology, Blood Cell Counter, Knowledge Discovery in Databases, Data Mining, Clustering, K-Means Clustering, Association Rule Mining.

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.

Complete Blood Count (CBC) or Full Blood Count (FBC) of the blood can be found using either a manual procedure or an automated procedure. A Blood Cell Counter is an automated system that generates the CBC blood test results.

Knowledge Discovery in Databases (KDD) [1], [2] is used to convert the raw data into a form that is appropriate for the Data Mining process and then to generate meaningful results from data and hence it is applied in this paper on the blood cell counter data to generate knowledge.

2. METHODS

2.1 Automated Blood Cell Counter Data
A Blood Cell Counter [3] is an automated machine that can be loaded with blood samples and Complete Blood Count of the given blood samples are generated as an excel report. The report also contains the patient id, hospital number, date and time of the test which are extracted from the barcode pasted on the blood sample container. The number of red blood cells, white blood cells and platelets are some of the blood counts generated by the Automated Blood Cell Counter.

2.2 Data collection
Twelve thousand cell counter data are collected from a Clinical Pathology department of a reputed hospital. The data is present as an excel file.

2.3 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 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 [1]. The 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) [4] uncovers relationships among data in a database.

2.5 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.

Some of the major clustering models are Centroid based clustering, Density based clustering, Connectivity based clustering and Distribution based 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.

2.6 Association Rule Mining

Association Rule Mining is used to find the associations or relationships between various attributes of a database.

Association Rule Mining involves the following two major processes:

- Identifying Frequent Patterns and
- Generating Association Rules

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 [5] to generate association rules.

Support of an item or item set X is the probability of X being present in the database. The item or item sets with support value greater than the minimum support are the frequent item or frequent item sets.

Confidence of a rule $X \rightarrow Y$ is the probability of items or item sets X and Y being present together in the same transaction in the database.

The Apriori algorithm is used to find the frequent item sets faster. The algorithm uses the Apriori property "Any sub set of frequent item set is also frequent". For example, if {A,B} is frequent item sets then {A} and {B} are also frequent item sets.

The Apriori property is used to reduce the number of candidate item sets as follows: "Any superset containing an infrequent item set is also infrequent". For example if an item set {C,D} is infrequent, then all the item sets containing them such as {B,C,D}. {A,C,D}, {A,B,C,D} and etc. are also infrequent. This eliminates the search space associated with the infrequent items from being used in the generation of association rules.

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.

2.7 Automated Blood Cell Counter Data Format

The Blood Cell Counter Data is given as an excel file. A sample of it is given in table 2. The Blood Cell Counter data consists of values for each sample of blood for the various attributes such as RBC, WBC, PId, SId, PAge, PGender, RDate, RTime, Hgb, MCH and so on. The list of attributes [6] along with a detailed description is shown in table 1.

Table 1. Automated Blood Cell Counter Data Attributes

Attribute Name	Attribute Description
PID	Patient Id
R1DATE	Run Date
R1TIME	Run Time
PAGE	Patient Age
PGENDER	Patient Gender
SID	Sample Id
RBC	Red Blood Cell Count
WBC	White Blood Cell Count
Hgb	Hemoglobin Concentration
Hct	Hematocrit
MCV	Mean Cell Volume
MCH	Mean Cell Hemoglobin
MCHC	Mean Cell Hemoglobin Concentration
RDW	Red cell Distribution Width
Plt	Platelet Count
Pct	Prothrombin Consumption Time
MPV	Mean Platelet Volume
PDW	Platelet Distribution Width
NE%	Neutrophil percent
LY%	Lymphocyte percent
MO%	Monocyte percent
EO%	Eosinophil percent
BA%	Basophil percent

3. RELATED WORK

A major source of error in Clinical Pathology is specimen mislabeling that could lead to wrong diagnosis of the diseases that will affect the patients as well as the hospital. A patient with a disease may be diagnosed as normal (false negative) and one who is normal may be identified positive (false positive) for a disease and the patient may be treated wrongly for that disease. Hence Quality control is used in all laboratories to check errors and it plays a vital role in Clinical Pathology.

Specimen mislabeling can be reduced by collecting and trending the data on mislabeled samples with timely feedback to patient care [7]. Auto verification of results [8] in a laboratory information system is used to verify the correctness of a result.

Various combinations of Data Mining classification algorithms are used on medical data for efficient classification of the data [9]. [10] presents some of the ways of using sequences of clustering algorithms to mine temporal data. Association Rule Mining is used to diagnose diseases [11], [12] and risk patterns [13] from medical data. Taxonomy is used in certain cases to establish associations between

Table 2. Sample Automated Blood Cell Counter Data

Patient ID	Gender	Run1 Date	Age	WBC	RBC	Hgb	Hct	MCV	MCH	MCHC	RDW	Plt	Pct	MPV
1103240058	<Unknown>	01/02/2011		7.81	4.927	17.63	52.32	106.2	35.78	33.69	18.12	172.5	0.134	7.77
1103240063	Female	01/02/2011	9"Years"	1.02	3.101	9.25	26.64	85.92	29.84	34.73	17.9	5	0.004	9.19
		01/02/2011		9.5	1.745	2.74	10.3	59.01	15.72	26.63	28.38	5.3	0.004	8.65
1103240072	Male	01/02/2011	51"Years"	27.76	4.501	10.21	34.49	76.63	22.68	29.59	23.75	60.3	0.06	10.05
		01/02/2011		0.47	2.727	8.46	25.42	93.23	31.04	33.29	19.46	24.4	0.028	11.49
1103240074	Male	01/02/2011	36"Years"	11.81	4.255	13.73	40.43	95.03	32.27	33.96	14.36	189.9	0.171	9.05
1103240075	Male	01/02/2011	36"Years"	2.97	2.598	7.63	22.46	86.44	29.36	33.97	14.83	38.2	0.032	8.48
1103240076	Male	01/02/2011	60"Years"	21.89	3.912	12.22	36.82	94.1	31.25	33.2	14.08	170	0.152	8.94
1103240077	Male	01/02/2011	53"Years"	22.78	4.086	11.49	34.6	84.67	28.11	33.2	16.99	417.2	0.375	8.99
		01/02/2011		27.17	3.577	10.2	29.97	83.79	28.5	34.01	22.7	11.1	0.01	9.14
1103240079	Male	01/02/2011	69"Years"	15.1	2.889	8.73	26.64	92.2	30.2	32.76	22	210.7	0.187	8.9
1103240080	<Unknown>	01/02/2011	72"Years"	11.53	3.612	11.67	34.58	95.75	32.32	33.75	17.54	8.8	0.009	11.26
1103240081	Male	01/02/2011	35"Years"	7.27	3.953	12.71	37.63	95.18	32.16	33.79	17.74	189.9	0.159	8.38
1103240082	Male	01/02/2011	11"Years"	5.06	2.779	10.26	30.36	109.3	36.93	33.8	18.57	55.7	0.063	11.42

different items in a data base [14]. Apriori algorithm is used to find frequent item sets in a database and to generate Association Rules from the frequent item sets [15].

A survey of various Data Mining Tools is presented in [16] and each of the tools is designed to handle a specific type of data and to perform a specific type of task.

Medical data is taken most of the times from medical records [17] and the data is found to be heterogeneous [18] in nature. The privacy issues [18] are to be finalized before handling medical data. The data that is taken from the Blood Cell Counter for our work is De-identified and the patient id and names are changed by the Clinical Pathology department before supplying the medical data for analysis.

The quality of Blood Cell Counter Data is ensured by the application of range checks and delta checks on the various attributes of the data [19]. A range check uses a normal range of values for the attributes to verify the correctness of a blood sample. The abnormal results are tested again using a manual procedure. Alternatively a delta check is used that can find the correctness of the results for both the normal and abnormal results. Delta check uses the previous results of a patient within a quantum of time to verify the correctness of a result. If the current and previous results are abnormal such as the blood counts of a chemotherapy patient, the current result need not be repeated manually. If the previous result of a patient is abnormal and the current one is normal, then the test is repeated manually. Association rules are generated [20] for the Automated Blood Cell Counter Data.

4. RESULTS AND DISCUSSION

The Cell Counter 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 Blood Cell Counter data cannot be replaced by any other value and hence those records were not considered for further processing.

The attributes RBC, WBC, Hg count, MCH, MCHC, MCV, MPV, PCT, RDW and WBC components percentages were required for analyzing the blood cell counter data and hence the records without these fields were removed. The resultant excel file contained the records with patient id, gender, age, date and time of results and the blood count fields were selected for further processing.

4.2 Data Selection

The cleaned blood cell counter data was taken as the data source for data selection process. Associations were to be generated as knowledge from the data and hence the attributes that were desired to be part of the association rules were identified. Patient ID, Run Date, Result Time, RBC, WBC, Hg, MCHC, MCH, MCV, MPV, PCT, RD were selected and a sample of the selected data is shown in table 3.

Table 3. Sample Preprocessed Data

Patient ID	Run1 Date	Gender	Age	WBC
1103240063	01-02-2011	Female	9"Years"	1.02
1103240072	01-02-2011	Male	51"Years"	27.76
1103240074	01-02-2011	Male	36"Years"	11.81
1103240075	01-02-2011	Male	36"Years"	2.97
1103240076	01-02-2011	Male	60"Years"	21.89
1103240077	01-02-2011	Male	53"Years"	22.78
1103240079	01-02-2011	Male	69"Years"	15.1
1103240081	01-02-2011	Male	35"Years"	7.27
1103240082	01-02-2011	Male	11"Years"	5.06

4.3 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 attributes are used to find out whether the value is normal or abnormal. Hence the individual values are replaced with normal and abnormal. A value 1 is stored for the normal values and 0 is stored for the abnormal values. The

flattened data is shown in table 4. Also the excel data is converted into a SQL Data base.

Table 4. Sample Transformed Data

Patient ID	**Run1 Date**	Gender	Age	WBC
1103240063	01-02-2011	Female	9"Years"	0
1103240072	01-02-2011	Male	51"Years"	0
1103240074	01-02-2011	Male	36"Years"	0
1103240075	01-02-2011	Male	36"Years"	0
1103240076	01-02-2011	Male	60"Years"	0
1103240077	01-02-2011	Male	53"Years"	0
1103240079	01-02-2011	Male	69"Years"	0
1103240081	01-02-2011	Male	35"Years"	1
1103240082	01-02-2011	Male	11"Years"	1

4.4 Association Rule Generation

The following table shows the set of association rules generated from the Automated Blood Cell Counter Data.

Table 5. Association Rules Generated

Association Rule	Confidence (%)
Hct ^ RBC → WBC	100
Hct ^ WBC → RBC	100
RBC ^ WBC → Hct	100
RBC → Hct ^ WBC	63
WBC → Hct ^ RBC	85
MCH ^ Hct → Hgb	100
MCH ^ Hgb → Hct	100
MCH → Hct ^ Hgb	100

The number of item sets generated using general Association Rule Mining algorithm and the Apriori algorithm are given in table 6.

Table 6. Candidate and Frequent Item Set Counts

	Number of Item Sets / Association Rules Generated	
	General	**Apriori**
Candidate – 1 Item Set	17	17
Frequent – 1 Item Set	11	11
Candidate – 2 Item Set	136	55
Frequent – 2 Item Set	24	24
Candidate – 3 Item Set	680	35
Frequent – 3 Item Set	2	2
Candidate Associate Rules	12	12
Association Rules	8	8

It is clearly seen that the Apriori algorithm reduces the number of candidate item sets at each level.

4.5 Clustering

The Automated Blood Cell Counter Data is clustered into 4 clusters as shown in figure 1. The attribute RBC is used for clustering.

The records are sorted on the RBC values and the first 4 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 in each of the clusters when the k value is 2, 3 and 4 are given as follows in table 7.

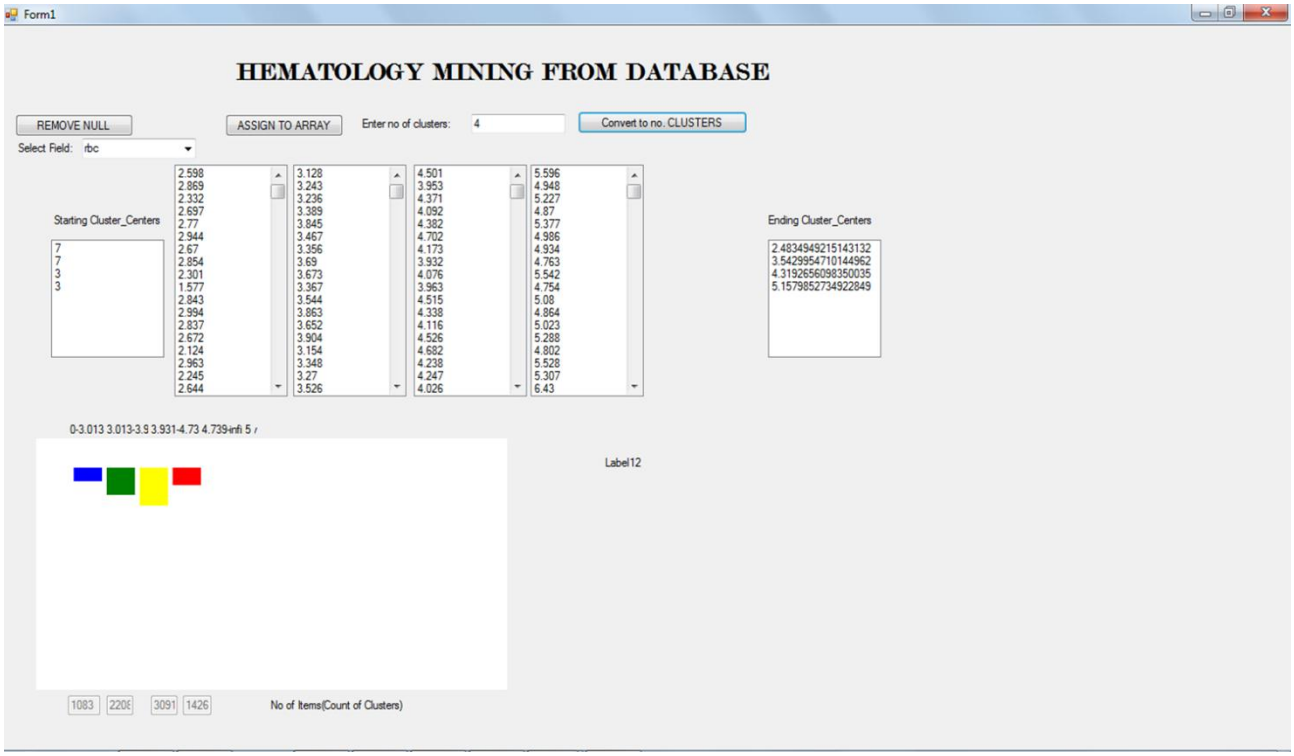

Fig 1: 4 Clusters formed using the RBC attribute of the Automated Blood Cell Counter Data

Table 7. Cluster Counts for Sorted Data

Cluster Number K	Number of elements in Cluster i			
	i = 1	i = 2	i = 3	i = 4
2	2847	4961	-	-
3	1631	3686	2479	-
4	1083	2208	3091	1426

The process is repeated for the unsorted data and the details of the cluster counts are given in table 8.

Table 8. Cluster Counts for Unsorted Data

Cluster Number K	Number of elements in Cluster i			
	i = 1	i = 2	i = 3	i = 4
2	2847	4961	-	-
3	1631	3685	2492	-
4	1077	2189	3106	1436

5. CONCLUSION

A brief study of Hematology, Automated Blood Cell Counter and Blood Cell Counter data is presented in the paper. The format of the blood cell counter 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 Blood Cell Counter Data to convert the raw data into a transformed data that was used for generating more knowledge from the system. Various association rules and various clusters are formed.

A single test result was considered here and the data is to be collected from various laboratories and various Blood Cell Counter machines for integrating the data.

6. 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 functions of the Blood Cell Counter and also for providing the De-identified blood cell counter data.

7. REFERENCES

[1] Jaiwei Han, Michelinne Kamber, Data Mining : Concepts and Techniques, Morgan Kaufmann Publishers, Second Edition, 2006

[2] Margaret H.Dunham, Data Mining: Introductory and Advanced Topics, Pearson Education, 2007.

[3] Automated Blood Cell Counter: www.medscape.com

[4] Dion H.Goh and Rebecca P.Ang. An Introduction to Association rule mining: An application in counseling and help-seeking behavior of adolescents. Behaviour Research Methods, 39(2), 2007, pp. 259-266.

[5] Rakesh Agrawal, T. Imielinski, A. Swami, Mining Associations between Sets of Items in Large Databases, Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data, Washington D.C., May 1993, pp. 207 - 216.

[6] Complete Blood Count (CBC) with Five-Part Differential NHANES 2003–2004, pp. 3 – 4.

[7] Karen Quillen and Kate Murphy, Quality Improvement to Decrease Specimen Mislabeling in Transfusion Medicine, Archives of Pathology and Laboratory Medicine, Vol 130, August 2006, pp. 1196 - 1198.

[8] Dale J. Duca, Auto Verification in a Laboratory Information System, Laboratory Medicine, January 2002, number 1, Volume 33, pp. 21 – 25.

[9] Alp Aslandogan Y. and Gauri A.Mahajani, Evidence Combination in Medical Data Mining, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), Volume 2, 2004, pp. 465 – 469

[10] Rakesh Agrawal, T. Imielinski, A. Swami, Database Mining: A Performance Perspective, IEEE Transactions on Knowledge and Data Engineering, Volume 5 Issue 6, December 1993, pp. 914 – 925.

[11] Massoud Toussi, Jean-Baptiste Lamy, Philippe Le Toumelin, and Alain Venot, 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

[12] Sengul Dogan and Ibrahim Turkoglu. Diagnosing Hyperlipidemia using Association rules, Mathematical and Computational Applications, Association for Scientific Research, Vol.13,No. 3, 2008, pp. 193-202

[13] Jiuyong Li, Ada Wai-chee Fu and Hongxing He Et. Al, Mining risk Patterns in Medical data, KDD'05, Chicago, Illinois, USA, 2005, pp. 770 – 775.

[14] Ramakrishnan Srikant and Rakesh Agrawal, Mining Generalized Association Rules, Proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Swizerland, September 1995

[15] Rakesh Agrawal and Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules, Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, September 1994

[16] Michael Goebel, Le Gruenwald, A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations, ACM SIGKDD, June 1999.

[17] Patricia Cerrito, John C. Cerrito, Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs, Proceedings of SUGI 31, March 26 – 29, 2006 paper 077-31, 2006

[18] Cios KJ, Moore GW, Uniqueness of Medical Data Mining, Artificial Intelligence in Medicine, 2002 Sep-Oct; 26(1-2): 2002, pp. 1- 24.

[19] Minnie D, Srinivasan S, 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, September 2011, pp 338 – 342.

[20] Minnie D, Srinivasan S, 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, April 2012, pp 27 – 32.