Feature Selection using K-Means Genetic Clustering to Predict Rheumatoid Arthritis Disease

B.Jayanthy, C.Senthamarai

Abstract:*In our Society, Aging society plays serious problems in health and medical care. When compared to other diseases in the real life Rheumatoid Arthritis disease is a common disease, Rheumatoid Arthritis is a disease that causes pain in musculoskeletal system that affect the quality of the people. Rheumatoid Arthritis is onset at middle age, but can affect children and young adults. If the disease is not monitored and treated as early as possible, it can cause serious joint deformities. Cluster analysis is an unsupervised learning technique in data mining for identifying or exploring out the structure of data without known about class label. Many clustering algorithms were proposed to analyze high volume of data, but many of them not evaluate cluster's quality because of inconvenient features presented in the dataset. Feature selection is a prime task in data analysis in case of high dimensional dataset. Optimal subsets of features are enough to cluster the data. In this study, Rheumatoid Arthritis clinical data were analyzed to predict the patient affected with Rheumatoid Arthritis disease. In this study, K-Means clustering algorithm was used to predict the patient affected with Rheumatoid Arthritis Disease. Genetic algorithm is used to filter the feature and at the end of the process it finds optimal clusters for k-Means clustering algorithm. Based on the initial centroid , K-Means algorithm may have the chance of producing empty cluster. K-means does not effectively handle the outliers or noisy data in the dataset. K-means algorithm when combined with Genetic Algorithm shows high performance quality of clustering and fast evolution process when compared with K-Means alone. In this paper, to diagnosis Rheumatoid Arthritis disease we use machine learning algorithm FSKG. A predictive FSKG model is explored that diagnoses rheumatoid arthritis. After completing data analysis and pre-processing operations, Genetic Algorithm and K-Means Clustering Algorithm are integrated to choose correct features among all the features. Experimental Results from this study imply improved accuracy when compared to k-means algorithm for rheumatoid disease prediction.*

Keywords: Clustering , Data mining, Feature selection , Genetic algorithm, K-Means, Machine Learning.

I. INTRODUCTION

Rheumatoid Arthritis is a chronic inflammatory disease that injures joints [9]. It produces pain, swelling, nodules in human body. In India, especially in Tamil Nadu people are not aware of this rheumatoid arthritis disease. Due to this lack of clinical knowledge, people are affected by heart disease, stiffness in body and sudden death that will entirely affect the economic status of the family. Lack of disease identification leads to wrong treatment that cause people survive their life with unexplained pain in their body. There are many factors are used to identify rheumatoid arthritis. In that only some factors are enough to correctly identify rheumatoid arthritis. So feature selection here plays an important role to cluster the rheumatoid arthritis patients from other patients having normal pain. Feature selection is one of the prime tasks in machine learning. Huge objects are characterized using lot of features. Feature selection plays an important role to reduce the dimension of dataset. Limited features are enough to discover pattern from database. Feature selection is act as a cost effective predictors to filter the optimal features. Research in the field of Cloud Computing [1], Text Mining [2], Big Data [3], and Image Mining [4] concentrate on feature selection. Cluster analysis [5-8] partitions the dataset into meaningful groups (clusters) based on similar properties of objects. Many algorithms [5-8] applied on large dataset but it is not in clear such that aim of clustering is not formulated. They reduced the intercluster variance but they are failed to handle noisy data.In this paper we used K-Means Clustering with Genetic Algorithm SFKG to predict rheumatoid arthritis patient in earlier stage using minimal features.

hronic inflammation of the joints is caused by Rheumatoid arthritis (RA) which is an autoimmune disease. Rheumatoid arthritis causes morning stiffness, severe pain in joint, pain in wrist, nodules, and swelling in bones. People affected with Rheumatoid arthritis under 16 years of age is referred to as Juvenile Idiopathic Arthritis(JIA) or Juvenile Rheumatoid Arthritis(JRA).

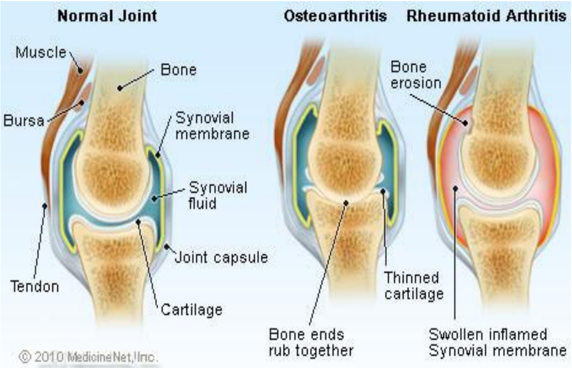


Fig-1 Rheumatoid Arthritis

A. Rheumatoid Arthritis Symptoms and Sign:

- Fatigue
- Loss of energy
- Lack of appetite
- Low-grade fever
- Muscle and joint aches and
- Stiffness

Manuscript published on 30 September 2019
* Correspondence Author

B.Jayanthy*, Research Scholar, Department of Computer Application, Government Arts College, Salem-636007, Tamil Nadu. Mobile: 9442396284, Email: jayanthyresearch@gmail.com

Dr.C.Senthamarai* , Assistant Professor, Department of Computer Application, Government Arts College, Salem-636007, Tamil Nadu

B. Diagnosing Rheumatoid Arthritis:

- Physical Exams
- Blood tests
- Anaemia
- Rheumatoid factor(antibody, blood protein=80% in later stage, 30% in earlier stage)
- Anti-CCP
- Elevated erythrocyte sedimentation- find the amount of inflammation in the joints
- X-Ray

The standard rheumatoid arthritis criteria were established in 1987 shown in Table-I[10].

Table-I: Classification of Rheumatoid Arthritis Criteria Revised-1987

S.No	Criteria	Definition
1	Morning stiffness	Stiffness at joints while wakeup upto 1 hour and present for atleast 6 weeks
2	Swelling at joints	Swelling occurs atleast three or more joints for atleast 6 weeks
3	Swelling at wrist	Swelling occurs at wrist for 6 or more weeks
4	Symmetrical joint swelling	Swelling occurs at same joint areas on both sides of the body
5	X-Ray changes	Erosions or unequivocal body decalcification
6	Rheumatoid nodules	nodules over bony prominences or extensor surfaces that was observed by physician
7	Serum Rheumatoid Factor	Abnormal amounts of serum rheumatoid factor that results positive in less than 5% of normal.

The method described in Table1 is not suitable for detection of rheumatoid arthritis at early stage. Anti CCP, SJC, ESR was used to predict rheumatoid arthritis in their earlier stage.

II. METHOD

A. Clustering:

Clustering is the method of grouping abstract objects into classes of similar objects. Cluster is a collection of objects that belongs to the same class.i.e., same type of objects areput in one cluster and different type of objects are put in another cluster.

B. The K-Means Method:

In this, each of the K clusters is represented by the means of objects (called centroid) within it. This method also called the ***centroid method.*** At each step the centroid point of each cluster is assumed to be known and each of the remaining points are allocated to the cluster whose centroid is closest to it. After first allocation is completed, the centroids of the clusters are recomputed using means and this process is repeated until there is no change in the clusters. The K-means method uses "Euclidean distance measure"

C. Algorithm and Steps in K-Means:

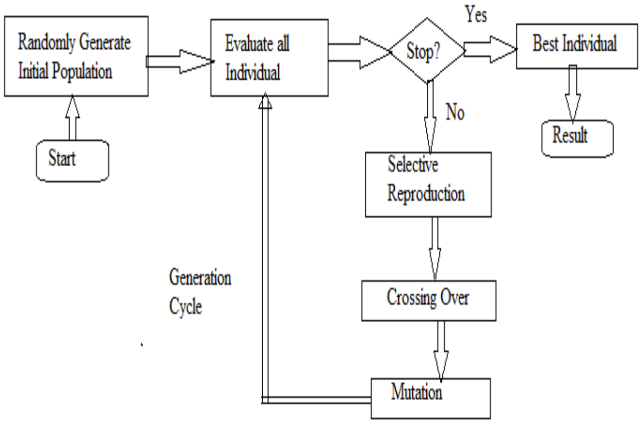
- Decide K , where K is the number of clusters.
- Select seeds for K clusters randomly as centroids of the K clusters.
- Calculate distance of each object from each of the centroids in the dataset.
- Based on the distances computed in step3, allocate each object to the cluster to its nearest.
- Compute the centroids of the clusters, by computing the means of the attribute values of the objects in each cluster.
- If cluster membership is unchanged, go to step7 else go to step3 Users may decide to stop at this stage or to split a cluster or combine two clusters until a stopping condition is met.

D. Genetic Algorithm:

Genetic algorithm is one of the search techniques that work on the principal of natural selection, for develop solution of large optimization problems. Genetic algorithm has the following elements.

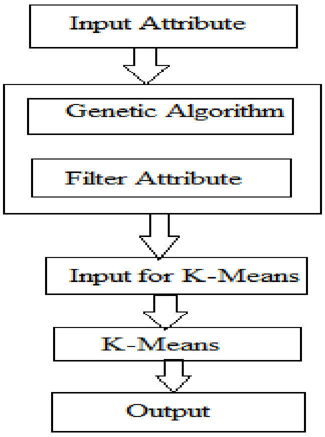
- *Population of chromosomes*
- *Selection based on fitness*
- *Crossover that produce offspring*
- *Random mutation*

This algorithm initialize with population of "individuals". Each individual represents a possible solution to a given problem. Possible solution within population of biological individual is coded in "chromosomes". Chromosome is a sequence of genes is assigned a "fitness" based on fitness function. Individual are selected based on fitness for reproduction by using cross over with other individual in the population. It is used to construct new individual as offspring. This offspring will be fit better than old individual and the generation is complete.



E. Feature Selection using Genetic K-Means

In the proposed algorithm FSKG, the initial seeds for the k-means are selected with the help of genetic algorithm. For predicting rheumatoid arthritis it is important to select proper attributes. All the attributes are not necessary to decide the rheumatoid arthritis disease. So with the help of Genetic Algorithm accurate attributes are selected for predicting the rheumatoid arthritis disease. After that the selected attributes are taken to k-means algorithm.



III. RESULT AND DISCUSSION

Sample of 140 rheumatoid arthritis disease patients data are used in this experiment. 30 features are listed in Table-II to predict the rheumatoid arthritis. Using Genetic algorithm only 4 feature such that ESR, Anti-CCP, RF factor and HLA are extracted that act as an optimal feature to predict rheumatoid arthritis at early stage. Parameters used in FSKG algorithm is listed in Table-III.

Table-III: Features consider for the research

S.No	Feature
1.	Fatigue
2.	Loss of Energy
3.	Lack of Appetite
4.	Low Grade Fever
5.	Muscle Pain
6.	Joint Pain
7.	Joint Redness
8.	Joint swelling
9.	Joint tenderness
10.	Joint warmth
11.	Joint deformity
12.	Rheumatoid nodules
13.	Stiffness
14.	Loss of joint range of motion
15.	Loss of joint function
16.	Limping
17.	Depression
18.	Anaemia
19.	Frustration
20.	Social Withdrawal
21.	Anti-CCP(Anti-cyclic citrullinated peptide)
22.	ANA(Antinuclear antibody)
23.	CRP(C-reactive protein)
24.	ESR(Erythrocyte sedimentation rate)
25.	HLA
26.	Lyme serology
27.	RF factor(Rheumatoid factor)
28.	Uric acid
29.	CBC
30.	SFA

Table-III: Parameters of FSKG algorithm

Parameter	Value
Population size	**140**
Crossover Probability	**0.8**
Mutation Probability	**0.2**
Maximum Number of successive Generation	**15**

During K-means genetic clustering isolated objects are treated as outliers or noises that are removed from dataset. At the end of Genetic Algorithm reduced features are selected. Table-IV displays object Reduction and percentage of feature reduction.

Table-IV: Feature and Object Reduction for Rheumatoid Arthritis dataset.

Algorithm	Dataset	Features Number	Feature Reduction in Percentage	Object Reduction in Precentage
FSKG	140	30	16.6	10.0
K-Means	140	30	0.0	9.0

Figure I show the feature and object reduction for Rheumatoid Arthritis dataset

Figure I: Feature and Object Reduction for Rheumatoid Arthritis dataset.

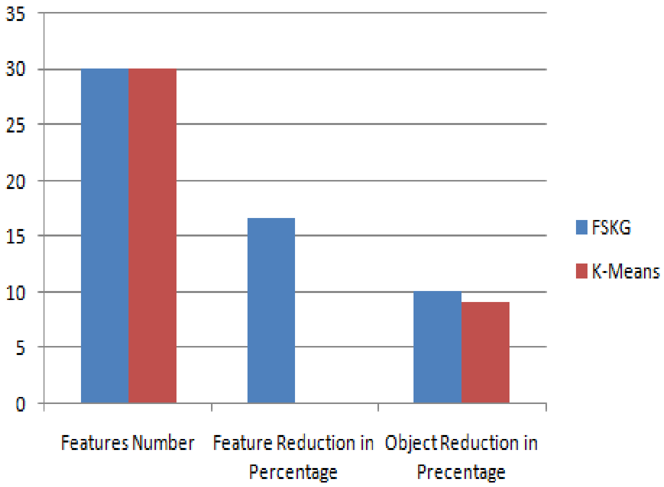


Table-V shows the accuracy comparisons among different combination and number of features in SFKG algorithm

Table-V: Accuracy comparison

Dataset	Features Number	Patient **having Rheumatoid Arthritis in percentage**	Patient not **having Rheumatoid Arthritis in percentage**	Accuracy
140	30	65	35	75.4
140	16	77	23	86.3
140	4	80	20	87.9

Figure-II shows the Accuracy comparison of FSKG with different number of features

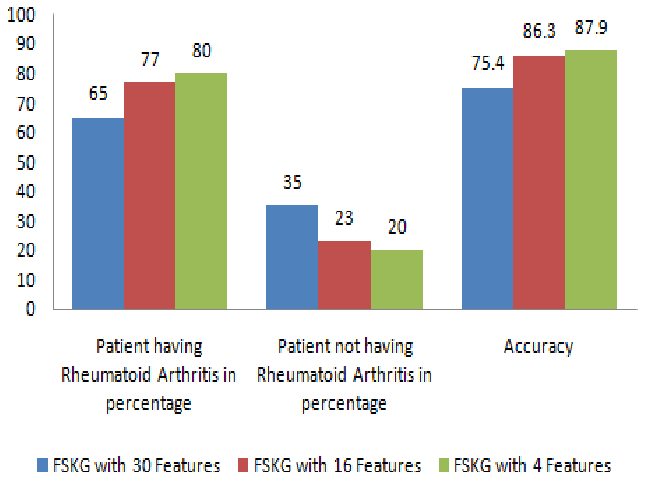


Figure-II: Accuracy comparison

IV. CONCLUSION

In this paper we presented new algorithm FSKG to improve accuracy and predict patients having rheumatoid arthritis at early stage. The algorithm FSKG shows best results than K-Means in case of handling noise. FSKG algorithm showed that rheumatoid arthritis can be predicted only with the help of 4 features. If doctors can early predict the rheumatoid arthritis based on ESR, Anti-CCP, HLA, Rh factor test it will save the life of many people and also quality of life is increased

REFERENCES:

1. Kaur, Pankaj Deep and Chana, Inderveer, "Cloud based intelligent system for delivering health care as a service", Computer methods and programs in biomedicine, Elsevier, 2014, 113, 1, 346-359
2. Jing, Li-Ping and Huang, Hou-Kuan and Shi, Hong-Bo, "Improved feature selection approach TFIDF in text mining. Book Machine Learning and Cybernetics", 2002. Proceedings. 2002 International Conference on, 2, 944-946, 2002, IEEE
3. Wu, Xindong and Zhu, Xingquan and Wu, Gong-Qing and Ding, Wei," Data mining with big data", IEEE transactions on knowledge and data engineering, 26, 1, 97-107, IEEE, 2014
4. Hong, Richang and Wang, Meng and Gao, Yue and Tao, Dacheng and Li, Xuelong and Wu, Xindong," Image annotation by multiple-instance learning with discriminative feature mapping and selection", IEEE transactions on cybernetics, 44, 5, 669-680, IEEE, 2014
5. Ziad Obermeyer, Ezekiel J.Emanuel, "Predicting the Future- Big Data, Machine Learning, and Clininal Medicine", The New England Journal of Medicine, Vol.375, No.13(2016), 1216-1219
6. Yong Gyu Jung, Min Soo Kang, Jun Heo, "Clustering performance comparison using K-means and expectation maximization algorithms", Biotechnology & Biotechnological Equipment, Vol. 28, No. S1(2014), 45-48
7. Min-Soo Kang, Yong-Gyu Jung, Du-Hwan Jang, "A study on the search of Optimal Aquaculture farm condition based on Machine Learning", The journal of The Institute of Internet, Broadcasting and Communication(IIBC) Vol. 17, No.2 (2017), 135-140.
8. Jae-Gyun Park, Eun-Soo Choi, Min-Soo Kang, Yong-Gyu Jung, "Dropout Genetic Algorithm Analysis for Deep Learning Generalization Error Minimization", International Journal of Advanced Culture Technology Vol.5 No.2 (2017), 74-81.
9. Chen Lin, Elizabeth W. Karlson, "Automatic Prediction of Rheumatoid Arthritis Disease Acitivity from the Electronic Medical Records", PLOS, Vol.8, Issue.8(2013) 1-10.
10. www.rheumatology.org

AUTHORS PROFILE:

Mrs.B.Jayanthy, at present working as an Assistant Professor in the Department of Computer Science, Jairam Arts & Science College, Chinnathirupathi, Salem. She did Bachelors of Computer Science from Government Arts College for Women, Salem-8(Periyar University) on 2004. She did Master of Computer Science and M.Phil Computer Science from Periyar University. She did Bachelor of Education(B.Ed) from TamilNadu Teachers Education University. She published papers in National journals. She has 10.5 years of experience in teaching field. Her area of speculation are in Data mining, Machine Learning, Java. She guided M.Sc,MCA, M.Phil students for doing projects and research. She developed websites for Government colleges also.

Dr.C.Senthamarai, at present working as an Assistant Professor in the Department of Computer Application, Government Arts College(Autonomous), Salem-7. Before that she worked as an Assistant Professor in the Department of MCA, K.S.Rangasamy College of Technology, Tiruchengode. She did Bachelors of Physics and Master of Computer Application. She completed Doctor of Philosophy in Computer Science. She has more than 25 years of experience in teaching field and research field. She published many papers in International journals and National journals. Her research interest includes Grid computing, Cloud computing, Fuzzy Logic, Genetic Algorithm, Machine Learning, Computer Network and Data mining. She guides many research scholars to.