A NOVEL PREDICTIVE DATA MINING TECHNIQUE FOR PREDICTING SLE USING ASSOCIATION RULES AND K-MEANS CLUSTERING (ARMKM)

Ms. A. Malarvizhi

Research Scholar and Assistant Professor, PG and Research Department of Computer Science,
H.H. The Rajah's College (A), Pudukkottai, Tamilnadu- 622001
Email:malarselvamtnau@gmail.com

Dr. S. Ravichandran

Assistant Professor and Head, PG and Research Department of Computer Science,
H.H. The Rajah's College (A), Pudukkottai, Tamilnadu- 622001
Email:rajahsravis@gmail.com

ABSTRACT **- Systemic Lupus Erythematosus (SLE) is an unpredictable disease and stays for a very long period of time in the human body. It has its own wading and waning periods. SLE in a person is confirmed after several. Diagnosis of SLE requires previous history of patients which can generate huge amounts of data. Hence the main objective of this paper is to predictSLE from patient's records and suggesting a non-invasive data mining model. This work uses a combination of data mining techniques namely association rule mining for reducing the complexity in historical disease data and K-Means Clustering for classifying and predicting SLE related information.**

Key words: Data mining techniques, ARMKM, Association Rule Mining, Apriori, K-Means Clustering

1. INTRODUCTION

SLE is a chronic disease which affects multiple organs or is a multi-system disorder. SLE affected organs include, artery, heart, kidney, nervous system, myocarditis, lungs etc. [1].SLE is an auto immune inflammatory and chronic disease producing antibodies on white blood cells (WBC) automatically. SLE occurs frequently in childbearing women and its severity levels vary. It may start with minor ailments and end up with renal failure or impaired central nervous system [2]. It affects females more than males [3].SLE is not a completely controllable disease and the therapies carry risks of weakening human organs during its wading periods. Further, SLE is characterized by changes in measurable test parameters.Previous studies have shown that more than 8% of SLE patients approach doctors after 5 years without knowing the presence of lupus [4]. Thus it is imperative to detect SLE very early mainly to save the patient and secondly for researching new methods of treatments. As SLE can affect multiple organs, each organ affected is treated as a separate disease, thus helping in misrepresentation of the actual cause. Analysis of lupus is conducted on multiple information parameters resulting in huge data sets. The data formed from several sources may have hidden patters of SLE, thus increasing the need to probe these hidden patterns. Data mining methodologies and techniques can help discover these hidden and dormant factors for predicting SLE in patients. K-means is used is mining information as it handles massive data and groups them efficiently and quickly. This technique selects data objects (Centroid) in the first step using random a selection method. The remaining data objects with similar distances are assigned to the selectedclusters. Working on the positives of K-means this work proposes a combination of association rule mining and K-Means Clustering (ARMKM) algorithm to predict SLE easily and efficiently. Thus, this research work proposes a predictive data mining model to predict SLE with clinical effectiveness.

2. RELATED STUDIES

A survey ondata mining techniques in predicting sequential patterns was done on several fields including health care in [5]. The authors defined an association rule algorithm for medical data sets using several data mining algorithms and found Naive Bayes better in performance. The study in [6] tested a parallel approach by using feed forward neural network and back propagation algorithm to diagnose breast cancer. The study found that neural networks could be efficiently implemented in biological data analysis. Heart disease was also classified using neural network in [7], which found Neural networks were fault tolerant and could be divided based on supervised or unsupervised training. Classification And Regression Trees (CART) algorithm was used in [8]to monitor diabeteswith above 95% accuracy. CART was used in pruning, cross validation and testing. Study in [9] exposed the effect of SLE on central nervous systemswith morphological and functional modalities. A review study conducted in Asia showed the prevalence of SLE amongst subjects[10].

3. ISSUES AND CHALLENGES

SLE is an autoimmune disease characterized by the production of about 100 auto antibodies[11]. This is an issue diagnosis and becomes a problem in defining an efficient line of treatment. Older treatment methods often implied a reduced life span for SLE patients, due to organ malfunctions or therapy toxic effects. Though current improvements in Medicare have enhanced SLE patient's survival rates, fear of mortality remains a major issue as patients suspect SLE treatment methods [12].Lupus patient's data in a multi-facility hospital can record disease information of the patient at various times. The issue is that many Hospital Information systems are not. Some hospitals use decision support system with limited facilities, while others use standalone. Even if the patient approaches the same hospital, the hospital needs a decision support system to predict the disease earlier.Further, analysis of lupus from vast amounts of information results in costly processing time which becomes a major problem. Also the data formed from several visits may be misinterpreted as an individual ailment by experts, thus creating an issue in early predictions of SLE.

4. ARMKM (ARM BASED K-MEANS ALGORITHM)

SLE predictions from patient'svisit or disease or lab reports information involves huge sets of data. A combination of data mining techniques are needed for analysis of such huge data sets, since one technique alone may not be sufficient to produce the desired results. The proposed work uses Apriori and K-Means Clusteringalgorithms, where Apriori reduces the complexity of the patient disease information dataset and k-means clusters symptoms data for SLE prediction. Patient's raw data is gathered, cleaned and grouped for SLE prediction in ARMKM. Figure 1 depicts the flow of ARMKM

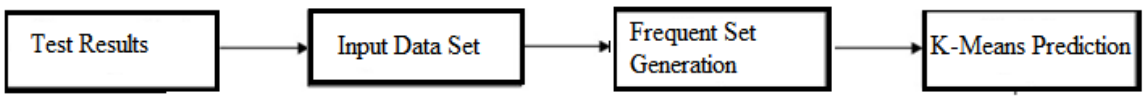

Fig. 1 – ARMKM Flow Chart

Lupus is a difficult disease to diagnose, because its symptoms can be vague and can not be determined or diagnosed with a single test. However, when certain clinical criteria are met, lab tests can help confirm the presence of lupus in diagnosis. On early prediction blood, urine and other tests can also help monitor the disease or show the effects of treatment. Table 1 lists the parameters taken in training by ARMKM.

Table 1. ARMKM Training Parameters

Parameter	Criteria
Antinuclear Antibody (ANA)	Up to 60% for people with lupus
Anti-dsDNA	+ anti-dsDNA test in 40-70% lupus cases
Anti-Ro(SSA)	Anti-Ro 24% to 60% in lupus patients.
Anti-La(SSB)	Presence is serious problem in newborns
C-Reactive Protein (CRP)	+ implies Kidney inflammation
Erythrocyte Sedimentation Rate (ESR)	gauge or monitor disease activity.
Complete Blood Cell Count (CBC)	decrease WBC in 50% of people with lupus.
Chemistry Panel	Abnormalities for lupus in kidney and liver
Urine- Protein/Creatinine Ratio	protein loss when lupus in kidney
Urinalysis	+protein, +RBC, +WBC – lupus in kidney

The prime objective of the proposed ARMKM algorithm is to overcome deficiencies in k-means clustering algorithm, which is sensitive to initial conditions and selection. Selection of different parameters in initial condition may realize different cluster results as it may be trapped in the local optimum. Results for the same medical dataset may lead to different sets of rule and accuracy is dependent on the local optimum becomes an issue. To improve the accuracy, efficiency and consistency in the cluster members, ARM is used to generate frequent sets from SLE test parameters. The proposed method generates a probable set of pre-defined clusters using Association Rule Mining (ARM), which is then passed as input to K-means for predicting SLE. Since, ARM generates frequent sets of data from the test results of patient's, selected test parameters that occur frequently, most recurring items are selected in the 2 item set again derived from 1 item set. This step reduces complexity while increasing the efficiency of K-means. These selected items output overcome the local minimum problem as they select the most recurring test parameters helping in accuracy with reduced prediction time. Though Apriori takes more time in processing two itemsets, the time lapse is covered in K-means as it works on therefined data set.

5. EXPERIMENTAL RESULTS

The execution time describes the amount of time needed for predicting SLE in ARMKM. Table 2 depicts the comparative performances of K-Means, Fuzzy C-Means and ARMKM in predicting SLE based on parameters from Table 1 in the patient data set.

Table 2 – Comparative Performances of Algorithms on Time Consumed

Lupus Criteria	K MEANS (in msec)	FUZZY C MEANS (in msec)	ARMKM (in msec)
Anti-dsDNA	2156	3076	2100
Erythrocyte Sedimentation Rate (ESR)	2054	2194	1999
Complete Blood Cell Count (CBC)	2632	2692	2200
Chemistry Panel	741	791	710
Urine- Protein/Creatinine Ratio	701	891	680

Figure 2 depicts the Performance of ARMKM in SLE Prediction and Table 3 list the clustering accuracy of ARMKM in comparison with other algorithms. It can implied from Table 2 and Figure 2 that the proposed ARMKM model takes less time when compared to other algorithms in SLE prediction.

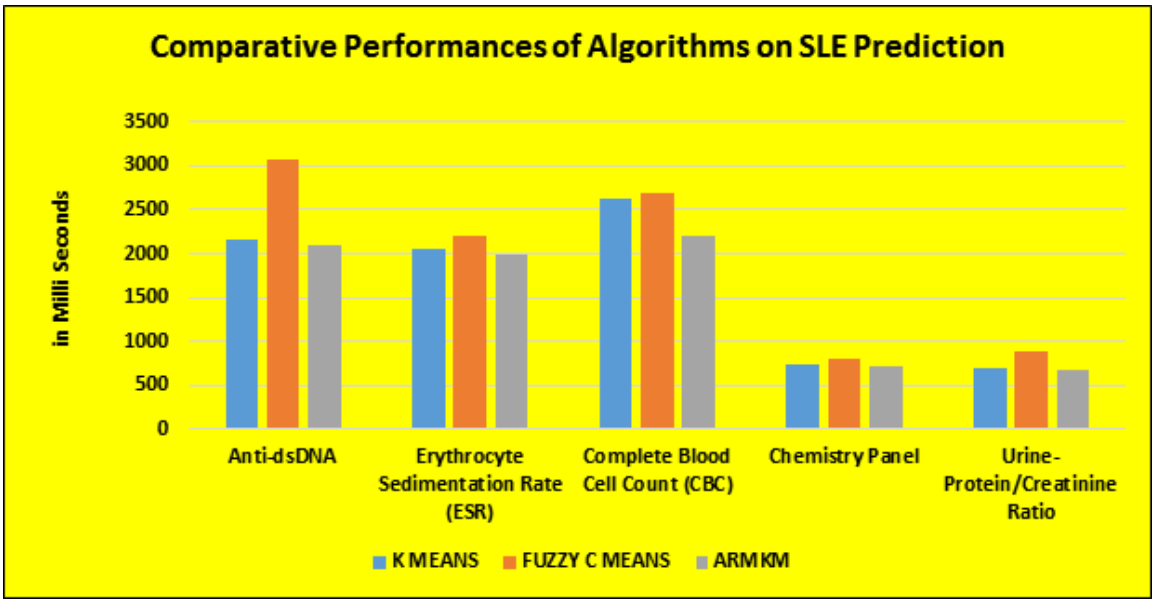

Fig. 2 – Performance of ARMKM in SLE Prediction

Table 3 – Comparative Performances of Algorithms on Clustering Accuracy

Lupus Criteria	K MEANS (%)	FUZZY C MEANS (%)	ARMKM (%)
Anti-dsDNA	74	76	83
Erythrocyte Sedimentation Rate (ESR)	84	81	88
Complete Blood Cell Count (CBC)	79	75	96
Chemistry Panel	87	89	91
Urine- Protein/Creatinine Ratio	87	84	92

Clustering accuracy can be treated as the verification ofaccurate grouping of required elements on the data set for desired outcomes or predictions. It is evident from Table 3 that clustering accuracy of ARMKM is better when compared to other two algorithms.

6. CONCLUSIONS

Clustering techniques are considered as an efficient and popular data mining technique as it groups similar data. Several similarity and dis-similarity measures as per clinician's advice are applied to find the relationships and patterns for predicting SLE using K-Means Clustering Algorithm in ARMKM. This research paper has discussed about the disease prediction using SLE patient's data set. The work has proposed and demonstrated the need for innovative technique to detect chronic SLE from test results. In future this technique can be extended to predictive data mining of other ailments, provided the criteria or parameters in test results are identified by clinicians. Thus this work concludes that ARMKM is a technique which can predict SLE or other complicated diseases in a noninvasive way and from clinical lab test results.

REFERENCES

[1] A.T. Sayad, P.P. Halkarnikar, Risk level prediction system in ischaemic heart disease", Proceedings of IRF International conf. 30th march 2014. ISBN: 978-93-82702-69-6.
[2] Binoy JP, Muhammed F, Kumar N, Razia MV. Clinical profile of systemic lupus erythematosus in North Kerala. J Indian RheumatolAssoc 2003;11:94-7.
[3] V. Manikantan and S. Latha, "Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods", International Journal on Advanced Computer Theory and Engineering (IJACTE), Vol 2, Issue 2, 2013, ISSN (Print) : 2319 – 2526.
[4] Chauraisa, V, and Saurabh Pal. "Early Prediction of Heart Diseases Using Data Mining Techniques." Carib. j. SciTech Vol 1 2013 p 208-217.
[5] Vijiyarani, S, and S Sudha. "Disease prediction in data mining technique–a survey", International Journal of Computer Applications & Information Technology, Vol II, Issue I Jan p 17-21 ISSN 2278-7720.
[6] Rani, K Usha. "Parallel approach for diagnosis of breast cancer using neural network technique." International Journal of Computer Applications Vol 10 issue 3 (2010): p 1-5
[7] Rani, K Usha. "Analysis of heart diseases dataset using neural network approach." International Journal of Data Mining and knowledge management process, Vol 1, No 5, Sep (2011).
[8] Kavitha, K, and RM Sarojamma. "Monitoring of Diabetes with Data Mining via CART Method." International Journal of Emerging Technology and Advanced Engineering Vol 2 Issue 11 (2012): p 157-162
[9] Gal, Y, Gilad Twig, OshryMozes, Gahl Greenberg, Chen Hoffman, Yehuda Shoenfeld, "Central nervous system involvement in systemic lupus erythematosus: an imaging challenge." The Israel Medical Association journal: IMAJ 15.7 (2013): 382-386.
[10] Osio-Salido E, Manapat-Reyes H. Epidemiology of systemic lupus erythematosus in Asia. Lupus 2010;19:1365-73.
[11] Hsieh SC, Yu CL. Autoantibody profiling in systemic lupus erythematosus. CurrBiomark Find 2013;3:55-65.
[12] Ippolito, A and M Petri, 2008, An Update on Mortality in Systemic Lupus Erythematosus, ClinExpRheumatol, 26(5 Suppl 51):S72-9.
[13] ChanchalYadav, Shuliang Wang, ManojJumar, "Algorithm and approaches to handle large Data- A Survey.", International Journal of Computer Science and Network, Vol2, Issue 3, March 2013 http://arxiv.org/pdf/1307.543.