An Optimized Algorithm based on ACO Integrated Association Rule Mining and Text Mining for Detecting Erythemato-Squamous Diseases

Anjali Saini
[1] Department of Computer Science and Engineering, Singhania University, Rajasthan,India

Abstract: In this research paper, a new method based on Association Rules and Ant Colony Optimization(ACO) is presented for the detection and classification of skin disease named as erythemato-squamous diseases.The presented work is the intelligent prediction approach based on the association rule mining under the weighted attribute analysis.In this work ,the ACO approach is defined to represent the dataset in form of network and each node is represented by an attribute where the weightage to these nodes is defined and based on the ANT path generated by covering the high weight attributes and values.The proposed system performance is compared with some other standard algorithms and results are obtained in terms of total rule generation and average estimated path length.The correct accuracy rate of the proposed system is 98.7%.This model can be used for obtaining different decision regarding disease detection and classification.

Keywords: Association Rule Mining, Text Mining, ACO, Erythemato-Squamous diseases.

I. INTRODUCTION

Data Mining has also been termed as data cleaning , data archeology, information finding or information harvesting depending upon the area where it is being used.Association rule mining and Classification Rule discovery are the two important techniques in data mining.Classification Rule Mining discovers a set of rules for predicting the class of unseen data.The aim of classification is to build a model of training data that can correctly predict the class of unseen or test objects in which input is a set of objects along with their classes.

Ant Colony Optimization (ACO), which was introduced in the early 1990s as a novel technique to solve hard combinatorial optimization problems[8]. Ant Communication is accomplished primarily through chemicals called pheromones. Ants exchange information with one another by laying down pheromones along their trails. Other ants notice or see the presence of pheromone and tend to follow paths where pheromone concentration is higher. Over time, on the other hand, the pheromone trail starts to evaporate, thus lowering its attractive strength. The greater time it takes for an ant to travel down the path and back again, the greater or more time the pheromones have to evaporate. A short path gets move over faster, and in this manner the pheromone density remains high as it is laid on the path as fast as it can evaporate. ACO is basically the optimization approach that is basically used to speed up the algorithmic process. In wireless network the ACO is basically used to optimize the communication process. According to this approach a node generate the ant to find the optimized path over the network. These ants place the pheromones on this located path so that all other nodes can follow these pheromones to communicate on this optimized path.

Text mining is all about extracting patterns and associations previously unknown from large text databases. The sequence mining technique is used to find a set of items across time or position in a given database. Association Rule Mining is another important method or approach in the data mining. An association rule can be expressed in the form X=>Y, where X and Y are disjoint sets of items. In a dataset D, existing of data instances, the rule X=>Y has support **s**, equal to the percentage of instance of D that contain both X and Y. Support count **supc** is the number of instances of D that contain both X and Y. The confidence **c** of the rule is the percentage of instances in D that contain Y among those that contain X. The need for finding association and patterns arose when the market felt the need to learn about customer behavior of purchasing, so as to enhance the sale, enhance business, attract customer, and to maximize the profit. So, in

1993, R. Agrawal et.al felt the need to design a new algorithm which was Apriori Algorithm, for discovering associations among the dataset. [1] Apriori algorithm is one of the classic algorithm designed to operate on databases containing transactions. It is one of the oldest algorithm in association rule mining defined by Aggarwal et al. In 1993. The main idea of this approach is to find a useful pattern in various sets of data .Association Rules are used for medical dataset is more critical application area. The work is defined on this area in this research. The consideration is given to the skin disease prediction. In such kind of datasets, the accuracy is one of the major criteria and to handle the multiple attributes simultaneously is a challenge. The next focus is on improvement of association mining algorithm and the using the new algorithm dataset of medical datasets. The problem with Apriori Algorithm is that it generates a huge number of candidate set and whenever a candidate is generated, there is a need to check whether it is frequent or not,which means higher requirement of memory space, time utilization and in short resource utilization.

So, a new algorithm is generated to improve the shortcomings of apriori algorithm. In this work, the ACO approach is defined to represent the dataset in the form of a network and each node is represented by an attribute. The weightage to these nodes is defined and based on this the ANT path is generated by covering the high weight attributes and values. As the ANT generates the path till the class attribute, the final result is achieved. If some blockage found, it is considered as the false recognition of diseases for particular record. The process is defined for all records and based on it the initial data limits are generated. These limits are then associated to generate the effective outcome based on support and confidence values. These values provides the satisfiability of results.The work is defined for skin dataset. The association rules are here integrated with ACO approach to perform the optimization.

II. RELATED WORK

Adepele Olukunle in 2002 performed a work on fast association rule mining algorithm which is suitable for medical image data sets. Author provide a flavour of Presented implementation environment. Author also give an instance, how presented proposed algorithm work to assess its suitability[10].

A hybrid method that combines Particle Swarm Optimization and Ant Colony Optimization(PSO/ACO) was introduced in (Holden and Freitas,2008) for mining rules that can be used for classification.The disadvantage of using PSO algorithm was that nominal values had to be converted into binary numbers before mining. The hybrid algorithm suggested eliminates the need for this preprocessing phase. The algorithm was compared to PART, which is an industry approved algorithm. Authors also compared performance of module handling only continuous data to another new classification algorithm based on differential evolution. Results indicate that this hybrid algorithm provides very good accuracy measure and outputs simpler (smaller) rule sets, thus achieving the aim of correctness and understandability.

Sunita Soni (2009) performed a work and propose a new framework (associative classifier) that uses weighted association rule mining (WARM) [5]. In any prediction model all attributes do not have same importance in predicting the class label. So different weights can be selected to different attributes according to their predicting capability. Author proposed a theoretical model to introduce new associative classifier that takes advantage of weighted association rule mining. The model can be used in any domain to enhance the prediction accuracy.

K. Zuhtuogullari (2011) performed a work," An Improved Itemset Generation method for Mining Medical Databases"[2]. In this study an extensible and improved itemset generation approach has been constructed and implemented for mining the relationships of the symptoms and disorders in the medical databases. The algorithm of the developed software discovers the frequent illnesses and generates association rules using Apriori algorithm.

In 2011, Pooia Lalbakhsh performed a work," Focusing on Rule Quality and Pheromone Evaporation to Improve ACO Rule Mining"[3]. In this paper an revised form of Ant-Miner algorithm is introduced and compared to the previously proposed ant-based rule mining algorithms. Presented algorithm modifies the rule pruning process and introduces a dynamic pheromone evaporation approach. The algorithm was executed on five standard datasets and the average accuracy rate and numbers of discovered rules were analyzed as two important performance metrics of rule mining.

K.Rameshkumar in 2013 proposed the n-cross validation technique to decrease association rules which are irrelevant to the transaction set. The expected approach used partition based methods which are supported to association rule validation[7].

III. PROPOSED APPROACH

In this present work an effective ACO integrated association mining approach is suggested for skin disease prediction. The presented work is the intelligent prediction approach based on the association rule mining under the weighted attribute analysis. In this work, the data is taken from external sources that contains the data records with the specification of disease class. The work is defined under certain limits so that effective detection of disease will be done. The presented work is divided in three main stages. At the first stage, each data record is capture for all the relative attributes. These attributes are defined in the form of a connected network architecture. This architecture is then defined with some ants that are distributed over the attribute network. These ants identify the relation between the attribute and value analysis so that the weightage to each attribute and relative value will be assigned.Here the distance between the nodes of connected network architecture is given by this formula-

$$d(i,j)= sqrt ((x(i)-x(j))^2 + (y(i)-y(j))^2) \qquad (1)$$

where d is the distance and x(i),y(i) and x(j),y(j) are the co-ordinates of the nodes

Once the weightage is assigned, the next stage is to process these ants over the attribute network. The attribute network is analyzed to generate the effective path till the class attribute not arrived. If the class attribute is arrived, the association path is terminated and considered as the effective prediction path for the arrived, the association path is terminated and considered as the effective prediction path for the weighted is here generated based on associativity of the attribute and the relative value with other. Once the prediction paths are identified, the final stage is to identify the valid paths over this set. For this, the pruning stage is defined over the network. This pruning stage is defined based on the support and confidence value analysis. The path with higher confidence value are considered as the adaptive acceptable values and the values that are not below confidence value are neglected and removed from the rule. After this pruning stage, the true identification of rules is done that are predicted under the confidence rule. Here, the results are obtained in terms of total rules generation and average prediction path length,where accuracy is calculated using this equation:

$$Accuracy = Total\ sum / total\ rules*100$$

A. Data Collection

The first work of the approach is to collect the data to implement the whole concept. We are very clear about the database selection. In this work, medical dataset will be used. The data will be able to present the disease analysis in a patient under different classes. This kind of data can be collected from the UCI respiratory easily. In this work, the dataset of skin disease is taken. The description of this dataset is given below

TABLE 1. DATASET DESCRIPTION

Attribute	Description and Frequency
File Name	Dermatology.arff
Number of attributes	34
Number of Instances	366
Number of Classes	6
Clinical Attributes	12
Histopathological Attributes	22
Class Attribute	1

B. Proposed Algorithm

1) Rule Generation Algorithm : The main task associated in this work is the generation of rule set over the available dataset. To generate this rule set the ACO approach is defined which is defined under-

```
Rule Generation Algorithm(DataSet, N)
/* here Dataset is the actual skin disease dataset of size N*/
    1. Set the ClassAttribute ClassAtt(:)=Dataset(N,:)
    2. Set the Parameters for Ant Network
        a. NumberOfAnts
        b. EvaporationVectors
        c. AntPositions
    3. Initialize the Ant Netowrk Under Defined Set Parameters
    4. Generate all the Ant Inspired Attribute Graph under Parameteric Specification.
    5. Define the Weightage to the Attribute node under the association rule specification.
    6. Obtain the Weighted Ant Network from the Attribute set.
    7. For i=1 to MAX_ITERATIONS
       /* Perform the Ant based rule generation for some specific number of iterations*/
```

8. Perform the ant network analysis under the weighted and associated attribute analysis.
9. Generate the Ant Effective Path over the weighted attribute Matrix based on ant parameters and Identify the cost of generated Path
10. Obtain the Attribute set over the Weighted Path and represent it as the Generated Rule(GR) and then return the Generated Ruleset

2) Rule Filteration using pruning method : Another associated task with this approach is the filteration of generated ruleset by performing the global pruning method. According to this ruleset, the identification of valid rule is defined. This stage is called pruning mechanism. The pruning is here defined under the cost analysis. In this method, the rules with lesser cost value are pruning and the new values are obtained.

```
RulePruning(Rules,N)

   /*Here the algorithm will get N Number of rules
     as input for pruning process*/
          Define the Support and Confidence Threshold
          Limits for Pruning Process
       1. For i=1 to N
          /*Process All rules *.
       2. Rule=Rules(i);
       3. L=Length(Rule(i))
       4. /* Obtain the Length of Each Rule*/
       5. For J=1 to L and For K=1 to L
       6. If (Rule(J), Rules(I,K))
       7. Increment Count By 1
       8. If (Count>Support_threshold value)
       9. If(Count>Confidence_Threshold value)
      10. Include Rule as Valid Rule in RuleList
      11. Else
      12. Remove Rule from Rules and return rules;
      13. /*Obtain the Prune Rule List*/
```

IV. RESULTS AND ANALYSIS

In this section we used three criteria to compare and analyze performance of rule mining algorithms under consideration. The first criterion is predictive accuracy which is defined in terms of accuracy rate.The number of rules in a rule set is second criterion and the number of attribute value combinations or conditions per rule is the final criterion.The experiment is performed on erythemato-squamous disease dataset.

The results of the proposed approach are compared with existing approaches. The results obtained from the work are tested on same dataset that is used by the earlier researchers.Here dermatology dataset was taken from UCI (University of California, Irvine) machine learning repository. There are 366 records in this database .The proposed algorithm has been implemented using MATLAB and the experiment is performed for 150 iterations.

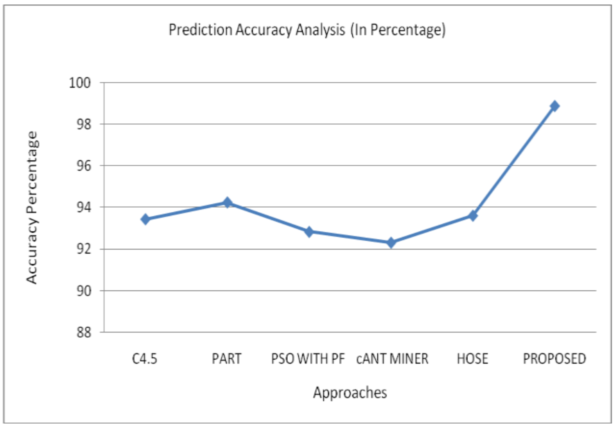

Figure 4.1 : Prediction Analysis Under Different Approaches

Here the analysis of proposed work is defined under different approaches. As shown in the figure, the result of proposed approach is better than all the existing approach. The result is here obtained for 85% confidence value. The quantitative results of these values is shown in table 4.1

TABLE 4.1:ACCURACY PREDICTION ANALYSIS UNDER DIFFERENT ALGORITHMS

Different Algorithms	Prediction Accuracy
C4.5	93.44
PART	94.25
PSO with PF	92.83
cANT MINER	92.33
PROPOSED	98.5

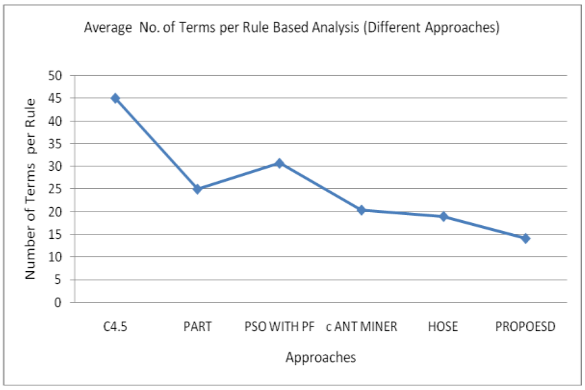

Figure 4.2 : Average Number of terms Under Different Approaches

Here the analysis of proposed work is defined under different approaches based on the number of terms. As shown in the figure, the result of proposed approach is better than all the existing approach. The result is here obtained for 85% confidence value. The results of these values is shown in table 4.2

TABLE 4.2 : AVERAGE NUMBER OF TERM ANALYSIS UNDER DIFFERENT ALGORITHMS

Different Algorithms	Number of Terms in Rule
C4.5	44
PART	26
PSO with PF	30.6
cANT MINER	20.3
PROPOSED	14.16

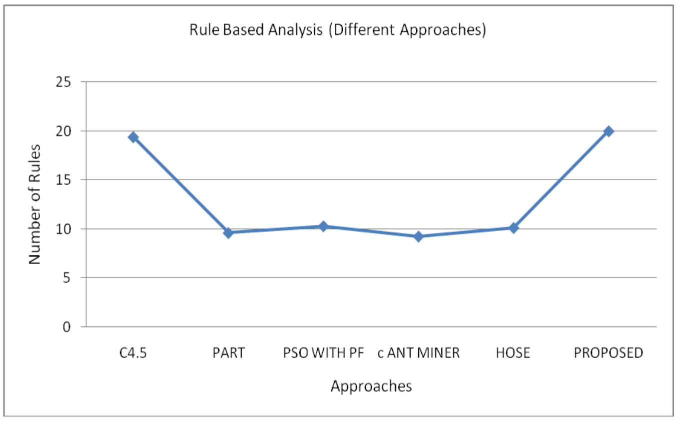

Figure 4.3 : Average Number of Rules Under Different Approaches

Here the analysis of proposed work is defined under different approaches based on the number of rules . The quantitative results of these values is shown below in table 4.3

Different Algorithms	Number of Rules
C4.5	19.7
PART	9.8
PSO with PF	10.28
cANT MINER	9.26
HOSE	10.13
PROPOSED	21

TABLE 4.3 : AVERAGE NUMBER OF RULES UNDER DIFFERENT ALGORITHMS

V. CONCLUSION AND FUTURE PROSPECTS

Discovering Association rules(AR) in the dataset is the important class of data mining and there is a large need of finding association rules in the current situations or scenarios. The association mining with text mining is helpful in different application area. One of such specific area is disease prediction that is based on medical data analysis. The presented work is defined in same field. Here the work is carried out on skin disease dataset. The results shows that the presented work has performed the accurate and predictive detection of disease. The result set shows the presented approach is far better than existing earlier approaches. In future scope, many other datasets can also be considered for the same work mentioned here and comparative analysis and prediction can be performed under some more parameters.

REFERENCES

[1] Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules (1994)

[2] K.Zuhtuogullari,"An Improved Itemset Generation Approach for Mining Medical Databases", 978-1-61284-922-5/11,2011 IEEE

[3] Pooia Lalbakhsh," Focusing on Rule Quality and Pheromone Evaporation to Improve ACO Rule Mining", 978-1-61284-691-0/11,2011 IEEE

[4] Jiuyong Li," Mining Causal Association Rules", 978-0-7695-5109-8/13, 2013 IEEE

[5] Sunita Soni," An Associative Classifier Using Weighted Association Rule", 978-1-4244-5612-3/09,2009 IEEE

[6] Adel Ardalan, Prof.Rahgozar,‖Ant Miner:Ant Colony-Based Association Rule Miner‖,spring 2006

[7] K.Rameshkumar,"Relevant Association Rule Mining from Medical Dataset Using Irrelevant Rule Elimination Technique"

[8] Marco Dorigo and Thomas Stutzle, ―Ant Colony Optimization , Edition 2004 and ISBN-81-203-2684-9.

[9] Kapil Bakshi, "Considerations for Big Data: Architecture and Approach", IEEE, 2012

[10] Carlos Ordonez," Mining Constrained Association Rules to Predict Heart Disease", 0-7695-1119-8,2011 IEEE