

*Working Paper Series*

HEARIN CENTER  
FOR  
ENTERPRISE SCIENCE

**HCES-03-99**

**Diversity Data Mining**

by  
**Gary Kochenberger**  
**Fred Glover**



***The University of Mississippi***

Director, Keith Womer  
School of Business Administration  
The University of Mississippi  
Post Office Box 1848  
University, MS 38677-1848  
(662) 915-5820  
<http://hces.bus.olemiss.edu>

---

# Diversity Data Mining

Gary Kochenberger<sup>a</sup> and Fred Glover<sup>a</sup>

<sup>a</sup> Hearin Center for Enterprise Science, School of Business Administration, University of Mississippi, University, MS 38677, USA .

Latest Revision: August , 1999

---

**Abstract** — Data mining has greatly assisted management in recent years, demonstrating particular value in the commercial sector by revealing insights that have led to profitable market opportunities. With all its emerging sophistication, however, data mining stops short of offering management an ability to optimize certain key decisions. Consequently, the full promise of data supported decision making is still not fully realized.

A special type of optimization capability, embodied in a model of *diversity data mining (DDM)*, provides a useful new tool to augment and expand the functionality of current data mining methods. We show that DDM has important applications in a wide range of areas in business, industry and government, and we describe a highly effective algorithm for solving such problems. We provide computational experience demonstrating the efficiency of our approach, and show it is able to successfully handle models more than an order of magnitude larger than those previously treated in the literature of diversity optimization.

---

This research has been supported in part by the National Partnership for Advanced Computational Infrastructure (NPACI) grant ACI – 9619020.

## 1. INTRODUCTION

Database management systems for storing and manipulating large quantities of data have become commonplace in recent years. The quality revolution of the 1980s has created a legacy that has imparted a strong customer and market orientation to data base activities. Firms in every segment of industry are collecting and warehousing unprecedented volumes of data. These activities, facilitated by modern IT/IS systems, are implemented for the purpose of capturing data that yield a store of useful information when subjected to appropriate analysis – a goal whose repeated realization is generating substantial rewards for companies in every sector of the economy. Public and private organizations alike are amassing ever-increasing amounts of data. Data warehousing applications are crossing the boundaries of organizational function and classification, demonstrating a relevance that is universal.

The usefulness of data warehousing has been significantly enhanced by the development of data mining techniques over the past decade. Data mining tools enable management to go beyond simple inquiries and descriptive statistical analysis to engage in pattern recognition and prediction. The quest to uncover “hidden information” and to assess its potential uses has provided valuable insights in medical research, geological exploration, market analysis, fraud detection and many other areas. (See, e.g., Fayyad et al., 1996; Kelly, 1997; Ramakrishnan and Grama, 1999.)

By discovering patterns and relationships that would otherwise go undetected, data mining makes it possible to derive greater value from data than can be obtained from simple query and descriptive analysis approaches. The current state of the art, however, stops short of offering management the ability to extract optimal collections of information – i.e., collections that maximize some criterion of merit – and thus these developments fail to provide a thoroughly adequate foundation for high level strategic planning. Because of this limitation, current data mining methods have not yet realized their full potential. It is noteworthy, however, that the step of adding an optimizing capability to the suite of data mining tools is technically possible today. Moreover, such a step offers the potential to greatly expand the managerial usefulness of data warehousing.

Our goal in this paper is to identify an optimization model for data mining that contributes to closing the gap that currently exists. This model gives rise to the realm we call *diversity data mining (DDM)*, which embodies the objective of identifying optimally diverse subsets of populations. The formulation of models for optimizing diversity is itself a domain of relatively recent vintage (Kuo, Dhir and Glover, 1993), and its applications are only beginning to be explored in depth. Within the realm of data mining, however, as manifested through DDM, these applications become particularly compelling. As we will demonstrate, diversity data mining addresses critical needs in a host of decision making areas.

In the sections that follow, we describe a general diversity maximization model as a foundation for DDM. In addition, we identify a highly effective algorithm for this model which reinforces the practical significance of DDM. We demonstrate the power of this method by computational testing on problems of sizes that are vastly larger than any previously considered in the domain of diversity optimization. We also show how the basic model can be expanded to include a variety of additional conditions and constraining relationships, enhancing its range of applications. Finally, we review the long term implications of diversity data mining and examine the potential for future applications.

## 2. THE GENERAL DIVERSITY MAXIMIZATION PROBLEM

The diversity maximization problem can be stated as follows. Consider a set of elements  $S = \{s_i : i \in N\}$ , defined over the index set  $N = \{1, 2, \dots, n\}$ , where each element,  $s_i$ , has  $r$  attributes denoted by  $s_{ik}$ ,  $k \in R = \{1, 2, \dots, r\}$ . The objective is to select a subset of size  $m$ , where  $m$  is strictly less than  $n$ , to maximize the *diversity* of the elements chosen.

To express the objective function formally, we associate a measure of diversity  $d_{ij}$  with each pair of elements  $s_i$  and  $s_j$ ; that is,  $d_{ij}$  is some function of the elements  $s_{ik}$  and  $s_{jk}$ ,  $k \in R$ , which is selected by the decision maker according to the context. Then the problem can be represented:

$$\text{Max}D : \max x_0 = \sum_{i \in N} \sum_{j \in N} d_{ij} x_i x_j = xDx$$

subject to

$$\sum_{j \in N} x_j = m$$

where  $d_{jj} = 0$  and  $x_j$  is a binary variable denoting whether or not element  $j$  is chosen to be a member of the selected subset.

We show in the following section that the *MaxD* model is quite general and is capable of representing problems from a wide variety of areas. In spite of the apparent simplicity of its formulation, the model contains an unsuspected capacity for handling multiple considerations of considerable complexity. Included among these are situations where the goal is not merely to select elements that are diverse, but that also satisfy required levels of quality along multiple dimensions.

## 3. DATA MINING APPLICATIONS OF DDM

We begin by sketching some of the general areas of diversity data mining to which the foregoing model can be applied, and then illustrate the manner in which such applications are expressed in the model. The overall goal of the *MaxD* model within the DDM setting is to extract information from a relevant data base to support decision making. The need to make decisions involving diversity issues conspicuously arises in many realms of business and government. Yet this type of application has been largely overlooked in the literature. A principal reason is that methods for solving diversity maximization models have previously been exceedingly limited in their ability to solve problems of practical size and complexity. We remove this obstacle with the solution method described in the paper.

To illustrate the types of applications possible, the following examples identify settings where the maximum diversity concept is critically important.

1. *Environmental Balance*: Ecological systems depend on diversity for survivability. Considerations of diversity maximization are crucial for establishing systems that are viable, robust and balanced.
2. *Medical Treatment*: Combating diseases, both by preventive planning and after the onset of illness, is enhanced by programs that offer more diverse lines of defense in order to combat the broadest spectrum of potential disease causing agents.
3. *Genetic Engineering*: Recombinant DNA and RNA applications yield a richer field of outcomes by designs that generate greater number of alternatives, and where those alternatives, in turn, embody greater diversity in their underlying structure.
4. *Molecular Structure Design*: The quest for improved molecular structures, which affect fields ranging all the way from medicine to metallurgy, depends on finding stable ways to fit molecular shells, and to appropriately position component molecules in available candidate locations. Processes to achieve this have so far

been limited by the range of diversity in the elements that are generated and interrelated by standard approaches.

5. *Agricultural Breeding Stocks*: In both animal and plant genetics, the goal of obtaining new varieties by controlled breeding strategies is aided by drawing on breeding stocks with desirable qualities of diversity. Better ways of characterizing and generating subsets of stocks with maximum diversity directly contribute to this goal.
6. *Right Sizing the Firm*: Organizations that need to engage in “downsizing” are at risk of creating critical skill and knowledge gaps when employees with very similar profiles are eliminated. The adverse loss of institutional knowledge can be mitigated by developing a downsizing plan designed to maximize the diversity of those who are retained with the firm.
7. *Composing Jury Panels*: The pursuit of a fair and complete hearing of the evidence brought against a defendant is best served when the case presented is viewed and analyzed from diverse points of view. This ideal is approached by selecting jurors from a pool of qualified citizens with the goal of maximizing the diversity of those chosen.

The applications mentioned above have a common theme of harvesting information from a data base to assist in selecting elements with the greatest variety of characteristics. In such applications, the *MaxD* model serves as an interface between the data base housing the raw data and the decision maker. Implementing the model constitutes an advanced form of data mining by revealing information in the form of optimal solutions – solutions not observable directly from the data in the absence of the model.

#### **4. SOLUTION METHODOLOGY**

The *MaxD* model belongs to a class of NP-hard problems, and thus no method is known to exist that is guaranteed to be able to find an optimal solution in “better than exponential” time. In fact, methods that can be proved to converge are unable to find and verify optimal solutions for many problems of realistic sizes within reasonable time limits. Consequently, except for small instances, such problems are preferably approached by heuristic methods rather than exact (theoretically finite) methods.

The literature on *MaxD* contains few papers of a computational nature and the methods reported have been tested on rather small problem instances only. In the original exposition of the model, Kuo, Glover and Dhir (1993) present an equivalent linear mixed integer zero-one formulation of *MaxD* and demonstrate its use on a small example problem. This approach has the advantage of lending itself to readily available, optimal seeking branch and bound algorithms. However, this approach is not viable for problems large enough to be of significant practical interest.

More recently, Ghosh (1996) presents a randomized greedy heuristic for the problem, and Glover, Kuo and Dhir (1998) give several constructive and destructive heuristics. All these methods have been shown to produce high quality solutions on small test problems (30 to 40 variables) where optimal solutions are known. The virtue of these previously reported methods lies in their simplicity. The downside is that for a problem instance of greater dimension, they lack the “intelligence” to navigate a complicated solution space characterized by strong local optima.

The approach we take in this paper employs a basic version of tabu search (TS) to guide our search of the solution space. The added search sophistication of TS notably enhances our ability solve problems of much greater dimension and difficulty than is possible by lower level heuristics alone. Our tabu search implementation to solve the *MaxD* model is a variation of the method we have developed and extensively tested for solving the unconstrained binary quadratic program. Detailed descriptions of the method can be found in Glover, Kochenberger and Alidaee (1998) and Glover, Kochenberger, Alidaee, and Amini (1999). Below we give a brief overview of the method.

##### **4.1 Tabu Search Overview**

Our method is centered around the strategic oscillation approach, which constitutes one of the primary strategies of tabu search. The variant of strategic oscillation we employ alternates between constructive phases that progressively set variables to 1 (whose steps we call “add moves”) and destructive phases that progressively set variables to 0 (whose steps we call “drops moves”). To control the underlying search process, we use a memory structure that is updated at *critical events*, which are identified by conditions that generate a subclass of locally optimal solutions. Solutions corresponding to critical events are called *critical solutions*. For the maximum diversity problem, we modify the definition of a critical event to stipulate that such an event occurs when an add move during a constructive phase or a drop move during a destructive phase yields a trial solution with exactly  $m$  variables equal to 1.

A parameter *span* is used to indicate the amplitude of oscillation about a critical event. We begin with *span* equal to 1 and gradually increase it to some limiting value. For each value of *span*, a series of alternating constructive and destructive phases is executed before progressing to the next value. At the limiting point, *span* is gradually decreased, allowing again for a series of alternating constructive and destructive phases. When *span* reaches a value of 1, a *complete span cycle* has been completed and the next cycle is launched.

Information stored at critical events is used to influence the search process by penalizing potentially attractive add moves (during a constructive phase) and inducing drop moves (during a destructive phase) associated with assignments of values to variables in recent critical solutions. Cumulative critical event information is used to introduce a subtle long term bias into the search process by means of additional penalties and inducements similar to those discussed above. A complete description of the framework for the method is given in Glover, Kochenberger, Alidaee and Amini (1999).

## 5. COMPUTATIONAL EXPERIENCE

Our method for solving *MaxD* was tested on randomly generated problems of size 100, 300, 500 and 1000 variables. These problems are substantially larger than those reported earlier in the literature. Prior to this study, the largest problem tested had just 40 variables.

For each problem size, five instances of the problem were considered, each with a different value of  $m$ . Our test problems were 100 % dense (off diagonal) and the  $d_{ij}$  values were randomly generated between 10 and 50. (Our method is not restricted to require the  $d_{ij}$  coefficients to satisfy sign conditions such as nonnegativity, however.) Results from our runs are shown in Table 1.

**TABLE 1.** Test Problem Results

ID	n	m	Best	Cycle	Run	Total
Div11	100	10	3824	1	20	2
Div12	100	15	8316	1	20	2
Div13	100	20	14280	1	20	2
Div14	100	25	21742	1	20	2
Div15	100	30	30648	4	20	2
Div21	300	30	32244	3	50	15
Div22	300	45	69920	10	50	15
Div23	300	60	120848	20	50	15
Div24	300	75	185268	44	50	15
Div25	300	90	262786	48	50	15
Div31	500	50	87172	7	100	58
Div32	500	75	189346	65	100	58

ID	n	m	Best	Cycle	Run	Total
Div33	500	100	329572	25	100	58
Div34	500	125	506994	2	100	58
Div35	500	150	720962	5	100	58
Div41	1000	100	335490	23	100	194
Div42	1000	150	734564	76	100	194
Div43	1000	200	1283750	7	100	194
Div44	1000	250	1982416	64	100	194
Div45	1000	300	2828982	79	100	194

Note: **Run Length** is the total # of Span cycles executed  
**Total Time** is the time in Pentium 200 seconds for the **Run Length**  
(rounded up to the nearest second)  
**Best Soln** is the best solution found during the entire search process  
**Cycle** is the cycle # at which the best solution was found

The solutions shown in Table 1 represent the best solutions found by our method in one run terminated by the limit on the number of span cycles allowed (as listed in the table). Neither iterated restarting nor parameter tuning were employed. Experience in other settings with a wide variety of problems (see Glover, Kochenberger, & Alidaee, 1998) suggests that the quality of the solutions generated by this approach is very high and often optimal. Thus, we expect the solution listed here to be of high quality, although no exact method exists for *MaxD* models of these sizes that is able to provide a basis for comparison.

Our solution times, even for the largest problems, are quite modest. We solve 100 variable problems in less than 2 seconds, and comfortably solve 1000 variable problems in less than 200 seconds on a Pentium 200 PC. Significantly, when these nonlinear problems are expressed as equivalent linear mixed integer models, the number of variables becomes  $n(n+1)/2$ . A problem with  $n = 1000$  thus translates into one with more than 500,000 variables, which is two orders of magnitude larger than a problem with  $n = 100$ , relative to a standard comparison based on a linear model. (Similarly, these problems are nearly three orders of magnitude larger than those with  $n = 40$ , the previous largest value of  $n$  tested.) Our on-going testing indicates that problems in which  $n$  receives a value of several thousand can be solved in a matter of a few minutes on a PC.

## 6. MODEL EXTENSIONS

The basic diversity model, *MaxD*, is robust enough to represent a wide range of problems. Nonetheless, further considerations can arise in certain applications that require additional constraints to be added to the model. If these new constraints are linear, they can be accommodated via quadratic penalties within the basic *MaxD* framework; i.e., the further constrained model can be re-cast into the form of *MaxD* and solved by the method illustrated in this paper. The paper by Kochenberger, Alidaee and Amini (1998) discusses such re-formulations in general. We illustrate some useful possibilities of this type by the examples below.

### 6.1 Reformulation Example 1.

Suppose that two elements, which may be quite different from each other on most attributes, are unacceptably close on some critical attribute – so much so that we want to require that not both elements be chosen. Denoting the elements by  $i$  and  $j$ , we can preclude both from being chosen by imposing the constraint

$$x_i + x_j \leq 1.$$

Such a constraint is not explicitly accommodated by the *MaxD* model. However, we can readily handle the constraint by introducing the penalty term

$$Px_i x_j$$

and subtracting it from the objective function, where P is a suitably chosen positive constant. Since we are maximizing, not both  $x_i$  and  $x_j$  will receive a value of 1 in an optimal solution.

Denoting by M the set of element pairs that require such mutually exclusive conditions, our modified problem can be written

$$\max x_0 = xDx - P \sum_{(i,j) \in M} x_i x_j = xQx$$

subject to

$$\sum_{j=1}^n x_j = m$$

By absorbing the quadratic penalty terms into the matrix D (to produce the matrix Q), we retain the form of the original model, *MaxD*, which enables our tabu search method to be applied without modification. The parameter P must be large enough to force the desired result. Any value greater than an upper bound on the original objective function will clearly work. However, much smaller values have proven successful in practice.

## 6.2 Reformulation Example 2.

The mutually exclusive relationships considered in the preceding example are a special case of a more general type of relationship encountered in a variety of application settings. For example, costs may be attached to the elements to be selected, and a budget limit may be imposed by means of a general linear inequality. Similarly, measures of quality may be associated with the elements, and a linear inequality can be introduced to assure the elements chosen will satisfy an overall quality level. (Multiple measures and inequalities can be introduced to handle definitions of quality of different types.)

A variety of other considerations may likewise lead to further constraints in the form of linear inequalities or equations. In general, linear inequalities over zero-one variables with integer coefficients can be transformed into equations by identifying appropriate bounds on associated slack variables, and then replacing these slack variables by equivalent expansions of zero-one variables.

Whenever the constraining relationships can thus be represented by a system of linear equations in the binary variables, a quadratic penalty (of slightly different construction than the one considered in the previous example) can be employed to incorporate the relationships into the form of the basic *MaxD* model.

To illustrate the approach, consider the further constrained model of the form

$$\max x_0 = xDx$$

subject to

$$\sum_{j=1}^n x_j = m$$

$$Ax = b$$

where the equality system  $Ax = b$  represents the additional relationships that need to be taken into account. Taking P, as before, to be a suitably chosen positive penalty, we can re-write the foregoing model as

$$\max x_0 = xDx - P*(Ax - b)'(Ax - b)$$

$$= xDx + xZx + c$$

$$= xQx + c$$

subject to

$$\sum_{j=1}^n x_j = m$$

where the matrix  $Z$  and the additive constant  $c$  result directly from the matrix multiplication indicated. Thus we are back once more to our basic *MaxD* model, disclosing its broad applicability to DDM problems. This reformulation again affords the opportunity to exploit these problems with our tabu search approach.

## 7. SUMMARY & CONCLUSIONS

The realm of diversity data mining (DDM) makes it possible to identify subsets of populations that maximize measures of diversity, and has important applications in a wide variety of areas. The *MaxD* diversity model we have identified as a foundation for DDM applications gives a means to extract information in a form and quality not otherwise available.

To give these applications practical significance, we have identified a special solution method based on tabu search. Our computational testing shows that this method quickly obtains high quality solutions to problems of dramatically greater size than previously tested in the literature. Such an ability is crucial for handling problems that arise in real world settings.

In addition, we have shown how extensions of the basic model can readily be reformulated to permit relationships of diverse structure and complexity to be captured within the same model framework. Our introduction of a model for diversity data mining, and the demonstration that it can be solved effectively to yield a useful practical tool, opens the door to studies for assessing the potential of DDM in a wide variety of applications.

## 8. REFERENCES

U.M. Fayyad et al., eds. (1996), *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Menlo Park, Calif.

J.B Ghosh (1996), "Computational Aspects of the Maximum Diversity Problem," *Operations Research Letters*, Vol. 19, pp. 175-182

F. Glover, C. C. Kuo, and K.S. Dhir (1998), "Heuristic Algorithms for the Maximum Diversity Problem," *Journal of Information & Optimization Sciences*, Vol. 19, No. 1, pp. 109-132.

F. Glover, G. Kochenberger, and B. Alidaee (1998), "Adaptive Memory Tabu Search for Binary Quadratic Programs," *Management Science*, Vol 44, pp 336-345.

F. Glover, G. Kochenberger, B. Alidaee, and M. Amini (1999), "Tabu Search with Critical Event Memory: An Enhanced Application for Binary Quadratic Programs", Research Report, Hearin Center for Enterprise Science, University of Mississippi.

S. Kelly (1997), *Data Warehousing in Action*, John Wiley & Sons Ltd, England.

G. Kochenberger, B. Alidaee, and M. Amini (1998) "Applications of the Unconstrained Binary Quadratic Program," Working Paper, University of Colorado.

C.C. Kuo, F. Glover, and K.S. Dhir (1993), "Analyzing and Modeling the Maximum Diversity Problem by Zero-One Programming," *Decision Sciences*, Vol.24, pp. 1171-1185.

N. Ramakrishnan and A. Grama (1999), "Data Mining: From Serendipity to Science," *Computer*, August 1999, 34-37.