World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:6, 2008

Cluster Algorithm for Genetic Diversity

Manpreet Singh, Keerat Kaur, and Bhavdeep Singh

***Abstract*—**With the hardware technology advancing, the cost of storing is decreasing. Thus there is an urgent need for new techniques and tools that can intelligently and automatically assist us in transferring this data into useful knowledge. Different techniques of data mining are developed which are helpful for handling these large size databases [7]. Data mining is also finding its role in the field of biotechnology. Pedigree means the associated ancestry of a crop variety. Genetic diversity is the variation in the genetic composition of individuals within or among species. Genetic diversity depends upon the pedigree information of the varieties. Parents at lower hierarchic levels have more weightage for predicting genetic diversity as compared to the upper hierarchic levels. The weightage decreases as the level increases. For crossbreeding, the two varieties should be more and more genetically diverse so as to incorporate the useful characters of the two varieties in the newly developed variety. This paper discusses the searching and analyzing of different possible pairs of varieties selected on the basis of morphological characters, Climatic conditions and Nutrients so as to obtain the most optimal pair that can produce the required crossbreed variety. An algorithm was developed to determine the genetic diversity between the selected wheat varieties. Cluster analysis technique is used for retrieving the results.

***Keywords*—**Genetic diversity, pedigree, nutrients.

I. Introduction

THE data mining process involves applying a data mining method to data so as to extract patterns of information. Clustering is the most important unsupervised data mining method. It involves finding structure in a collection of unlabelled data so that inter cluster similarity is minimized and intra cluster similarity is maximized. Clustering can be Exclusive or Overlapping, Distance based or Conceptual based on the way data items are bound together.

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects, which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Manpreet Singh is with the Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana (e-mail: mpreet78@yahoo.com).

Keerat Kaur is M.Tech Student in the Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana (e-mail: keerathere@yahoo.com).

Bhavdeep Singh is with Logan, Britton, Shikago, USA (e-mail: singhbhavdeep@hotmail.com).

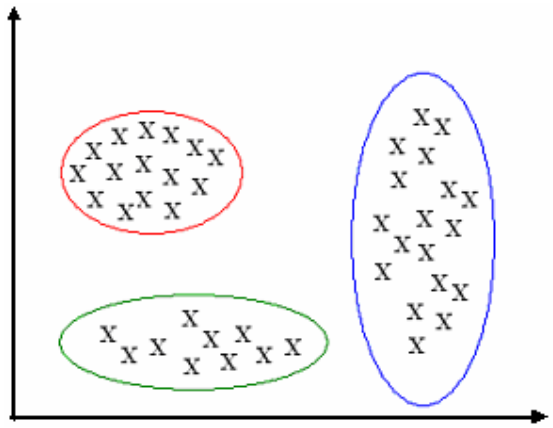

Fig. 1 Clustering

Thereby, clustering is the grouping of data items based on their similarity. [2][3]

A. Types of Clustering

- **Distance Based Clustering**: Two or more objects belong to the same cluster if they are "close" according to a given distance. This is called distance-based clustering.
- **Conceptual Clustering:** Two or more objects belong to the same cluster if they are defined by a common concept. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

B. Goals of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion, which would be independent of the final aim of the clustering. Consequently, it is the user, which must supply this criterion, in such a way that the result of the clustering will suit their needs.

For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection).

C. Clustering Algorithms

Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering

- Hierarchical Clustering
- Probabilistic Clustering

In the first case, data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. A simple example of that is shown in the Fig. 2, where a straight line on a bi-dimensional plane achieves the separation of points. On the contrary the second type, the overlapping clustering (Fig. 1), uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.

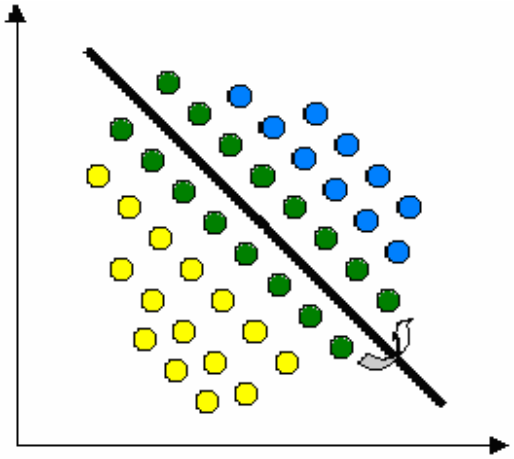

Fig. 2 Overlapping Clustering

Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Finally, the last kind of clustering uses a completely probabilistic approach.

II. Crop Hybridization

In traditional terms, hybridization refers to the union of the male and the female gamete to produce a zygote. In plant science, hybridization also refers to the crossing or mating of two plants.

In his quest to find more variability, man started experimenting with hybridization of plants so as to achieve the perfect plant type. This process was actually the beginning of expedited evolution since it led to the formation of new plant types artificially or due to human intervention at a much faster pace than it would have happened in nature. For example, the bread wheat that we eat today has taken about 500 years to evolve to its present form through human intervention. This form of wheat would have taken thousands of years to evolve had it been left to the natural evolution process.

III. Genetic Diversity

Genetic diversity is the variation in the genetic composition of individuals within or among species. Genetic diversity depends upon the pedigree information of the varieties. The genetic diversity of a species, an ecosystem, or in fact anything living is a crucial indicator as to how life is coping over time in the environment that it exists in.

A. Role of Genetic Diversity in Agriculture

Genetic diversity is the basis of the ability of organisms to adapt to changes in their environment through natural selection. Populations with little genetic variation are more vulnerable to the arrival of new pests or diseases, pollution, changes in climate and habitat destruction due to human activities or other catastrophic events. The inability to adapt to changing conditions greatly increases the risk of extinction. Gene conservation management aimed to save adaptive genetic diversity should be based on the knowledge of the genetic basis of adaptation.

Crop genetic diversity is not just a raw material for industrial agriculture; it is the key to food security and sustainable agriculture because it enables farmers to adapt crops suited to their own site specific ecological needs and cultural traditions. Without this diversity, options for long term sustainability and agricultural self-reliance are lost. Genetic variability is required to achieve genetic gains in a breeding program. Monitoring of genetic diversity can form a basis for rational correction of breeding programs and the strategies in plant industry.

The characterization of genetic variability and an estimate of the genetic relationship among varieties are essential to any breeding program. Obtaining accurate estimates of genetic diversity levels among germ-plasm sources many increase efficiency of breeding efforts to improve crop species.

Plant breeding deals with high-yielding genotypes and parental selection is the first step in any plant-breeding program. However, how best to choose parents of these genotypes, remains an unsolved question. Research on parent selection may be approached in two ways: *a priori* and *a posteriori* choice. The former consists of selection methods based on *per se* parent performance, such as mid-parental value, divergence according to coefficient of parentage, character complementation, multivariate analysis and parental distances, least squares, parental complementation, and ideal genotype. A long period of time is necessary to choose parents by the second way, especially in perennial plants. [1]

Theoretical arguments and empirical results with wheat indicate that the probability of recovering a superior progeny genotype is greater if both parents are similar in performance as opposed to one parent being inferior. Yet, genetic diversity between parents is necessary to derive transgressive segregates from a cross [4][5].

IV. Genetic Diversity Measures

Evolutionary or ecological measures of genetic diversity focus particularly on genetic similarity or difference between different species. Most studies of crop genetic diversity are based on the similarity or difference between different crop populations within the same crop species.

A. Spatial Diversity Measures

Spatial diversity i.e. the diversity within a given geographical area maybe " the most commonly recognized concept of diversity " Two concepts are often used in spatial measures of genetic diversity. "Richness" refers to a simple count measure, for example of the number of varieties of a particular crop species planted in a given area. "Abundance" is a measure of the evenness of the spatial distribution of elements of the set being considered. For example, suppose the same ten crop varieties are planted in two identical regions. In one region, each variety is planted on one-tenth the area, but in the other region one variety is planted on 91 percent of the area and the other nine varieties occupy one percent each. By a simple count measure (such as richness), the two regions are equally diverse, but introducing abundance would suggest the first region is more diverse than the second. This, along with the fact that named varieties may be very similar genetically, is why simply counting numbers of varieties is likely to be an inadequate measure of crop genetic diversity. Simple diversity indices that reflect varietal distribution (thus partially capturing the concepts of richness and abundance), include the proportion of area planted to the most popular variety or given number of varieties (equivalent to concentration measures used in the industrial organization literature.) A related index is the number of varieties covering a given percentage of total crop area.

B. Temporal Diversity

It was observed that in a number of scientifically bred crops, temporal diversity (or diversity through time) has replaced spatial diversity as one means of maintaining or even raising resistance or tolerance to pests and diseases. Temporal diversity depends on maintaining breeding effort by humans. Faster varietal turnover might be expected to be associated with increased temporal genetic diversity, but like pedigree-based measures, varietal turnover is more a measure of the output of a plant-breeding program than of genetic diversity. Newly released varieties might be genetically somewhat dissimilar to older varieties, or they might be very closely related genetically. Time-series of spatial diversity measures could provide useful information about temporal change in diversity, but such a series would not strictly measure "temporal diversity." More formal assessment of temporal genetic diversity could be made by statistically testing differences between genetic distance measures over temporal samples.

C. Measures of Relationships between Varieties

Other indices of genetic diversity are built up from measures of "genetic distance," i.e., the degree to which varieties or species differ genetically. To a certain extent such measures address the problem raised by simply counting named varieties that may be very similar genetically. Genetic distance indices can be calculated based on observations of different crop characteristics, including morphological indicators such as plant height, grain weight, and so on. As indicated, morphological indicators have the advantage that they may be closely linked to the traits on which farmers base their decisions, but the disadvantage that they are often influenced by environment and multiple genes, and therefore not reflective of genetic distance at the chemical (enzyme) or molecular (DNA) level. Genetic distance indices have perhaps most commonly been applied to this biochemical information. The use of biochemical and molecular markers requires systematic physical sampling as well as laboratory time and materials, and as a result can be quite costly. An alternative approach to measuring genetic distance between varieties, at least for scientifically bred crops with documented pedigrees, is based on comparison of the heritage of pairs of varieties i.e. using pedigree information. This approach uses the coefficient of diversity (COD), which equals 1 – the coefficient of parentage (COP). The COP is a pair-wise comparison based on pedigree analysis. COD/COP analysis is less costly than analysis of proteins or molecular methods, but it also has some disadvantages.

D. Building Diversity Indices

Genetic distance indices measure differences between different crop varieties or species, but they themselves do not measure overall genetic diversity. Some tree-based measures and other measures based on matrices of similarity coefficients, permit weighting to reflect the distribution of crop varieties.

E. Measures of Plant Breeding Activity Using Genetic Resources

A number of other measures have been applied to the study of genetic resources, but they usually refer to aspects of a scientific plant breeding program or the development of such a program from initial crosses involving landraces, rather than to direct measures of genetic diversity. These include numbers and origin of landraces in the ancestry of the varieties being studied, or the number of breeding generations since the initial cross numbers of distinct parental combinations and numbers of unique landrace ancestors per pedigree or coefficient of parentage (COP) based measures. Note that all of these pedigree-based measures are less useful in a crop, such as corn, that may not always follow a strict pedigree breeding system, or in crops for which pedigrees are partially or completely private for proprietary reasons.

V. Predicting Genetic Diversity

Various methods, including pedigree and DNA marker analyses, have been used to quantify genetic diversity among genotypes. Coefficient of parentage (COP) indirectly measures genetic diversity among cultivars by estimating, from pedigree records, the probability that alleles at a locus are identical by descent; however, assumptions made when calculating COP regarding relatedness of ancestors, selection pressure, and genetic drift are generally not met. Although molecular marker analyses directly measure DNA sequence variation among genotypes, results may be confounded by biased or incomplete genome coverage, detection of co migrating non homologous fragments, or high cross- over frequency between markers used in the evaluation and linked genetic material. In addition, low polymorphism levels are

typically detected among wheat cultivars therefore; obtaining accurate DNA marker-based diversity estimates may require intensive screening efforts.

Despite concerns about accuracy of diversity estimates generated by both methods, pedigree information and DNA marker data have been used to assess genetic relationships among inbred cultivars of several crops including wheat.

Genetic Diversity can be predicted:

- Using Pedigree Information
  - COP
- Using Genetic Information
  - Molecular Markers

A. Genetic Diversity Determination Using Pedigree Information

Pedigree: A representation of the ancestry of an individual or family; a family tree.

- **Morphological features:** Properties related to the external structure of soil (such as color and texture) or of plants.

An approach to measuring genetic distance between varieties, at least for scientifically bred crops with documented pedigrees, is based on comparison of the heritage of pairs of varieties This approach uses the coefficient of diversity (COD), which equals 1 – the coefficient of parentage (COP). The COP is a pair wise comparison based on pedigree analysis.

COD/COP analysis is less costly than analysis of proteins or molecular methods, but it also has some disadvantages:

- It ignores the possibility that alleles could be identical even without common heritage.
- It relies on the assumption that the ultimate ancestors that are recorded in a pedigree are unrelated, which may resources, but they usually refer to aspects of a scientific plant breeding not be true; and
- It assumes that "each parent contributes equally to offspring, despite the effects of recurrent selection and random genetic drift"

VI. Problem Formulation

This project is undertaken to cover the data mining applications in existing knowledge buried in large biological texts and to infer results from them. Morphological characters are the various parameters related to the wheat varieties. If we take the filter value any desired morphological characteristics as input we get a list of varieties satisfying the conditions. This list is processed, taking a pair of varieties at a time to find out the most optimal and probable pair of genetically diverse varieties. The results are shown graphically, depicting the genetic diversity among the varieties based on the pedigree levels. We also get a percentage probability of getting the required hybrid breed as an output.

VII. Solution Methodology

Database was created for the different varieties of wheat. It contains the pedigree information and the morphological characters for the different varieties of wheat. The model makes use of Clustering as the data mining method and is based on the concept of overlapping and conceptual clustering. First of all, the varieties are selected on the basis of morphological characters, climatic conditions, nutrients etc. This information is important to develop a variety with particular useful characters. Now we have to determine the genetic diversity between the varieties. For this, we need to compare the pedigree information (parentage) of the varieties.

Fig. 3 shows the family trees originating from wheat varieties C306 and BW11.

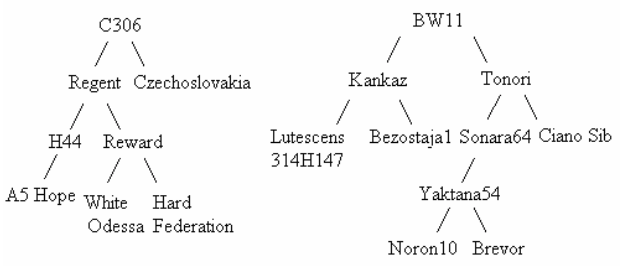

Fig. 3 Pedigree trees

The varieties are selected form a list. Different pairs of varieties are analyzed to calculate the probability percentage of obtaining the desired variety.

The formula used for calculating the results is given below:

$$P_{i+1} = P_i + \left(\frac{\eta_i + 50/L_{i+1}}{C_{i+1}}\right) D_{i+1}$$

Where

η_i is constant for pedigree level 'i' indicating the effect of that level on the genetic diversity.

L_i indicates the hierarchical level under observation.

C_i and D_i correspond to the number of varieties in a level and the number of distinct varieties in a level respectively.

P_i is the percentage probability for the varieties to be genetically diverse upto level 'i'.

So, higher the value of P_i, greater will be the genetic diversity between the crop varieties.

A graph is plotted between pedigree levels and genetic diversity utilizing the formula for P_i.

VIII. Results and Conclusion

This model was developed so as to incorporate knowledge discovery from large databases in the field of bioinformatics. The project is mainly designed to find out the most optimal and probable parent varieties for a desired crossbreed variety. Fig. 4 shows the representation of Genetic Diversity in percentage and also in graphical form. The filter values are entered as height greater than or equal to 45, yield greater than or equal to 45 and DTF as less than 45. These input values generate the lists as shown. The user selects the varieties *Pegora* and *H44.* The result is shown in percentage and also shown graphically for the different levels of parentage. The database is provided for the morphological characters, climate conditions and nutrients for the given varieties. The

parameters for the varieties are shown in the form of list. Every time user makes an entry or change in the database, the list is updated accordingly.

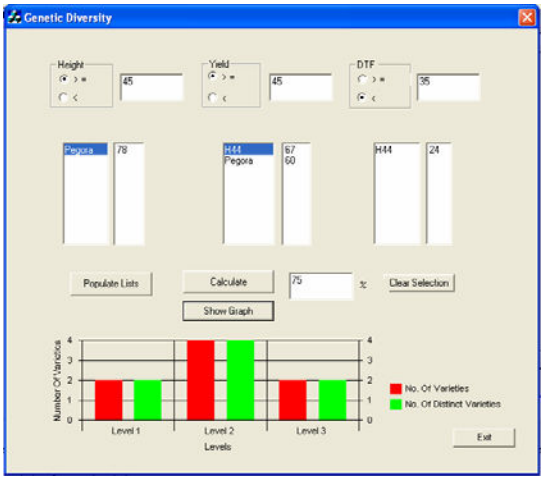

Fig. 4 Representing Genetic Diversity

REFERENCES

[1] Dias, Picoli, Rocha and Alfenas "*A priori choice of hybrid parents in plants*", Genetics and Molecular Research. Vol. 12, 2004, pp 116-130.

[2] Fan, Jianhua and Li, Deyi "*Overview of data mining and knowledge discovery*" Journal of Computer Science and Technology. 13 (4), 1998, pp 348-368.

[3] Fayyad, Usama; Stolorz and Paul "*Data mining and KDD: Promise and challenges*" Generation Computer Systems. 13 (2-3), 1997, pp. 99-115

[4] Jagdeep Singh "*Development of Biotechnology Information System using a Web Server*" M.Tech Thesis PAU, Ludhiana, Punjab, India, 2002, pp 1-40.

[5] Manpreet Singh "*Development of Data Mining model for bioinformatics system*" M.Tech Thesis PAU, Ludhiana, Punjab, India, 2003, pp 1-30.

[6] Raghavan, Vijay V.; Deogun, Jitender S. and Server, Hary "*Introduction to Data Mining*" Journal of the American Society for Information Science 49 (5), 1998, pp 397-402.