Abstract—Data mining is a technique of analyzing the dataset such that the final conclusion can be accessed easily and quickly from the dataset. Here in this paper association rule mining technique is implemented for the analysis of agricultural dataset. The idea is to apply association rule mining technique for generating rules and establishing a relationship between them. So as to enhance crop productivity. Also a performance comparison is done between Apriori and FP Growth algorithm.

Keywords— Agricultural data , Apriori algorithm, Association Rule Mining, FP Growth Algorithm, Data mining

I. INTRODUCTION

Data mining is a technique used to presume valuable and appropriate information to conduct proficient choices and other scientific investigations [1]. This is also known as “Knowledge Discovery”. The beginning of data mining systems for discovering usage pattern from Web data (Web Usage Mining) indicates that these techniques can be a viable alternative to traditional decision making tools. The expression Data Mining, or Knowledge Discovery in Databases, has been accepted for a field of research agreement with the habitual discovery of understood information or knowledge within databases [3].

The implicit information surrounded by databases, and generally the remarkable association bonding among sets of objects, that led to association rules, may reveal useful patterns for decision support, financial and many applications. Association Rule Mining is an important area of data mining research and comparatively a younger member of data mining community. This actually fascinated a lot of awareness in current data mining research. A very dominant association rule mining algorithm, Apriori [4], has been build up for rule mining in large transactional databases. Many other algorithms developed are derived and/or conservator of this algorithm. A major walk forward in improving the presentations of these algorithms was made by the preface of a work of fiction, compact data structure, passed on to as frequent pattern tree, or FP-tree [5], and the associated mining algorithm, FP-growth. The main difference between these two approaches is that the Apriori like methods are based on bottom-up creation of frequent item set arrangements and the FP-tree based ones are partition-based, divide-and-conquer methods.

II. LITERATURE REVIEW

Though varieties of data mining techniques, like association rule mining, cluster analysis etc. are being used so often in different applications, in this section we have specified most of the work related to data mining and related techniques. AIS Algorithm was the first association rule mining algorithm [6] where generation of candidate item sets is done at the time of scanning the database through extending known-frequent item sets with items from each transaction. It is a multi-pass algorithm. An estimate of the supports of these candidates is used to steer whether these candidates are needed to be extended further for production of more candidates. The chief drawback of this algorithm is that it generates too many candidates which are small and requires much more space and is thereby inefficient.

2.1. Apriori Algorithm

Apriori algorithm which is introduced in [7] that is more efficient than AIS by an order of magnitude. The foremost advantage of Apriori is that it incorporates the subset frequency based pruning optimization that means, it only process any item-set whose subsets are frequent also. It utilizes a data structure that is known as hash tree which is used for storing the counters of candidate item-sets. The main drawbacks of Apriori are: i) It performs n passes in excess of the database, where n is the distance end to end of the greatest frequent item-set. The counts of candidate item-sets of length k are obtained in kth pass, ii) it follows a tuple-by-tuple approach where counters of candidate item-sets are updated after reading individual transaction of whole database so that much redundant work is performed after each individual transaction. Based on this algorithm, lots of new algorithms are deliberated with enhancements and modifications. Apriori algorithm has been developed for rule mining in large transactional databases. This algorithm divides the difficulty of mining patterns into two steps:

In first step find all the combinations of items that have transaction support beyond minimum support, and then generate frequent patterns. In second step a subsequent pass, say pass k, consists of two phases. First, the frequent pattern Lk-1 (the set of all frequent (k-1) - patterns) found in the (k-1)th pass are used to generate the candidate sets Ck, using the apriori-gen() function. The algorithm ends when Lk turns out to be blank.
2.2 FP-Growth Algorithm

In FP-Tree/Growth Algorithm, after a preprocessing scan over the database, it creates a reduced characterization of the database which is known as FP-tree and then data mining is executed over the FP-tree. Frequent patterns are pattern (such as item sets, subsequences or substructures) that become visible in a data set frequently. FP-growth algorithm which is enhanced version of Apriori algorithm gives better performance while scanning the data base. FP-growth algorithm takes less time to search the each level of the tree. The FP-growth method transforms the problem of finding long frequent patterns to searching for shorter ones recursively and then concatenating the suffix. It uses the smallest amount of frequent items as a suffix, recommending good selectivity. The scheme considerably condenses search expenditure.

The FP Tree Algorithm is to maintain a frequent pattern tree (FP-Tree) of the database. It is a make bigger prefix-tree structure, storing essential quantitative information about frequent sets. The tree nodes are frequent items and are design in such a way that more commonly occurring nodes will have improved possibility of sharing nodes than the less regularly happening ones. The algorithm constructs the conditional FP-tree and performs mining on this tree. The FP-tree Algorithm will take more time for recursive calls in the algorithm. If the user access paths are common then FP-tree algorithm is very efficient. The existing All offered Algorithms have their own advantages and disadvantages. If all the transactions are dissimilar then Apriori algorithm is good or else FP-tree Algorithm is excellent. If the user rights to use paths are frequent then FP-tree algorithm is very proficient.

2.3 Literature Survey

Sally Jo Cunningham and Geoffrey Holmes [8] have described a WEKA (Waikato Environment for Knowledge Analysis) system which provides a complete suite of facilities for applying data mining techniques to large data sets.

D Ramesh and B Vishnu Vardhan [9], have shown different Data Mining techniques, such as K-Means, K-Nearest Neighbour(KNN), Artificial Neural Networks(ANN) and Support Vector Machines(SVM) were adopted to estimate crop yield analysis with existing data.

Shoemaker C.A. et al. [10] offered an association rule mining system that was capable of handling set-valued attributes. They introduced two algorithms for mining association rules directly from set-valued data and compared their performance as well.

Piao et al. [11] also conducted research on the same type of algorithm. The research was based on mining negative and positive association rules, based on dual confidence.

Demetrijevic et al. [12] developed a web usage mining system. They implemented a system for the discovery of association rules in web log usage data as an object-oriented application and used it to experiment on a real life web usage log data set.

Data mining in Agricultural datasets is a comparatively novel research field. Association Rule Mining can play an important role in discovering knowledge from agricultural datasets, data about soil and cultivation, survey data from agricultural research, data containing information on geographical conditions and crop production to name a few. Agriculture data [13],[14],[15] is available in an increasing amount of formats and portals. The dataset used in our work is a collective study of the agricultural portals available on web these days. These portals provide information on various crops from grape to pumpkin and gram to soybean.

But our interest of research is the major crops of Madhya Pradesh namely- Wheat, Rice, Jowar, Bajra and Soyabean. Where we have tried to figure out the relationship between the above mentioned crops on the basis of the soil type, ph level of the soil, season and the crop yield. Such research can assist in making decisions regarding selection of crops to be grown in particular area, based in geographical conditions, increasing production of crop by selecting proper resources, selecting proper weather conditions for crop production etc.

III. PROPOSED METHODOLOGY

Agriculture, with its allied sectors, is unquestionably the largest livelihood provider in India, more so in the vast rural areas. It also contributes a significant figure to the Gross Domestic Product (GDP). Sustainable agriculture, in terms of food security, rural employment, and environmentally sustainable technologies such as soil conservation, sustainable natural resource management and biodiversity protection, are essential for holistic rural development.

In this paper association rule mining has been applied on agricultural dataset for mining out the frequent patterns from data and generate rules to portray the relationship amongst the crops based on their yield. This is an effort in the direction to help understand, enhancement of crop yields and the reduction of crop losses. Here, Apriori Algorithm is considered for generating results. Further these results are compared with the results generated by FP-Growth Algorithm. This comparison will give us an insight, which algorithm gives best results with the dataset.
The data considered for the research work is mainly comprising of major crops grown in Madhya Pradesh. The name of these crops are:


The productivity of the above crops is then calculated upon various parameters. The parameters that have been considered are as follows:

- Soil type used for crop production.
- Ph value of the soil used for the crop.
- Season in which the crop is grown.

The experiments are performed on agricultural dataset which contains attributes of various products as shown in table:

<table>
<thead>
<tr>
<th>No.</th>
<th>Crop</th>
<th>yield</th>
<th>soil_type</th>
<th>ph_value</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bajra</td>
<td>Excellent</td>
<td>shallow_black</td>
<td>high</td>
<td>kharif</td>
</tr>
<tr>
<td>2</td>
<td>Bajra</td>
<td>Poor</td>
<td>mixed_red black</td>
<td>high</td>
<td>kharif</td>
</tr>
<tr>
<td>3</td>
<td>Bajra</td>
<td>Poor</td>
<td>coastal_alluvial</td>
<td>low</td>
<td>kharif</td>
</tr>
<tr>
<td>4</td>
<td>Bajra</td>
<td>Good</td>
<td>deltaic_aluvium</td>
<td>high</td>
<td>kharif</td>
</tr>
<tr>
<td>5</td>
<td>Bajra</td>
<td>Excellent</td>
<td>deltaic_aluvium</td>
<td>medium</td>
<td>kharif</td>
</tr>
<tr>
<td>6</td>
<td>Bajra</td>
<td>Good</td>
<td>deltaic_aluvium</td>
<td>low</td>
<td>kharif</td>
</tr>
<tr>
<td>7</td>
<td>Jowar</td>
<td>Excellent</td>
<td>sandy</td>
<td>high</td>
<td>rainy</td>
</tr>
<tr>
<td>8</td>
<td>Jowar</td>
<td>Poor</td>
<td>laterite</td>
<td>high</td>
<td>rainy</td>
</tr>
<tr>
<td>9</td>
<td>Jowar</td>
<td>Poor</td>
<td>red_yellow</td>
<td>high</td>
<td>rainy</td>
</tr>
</tbody>
</table>

3.1 Apriori Algorithm

Apriori (T, minSupport) {//T is the database and minSupport is the minimum support

F1= {frequent items};
for (k= 2; F_k-1 =∅; k++) {
C_k= candidates generated from F_k-1
//that is cartesian product F_k-1 x F_k-1 and eliminating any k-1 size itemset that is not frequent
for each transaction t in database do{
#increment the count of all candidates in C_k that are contained in t
F_k = candidates in C_k with minSupport
}//end for each
}//end for
return \bigcup_k F_k;

1. Generate all frequent item sets using specified minSupport.
2. Count supports of each individual item.
3. Create a set F with all individual items with min support.
4. Create "Candidate Set" C_k based on F_k-1.
5. Check each element c in C_k to see if it meets minSupport.
6. Return set of all frequent item sets.
7. Create two sets differing only in the last element, based on data set.
8. Join those item sets.
9. Compare each subset S of C to F_k-1, if s is not in F_k-1, delete it.
11. Take Frequent Item Set F
12. If \{F_1, F_2,...F_{k-1}\} \Rightarrow \{F_k\} meets some min confidence, make it a rule.
13. Remove last element from antecedent, insert into consequent, check again.
14. Generate all confident Association Rules from frequent item sets.

3.2 FP Growth Algorithm
procedure FP_growth(Tree,α)

(1) if Tree contain a single path P then
(2) for each combination (denoted as β) of nodes in the path P
(3) generate pattern βUα with support_count = min_support count of nodes in β;
(4) else for each a_i in the header of Tree {
(5) generate pattern β = a_i U α with support_count = a_i.support_count
(6) construct β’s conditional pattern base and then β’s conditional FP_tree Tree_β.
(7) if Tree_β ≠ ∅ then
(8) call FP_growth (Tree_β,β) ; 

IV. RESULT ANALYSIS

These experiments are performed on computer with i5 Processor, 2.0 GB RAM and hard disk 320 GB. The proposed approach has been developed in Net beans and JDK with WEKA libraries. Time has been taken in seconds.

4.1 Rules Generated with Apriori Algorithm:
1. Rice=Excellent => Wheat=Excellent  conf:(1)
2. Bajra=Good Jowar=Poor => Soyabean=Poor  conf:(1)
3. Jowar=Poor Wheat=Excellent => Bajra=Good conf:(1)
4. Bajra=Good Wheat=Excellent => Jowar=Poor conf:(1)
5. Bajra=Good Jowar=Poor => Wheat=Excellent conf:(1)
6. Soyabean=Poor Wheat=Excellent => Bajra=Good conf:(1)
7. Bajra=Good Wheat=Excellent => Soyabean=Poor conf:(1)
8. Soyabean=Poor Wheat=Excellent => Jowar=Poor conf:(1)
9. Jowar=Poor Wheat=Excellent => Soyabean=Poor conf:(1)
10. Jowar=Poor Soyabean=Poor Wheat=Excellent => Bajra=Good conf:(1)

Table 2. Inferences drawn from the generated rules.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>If the yield of rice is excellent then the yield of wheat will be excellent.</td>
</tr>
<tr>
<td>2</td>
<td>If the yield of bajra is good and jowar is poor then the yield of soyabean will be poor.</td>
</tr>
<tr>
<td>3</td>
<td>If the yield of jowar is poor and wheat is excellent then the yield of bajra will be good.</td>
</tr>
<tr>
<td>4</td>
<td>If the yield of bajra is good and wheat is excellent then the yield of jowar will be poor.</td>
</tr>
<tr>
<td>5</td>
<td>If the yield of soyabean is poor and wheat is excellent then the yield of jowar will be poor.</td>
</tr>
<tr>
<td>6</td>
<td>If the yield of soyabean is poor and wheat is excellent then the yield of soyabean will be poor.</td>
</tr>
<tr>
<td>7</td>
<td>If the yield of bajra is good and wheat is excellent then the yield of soyabean will be poor.</td>
</tr>
<tr>
<td>8</td>
<td>If the yield of soyabean is poor and wheat is excellent then the yield of soyabean will be poor.</td>
</tr>
<tr>
<td>9</td>
<td>If the yield of jowar is poor and wheat is excellent then the yield of soyabean will be poor.</td>
</tr>
<tr>
<td>10</td>
<td>If the yield of jowar is poor, soyabean is poor and wheat is excellent then the yield of bajra will be good.</td>
</tr>
</tbody>
</table>

4.2. Results after applying Apriori Algorithm

Table 3. Result analysis of Apriori on Agricultural Data

<table>
<thead>
<tr>
<th>Support</th>
<th>50</th>
<th>35</th>
<th>55</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence</td>
<td>30</td>
<td>20</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>Frequent Sets in Apriori</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Rules generated in Apriori</td>
<td>8</td>
<td>10</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Generation time</td>
<td>0.017 s</td>
<td>0.015 s</td>
<td>0.018 s</td>
<td>0.067 s</td>
</tr>
</tbody>
</table>

Figure 1. Result for Support with frequent itemsets and rules generated

Where x axis=Support and y axis= Frequent itemsets and rules generated.

Figure 2. Result for confidence with frequent itemsets and rules generated
Where x axis = confidence and y axis = frequent itemsets and rules generated.

Number of cycles performed: 13
Generated sets of large itemsets:
Size of set of large itemsets L(1): 13
Size of set of large itemsets L(2): 8
Size of set of large itemsets L(3): 4
Size of set of large itemsets L(4): 1

4.3. Results after Applying FP Growth Algorithm
1. Rice=Excellent ==> Wheat=Excellent conf:(1)
2. Bajra=Good Jowar=Poor ==> Soyabean =Poor conf:(1)
3. Jowar=Poor Wheat=Excellent ==> Bajra=Good conf:(1)
4. Bajra=Good Wheat=Excellent 5 ==> Jowar=Poor conf:(1)
5. Jowar=Poor Soyabean =Poor Wheat=Excellent ==> Bajra=Good conf:(1)

Table 4. Result analysis of FP Growth on Agricultural Dataset

<table>
<thead>
<tr>
<th>Support</th>
<th>50</th>
<th>35</th>
<th>55</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence</td>
<td>30</td>
<td>20</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>Frequent Sets in FP Growth</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Rules generated in FP Growth</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Generation Time FP Growth</td>
<td>0.007 s</td>
<td>0.008 s</td>
<td>0.008 s</td>
<td>0.023 s</td>
</tr>
</tbody>
</table>

4.4. Performance Comparison between Apriori and FP Growth

<table>
<thead>
<tr>
<th>Support</th>
<th>50</th>
<th>35</th>
<th>55</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence</td>
<td>30</td>
<td>20</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>Frequent Sets in Apriori</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Rules generated in Apriori</td>
<td>8</td>
<td>10</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Generation Time Apriori</td>
<td>0.017 s</td>
<td>0.015 s</td>
<td>0.018 s</td>
<td>0.067 s</td>
</tr>
<tr>
<td>Rules generated in FP Growth</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Generation Time FP Growth</td>
<td>0.007 s</td>
<td>0.008 s</td>
<td>0.008 s</td>
<td>0.023 s</td>
</tr>
</tbody>
</table>

Table 5. Result analysis of Apriori and FP Growth on Agricultural Dataset

V. CONCLUSION

The Apriori and FP growth algorithms implemented here for the agricultural dataset provides a new way of analyzing the various crops of Madhya Pradesh and their productivity ideas. Such as wheat gives good yield with medium pH level and laterite soil. Similarly, Soyabean and Bajra gives excellent results with a combination of shallow black soil and high pH level.
A comparison between the Apriori and FP Growth Algorithm shows that FP Growth is a better candidate for such type of analysis.

REFERENCES


[5] J. Han, J. Pei, Y. Yin, “Mining Frequent Patterns without Candidate Generation”. Proc. of ACM-SIGMOD, May 2000.


[14] uasr.agropedia.in/content/climate-and-soil-requirements-bajra-crop