International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

Conceptual Overlapping Clustering for Effective Selection of Parental Rice Varieties

Madhavi Dabbiru, Shashi Mogalla , PRM Rao
Department of IT,GMRIT, Rajam, Srikakulam Dist, A.P, India.
madhavicris@yahoo.co.in
Department of CSSE, COE, Andhra University, Visakhapatnam A.P, India.
smogalla2000@yahoo.com
ARS Ragolu, Srikakulam, A.P, India.
arsragolu@yahoo.co.in

ABSTRACT

The process of rice breeding involves producing new rice varieties as a cross of parental rice varieties followed by rigorous testing and examination phase for purity, productivity and resistivity to regional climatic conditions. The selection of appropriate parental seeds to produce new rice variety with more or less predictable characteristics is highly desirable as it reduces expensive experimental evaluation efforts. Authors suggest the applicability of Conceptual Overlapping Clustering Algorithm [7] developed by them for conceptual overlapping clustering to aid proper selection of rice varieties and demonstrated the performance of this algorithm on a real world dataset.

General Terms

Pattern Recognition, Information Retrieval.

Indexing terms

Associative clustering, Conceptual clustering, Overlapping clustering, Rice breeding methods, Colossal pattern mining.

Academic Discipline And Sub-Disciplines

Computer Science;

SUBJECT CLASSIFICATION

Data Mining

COVERAGE

In general data mining tasks can be broadly classified into two categories: Predictive and Descriptive. Predictive mining tasks aims to forecast the value of unknown attributes while descriptive mining tasks characterize general properties of the data in the database. Examples of descriptive data mining include Characterization, Association rule discovery and Clustering. In the recent times, association rules are used for clustering tasks especially for databases that are large in size. This work focuses on descriptive data mining techniques and their hybridization.

TYPE (METHOD/APPROACH)

Yen et.al, proposed and introduced the concepts of overlapping clusters. Overlapping clustering algorithms form clusters that are non-exclusive and non-exhaustive so that a data object may participate in more than one overlapping clusters simultaneously and another data object may not participate in any of the overlapping clusters. For example, we can use overlapping clustering for document clustering to form indexes which enable us to identify list of documents covering selected topics. A document may exist in more than one such list in accordance to its relevance to multiple topics as the documents are formed by overlapping clustering. A limitation of all the methods discussed above is that there is no emphasis on associating semantics to the resultant clusters. Conceptual clustering insists on cluster formation based on the semantics of data objects. Efforts are on in this direction and associative mining results are used for generating conceptual clustering. Apriori based associative clustering [4], forms more number of clusters based on a large number of frequent itemsets mostly related through subset superset relationship. However, using all frequent itemsets for the formation of conceptual clusters may result in large number of insignificant clusters. Hence the authors suggest formation of conceptual clusters using a compressed set of significant frequent itemsets namely colossal patterns.

1. INTRODUCTION

The fundamental goal of rice breeding research is to develop rice varieties with high yielding potential, high grain quality and resistance to biotic and abiotic stresses. Success in developing improved rice varieties depends on the adoption of proper breeding methodologies and hence is identified as a research priority.

Each rice variety is resistant to certain diseases and insect pests along with their biotypes. For example, the insect pest Gall midge has six biotypes in India [10]. The rice breeding techniques are slow and expensive. After the four stages of cross, pedigree nursery, yield trial and local adaptability test then that rice variety is released to the market and become

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

available to farmers. Based on the purity, productivity and resistance tests to regional pests the performance of newly developed seeds is analyzed through two to six generations before its commercial use. Thus the whole process of development of a new rice variety takes 2 to 3 years.

The new rice variety developed is expected to share the characteristics from the parental rice varieties. Hence proper selection of parental rice varieties helps in the development of new rice variety with more or less desirable characteristics. Authors suggest conceptual clustering of rice varieties for effective selection of parental rice varieties. Authors propose to form highly cohesive conceptual clusters based on colossal patterns applying association mining techniques. The members of such clusters possess a set of common characteristics represented by a frequent itemset. Clusters formed by all frequent itemsets are enormous in number with less number of common characteristics which indicates less cluster cohesion. Ke wang et al, proposed the concept of large items to cluster transactional data. This strategy is based on two criterion functions; inter cluster cost and intra cluster cost [11]. Conceptual clusters formed from the lengthier frequent itemsets (colossal patterns) have better cluster cohesion and hence preferred.

A rice variety is represented as a transaction with a listing of its characteristics like its resistivity to different pests and diseases represented as constituent items. The resultant data set contains less number of transactions which are lengthy and hence expected to result in lengthy patterns in turn. Conventional association mining algorithms cannot deal with extraction of very lengthy or colossal patterns. Authors propose to use a specifically designed colossal pattern mining algorithm for formation of conceptual clusters with high cohesion. Certain rice varieties may have compatibility with more than one cluster as they possess characteristics common to the elements of more than one cluster. The associative clustering techniques naturally support formation of overlapping clusters as they allow certain transactions to cover more than one colossal pattern and hence become members of their corresponding clusters. D.C.Wimalasuriya et al., [4] applied association mining results to cluster zebra fish genes. The transactions contained the EST classes of each gene. Minimum support threshold is taken as input for finding frequent itemsets using Apriori. The transactions that cover a frequent itemset constitute a cluster.

This paper presents a novel algorithm for conceptual overlapping clustering and applies it on agricultural data for effective selection of parental rice varieties for developing new rice varieties with predictable characteristics. The rest of the paper is organized as follows: section 2 presents concepts related to conceptual overlapping cluster. Section 3 discusses the methodology for formation of conceptual clusters using colossal patterns. Section 4 discusses evaluation metrics followed by section 5 depicting the experimentation and results. Section 6 suggests application of COCA for effective selection of parental rice varieties during rice breeding process followed by conclusions.

2. FORMATION OF CONCEPTUAL OVERLAPPING CLUSTERS

A novel clustering technique COCA [7] developed by the authors, is used to pool up rice varieties which are resistant to certain diseases and pest communities. COCA insists on cluster formation based on the semantics of data objects. COCA performs associative mining for formation of conceptual clusters each containing members with similar characteristics represented by colossal patterns. The clusters identified by COCA are overlapping and non exhaustive. At the same time, COCA forms highly distinct clusters while ensuring maximum coverage of data objects.

Definition 1 (Overlapping Clustering)

Overlapping clustering distributes a set of data objects into various clusters based on their similarity such that an object which is similar to the members of more than one cluster becomes a member of all those clusters. For example, let O= {o1, o2 … o9} be data objects. Based on the similarity among the objects, the objects of O may be distributed into two overlapping clusters C1 and C2, where C1= {o1, o2, o4, o5, o8}, C2= {o1, o3, o5, o7, o9} with o1 and o5 common to both clusters. This clustering solution is non-exhaustive as there is an object o6 which is not covered in any one of the clusters.

Definition 2 (Conceptual Clustering)

Let O= {o1, o2 … on} be the set of data objects. Conceptual clustering finds a set of descriptions {d1, d2… dk} where k < n such that

$$\bigcup_{i=1}^{k} g(di) = S \subseteq O$$

Where g (di) represents the objects of i^{th} conceptual cluster described by di . Certain descriptions resulted by conceptual clustering may involve common characteristics and is referred to as pattern overlap which is to be minimized.

Definition 3 (Pattern Overlap) :

di, dj are non null sets of items constituting the i^{th} and i^{th} pattern/description. Their overlap is defined as,

$$\text{Pattern Overlap}(di, dj) = \frac{|\ di \cap dj\ |}{\text{Min}\{|\ di\ |, |\ dj\ |\}}$$

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

Numerator is the number of common items shared by i^{th} and j^{th} patterns. Accordingly the overlap of i^{th} pattern to itself is 1, representing maximum overlap and 0 is the minimum overlap.

A pattern with more than 50% pattern overlap with an existing pattern is considered useless as it is expected mostly to confine to objects that are already covered by the previous descriptions. Conceptual clusters are possibly overlapping as there are objects which satisfy more than one description corresponding to multiple clusters. However, overlapping clustering solution is desired to have maximum coverage with minimum number of descriptions. Hence a description is considered useful only if it covers a distinct group of objects. COCA algorithm achieves conceptual overlapping clustering using association mining techniques. The frequent itemsets/patterns identified are taken as descriptions for formation of conceptual overlapping clusters. The transaction ID lists for each of the pattern forms a conceptual cluster corresponding to the description of a pattern.

Definition 4 (Resistivity list:)

The resistivity of a rice variety to various diseases and pests expressed as a list of pests to which the rice variety can withstand is referred to as resistivity list of a rice variety. In Association mining [1] terminology, a transaction refers to a resistivity list, an item refers to a pest or disease and frequent itemset/pattern refers to resistive behavior of a significant group of rice varieties.

3. METHODOLOGY

Given the transactional database TD, maximum number of clusters M and tolerable noise percentage, the Conceptual Overlapping Clustering Algorithm (COCA) forms conceptual overlapping clusters as shown in figure 1. The transaction database and the patterns are represented in the vertical format to avoid multiple database scans to gather the transactions covering an itemset or pattern or description as it grows. Frequent 2-itemsets along with their Tid-lists form the initial pool of patterns which will be merged for forming colossal patterns, using colossal pattern mining algorithm developed by the authors.

The algorithm takes four times the noise percentage as the minimum support threshold. Significant descriptions are formed by selecting the colossal patterns based on their length and pattern overlap as described in lines 9 to 32 of algorithm 1. While the lines 9 to 16 forms the first description, the loop described in lines 17 to 32 adds successive descriptions and reduces the transaction database. Overlapping clusters are formed from each significant pattern as described in algorithm 2. If the size of the reduced transaction database is greater than 25% of its original size, repeat the whole process on reduced transaction database to form initial pool of patterns, colossal patterns and thereby new descriptions and add them to D as shown in line number 34. If it is less than 25% of its original size, the minimum support threshold coincides with noise threshold. Hence the resulting colossal patterns will not suggest any significant descriptions and the process terminates. The listing of COCA is shown as algorithm 1.

Algorithm 1 (For formation of descriptions D)

Input: minsuppercent, Dbsize, noise, tow (0<tow<1), M Maximal number of patterns to mine.
Output: Clusters representing groups of rice varieties and frequent itemsets representing common characteristics shared by rice varieties grouped into a cluster.

```
1: IP = Φ; size=0; trans_count= Dbsize ;cp = Φ;
2: minsup = (minsuppercent/100) * trans_count;
3: do while (minsup > noise)
4:  { IP = gettwoitems (minsup);
5:     size = checkIpoolsize ();
6:     if (size<2) break;
7:     else
8:     {  cp← getcolossalpatterns (minsup, tow, M);
9:     i=1; D= Φ;
10:   di ← MaxlenPattern (cp);
11    di.frequency = di.TID.size ();
12:   deletefromcp (di);
13:   D ← D U di ;
14:    trans_count = trans_count - di.frequency;
15:   minsup = (minsup/100) * trans_count;
16:   updatetransactiontable (di, di TID);
17:    do while (cp != Φ)
18:    {   i=i+1;
19:   di ← MaxlenPattern (cp);
20:   di.frequency = di.TID.size ();
21:   deletefromcp (di);
22:    j=1;
23:    Repeat for each pattern ISj Є D,
24:  { Patternoverlap (ISj,di) =      |  ISj ∩ di     |
                                     Min {|ISj|, | di |};
```

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

25: j=j+1
26: Until Patternoverlap (ISj,di) >50% }
27: if (j==i)
28: { D ← D U di ;
29: trans_count = trans_count - di.frequency;
30: minsup = (minsup/100) * trans_count;
31: updatetransactiontable (di, di TID);
32: } } }//end of whiles
33: formOverlapclusters (D);
34: if trans_count ≥ (¼ * Dbsize) then goto step 1.
35: End.

Finding significant descriptions:

Figure 2 shows the method of finding significant descriptions. The step-wise process of finding significant descriptions from the colossal patterns is described below:

Find the lengthiest pattern and initialize set of descriptions 'D' to contain it. The transaction database is reduced by removing the transactions covering newly discovered descriptions. The new minimum support threshold is calculated for the updated data. Successive descriptions are found by examining the colossal patterns one by one in the descending order of their lengths as described below:

Select a pattern which is distinct from the descriptions in D such that its overlap is less than fifty percent for all di in D and insert it into D. The transaction database is reduced by removing the transactions covering newly discovered descriptions. The new minimum support threshold is calculated for the updated data. Repeat above process until minimum support threshold is less than noise. When the support threshold falls below tolerable noise level, no more significant descriptions can be formed from the same set of colossal patterns. A new set of colossal patterns are required to cover the remaining transactions in the reduced transaction database if it is more than 25% of its original size.

Algorithm 2 (Cluster generation)

Once the descriptions are formed, transactions covering each description constitute a cluster.
The i^{th} cluster is the intersection of TID lists of items constituting description di which may be written as below:

$$Ci = \bigcap_{j=1 \text{to } n} TID\ ij$$

Where TID ij is the list of transaction id's of j^{th} item in i^{th} description.
Transactions covering more than one description are made members of multiple clusters and hence form overlapping clusters. It may be observed that, the vertical format representation of items, patterns and descriptions simplifies the computation of intersection while cluster formation.

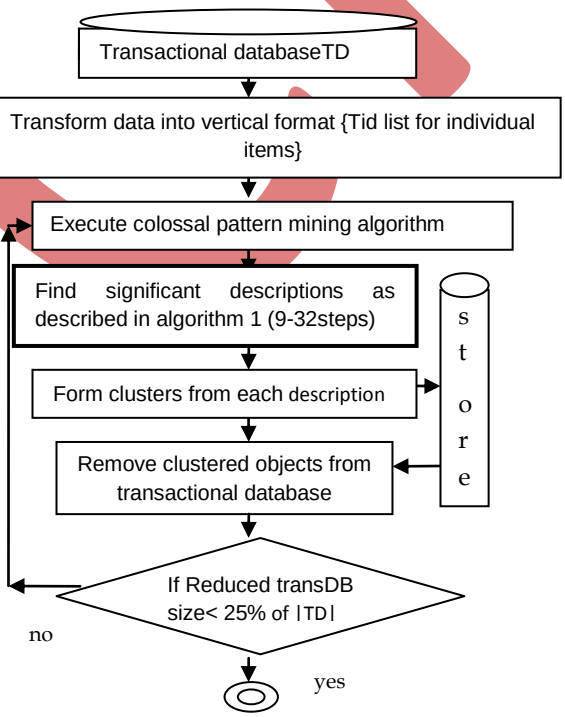

Figure 1: Block diagram of COCA process

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

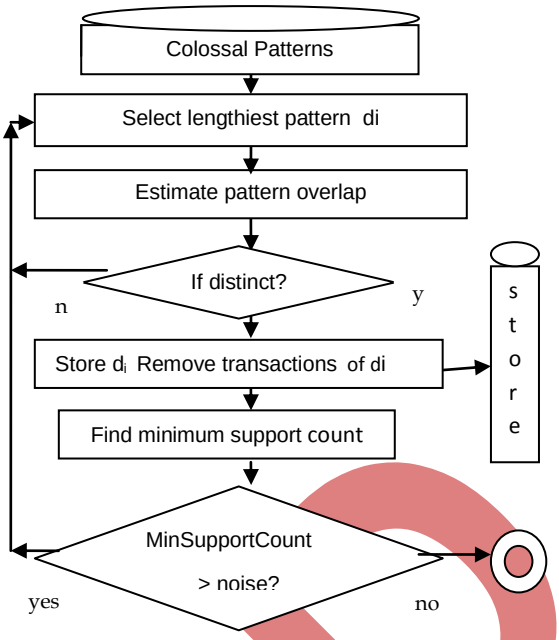

Figure 2: Finds significant descriptions

4. EVALUATION METRICS

For unsupervised learning, there are a variety of cluster evaluation measures like cohesion and separation which are estimated in terms of proximity measures. These traditional cluster quality measures are not suitable to assess the clustering quality of conceptual overlapping clusters formed from the results of association mining. The following metrics are suggested by the authors to estimate the clustering quality.

Coverage (%):

The percentage of participated data objects out of the total number of data objects gives the coverage percentage of a clustering solution.
It is desirable to have a coverage percentage is nearing 100.

Average pattern length:

The average pattern length is a weighted sum of lengths of various patterns constituting the clustering solution. Let NCCj represents the length of jth pattern and Sj represents the size of the j^{th} cluster. The average pattern length is given as follows:

$\Sigma j (Sj / \Sigma i Si) * NCCj$

For all i and j, starting from 1 to n, where n is the number of patterns.
As the minimum support threshold increases, the coverage increases at the cost of cohesion estimated in terms of average pattern length. A good clustering solution is selected as a tradeoff between the coverage and the average pattern length.

5. EXPERIMENTATION AND RESULTS

Experiments are conducted on a 2.00GHz Intel processor, 1 GB memory, running Windows XP. The algorithm is implemented in Java. The dataset contains rice accessions of National screening nurseries [2], [3] and their resistivity to different pests and diseases in different locations of India. Conceptual Overlapping Clustering Algorithm (COCA) developed by the authors is applied on this agricultural data. The rice varieties which are commonly resistant to certain set of diseases and pests are successfully grouped into clusters, so that the pooled up data could be studied and analyzed in a short time. Each of these clusters form the population from which parent grains can be selected in order to generate an offspring with more or less predictable resistive behavior. We have taken 108 rice varieties and considered the reaction to various diseases and the reaction to insect pests along with their biotypes. Transactions designate rice varieties; items designate various diseases and pests at various locations bearing similar raising beds. From table 1, it can be observed that as minimum support threshold increases, coverage of rice varieties can be maximized and also the runtime is reduced considerably. However, the increased support thresholds results in elimination of most of the colossal patterns which are essential for forming conceptual clusters with high cohesion. It was also observed that clusters formed from colossal patterns contain lesser number of elements compared to those of short patterns.

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

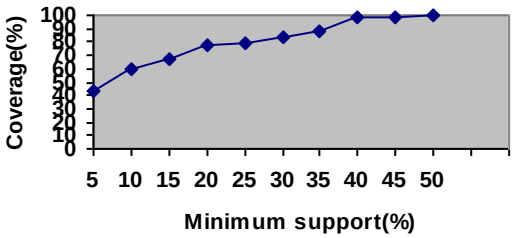

Figure 3: Graph showing the coverage of rice clusters for various minimum support thresholds.

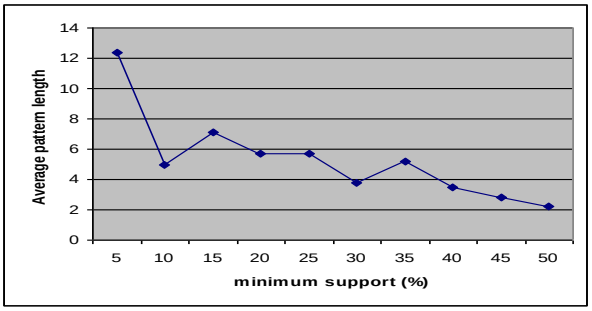

Figure 4: Graph showing the average pattern length for various minimum support thresholds.

The graph shown in figure 3 depicts the coverage % for various rice clusters. The graph shown in figure 4 depicts the average pattern length for varied minimum support thresholds.

6. PROPOSED BREEDING PROCESS

The COCA algorithm is used to cluster rice varieties into groups with common characteristics such as grain quality, resistivity to specific pests and diseases predominantly considered in different regions. Figure 5 gives an overview of breeding process for agricultural products like rice. F1 in the figure represents the first generation seed which is developed in the process of breeding from two different parent seeds say A and B and hence may possess new characteristics set which is a combination of those of A and B. The characteristics of F1 seed become more or less predictable when most of the characteristics of A and B are common except for a few. For example, when the goal is to generate a seed with description {c2,c3,…,c9} when there is no proven rice variety with all these characteristics, you may experiment by taking parent A from cluster1 with description {c1,c2,…,c8} and parent B from cluster2 with description Simultaneously several experiments can be conducted with the cross product of members of cluster1 and cluster2 to increase the chances of developing more than one off springs with the desired characteristics set. F2-F6 represents the second generation to sixth generation seeds which will be developed and tested to check the adaptability and sustenance of the newly developed seeds to regional conditions. Hence COCA facilitates selection of parental rice varieties for forming new breeds with predictable characteristics and avoids the expensive and time consuming trial and error selection methodology.

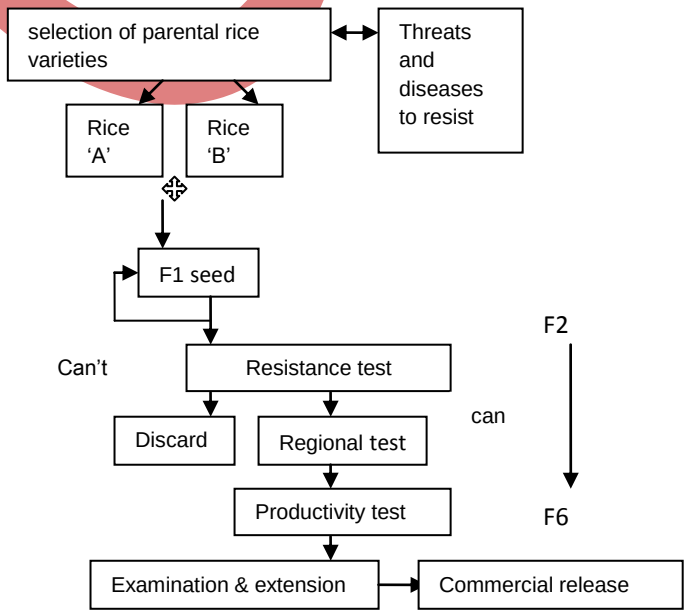

Figure 5: Proposed breeding process

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

Another advantage of maintaining a repository of rice varieties grouped in terms of clusters is to facilitate selection of optimal rice variety from a group of rice varieties that can withstand a set of local environmental conditions.

Table 1: COCA clustering results.

Minimum support (%)	Avg length of pattern	No. of clusters formed	Size of clusters Max Min	Coverage (%)	Run time
50	2.2	5	66 31	100	4 sec
45	2.8	6	50 27	99	5sec
40	3.5	6	59 31	98.1	3sec
35	5.2	6	44 3	88	28 sec
30	3.8	6	39 9	84.2	26 sec
25	5.7	8	27 3	78.7	1min 21 sec
20	5.7	7	27 10	77.7	2min
15	7.1	9	17 3	66.6	5.6min
10	8.8	15	14 4	57	13.7min
5	12.4	13	6 3	43.5	26.36min

7. CONCLUSION

All over the world, rice breeding researchers work to develop high quality and more productive rice varieties. In this paper, the authors have proposed an algorithm for conceptual overlapping clustering (COCA) of rice varieties in order to segregate them into groups with common resistive behavior. COCA identifies the cluster descriptions making use of colossal patterns extracted from rice varieties represented as transactions in association mining terminology. Clusters formed from colossal patterns are highly cohesive with possibly lesser cardinality. Hence authors suggest to apply this COCA algorithm for intelligent parental selection in the rice breeding process instead of random shuffling trials which saves time and laboratory expenses.

ACKNOWLEDGMENTS

We are thankful to Directorate of Rice Research, Hyderabad, A.P for providing us the data [2 & 3]

REFERENCES

[1] Agrawal R., T.Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In SIGMOD'93, pp 207-216, Washington, D.C., 1993.

[2] Anonymous 2008 Progress report 2008 Volume 1: Varietal improvement, by All India Coordinated Rice Improvement Programme, DRR HYD-30 A.P

[3] Anonymous 2008 Screening Nurseries. All India Coordinated Rice Improvement Programme, DRR HYD-30 A.P

[4] Daya C. Wimalasuriya, Sridhar Ramachandran, and Dejing Dou1 Clustering Zebrafish Genes Based on Frequent-Itemsets and Frequency Levels PAKDD 2007, LNAI 4426, pp. 912–920, 2007.

[5] D. Madhavi & M. Shashi, An Efficient Algorithm to Mine Prodigious Frequent Patterns In: Proceedings of the 2008 international international conference IISA'08, UCONN, New York.

[6] D. Madhavi & M. Shashi, Mining Colossal Patterns from Gene Expression Data, "International Journal of Intelligent Information Processing" Serial publications, Vol. 3 No l. June 2009 issue of the journal.

[7] D. Madhavi & M. Shashi, Conceptual Overlapping Clustering Algorithm, In proceedings of International conference IISA 2010, A.U, Visakhapatnam, A.P, India.

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

[8] D. Madhavi Ph.D Thesis titled “An Efficient Algorithm for Colossal Pattern Mining to generate conceptual overlapping clusters “, 2011 submitted to Acharya Nagarjuna University, Guntur Dt., A.P, India.

[9] Guillaume Cleuziou, OKM, 978-1-4244-2175-6/08/IEEE.

[10] J.S.Bentur, C.Cheralu & P.R.M. Rao, Monitoring virulence in Asian rice gall midge populations in India, Entomologia Experimentalis et Applicata 129:96-106 2008, The Netherlands Entomological society.

[11] Ke wang et al, clustering transactions using large items 1999.

[12] Zhu.F, Han, Yu Mining colossal Frequent patterns by core Pattern Fusion In: Proceedings of the 2007 international conference on DataEngineering, Istanbul, Turkey.

D.Madhavi received her M.E. Degree in 1996 (Computer Engineering) with distinction ndhra University. She received her Ph.D in 2011 from Acharya Nagarjuna University. She is presently working as essor in the department of Information Technology, GMRIT, Rajam, Srikakulam Dt, Andhra Pradesh, India. Her ch interests include Data Mining, Information Retrieval, and Software Engineering.

shi received her B.E. Degree in Electrical and Electronics and M.E. Degree in Computer Engineering with distinction ndhra University. She received Ph.D in 1994 from Andhra University and got the best Ph.D thesis award. She is working as a professor in the department of Computer Science and Systems Engineering at Andhra University, Andhra Pradesh, India. She received AICTE career award as young teacher in 1996. She is a co-author of the Indian Edition of text book on “Data Structures and Program Design in C” from Pearson Education Ltd. She published technical papers in National and International Journals. Her research interests include Data Mining, Artificial intelligence, Pattern Recognition and Machine Learning. She is a life member of ISTE, CSI and a fellow member of Institute of Engineers (India).

P.Ram Mohan Rao received his Ph.D in 1990 from Tamilnadu Agricultural University, Coimbatore. He worked as principal scientist ARS Ragolu, Srikakulam district, A.P, India. He published echnical papers in National and International Journals. He received ‘Sankranthi puraskaralu’state level Best Scientist award from Government of Andhra Pradesh for his contribution to rice research.