IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Generating Recommendations using an Association Rule Mining and Genetic Algorithm Combination

L.M.R.J Lobo[1], R. S. Bichkar[2]

Research Scholar, Department of C.S.E, SGGS Institute of Engg & Technology, Nanded, India[1]

Associate Professor, Department of C.S.E, Walchand Institute of Technology, Solapur, India[1]

Professor, Department of E & TC, G. H. Raisoni College of Engineering and Management, Pune, India[2]

Abstract: A Recommender system is a subclass that seeks to find out the rating or preference that a user would give to an object of interest. The set of available recommender systems generate a recommendation based on historical information available. The information is based on a user`s taste, but not intend. Genetic recommendation offer real-time recommendations in a specific order, thereby overcoming drawbacks of existing systems. In this paper, an Association Rule Mining technique combining features of Eclat algorithm and Genetic Algorithm is proposed. The idea is to apply association rule mining technique Eclat for generating rules and further use genetic algorithm to optimize these rules. A performance comparison is done between results achieved by another popular Association Rule Algorithm, The Apriori algorithm and the results of Eclat-Genetic algorithm. It is observed experimentally that the Eclat-Genetic model gives 28.31 % better result than the existing Apriori algorithm in terms of accuracy.

Keywords: Recommender, agriculture, Eclat-Genetic, Association Rule Mining, rules.

I. INTRODUCTION

The presently existing recommender systems generate recommendation based on user's information collected in the past. This information reflects a user`s taste, but it doesn`t include intend at a particular time. So, existing systems sometimes do not generate suitable recommendation. General recommender systems work on information collected using explicit and implicit methods [1], so do the existing systems. Genetic recommendation offer real-time recommendations using a fitness function that estimates the suitability of recommended lists.

An association rule mining technique, which combines features of Eclat and Genetic algorithm, is implemented for the analysis of an agricultural dataset to generate recommendations required by farmers based on this dataset for the future.

The idea is to apply an Association Rule Mining technique, Eclat for generating rules and further to use Genetic Algorithms to optimize these rules and establishing a relationship between them. Results show that the Eclat-Genetic Algorithm model produces 28 % more accurate results as compared to the popular Apriori algorithm**.**

A. RECOMMENDER SYSTEMS

Recommender systems emerged and became increasingly popular throughout the past decade in its application of E-Commerce. They use implicit and explicit data collections, and various algorithms to yield recommendations. A recommender system predicts a particular user’s likelihood to give a certain item high rating according to the characteristics of the item or what other people with similar taste think about the item.

Correspondingly, the systems use content-based approach and collaborative filtering approach.

- **Content-based recommender systems (CBRS)**

A content-based recommender system assumes each user to operate independently, and each item is represented by features or characteristics. Each user’s profile is created or updated according to each feedback on the desirability of the items that were listed in a list of recommendations.

- **Collaborative filtering recommender systems (CFRS)**

CFRS works perfectly with non describable items which CBRS has difficulties in recommending.

DOI10.17148/IJARCCE.2017.6831

CFRSs make Similarity decisions, and the cosine angle computation or Pearson\’s correlation is mainly used in clustering users and items.

- **Hybrid recommender systems**

In some new recommender systems, for compensating collaborative filtering shortcomings, content-based filtering is simultaneously involved to increase system accuracy on user preference. And such a system that utilized both techniques is called a hybrid recommender system. A hybrid recommender system solves the problem with extreme cases coverage that a simple CFRS can’t handle.

B. DATA INPUTS AND OUTPUTS IN RECOMMENDER SYSTEMS

The input in a recommender system depends hugely on the filtering algorithm deployed.

The inputs for content-based recommender systems fall into one of these categories

- Ratings reflect the opinion of users on the items, and most commonly ratings are either in the form of a place on a scale or a binary value, that means 1 or 0, yes or no.
- Demographic data refer to information of the users such as gender, education, address, a list of personal preferences that require user input, and these can be hard to obtain unless certain incentives were given.
- Content data is the textual analysis of the document that contains the item’s physical dimensions, composing components, functionalities, etc.
- A recommender System should also be able to provide user with useful information about the items that might interest them.

C. RECOMMENDER SYSTEMS FOR AGRICULTURE

With the evolutions in computer based data storage systems, the amount of stored data is tremendously increasing in every field. Each and every sector in this digital world is undergoing a dramatic change due to the influence of Information Technology being used in it. The agricultural sector needs more support for its development in developing countries like India which rely to a great extent on agricultural produce.

Huge data in the agricultural context is available. However, this huge amount of data is totally wastage of storage unless we know what to do with this huge amount of data. Nowadays, recent technologies are able to provide a lot of information on agricultural-related activities, on which we are able to analyse agricultural data in order to find out important information and knowledge which is extracted with the goal of increasing profitability in agriculture by applying data mining techniques to give timely recommendations to farmers.

D. ASSOCIATION RULE MINING

Association Rule Mining is the process of finding new interesting Correlations , frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [2].In data mining, association rule mining is an important and easy method to find frequent item sets from large dataset. It is intended to identify strong rules discovered in databases using two different measures of interestingness. The first one is support which generates frequent item set from the provided database and the other one is confidence which is focuses on rule generation.

Frequent Item Sets- A set of attributes is termed as frequent item set if the occurrence of the set within the database is more than a user given threshold.

Support- Support determines how often a given rule is applicable to a given data set.

Confidence- Confidence determines how frequently items in Y appear in transactions that contain X.

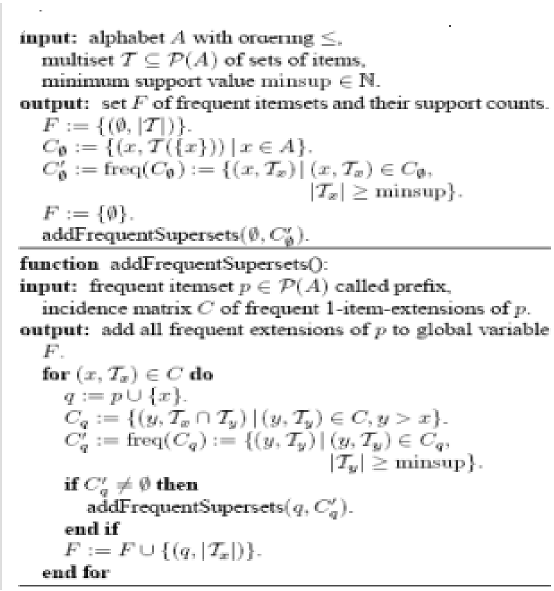
$$\text{Support} \ , \quad S(X\text{->}Y) = \frac{\Sigma(X \cup Y)}{N}$$

$$\text{Confidence,} \quad C(X\text{->}Y) = \frac{\Sigma(X \cup Y)}{\Sigma(X)}$$

Where, X and Y disjoint item set.

E. ECLAT ALGORITHM

Most frequent item-set mining algorithms as Apriori [3] and Eclat [4] use a total order on the items A of the alphabet and the item-sets P(A) to prevent that the same item-set, called candidate, is checked twice for frequency. Items orderings $\leq$ are in one-to-one-correspondence with item coding. The basic Eclat algorithm is shown in Fig 1 below



```
input: alphabet $A$ with ordering $\leq$,
   multiset $\mathcal{T} \subseteq \mathcal{P}(A)$ of sets of items,
   minimum support value minsup $\in \mathbb{N}$.
output: set $F$ of frequent itemsets and their support counts.
   $F := \{(\emptyset, |\mathcal{T}|)\}$.
   $C_\emptyset := \{(x, \mathcal{T}(\{x\})) \mid x \in A\}$.
   $C'_\emptyset := \text{freq}(C_\emptyset) := \{(x, \mathcal{T}_x) \mid (x, \mathcal{T}_x) \in C_\emptyset,$
                                        $|\mathcal{T}_x| \geq \text{minsup}\}$.
   $F := \{\emptyset\}$.
   addFrequentSupersets($\emptyset, C'_\emptyset$).
function addFrequentSupersets():
input: frequent itemset $p \in \mathcal{P}(A)$ called prefix,
   incidence matrix $C$ of frequent 1-item-extensions of $p$.
output: add all frequent extensions of $p$ to global variable
   $F$.
   for $(x, \mathcal{T}_x) \in C$ do
      $q := p \cup \{x\}$.
      $C_q := \{(y, \mathcal{T}_x \cap \mathcal{T}_y) \mid (y, \mathcal{T}_y) \in C, y > x\}$.
      $C'_q := \text{freq}(C_q) := \{(y, \mathcal{T}_y) \mid (y, \mathcal{T}_y) \in C_q,$
                                        $|\mathcal{T}_y| \geq \text{minsup}\}$.
      if $C'_q \neq \emptyset$ then
         addFrequentSupersets($q, C'_q$).
      end if
      $F := F \cup \{(q, |\mathcal{T}_x|)\}$.
   end for
```

Fig.1: Basic Eclat Algorithm

F. GENETIC ALGORITHMS

Genetic Algorithms (GA) are direct, parallel method for global search and optimization. GA is one of the most commonly used Evolutionary Algorithms (EA). They use populations with allowed number of solutions (individuals), they are added in the group of parallel algorithms.

A genetic algorithm (or GA) is a search technique used in computing to find true or approximate solutions to optimization and search problems. As such they represent an intelligent exploitation of a random search used to solve optimization problems. Although randomized, GA's are by no means random; instead they exploit historical information to direct the search into the region of better performance within the search space.

The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution; especially those follow the principles first laid down by Charles Darwin of survival of the fittest. The evolution usually starts from a population of randomly generated individuals and happens in generations.

In each generation, the fitness of every individual in the population is evaluated, multiple individuals are selected from the current population (based on their fitness), and modified to form a new population. The new population is used in the next iteration of the algorithm. The algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population.

GA is based on analogy with the genetic structure and behaviour of chromosomes within the population of individuals having the foundation that individuals in a population compete for resources and mates. Those individuals most successful in each 'competition' will produce more offspring than those individuals that perform poorly. Genes from `good' individuals propagate throughout the population so that two good parents will sometimes produce offspring that are better than either parent. Thus each successive generation will become more suited to their environment. A population of individuals is maintained within search space for a GA, each representing a possible solution to a given problem. Each individual is coded as a finite length vector of components, or variables, in terms of some alphabet, usually the binary alphabet {0, 1}. To continue the genetic analogy these individuals are likened to chromosomes and the variables are analogous to genes. Thus a chromosome (solution) is composed of several genes (variables).

A fitness score is assigned to each solution representing the abilities of an individual to `compete'. The individual with the optimal (or generally near optimal) fitness score is sought. The GA aims to use selective `breeding' of the solutions to produce `offspring' better than the parents by combining information from the chromosomes.

The main ingredients of GA are Chromosomes, Selection, Recombination and Mutation.

- **Selection**

During each successive generation, a proportion of the existing population is selected to breed a new generation. Fitness-based process is used to select individual solutions where fitter solutions (as measured by a fitness function) are typically more likely to be selected. At this stage elitism could be used – the best n individuals are directly transferred to the next generation. The elitism ensures, that the value of the optimization function cannot get worst (once the extremis is reached it would be kept).

DOI10.17148/IJARCCE.2017.6831

- **Crossover**

The most common type is single point crossover. In single point crossover, we choose a locus point at which you swap the remaining alleles from one parent to the other. The children take one section of the chromosome from each parent. Chromosome is broken based on the randomly selected crossover point. This particular method is called single point crossover because only one crossover point exists. Sometimes only one child is created, but generally both offspring are created and put into the new population. Crossover does not always occur. Sometimes, based on a set probability, no crossover occurs and the parents are copied directly to the new population.

- **Mutation**

After selection and crossover, we have a new population full of individuals where some are directly copied, and others are produced by crossover. In order to ensure that the individuals are not all exactly the same, we allow a small chance of mutation. We go through all the alleles of all the individuals, and if that allele is selected for mutation, we either change it by a small amount or replace it with a new value. Mutation is fairly simple. Mutation is, however, vital to ensuring genetic diversity within the population. Genetic Algorithm is a randomized algorithm that could be run for a very long time to obtain an optimal solution.

II. RELATED WORK

Today, Production agriculture is not the kind of farming our grandfathers would recognize. Global warming affects climatic conditions. This makes it very difficult to predict the climatic change which is very bewildered [5]. In the era of high technology, we are enriched by the number of new data mining techniques and services. Data mining in agriculture is a very recent research topic [6]. By better understanding the condition of their soil and other related factors, farmers are able to maximize the use of their land, by planting the right crops in the right places, and intervening with pesticides, fertilizer when conditions demand it which causes a highly improve yield [7]. In this technical era, almost all fields are computerized. Agriculture also having a large amount of data becomes a candidate for data mining. If we apply the proper technique on it this data no longer remains only pieces of data but, it gives lot of pattern from it. In this we need to make inferences from immense data so that we can make decisions driven by knowledge [8].

Various factors which affect the production of crops like soil type, crop price and other factors are taken into consideration. Data mining [9] is an art and science of intelligent analysis of large data sets for meaning and previously unknown insights. With the help of Knowledge Discovery in Databases and data mining we extract the meaningful data sets from the large amount of data [10 – 15]. Data mining in agriculture is an important research field [16]. Data mining has tools that powerfully generate rules from vast & diversified data such as in agriculture datasets which are large. Data mining also is a process of analysing [17, 18] this data from different prospective and summarizing it into useful information.

There are various techniques of data mining which are applied over the huge amount of data and we get some pattern or knowledge from it [19]. For optimization of solution or result we use Genetic algorithm. Genetic Algorithm is a randomized algorithm that could be run for a very long time to obtain an optimal solution for agricultural problems [20], [17], [13].

Lida et.al [7] proposed a systematic approach based on integrated information systems (IISs) for agricultural ecosystem management. They extracted data on terrain, land use, planting, and others, and integrated them for the purpose of agricultural and ecosystem management. They concluded that, for effective management of agriculture and ecosystems, a systematic approach was essential in which Integrated Information Systems played a crucial role.

Nasira et. Al [21] worked on price prediction of vegetables. This helped the farmers and also Government to make effective decision about that specific vegetable based on the complexity of vegetable price prediction, making use of the characteristics of neural networks such as self-adapt, self-study and high fault tolerance, to build up the model of Back-propagation neural network to predict vegetable price. This price prediction model was set up by applying the neural network. They took an example of tomato as a vegetable. The parameters of the model were analysed through experiment. At the end of the result of Back-propagation neural network showed absolute error percentage of monthly and weekly vegetable price prediction and analysed the accuracy percentage of the price prediction.

Veenadhari et.al [22] presented Data mining techniques for predicting crop productivity. They attempted to review the research studies on application of data mining techniques in the field of agriculture. Their publication serves as a review article which gives various techniques which predict crop productivity. Sumitha and Kirubakaran [5] worked on E-Agriculture Information Management System. They proposed some news about agriculture SMS passed on daily basis. They also proposed system for information passed on seasonal basis as well as other details information regarding agriculture. This system benefited Indian farmers since market and weather information was delivered to their mobile phones.

DOI10.17148/IJARCCE.2017.6831

III.METHODOLOGY

We have developed the optimized agricultural crop recommender system using data mining technique for agriculture application. For giving recommendations we have used association rule mining and genetic algorithm approach for optimization of the result depending on the agricultural historical database.
A recommender system considers a soil profile of the field and gives an interesting pattern and provides a recommendable crop to the farmer or user. As degree of crop recommendation can vary from one to another farm field.
In association rules it is intended to identify strong rules discovered in databases using two different measures of interestingness. The first one is support which generates frequent item set from the provided database and the other one is confidence which is focuses on rule generation.
As shown in Figure 2 below, our agricultural recommender system for the farmer who has to test their soil from soil test centre after that depend on the soil test result we recommend the crop on the basis of association rules. Our methodology is a two stage model.

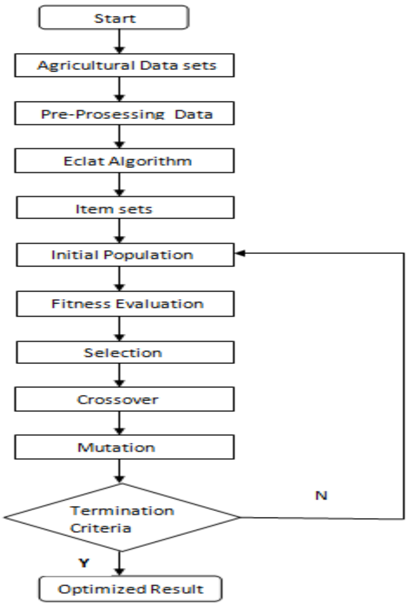


Fig 2 : Proposed agricultural recommender system

- **Applying Eclat**

We apply Eclat on the agriculture historical data to generate frequent item sets by applying the proper support for each rule we get good item sets. Data for a long period of time for the crops in Solapur district of Maharashtra was stored in a dataset. This is called historical data. This agricultural historical data is in the raw form. It has to be first converted into a proprietary form by pre-processing technique.

TABLE I String values for parameters

Parameter	Range	String value
Ph(i)	3.5 - 6.51	Low
..	7.50 - 9.1	High
E.C(i)	0.41 - 0.81	Low
..	1.60 - 2.6	High
O.C(i)	0.21 - 0.40	Low
..	0.81 - 1.00	High
N(i)	141 - 281	Low
..	281 - 701	High
P(i)	7.0 - 14.0	Low
..	14.1 - 36.0	High
K(i)	101 - 150	Low
..	151 - 300	High

DOI10.17148/IJARCCE.2017.6831

For pre-processing this raw data, we convert historical data into String values. In the pre-process, we classify that value depend on the range of the soil parameter as High and Low. We apply the Eclat algorithm for obtaining the item sets. To get association rules, on each item set we apply support and confidence. At this point we apply a threshold value of support to get item sets. So that, we get good association rules.
We convert those six values into String value as shown in Table 1

Transaction from transaction set F_k for each item is then fetched. For each item {a} the list of transactions containing {a} is detected for all I_j € F_k where i < j,
N = $I_i \cap I_j$

We then calculate the value of support S
S (X → Y) = σ (X U Y) / N

If N. sup ≥ min – sup then
F_{k+1} = ((F_{k+1}) U (N));
Repeat from 1 on {a} conditional database. On generated item sets, we apply section operation.

- **Applying Genetic Algorithm**

In the second stage, we apply Genetic algorithm to optimize the initial population (Eclat generated item sets as initial population for Genetic Algorithm) from which we get association rules. We use item set as initial population for the Genetic Algorithm. Then, we filter the best chromosomes and pass to fitness evaluation.
We apply various operation of the Genetic Algorithm like Selection, Crossover and Mutation. After mutation operation we check for the threshold value and get the optimized result. If we do not meet up with threshold value then we repeat the Genetic Algorithm operations until we get best rules that predict output as an optimized agriculture crop. For Fitness Evaluation we check whether fitness is greater than given threshold then add that transaction to the rules. If (fitness > i), where, i is threshold value range between 0 to 1. For the termination criteria, if generation is less than maximum generation or fitness value is maximum fitness then the termination criteria satisfy. Otherwise, repeat step from selection of initial population.

IV. EXPERIMENTAL SET-UP AND RESULTS

The system was developed using Java platform and R programming tools. The frequent set items were arrived at using Eclat algorithm of 'Apriori' package of R. The Genetic Algorithm was developed in JAVA language which was interfaced with the Eclat frequent item sets to generate best association rules. Testing was performed on the agricultural datasets with different crop. The chart of 4 crops with Apriori Algorithm Vs Eclat-Genetic Algorithm is shown in Figure 3. It can be seen from the chart that the prediction accuracy has improved by 28.31 %. On agricultural datasets we performed testing by giving different crop rules and result is compared with exiting algorithm (Apriori).

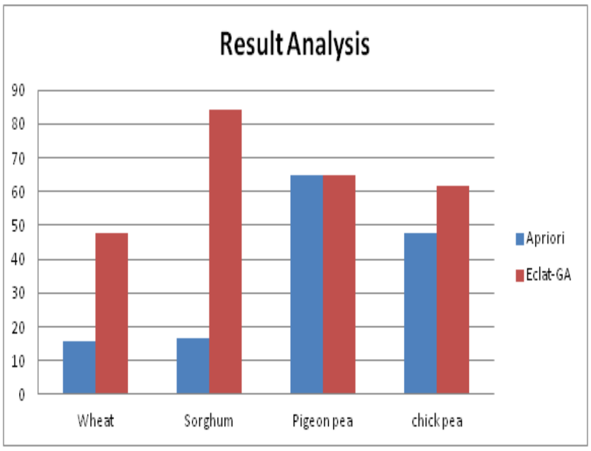
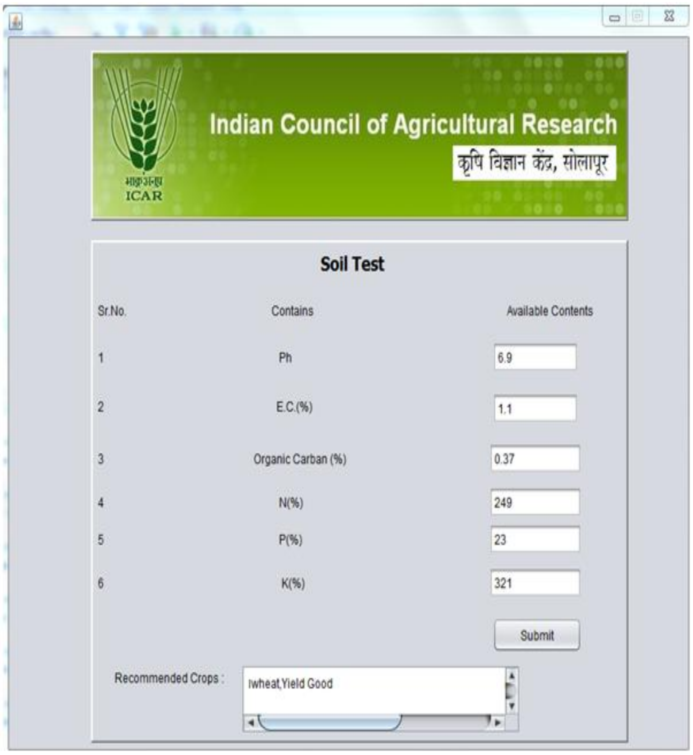


Fig 3: Result Analysis

On the basis of above chart, we can say that Eclat-Genetic Algorithm perform well as the prediction accuracy is 28.31 % more satisfactory. We also show for a given set of parameters a particular crop is recommended for a farmer to grow. An example of this is shown in a snapshot in Figure. 4



Indian Council of Agricultural Research

कृषि विज्ञान केंद्र, सोलापूर

भाकृअनुप
ICAR

Soil Test

Sr.No.	Contains	Available Contents
1	Ph	6.9
2	E.C.(%)	1.1
3	Organic Carban (%)	0.37
4	N(%)	249
5	P(%)	23
6	K(%)	321

Recommended Crops : Iwheat,Yield Good

Fig.4: Recommendation for crops

V. CONCLUSION

Although a lot of work is already published in this field, but in this work we have tried to use the enormous robustness of Association rule by applying GA on frequent item sets in an Eclat-GA form. The experimental result shows that, the proposed model gives 28.31\% better result than the existing technique (Apriori) in terms of accuracy. We also permit recommendation for crops based on parameters given. We believe that the toolkit can also handle other databases, after minor modifications. As an extension of this work, we can work on the complexity reduction of Genetic Algorithms by using distributed computing.

As a future scope, we can extend our work by providing recommendation of fertilizers and pesticides for a specific recommended crop. Along with these recommendations, we also suggest side business for farmers on his/her mobile by providing an SMS service or an android application which will be helpful to improve his economical status. It is also possible to provide daily recommendations, fertilizers for crop depend on the weather forecast, humidity etc. As for future work, the author is currently working on the complexity reduction of Genetic Algorithms by using a distributed computing environment.

ACKNOWLEDGMENT

The authors would like to thank their affiliated institutions for all the support given to them during their research and preparing of this publication.

REFERENCES

[1]. G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, 2005, pp. 734–749.

[2]. F. Khan, “Knowledge discovery on agricultural dataset using association rule mining,” in International Journal of Emerging Technology and Advanced Engineering Webs, vol. 4, no. 5, 2014, pp. 925-930.

[3]. R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proceedings of the 20th VLDB Conference Santiago, Chile, 1994.

[4]. M. J. Zaki, “Scalable Algorithms for Association Mining,” In IEEE Transactions on Knowledge and Data Engineering, Vol. 12, 2000, Pp. 372–390.

[5]. S. Thankachan and S. Kirubakaran, “E-agriculture information management system,” in International Journal of Computer Science and Mobile Computing, vol. 3, no. 5, 2014, pp. 599–607.

[6]. R. Agrawal, T. Imielinski, and A. Swami, “Database mining: A performance perspective,” in IEEE Transactions on Knowledge and Data Engineering, vol. 5, 1993, pp. 914–925.

DOI10.17148/IJARCCE.2017.6831

[7]. L. Xu, N. Liang, and Q. Gao, "An integrated approach for agricultural ecosystem management," in IEEE Transactions On Systems, Man, And CyberneticsPart C: Applications And Reviews, vol. 38, no. 4, 2008.

[8]. P.Bhargavi and S.Jyothi, "Applying naive bayes data mining technique for classification of agricultural land soils," in International Journal of Computer Science and Network Security, vol. 9, no. 85, 2009, pp. 117–122.

[9]. A. Abdullah and A. Hussain, "Data mining a new pilot agriculture extension data warehouse," in Journal of Research and Practice in Information Technology, vol. 38), number = 3, year = 2006..

[10]. S. Y. Sung, Z. Li, C. L. Tan, and P. A. Ng, "Forecasting association rules using existing data sets," in IEEE Transactions On Knowledge And Data Engineering, vol. 15, no. 6, 2003.

[11]. H. Toivonen, "Sampling large databases for association rules." Morgan Kaufmann, 1996, pp. 134–145.

[12]. Z. Diriba and B. Borena, "Application of data mining techniques for crop productivity prediction," in HiLCoE Journal of Computer Science and Technology, vol. 1, no. 2, 2011, pp. 51–55.

[13]. J. R. Prasad, P. R. Prakash, S. S. Kumar, M. S. Babu, and K. S. Rani, "Identification of agricultural production areas in Andhra Pradesh," in International Journal of Engineering and Innovative Technology (IJEIT), vol. 2, no. 2, 2012, pp. 51–55.

[14]. A. Mucherino, P. J. Papajorgji, and P. M. Pardolos, "Data mining in agriculture," in International Journal of Engineering and Innovative Technology (IJEIT), vol. 2, no. 2. Springer, 2009.

[15]. G. Ru, R. Kruse, M. Schneider, and P. Wagner, "Visualization of agriculture data using self-organizing maps," in Applications and Innovations in Intelligent Systems,Proceedings of AI-2008, BCS SGAI, Springer,vol. 16, 2009, pp. 47–60.

[16]. D. Miller, J. McCarthy, and A. Zakzeski, "A fresh approach to agricultural statistics: Data mining and remote sensing," in Section on Government Statistics at JSM, 2009.

[17]. K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in ACM, 2003, pp. 519–528.

[18]. D. Rajesh, "Application of spatial data mining for agriculture," in International Journal of Computer Applications, vol. 15, no. 2, 2011.

[19]. S. T. Gaikwad, S. B. Desai, and A. B. Kolekar, "Adoption of information and communication technology (ict) for development of Indian agriculture," in International Journal for Research in Applied Science Engineering Technology, vol. 4, no. 4, 2016, pp. 761–765.

[20]. R. Swaminathan, "Analysis of self organizing maps using visual dm techniques in agro database for prediction of yield," in International Journal of Advanced Computer Science,, vol. 3, no. 10, 2013, pp. 508–511.

[21]. G. M. Nasira and N. Hemageetha, "Vegetable price prediction using data mining classification technique," in International conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012.

[22]. S.Veenadhari, B. Misra, and C. Singh, "Data mining techniques for predicting crop productivity a review article," in IJCST, vol. 2, no. 1, 2011.