Learning Dynamics of Pesticide Abuse through Data Mining

Ahsan Abdullah
National University of Computers & Emerging Sciences, Islamabad, Pakistan,
ahsan@nu.edu.pk

Stephen Brobst
Teradata Division, NCR, Dayton, OH, USA

Ijaz Pervaiz
Directorate of Pest Warning & Quality Control of Pesticides, Punjab, Multan

Muhammad Umer, Azhar Nisar
National University of Computers & Emerging Sciences, Islamabad, Pakistan

Abstract

Recent studies by agriculture researchers in Pakistan have shown that attempts of crop yield maximization through pro-pesticide state policies have led to a dangerously high pesticide usage. These studies have reported a negative correlation between pesticide usage and crop yield in Pakistan. Hence excessive use (or abuse) of pesticides is harming the farmers with adverse financial, environmental and social impacts. In this work we have shown that how data mining integrated agricultural data including pest scouting, pesticide usage and meteorological recordings is useful for optimization (and reduction) of pesticide usage. The data used in this work has never been utilized in this manner ever before. We have performed unsupervised clustering of this data through Recursive Noise Removal (RNR) heuristic of Abdullah and Brobst (2003). These clusters reveal interesting patterns of farmer practices along with pesticide usage dynamics and hence help identify the reasons for this pesticide abuse.

Keywords: Data Mining, Agriculture, Pesticide Abuse, cotton, Unsupervised Clustering.

1. Introduction

Pakistan is one of the five major cotton growing countries in the world. Almost 70% of world cotton is produced in China (Mainland), India, Pakistan, USA and Uzbekistan (Chaudhry 2000). Pakistan is worlds 7th most populous country; anticipating population growth, in 60's and 70's pesticides were identified as means for increasing production, as a positive correlation is believed to exist between yield and pesticide usage. However, FAO (2001) have reported the existence of a negative correlation between pesticide usage and yield in Pakistan (Figure 1). A marked increase in yield loss while the pesticide usage is on the rise has created a complex situation.

Excessive use of pesticides is harmful in multiple ways. On one hand, farmers have to pay more for the pesticides, while on the other, increased pesticide usage develops immunity in pests, thus making them more harmful to the crops.

Excessive usage of many pesticides is also harmful for the environment and hazardous to human health.

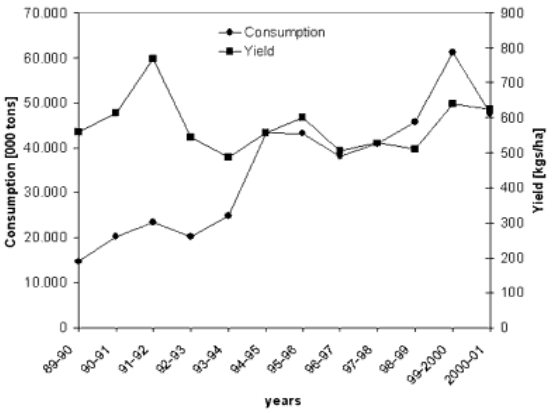

Figure 1: Yield and Pesticide Usage in Pakistan (FAO 2001)

Pesticide usage can be reduced by looking for the conditions in which the usage is optimum and trying to dig out for the circumstances that lead the farmers to an excessive pesticide usage. This can best be done by looking for patterns in the past happenings. In this paper we have shown how data mining can be successfully applied for this purpose. We applied an indigenously developed data mining tool based on our "Clustering by Recursive Noise Removal" technique from Abdullah and Brobst (2003) to the pest scouting, pesticide usage and meteorological data from Pakistani cotton fields.

Rest of the paper is organized as follows; Section 2 gives background of our work, Section 3 presents a brief review of related work, Section 4 and 5 describe structure and working of RNR algorithm, RNR application and discussion on results are presented in Section 6, while conclusions are summed up in Section 7.

2. Working Scenario

To learn from the past one needs a detailed record of the past. In our case details of past pest situations, pesticide usage history and farmer demographics was required i.e. pest scouting data. Pest scouting is a systematic field sampling process that provides field specific information on pest pressure and crop injury.

This data was obtained from the Directorate of Pest Warning and Quality Control of Pesticides (DPWQCP) Government of Punjab. Since 1984 the said directorate has been collecting and recording pest scouting data on a weekly basis from mostly 1800 random locations. For our study we have selected the province of Punjab, because it is a major producer of the cotton crop (Federal Bureau of Statistics 2002).

Pest scouting data by itself can be termed as a "Gold mine" of data and coupling it with pesticide usage and meteorological data can provide an excellent insight into the dynamics of past situations and their outcomes. Looking at the history of agriculture research in Pakistan, this is the first work that has utilized the true potential of this data. In Pakistan the Pest scouting data has never been digitized and until now it was impossible for any researcher to use it for an in depth analysis. As a pilot project we implemented a data warehouse using two years of pest scouting, pesticide usage and meteorological data consisting of 200 typed sheets and each record consisting of 40 attributes. The data warehouse was implemented after digitization, cleansing and integration of data generated by multiple disparate sources. In the first phase of implementation we covered district Multan only, which is one of the thirty four districts of Province of Punjab. District Multan is the hub of cotton production and cotton related activities in the Province (Federal Bureau of Statistics 2002).

Figure 2: Map of Pakistan

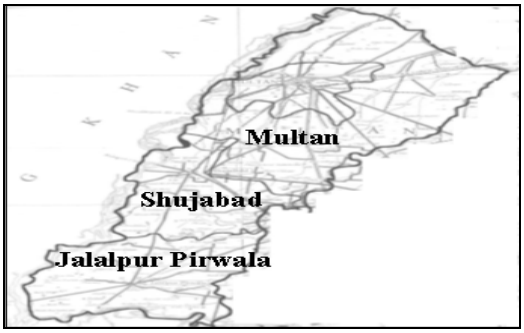

Figure 3: Area Under Study; district Multan.

Quality and validity of the underlying data is the key to meaningful and authentic analyses. After ensuring a satisfactory level of data quality (based on cost-benefit trade-off analysis) it is extremely important to somehow judge the validity of data that a data warehouse constitutes. We applied some very natural checks for this purpose i.e. checking if the data is in agreement to certain well known agricultural phenomenon such as relation between pest and predator populations or relation between predator population and pesticide sprayed etc. We found this data to be completely in agreement to such phenomenon. For instance figure 4 shows inverse relationship between predator population and pesticide spray.

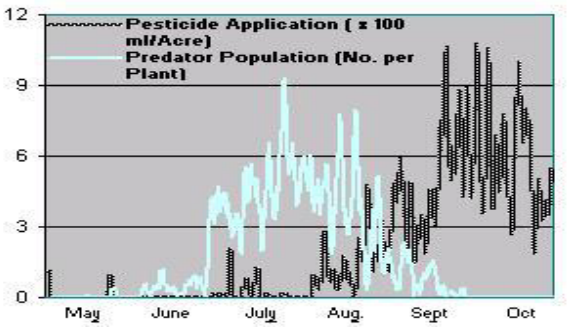

Figure 4: Y 2001 Pesticide usage vs Predators

3. Related Work

Data mining applications have quite recently found their way into agricultural research and a lot of activity can be seen in this area such as (Bertis et al 2001, Cunningham and Holmes 2001, Harms et al 2001, Scherte 2002, Yang et al 1999).Voss et al (2003) gives the detail of GIMMI project which is aiming at providing a one-stop and integrated access to the assessment of pesticide leaching into soil and groundwater. Several IT tools including data mining are to be implemented as part of this project. Avesani et al (2001) describes a new approach for acquisition and preprocessing of agricultural data mining.

Christensen and Di Cook (1998) describe simple numerical methods to establish the relationship between 10 soil characteristic variables and corn yield. Yang et al (1999) used remote sensing techniques in conjunction with AI neural networks to identify weeds in corn fields. Nuyen et al (2001) demonstrates use of data mining techniques on images to identify trash in the ginned cotton. Scherte (2002) uses a case study approach to help understand how data mining could be used in the manufacturing of textiles using SAS. Harms et al (2001) shows the integration of spatio-temporal knowledge discovery techniques into a Geo-Spatial Decision Support System (GDSS). They have used a combination of data mining techniques to find relationships between user-specified target climatic episodes and other climatic events and to predict the target climatic episodes. John D. Hosting et al (2002) have developed a case based advisory system CARMA for range land grass hoper infestation. Though the system uses expert knowledge, but does not perform data mining.

	V1	V2	V3
R1	1	2	3
R2	4	2	3
R3	3	4	1
R4	3	2	1
R5	1	5	3
R6	1	2	3
R7	5	7	8
R8	1	2	4
R9	5	7	8
R10	3	2	1

Figure 5(a): Input Data Set

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
R1	1	-0.5	-0.32	0.65	-0.5	0.5	0.98	0.92	0.92	-0.98
R2	-0.5	1	-0.32	0.65	-0.5	0.5	0.98	0.92	0.92	-0.98
R3	-0.33	-0.32	1	0.65	-0.5	0.5	0.98	0.92	0.92	-0.98
R4	0.65	0.65	0.65	1	-0.5	0.5	0.98	0.92	0.92	-0.98
R5	-0.5	-0.5	-0.5	-0.5	1	0.5	0.98	0.92	0.92	-0.98
R6	0.5	0.5	0.5	0.5	0.5	1	0.98	0.92	0.97	-0.98
R7	0.98	0.98	0.98	0.98	0.98	0.98	1	0.92	0.92	-0.98
R8	0.93	0.92	0.92	0.92	0.92	0.92	0.92	1	0.92	-0.98
R9	0.93	0.92	0.92	0.92	0.92	0.92	0.92	0.92	1	-0.98
R10	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	1

Figure 5(b): Similarity Matrix Corresponding to Input Data

4. RNR Framework

Clustering by Recursive Noise Removal (RNR) heuristic is an unsupervised clustering solution based on recursively removing noise and using different heuristics such as the Median Heuristic, due to Eades and Wormald (1986), MaxSort heuristic of Abdullah (1993) and other heuristics from the domain of crossing minimization. In this section we give some mathematical preliminaries and definitions that require understanding for appropriate appreciation of our work.

4.1. Cluster

A *cluster* is a collection of data elements that are highly similar to one another within the same cluster, but weakly similar from the data elements in other clusters. More formally, let $O = \{O_1, O_2, O_3, \ldots O_n\}$ be a set of n objects and let $C = \{C_1, C_2, \ldots C_k\}$ be a partition of O into subsets; such that $C_i \cap C_j = \varnothing$, $i \neq j$ and $\bigcup_k C_k = O$. Each subset is called a *cluster*, and C is a clustering solution.

4.2. Input data

The input data for a clustering problem is typically given in one of the two forms, as suggested by Han and Kamber (2000):

- Data matrix (or *object-by-variable structure*) is an $n \times p$ matrix, where corresponding to each of the n objects there are p variables, also called measurements or attributes. Usually $n >> p$. Figure 5(a) shows a data matrix.

- Similarity (or dissimilarity) matrix (or *object-by-object structure*) is an $n \times n$ matrix, which contains the pair-wise similarity (or dissimilarity) that is usually computed from the profile data for all pairs of n objects. Fig 5(b) shows the similarity matrix corresponding to the data matrix of Figure 5(a). Note all 1's on the diagonal.

4.3. Model Formulation

RNR algorithm works on a bipartite graph or bi-graph model. First step in its application is to transform given data set into a bi-graph $G_B(V_0, V_1, E)$, where V_0, V_1 are the bipartitions of vertices such that $V_0 \cap V_1 = \varnothing$. E is the edge set such that $e = |E|$, $n = |V_1| = |V_0|$ and density of G_B i.e. $\delta(G_B) = e/n^2$. This transformation is performed by first creating a similarity matrix from the given data i.e. by computing Pearson correlation between every pair of objects. Next step is to *discretize* the input similarity matrix. This is achieved by comparing every cell with a threshold value and replacing with 1 all values which are greater than or equal to the threshold. Remaining cells are set to 0.

We present an example to explain this model. Figure 5(a) presents a simple input data set i.e. a table consisting of 10 rows (R1 to R10) and 3 columns (V1 to V3). Figure 5(b) shows the corresponding similarity matrix ***S*** generated using Pearson's correlation between every pair of rows. Application of *discritization* step on similarity matrix of figure 5(b) with discritization threshold set to 0.9 results in Figure 5(c).

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
R1	1	0	0	1	0	0	1	1	1	0
R2	0	1	0	1	0	0	1	1	1	0
R3	0	0	1	0	0	0	1	1	1	0
R4	1	1	0	1	0	0	1	1	1	0
R5	0	0	0	0	1	0	1	1	1	0
R6	0	0	0	0	0	1	1	1	1	0
R7	1	1	1	1	1	1	1	1	1	0
R8	1	1	1	1	1	1	1	1	1	0
R9	1	1	1	1	1	1	1	1	1	0
R10	0	0	0	0	0	0	0	0	0	1

Figure 5(c): Discritized Similarity Matrix

Considering Fig 5(c) is an array representation of a bipartite graph, such that a 1 in a cell S_{ij} shows presence of an edge between vertices i and j, similarly 0 represents absence of an edge we get the bipartite graph (bi-graph) drawing shown in Fig 6. In this example $|V_0| = |V_1| = 10$, $e = 56$ and $\delta = 0.56$.

A bi-graph drawing (or layout) of G_B may be obtained by placing the vertices of V_0 and V_1 on distinct locations on two horizontal lines in the XY-plane. Now drawing each edge with one straight-line segment which connects the points where the end vertices of the edges are placed.

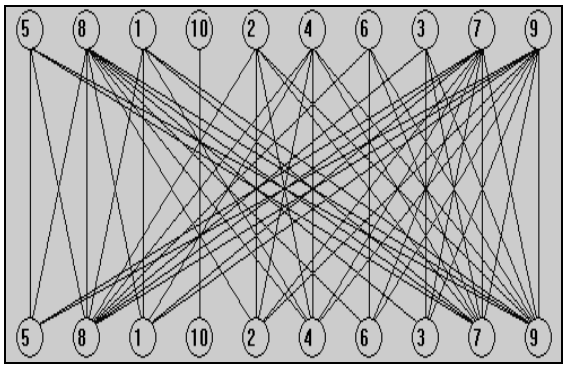

Figure 6: Bi-Graph Drawing

5. Applying RNR

Crossing minimization by permuting the vertices of a graph is an NP-Complete Problem Garey and Johnson (1979) and have a well established following in fields such as graph drawing, and VLSI design automation. Abdullah and Brobst (2003) for the first time used crossing minimization for data mining using Recursive Noise Removal (RNR) heuristic. Here we only present the results of RNR's application.

Applying RNR to the example described above identifies two clusters i.e. C_1 consisting of rows 5, 8, 7 and 9 and a slightly weak cluster C_2 consisting of rows 1, 2 and 4 and some singletons.

Figure 7 gives the resulting re-ordering of input matrix while figure 8 gives the corresponding bi-graph drawing with clusters identified by dotted boxes.

	R5	R8	R7	R9	R1	R2	R4	R6	R3	R10
R5	1	1	1	1	0	0	0	0	0	0
R8	1	1	1	1	1	1	1	1	1	0
R7	1	1	1	1	1	1	1	1	1	0
R9	1	1	1	1	1	1	1	1	1	0
R1	0	1	1	1	1	0	1	0	0	0
R2	0	1	1	1	0	1	1	0	0	0
R4	0	1	1	1	1	1	1	0	0	0
R6	0	1	1	1	0	0	0	1	0	0
R3	0	1	1	1	0	0	0	0	1	0
R10	0	0	0	0	0	0	0	0	0	1

Figure 7: RNR Output (Note Row and Column re-order)

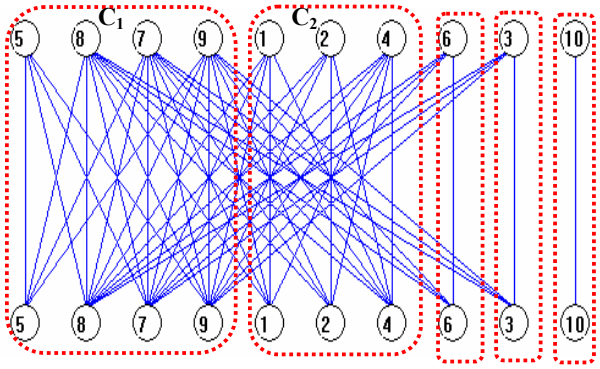

Figure 8: Output Bi-Graph Drawing

6. Discovering Patterns in Actual Data

We applied RNR on different groups of attributes in our pest scouting data and each experiment resulted in a unique finding. In this section we describe two of such experiments.

6.1. Experiment 1

Initiative behind this experiment was same as one of the goals set for this work i.e. finding the conditions in which the pesticide usage will be optimal. A typical question of a cotton grower would be "*which* pesticide should be used? And *when* ?"

We attempted to model these questions by looking for a pattern and relationship between pest population and meteorological data elements, and to find out (if possible) temperature and humidity thresholds at which population of a certain pest booms (or declines).

We randomly choose 20 records for year 2001 and noted populations of cotton pests such as Jasid (*Amrasca*), Thrips (*Thrip Tabaci*) and Spotted Boll Worm (*Earias Vitella*). Subsequently for each record retrieved the Min, Max temperature and % humidity from the daily weather database. This resulted in a 20×21 table. Subsequently we proceeded by creating a (20×20) similarity matrix based on calculating pair-wise Pearson's correlation.

Discritization of the similarity matrix was performed by leaving only those relationships in the similarity matrix that corresponded to the correlation threshold of 0.995 or above, all else were set to 0.

Figure 9 gives the random input. Two distinct clusters were identified by RNR heuristic as shown in figure 10. Clear grouping is found on the basis of pest populations where cluster 1 (C_1) have low populations (much below ETL[1]) and cluster 2 (C_2) has quite high pest populations. Average values shown in Table-1.

These clusters provide us with a good starting point and next we try to establish certain rules on the basis of this clustering. For each record, we look back in time for seven days and note meteorological recordings against minimum and maximum temperatures and

[1] Economic Threshold Level

humidity. Figure 11 shows the resulting graph of average values. Now on the basis of these graphs we establish two simple rules for pest population thresholds i.e.

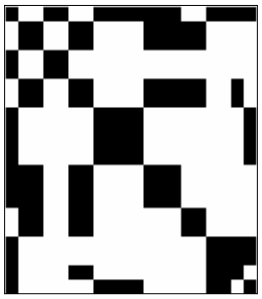

Figure 9: Input similarity matrix[2]

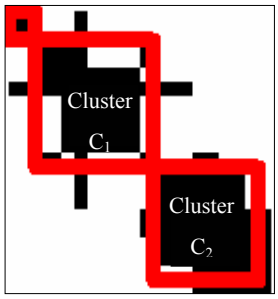

Figure 10: Clusters Identified by RNR

Cluster	Jassid	Thrip	SBW*
C_1	0.1	2.22	0.88
C_2	0.65	4.11	2.44
ETL	1	8-10	3

*Spotted Boll Worm

Table-1: Cluster Demographics

Temp > 29 AND Humidity > 70 then high pest incidence
Temp < 27 AND Humidity < 67 then low pest incidence

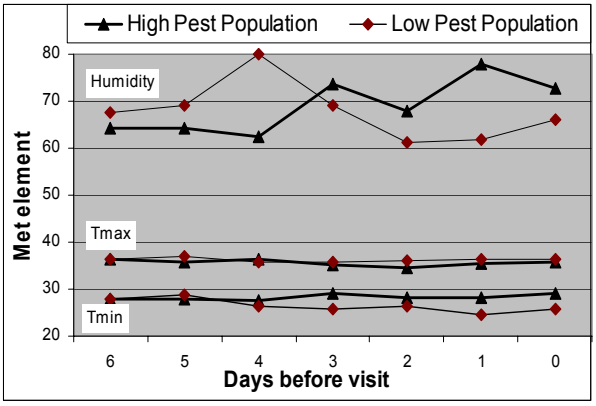

Figure 11: Meteorological recordings for the two clusters

Checking these rules against the data (325 matching records retrieved out of 2,000+) shows some very exciting results as shown in Figure 12.

This experimentation presents a very credible case that common farmer questions can be modeled through this data mining technique and answers can be given based on evidence present in the data **before** the pest attack occurs.

[2] Figure 9 presents a *structure plot* (Demetrescu, D. and Finocchi, I. (2001)) representation of a bi-graph where top vertex layer is placed on y-axis and bottom vertex layer on x-axis, an edge between two vertices is represented by shading corresponding point(s) in XY plane.

Strength of this method lies in clustering the evidence scattered in the data and hidden from the bare human mind.

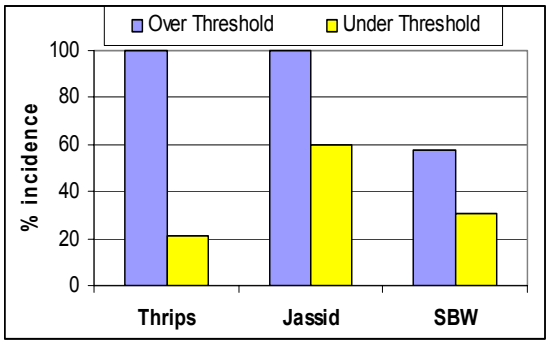

Figure 12: Experiment 1 Findings

6.2. Experiment 2

In this experiment we wanted to find out the possible reason(s) behind the variable efficacy of different pesticides i.e. why same pesticide used in similar quantities in *different* time period performs *differently*. For this we identified those records (68 of them) where pest populations were noted by the scouts one day after the pesticide was sprayed. We used seven variables i.e. day of sowing, day of spray, age of plant, pesticide used and population of different cotton pests such as SBW, Jassid, and Thrip. *Discritization* was performed much in the same way as in Experiment 1.

Figure 13 shows the discritized input similarity matrix. Running RNR on the input matrix results in the output matrix as shown in Figure 14 where strong grouping has occurred and two main clusters have been extracted. Table 2 gives the average cluster demographics.

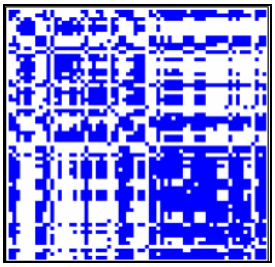

Figure 13: Input Similarity matrix

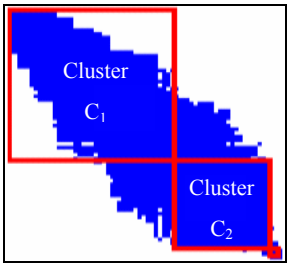

Figure 14: Clusters identified by RNR

	Sowing	Spray	Age	SBW	Jassid	Thrip
I/P	147.02	242.55	95.52	1.38	0.19	3.39
C_1	144.63	255.80	111.17	1.43	0.13	3.90
C_2	150.41	225.29	74.87	1.41	0.30	2.66

Table-2: Cluster demographics

RNR generated clustering are clearly distinctive on the basis of age of plant, Jassid and Thrip population change. Analysis of cluster details show that in C_1 age of all plants is above 100 days, and for C_2 age is at

most 90 days. There is grouping based on pesticides used also, the most interesting observation being presence of pesticide Cypermethrin (or Profinofos) in both clusters, but with different efficacy with respect to the pest Jassid, depending on the age of the plant. On further exploration it was strange to observe that peak usage of this pesticide some what coincides with an already declining population of Jassid, which shows delay in its usage and hence an increased population during early days of plant lifecycle; Thus resulting in two way loss to the farmer.

7. Conclusions

Unsupervised clustering of pest scouting, pesticide usage and meteorological data using Recursive Noise Removal heuristic is a new idea. In this work we have used this data for the first time in Pakistani agriculture and IT sector to dig out the answers for a complex scenario where pesticide usage is increasing with a simultaneous decrease of yield. This work shows that common farmer questions can be modeled quite easily into a data mining problem where one answers such questions by revealing patterns of interest in otherwise unorganized data. Through data intensive experiments we have discovered many interesting patterns.

8. References

- Abdullah, A. and Brobst, S. (2003): Clustering by recursive noise removal. *Proc. Atlantic Symposium on Computational Biology and Genome Informatics, USA,* pp. 973-977.
- Abdullah, A. (1993) "On placement of logic schematics", in *Proc. IEEE TENCON'93, Beijing, China,* pp. 885-888.
- Avesani, P., Olivetti, E., and Susi, S. (2002): Feeding Data Mining. *IRST Technical Report #0207-01, Istituto Trentino di Cultura, Povo (Trento), Italy.*
- Bertis, B., Johnston W. L., et al (2001): Data Mining in U.S. Corn Fields. *Proceedings of the First SIAM International Conference on Data Mining.*
- Chaudhry, M. R. (2000): New Frontiers in Cotton Production. *International Cotton Advisory Committee, USA.*
- Christensen, W. F., and Di Cook (1998): "Data Mining Soil Characteristics Affecting Corn Yield", citeseer.nj.nec.com/christensen98data.html
- Cunningham, S. J., and Holmes, G. (2001): Developing innovative applications in agriculture using data mining. *Department of Computer Science, University of Waikato, Hamilton, New Zealand.*
- Demetrescu, D. and Finocchi, I. (2001) "Removing cycles for minimizing crossings", Alcom-FT Technical report series, ALCOMFT-TR-01-148.
- Eades, P., and Wormald N. (1986): The median heuristic for drawing 2-layers networks. *Technical Report 69, Department of Computer Science, University of Queensland.*
- FAO (2001), Policy and Strategy for Rational use of Pesticides. *Food and Agriculture Organization. UNDP*
- Federal Bureau of Statistics (2002): Statistics Year Book of Pakistan. *Federal Bureau of Statistics, Govt. of Pakistan*
- Garey, W. R. and Johnson, D. S. (1979): *Computers and Intractability: A Guide to the Theory of NP-Completeness*. *Freeman, San Francisco.*
- Harms, S., et al (2000): Data Mining in a Geospatial Support System for Drought Risk Management. *Proc. First National Conference on Digital Government Research, Los Angeles, California,* pp. 9-16
- Han and Kamber (2000): Data Mining: Concepts and Techniques. *Morgan Kaufmann Publishers.*
- Hastings, J. et al (2002): CARMA: A Case-Based Rangeland Management Adviser, *AI Magazine*, 23(2): pp. 49-62.
- Nuyen, H. T., et al (2001): Some Practical Applications of Soft Computing and Data Mining. *A. Kandel, H. Bunke, and M. Last (eds.), Data Mining and Computational Intelligence, Springer-Verlag, Berlin,* pp. 273—307
- Scherte, S. L., PhD dissertation (2002): Data mining and its potential use in textiles: A spinning mill. *North Carolina State University.*
- Voss, H., et al. (2003): Simulation, Visualization, and Decision Support in GIMMI. *9 th EC GI & GIS Workshop, ESDI Serving the User, A Coruña, Spain*.
- Yang, C., Prasher S. O. and Landry J. A (1999) : Use of artificial neural networks to recognize weeds in a corn field. *Journée d'information scientifique et technique en génie agroalimentaire, Saint-Hyacinthe QC, Canada,* p. 60-65.