Journal of Medical Sciences

ISSN 1682-4474

science alert

ANSInet
an open access publisher
http://ansinet.com

Research Paper

J. Med. Sci., 13 (2): 95-102
15th February, 2013
DOI: 10.3923/jms.2013.95.102

A Novel Approach to Uncover the Patient Blood Related Diseases using Data Mining Techniques

[1]K. Dinakaran and [2]R. Preethi

A medical record in general is a systematic documentation of a single patient's long-term individual medical history and treatment. In medical field, patient records are used for analysing their health problem. Clinical dataset is the essential medical record which deals with patient's health details. In the traditional method, medication can be provided to only one patient at a time and it is difficult to identify group of the people having similar symptoms. Multiple health assessment is time consuming and impractical. The present study proposes a new methodology to find potential information related to blood oriented diseases. Generally, real world Complete Blood Count data are susceptible to noise and not suitable for computation. So, there is a need for data pre-processing. Among the refining techniques, data transformation method such as normalisation and data recoding are applied on relevant attributes of Complete Blood Count. Grouping of people having similar health problems can be done by unsupervised learning. Expectation Maximization Clustering algorithm and k-Means clustering algorithm together clusters effectively the patients based on the attributes. It is shown that refined data produces optimum result and may be useful for medical community to diagnose a group of patients.

Key words: Complete blood count, data mining, data pre-processing, normalization, data recoding, clustering analysis

JMS (ISSN 1682-4474) is an International, peer-reviewed scientific journal that publishes original article in experimental & clinical medicine and related disciplines such as molecular biology, biochemistry, genetics, biophysics, bio-and medical technology. JMS is issued eight times per year on paper and in electronic format.

For further information about this article or if you need reprints, please contact:

K. Dinakaran
Department of Computer Science and Engineering,
R.M.D Engineering College,
Kavaraipettai, Thiruvallur District, India

ANSInet
Asian Network for Scientific Information

[1]Department of Computer Science and Engineering, R.M.D Engineering College, Kavaraipettai, Thiruvallur District, India
[2]Department of Information Technology, R.M.K College of Engineering and Technology, Puduvoyal, Thiruvallur District, India

INTRODUCTION

Medical field plays a major role in day-to-day life. Clinical data set is a collection of patient records. They consists of health assessment, test reports, diagnosis and medication details which is maintained by a medical institution. These data are authentic and sensitive. A fundamental need to accurately maintain those data is to group patients having similar health problems and to help identify diseases. Clinical data set considered for this work is Complete Blood Count (CBC) obtained from an hospital.

Real time clinical databases are highly susceptible to noise and inconsistent data due to typically huge size and their likely origin from different sources and varied causes like faults in data collection instruments, human errors, occurring during data entry.

Inconsistent data will naturally lead to inaccurate results; hence it is not prudent to make further analysis. So, there is an urgent need to pre-process CBC data before taking up further processing.

Pre-processing: Pre-processing is a process of refining the data into appropriate format before carrying out any computations. Pre-processing techniques remove data inconsistency, incompleteness, unreliability. There are available a number of data pre-processing techniques in the literature. Data cleaning can be applied to remove the irrelevent data. (Sembiring and Zain, 2011; Ferreria *et al*., 2011; Al Jarullah, 2011). Data integration combines the data residing in different sources into a single source. In data transformation, the data are changed into appropriate format for analysis (Al Shalabi and Shaaban, 2006). Data reduction can reduce the data size much smaller in volume, yet closely maintain the integrity of original data (Han and Kamber, 2006; Sufi and Khalil, 2011).

Clustering: Clustering is grouping of similar data objects. Data object in the same clusters are similar to one another and dissimilar to other cluster. Clustering is a form of learning by observation. There are many clustering methods like: Partitioning method, Hierarchical method, Density based methods, Grid based methods, Model based methods, High dimensional clustering and constraint based clustering (Popchev and Peneva, 1988; Sun *et al*., 2008; Jiang *et al*., 2004).

Partitioning methods classify the data into k groups and they use iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. k-Means and k-medoids algorithms follow partitioning methods to cluster the objects.

Table 1: Symbols used in this study and their meaning

D	Data set containing n object
k	No. of cluster
E	Sum of squares of all objects in the data set
x	Point in the space representing a object
m_i	Mean of the cluster C_i
$p(c_k/x_i)$	Probability of cluster membership of object x_i, for each cluster
m_k	Model parameter

For grouping the patients having similar kind of diseases, we need fast computation and tight clusters and k-Means algorithm fulfil this needs. Expectation Maximization (EM) Clustering and k-Means algorithm are used jointly to address the present problem (Suresh *et al*., 2009). EM is a statistical clustering to cluster CBC data because it can be used to find the correct number of clusters automatically. The main concept behind EM is fitting the parameters of a distribution model by using training data (Sufi and Khalil, 2011; Zhong *et al*., 2005). k-Means produces k clusters so that, the resulting intra-cluster similarity is high but the inter-cluster similarity is low (Gupta *et al*., 1999). Cluster similarity is measured with respect to the mean value of the objects in a cluster which can be viewed as the cluster's centroid or centre of gravity (Na *et al*., 2010). Table 1 shows symbols used in this paper and their meaning.

DIMENSIONS OF CBC CLINICAL DATA

Complete Blood Count (CBC) is also known as Full Blood Count (FBC)/Full Blood Exam (FBE) or blood panel which gives information about the cells in a patient's blood. The cells that circulate in the blood stream are generally divided into three types: White Blood Cells (Leukocytes), Red Blood Cells (Erythrocytes) and Platelets (Thrombocytes) (Ghai, 2010). Abnormally high or low counts may indicate the presence of many forms of diseases. A major portion of the CBC is the measure of concentration of WBC, Red Blood Cells and Platelets in the blood which is shown in the Table 2. Complete Blood Count includes:

Leukocyte count/total count (TC): Total Count is the number of white blood cells is a volume of blood. Normal range lies between 4,300 and 10,800 cells per cubic millimetre (cmm). Leukocyte Count (LC) is expressed in international units as 4.3 to 10.8×109 cells per litre. Low number of WBC is called Leukopenia. High number of WBC is called leukocytosis.

Differential count (DC): White blood count is comprised of several different types of cells that are differentiated based on their size and shape. These data are expressed in percentage. The cells in DC are: Neutrophil Granulocytes: Normal range 55-60%, Lymphocytes: 32-35%, Monocytes: Very low (less than 1%), Esinophil: less than 4%, Basophil: Very low (less than 1%). Summation of content of all cells should be 100%.

Erthrocyte segmentation rate (ESR): Normal range is 11-15. Ratio of red blood cells to the volume of white blood cells is 45-52% normal for Men and 37-48% for women.

Hemoglobin (Hb): Hb is the amount of Hemoglobin in a specified volume of blood. Haemoglobin is the protein molecule with RBC that carries oxygen and gives blood its red color. Normal range of for haemoglobin is different between the sexes and is approximately appropriately 13-18 g dL^{-1} for men and 12-16 for women (g dL^{-1}). Low Hb leads to polycythemia.

Pack cell volume (PCV): PCV is the average volume of RBC and WBC. This is calculated approximately from values derived from the haemoglobin and red cell count. PCV should be approximately equal to 3*Hemoglobin.

Table 2: Sample complete blood count (CBC) data

	DC								
TC	P	L	E	ESR	Hb	PCV	PLT	RBC	BG
7200	75	21	4	50	6.0	18	85000	3.00	A+
7000	65	31	4	42	8.8	29	472000	3.60	B+
6700	57	28	15	5	10.2	32	95000	3.72	B+
4500	69	27	4	15	11.8	36	217000	4.63	B-
7200	50	43	7	18	11.4	35	227000	4.13	A+
12100	54	34	12	46	9.4	30	156000	4.20	B+
9600	62	34	4	16	9.8	24	240000	3.16	AB+
27200	83	15	2	20	10.5	31	284000	4.30	A+
9600	58	35	6	18	12.0	36	269000	4.36	B+
7600	70	23	7	24	8.5	24	138000	3.72	B-
7600	72	24	4	28	11.4	33	258000	4.19	O+

Platelets (PLT): PLT is the number of platelets in a specified volume of blood. Platelets are not complete cells, but actually fragments of cytoplasm from a cell found in bone marrow called Megakaryocyte. Platelets play a vital role in blood clotting. It is measured in millions per cubic mm. Normal range lies in between 1.5-4 millions. Low number of platelets leads to Thrombocytopenia and High number of platelets leads to Thrombocytosis.

Red blood cells (RBC): RBC count stands for the number of red blood cells in a specified volume of blood. Normal range of RBC is different between sexes and is approximately 4.7-6.1 million cells/cubic mm for men and 4.2-5.4 million cells/cubic mm for women. High number of RBC leads to anemia or erythroblastopenia.

Blood group (BG): Blood type is a classification of blood based on the presence or absence of inherited antigenic substances on the surface of RBC and anti-bodies all in the blood plasma. These antigens may be proteins, carbohydrates, glycoproteins, or glycolipids depending on blood group system. According to ABO blood group system there are four different kinds of blood groups: A, B, AB and O. Based on the presence of D antigen BG is classified into eight types.

ATTRIBUTE REFINEMENTS

Data obtained through complete blood count test may be inconsistent, since it is obtained from different patients. These attributes should undergo pre-processing so that one gets refined data in suitable format convenient for further operations. Figure 1 provides a schematic view of present approach.

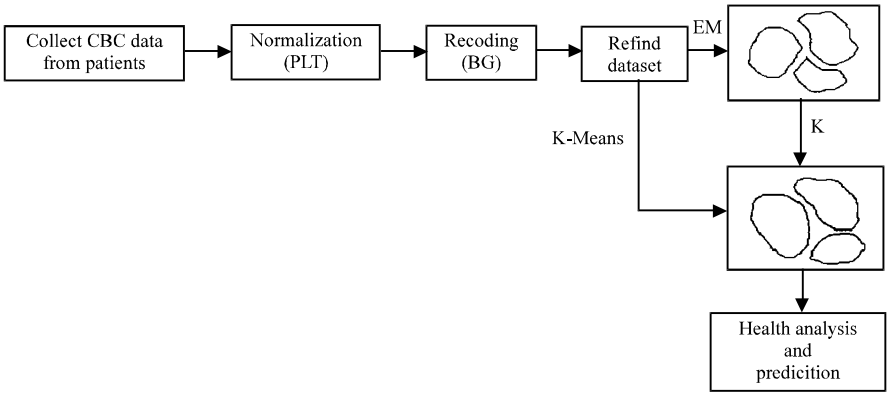

Fig. 1: Schematic diagram of proposed method

FBC data are collected from multiple patients to form a dataset. It consist of various dimensions such as TC, DC (P, L, E), ESR, Hb, PCV, PLT, RBC, BG. PLT attribute contains sparse data. Sparse data does not provide accurate results during clustering which will lead to incorrect diagnosis.

PLT attribute should undergo min-max normalization, a pre-processing transformation technique which scales the data to fall within a small specified range. Thus normalized data will improve efficiency during clustering and help produce optimum result. Numeric attributes result in good performance when compared to character string. In CBC, attribute BG is the combination of ABO blood system and Rh factor. These data are represented in the form of string in test results. Recoding technique converts BG format into numeric format. Recoded data reduces inconsistency in further operations and also helps in grouping patients. This pre-processing technique is applied recursively until CBC data are completely refined. Then k-Means method is used to cluster the refined data. Before clustering , the precise value of k is identified using Expectation Maximization algorithm, a model based method. Based on the cluster result, the health assessment of patients is made and the disease diagnosed. It is expected that the patients in the same cluster will have same kind of health problem. It will help to interpret the relationship among the patients in same cluster.

Data pre-processing: Complete Blood Count (CBC) data are collected from different patients. So, it may not be in suitable format for processing the data. It should undergo some refinement process such as data cleaning, data transformation, data reduction etc. (Ang *et al*., 2010). The primary objective is to identify an appropriate pre-processing technique for efficient clustering.

Normalization: Normalization is a kind of data transformation which may improve the accuracy and efficiency of mining algorithm. If the data to be analysed have been normalized, that is scaled to specific range such as [-1.0, 1.0], [0.0, 1.0]. There are many methods for data normalization including min-max normalization, z-score normalization and normalization by decimal scaling (Han and Kamber, 2006).

Min-max normalization: Accuracy of operational results will be high in min-max normalization when compared with others (Sufi and Khalil, 2011; Al Jarullah, 2011). In the CBC dataset considered for this study, platelets counts lie in the range between thousands and millions. If the range exceeds the limit the data become sparse and this in turn leads to inaccurate results. Hence, min-max normalization method should be applied on the attribute to scale into a small range [0.0, 1.0]. The results produced by this method is shown in the Fig. 2.

Platelets	
85000	269000
472000	138000
95000	258000
217000	50000
227000	479000
156000	189000
240000	98000
284000	270000

Normalization

N.PLT	
0.08	0.51
0.98	0.21
0.1	0.48
0.39	0
0.47	1.00
0.25	0.32
0.44	0.11
0.55	0.51

Fig. 2: Result of pre-processed platelets data

Table 3: Recoding of ABO blood data with suitable numeric values

ABO blood system	Values
A	2
B	3
O	4
AB	5

Table 4: Recoding of Rh factor with suitable numeric values

Rh factor	Values
+	1
-	0

Data recoding: Data recoding is a kind of data transformation which converts character string to numeric format. BG in CBC contains ABO blood system and Rh factor.

If it is not recoded then clustering becomes difficult. Data recoding of BG is a kind of substitution technique. Totally there are eight possible combination of BG (A+, B+, O+, AB+, A-. B-, O-, AB-) which should be converted into numeric format for optimized clustering. Substitution technique for ABO Blood system and Rh factor are shown in the Table 3 and 4, respectively.

For example, A+ is represented as 21 and AB- is recoded as 50. Simulated recoded result is shown in Table 5.

Clustering: Clustering is a process of identifying data objects having similar characteristics. Based on CBC data, we can cluster the patients having same kind of health condition and diseases. Using those results physicians can provide treatment.

Table 5: Pre-processed results of full blood count

	DC								
TC	P	L	E	ESR	Hb	PCV	PLT	RBC	BG
7200	75	21	4	50	6.0	18	0.08	3.1	21
7000	65	31	4	42	8.8	29	0.98	3.6	31
6700	57	28	15	5	10.2	32	0.10	3.72	41
4500	69	27	4	15	11.8	36	0.39	4.63	31
7200	50	43	7	18	11.4	35	0.41	4.13	11
12100	54	34	12	46	9.4	30	0.25	4.2	31
9600	62	34	4	16	9.8	24	0.44	3.16	51
27200	83	15	2	20	10.5	31	0.55	4.3	21
9600	58	36	6	18	12.0	36	0.30	4.36	31
7600	70	23	7	24	8.5	24	0.51	3.70	30
7600	72	24	4	28	11.4	33	0.21	4.19	31
1900	36	30	34	40	8.4	28	0.48	3.84	31
11300	76	18	5	37	9.8	30	1.00	3.93	31
400	68	27	5	20	10.3	31	0.32	4.9	30
7400	60	38	2	11	7.2	26	0.11	3.74	41

k-Means clustering technique has certain shortcomings like requiring predefined the number of clusters, being sensitive to noisy data and outliers and it is also difficult to compare the quality of clusters(Wilkin and Huang, 2007). The first shortcoming is overcome by using EM method which provides optimum number of cluster.

EM clustering algorithm:

- Make initial guess of the parameter vectors which involves randomly selecting k objects to represent the cluster means as well as additional parameter
- The EM algorithm seeks to find the Maximum Likelihood Estimate (MLE) of the marginal likelihood by iteratively applying the following two steps:
- **E step:** Given the current cluster centers, each object X_i is assigned to the cluster C_k with a center that is closest to the object. Here, an object is expected to belong to the closest cluster:

$$P(X_i \in C_j) = P(C_j/X_i) = \frac{P(C_j)\,P(X_i/\overrightarrow{c_k})}{P(X_i)}$$

- **M step:** Given the cluster assignment, for each cluster, the algorithm adjusts the center so that, the sum of the distance from the objects assigned to this cluster and new center is minimized. (i.e.) The similarity (Likelihood ratio) of objects assigned to the clusters is maximized:

$$m_k = \frac{1}{N}\sum_{i=1}^{N}\frac{X_i\,P(X_i \in \overrightarrow{c_k})}{\sum_j P(X_i \in C_j)}$$

EM algorithm is applied on CBC data to cluster the dataset based on the weighted probability distribution of the each dimension of data. The number of clusters obtained from EM is given as input to the k-Means algorithm.

- **K-means clustering algorithm:** k-Means clustering algorithm is used to cluster the CBC data based on certain dimensions. This algorithm requires number of cluster

Algorithm: k-Means partition algorithm

Input: The number of cluster k and a dataset D containing n objects
Output: A set of k clusters which minimizes criterion function E.
Method:

- Select an initial partition with k clusters containing randomly chosen samples and compute the centroid of the clusters (m_i).
- Generate a new partition by assigning each sample to the closest cluster centre.
- Compute new cluster centres as the centroid of the cluster.
- Repeat steps 2 and 3 until an optimum value of the criterion function is found or until the cluster membership stabilizes.

Main objective of k-Means algorithm is to minimize mean square error. The objective function is:

$$E = \sum_{i=1}^{k}\sum_{x \in Ci}\left|x - m_i \rightarrow\right|^2$$

In FBC data, clustering is performed by selecting the dimensions mentioned in the Table 5 In this analysis group the people having normal platelets counts, people suffered by thrombocytopenia and also patients suffered by thrombocytosis. Same method can be applicable for all dimensions. Based on attributes selected for clustering, prediction and analysis of disease differ as explained in section 2.

RESULTS AND DISCUSSION

Normalization and data recoding pre-processing technique for FBC dataset contains 1500 patient records which were collected from hospitals. Normalization was applied to platelet attribute and recoding was done for BG. On that pre-processed data, EM method was applied to identify number of clusters. The pre-processed results are shown in the Table 5.

Efficiency of EM cluster is measured by the log likelihood value. Increasing in log likelihood value increases efficiency and effectiveness of the clusters. While performing EM clustering method in:

- Original data's Log likelihood: -47.1678 and number of clusters identified is 19
- Pre-processed data's Log likelihood: -33.9952 and number of clusters identified is 9

Clustered Instances of complete blood count data are showed in Table 6.

Result obtained shows that pre-processed data produces efficient and optimized result when compared with original raw data.

K-means clustering results analysis depends upon the mean square error between the instance and centroid

of the cluster. Decreasing the sum of square error among the clusters increases the clusters efficiency. While performing k-Means clustering method in:

- Original data Sum of squared error is 318.86 and total number of iteration is 15. b) Pre-processed data Sum of squared error is 146.35 and iteration is 14

Effectiveness of the clusters was improved while performing in pre-processed data had proved by cluster's log likelihood ratio and sum of square error value obtained by EM clustering and k-Means respectively. In EM clustering log likelihood value is more for pre-processed data than original CBC data by 13. Similarly, in k-Means sum of square error was reduced to half with pre-processed data. This outcome shows that pre-processing is mandatory for real time data. Figure 3 shows how the normalized platelet objects are grouped into clusters and the cluster quality.

Table 6: Cluster instances of complete blood count data with percentage of objects in each cluster

Clusters	No of instances in the cluster	Percentage
0	268	18
1	207	14
2	208	14
3	39	2
4	67	4
5	190	13
6	174	12
7	202	13
8	145	10

Main constitutes in blood are RBC, WBC and platelets discussed in section 2. RBC is a container for Hb and PCV is calculated from Hb. Thus RBC is interrelated with Hb and PCV. If the patients have low Hb count they will be anemic then their PCV, RBC must be low. Those group of patients have diseases related RBC count. Similarly, total count and differential count are related with count of WBC and its constituents. Blood clotting agent is platelets. Table 7 shows that the patients having same diseases are grouped into single cluster. Grouping process helps the doctors to provide medication to many patients at a time. For example 100 male patients are grouped as the cluster 0 because of having neutrophilia, lymphocytophenia, eosionophilia, high ESR rate, anemia, low PCV value, erythroblastophenia diseases based on their blood count under the age of 46. Cluster 0 people have deficiency in both RBC and WBC counts but not in platlets. Hence, for these patients medication should be provided to improve RBC and lymphocytes count.

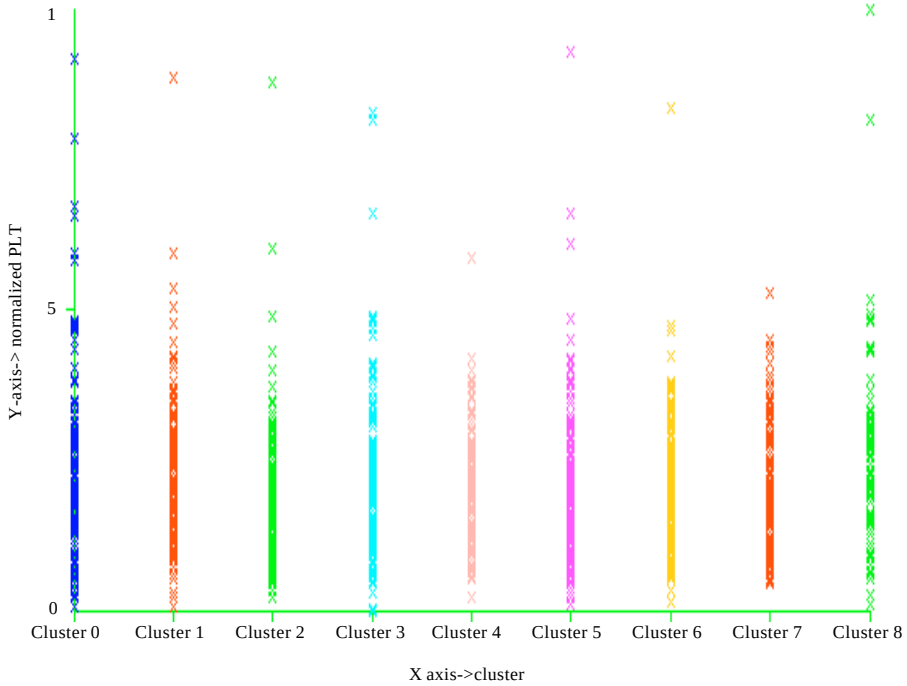

Fig. 3: Graphical results of k-Means clustering based on normalized platelets

Table 7: Report on patients having same problem based on their blood sample

Cluster	0	1	2	3	4	5	6	7	8
Patient count	100	145	251	206	164	182	177	203	72
Age	≤46	≤60	≤39	≤45	≤31	≤59	≤28	≤55	≤47
Gender	M	F	M	F	F	M	M	M	F
TC	TC-normal	TC-normal	TC-normal	TC-normal	TC-normal	TC-normal	TC-normal	TC-normal	TC-normal
DC-P	Neutrophilia	Neutrophilia	Neutrophilia	Neutrophilia	Neutrophilia	Neutrophilia	Neutrophilia	Neutrophilia	Neutrophilia
DC-L	Lumphocytophenia	L-normal	Lumphocytophenia	Lumphocytophenia	Lumphocytophenia	Lumphocytophenia	L-normal	Lumphocytophenia	Lumphocytophenia
DC-E	Eosinophilia	Eosinophilia	E-normal	E-normal	Eosinophilia	E-normal	Eosinophilia	Eosinophilia	E-normal
Hb	Anemia	Anemia	Hb-normal	Anemia	Anemia	Anemia	Anemia	Anemia	Anemia
PCV	Low PCV	Low PCV	PCV-normal	Low PCV	Low PCV	Low PCV	Low PCV	Low PCV	Low PCV
PLT	PLT-normal	PLT-normal	PLT-normal	PLT-normal	PLT-normal	PLT-normal	Thrombocytophenia	Thrombocytosis	PLT-normal
RBC	Erthroblastophenia	Erthroblastophenia	Erthroblastophenia	Erthroblastophenia	Erthroblastophenia	Erthroblastophenia	Erthroblastophenia	Erthroblastophenia	Erthroblastophenia

CONCLUSION

This study has presented techniques refining CBC clinical data before grouping of patients having same health condition and predicting the nature of diseases. Pre-processing on FBC includes normalization on PLT and data recoding on BG. These transformations refine the data into a suitable format for clustering. Pre-Processed data undergoes clustering using EM and k-Means methods. EM method produces weight based cluster result which in turn results in accurate k value. Results of clustering algorithm groups patients according to their health condition and disease they have. It helps the physicians to diagnose groups of people at a time and may be useful for drug analysts to design new drugs for predicted diseases. It also identifies the diseases that may lead possibly to mass death.

REFERENCES

Al Jarullah, A.A., 2011. Decision tree discovery for the diagnosis of type II diabetes. Proceedings of the International Conference on Innovations in Information Technology, April 25-27, 2011, Abu Dhabi, UAE., pp: 303-307.

Al Shalabi, L. and Z. Shaaban, 2006. Normalization as a preprocessing engine for data mining and the approach of preference matrix. Proceedings of the International Conference on Dependability of Computer Systems, May 25-27, 2006, Szklarska Poreba, Poland, pp: 207-214.

Ang, Q., W. Wang, Z. Liu and K. Li, 2010. Explored research on data preprocessing and mining technology for clinical data applications. Proceedings of the 2nd IEEE International Conference on Information Management and Engineering, April 16-18, 2010, Chengdu, China, pp: 327-330.

Ferreria, J., S. Ramos, Z. Vale and J. Soares, 2011. A data-mining-based methodology for transmission expansion planning. IEEE Intell. Syst., 26: 28-37.

Ghai, C.L., 2010. A Textbook of Practical Physiology. 7th Edn., Jaypee Publications, India.

Gupta, S.K., K.S. Rao and V. Bhatnagar, 1999. K-means clustering algorithm for categorical attributes. Proceedings of 1st International Conference on Data Warehousing and Knowledge Discovery, (ICDWDKD'99), Florence, Italy, pp: 203-208.

Han, J. and M. Kamber, 2006. Data Mining Concepts and Techniques. 2nd Edn., Morgan Kaufmann, San Francisco, CA, USA.

Jiang, D., C. Tang and A. Zhang, 2004. Cluster analysis for gene expression data: A survey. IEEE Trans. Knowledge Data Eng., 16: 1370-1386.

Na, S., L. Xumin and G. Yong, 2010. Research on k-means clustering algorithm: An improved k-means clustering algorithm. Proceedings of the 3rd International Symposium on Intelligent Information Technology and Security Informatics, April 2-4, 2010, Jinggangshan, China, pp: 63-67.

Popchev, I. and V. Peneva, 1988. CLUSTER-a package for cluster analysis. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Volume 3, November 4-7, 1988, New Orleans, LA., USA., pp: 1466-1467.

Sembiring, R.W. and J.M. Zain, 2011. The design of pre-processing multidimensional data based on component analysis. Comput. Inform. Sci., 4: 106-115.

Sufi, F. and I. Khalil, 2011. Diagnosis of cardiovascular abnormalities from compressed ECG: A data mining-based approach. IEEE Trans. Inform. Technol. Biomed., 15: 33-39.

Sun, J.G., J. Liu and L.Y. Zhao, 2008. Clustering algorithms research. J. Software, 19: 48-61.

Suresh, R.M., K. Dinakaran and P. Valarmathie, 2009. Model based modified K-means clustering for microarray data. Proceedings of the International Conference on Information Management and Engineering, April 3-5, 2009, Kuala Lumpur, Malaysia, pp: 271-273.

Wilkin, G.A. and X. Huang, 2007. K-means clustering algorithms: Implementation and comparison. Proceedings of the 2nd International Multi-Symposiums on Computer and Computational Sciences, August 13-15, 2007, Iowa City, IA., USA., pp: 133-136.

Zhong, W., G. Altun, R. Harrison, P.C. Tai and Y. Pan, 2005. Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property. IEEE Trans. NanoBiosci., 4: 255-265.